A Statistical analysis of algorithms dedicated for rare events

Anass Aghbalou

INSTITUT POLYTECHNIQUE DE PARIS

NNT : 2024IPPAT004

Thèse de doctorat

A Statistical Analysis of Algorithms Dedicated for Rare Events

Thèse de doctorat de l'Institut Polytechnique de Paris
préparée à Télécom Paris

École doctorale n°574 Ecole Doctorale de Mathématiques Hadamard (EDMH)
Spécialité de doctorat : Mathématiques appliquées

Thèse présentée et soutenue à Palaiseau, le 14 Février 2024, par

ANASS AGHBALOU

Composition du Jury :

Pavlo Mozharovskyi Associate Professor, Télécom Paris (LTCI)	Président/Examinateur
Stanislav Volgushev Professor, University of Toronto	Rapporteur
Olivier Wintenberger Professor, Sorbonne Université	Rapporteur
Maud Thomas Assistant professor, Sorbonne Université	Examinatrice
Anne Sabourin Professor, Université Paris Cité	Directrice de thèse
François PORTIER Associate Professor, ENSAI-CREST	Invité
Patrice Bertail Professor, Université Paris Nanterre	Invité

Abstract

This thesis focuses on the statistical analysis of algorithms dedicated to rare events, a critical area in contexts where data is scarce, such as financial extremes, environmental risks, and rare diseases. Standard statistical tools, like Hoeffding's inequality and empirical risk minimization, lose effectiveness in environments with limited data. This situation requires the development of new probability bounds and specific algorithms designed to enhance performance in scenarios of data scarcity.

The study first explores the challenges inherent in extreme values. The goal is to improve prediction capabilities in extreme zones where events are rare, and data is sparse. The theoretical framework includes dimension reduction techniques based on approaches like Sliced Inverse Regression (SIR), which optimizes estimation by reducing biases from high-dimensional data. The concept of Tail Conditional Independence (TCI) is introduced to assist in modeling extreme regions. This concept is based on the idea that certain linear combinations of covariates are sufficient for effectively predicting extreme values. This statistical framework is applied in fields such as finance, where precise predictions are essential for rare but critical events.

The second part of the thesis addresses cross-validation in data-scarce regions. It demonstrates that traditional cross-validation schemes, such as the K-fold method, can induce high biases in data-scarce scenarios. An alternative is proposed with polynomial and exponential probability bounds to evaluate the generalization risk. These new bounds enable better control of prediction errors and enhance the robustness of algorithms in rare regions. Specific adjustments to cross-validation are explored to minimize risk based on stability algorithms.

The thesis also tackles the challenge of imbalanced classification, where a minority class of interest is underrepresented, a common issue in areas like fraud detection and medical diagnostics. The cost-sensitive classification technique is favored. This method adjusts the learning algorithm by assigning different misclassification costs to each class, helping to avoid bias toward majority classes. By incorporating adapted deviation bounds, the results show that these techniques can ensure reliable performance even when the probability of rare events is very low.

In conclusion, this thesis proposes an innovative theoretical framework for scenarios where data scarcity renders classic approaches inadequate. The probabilistic bounds developed are validated by empirical experiments, demonstrating their ability to certify algorithm effectiveness in environments with limited data. This research thus contributes to extending the use of machine learning and statistical analyses in critical contexts, such as financial forecasting and medical diagnostics, where prediction accuracy and reliability are essential.

Résumé

Cette thèse se concentre sur l'analyse statistique des algorithmes dédiés aux événements rares, un domaine crucial dans des contextes où les données sont limitées, comme les extrêmes financiers, la gestion des risques environnementaux ou l'étude des maladies rares. Les outils statistiques standards, tels que l'inégalité de Hoeffding et la minimisation du risque empirique, perdent en efficacité dans ces environnements de rareté de données. Cela exige le développement de nouvelles bornes de probabilité et d'algorithmes spécifiques permettant d'améliorer la performance pratique dans ces scénarios.

Dans un premier temps, l'étude explore les défis posés par l'analyse des valeurs extrêmes. L'objectif est d'améliorer les capacités de prédiction dans les zones extrêmes où les événements sont rares et les données peu nombreuses. Le cadre théorique inclut des techniques de réduction de la dimension basées sur des approches comme la Sliced Inverse Regression (SIR), qui optimise l'estimation en réduisant les biais dus à la grande dimension des données. Le concept d'indépendance conditionnelle en queue (TCI) est introduit pour aider à la modélisation des régions extrêmes. Cette notion repose sur l'idée que certaines combinaisons linéaires de covariables suffisent pour prédire efficacement les valeurs extrêmes. Ce cadre statistique est appliqué à des domaines tels que la finance, où des prévisions précises sont essentielles pour des événements rares mais cruciaux.

L'analyse des valeurs extrêmes s'appuie sur des modèles probabilistes permettant de prévoir la probabilité d'occurrences futures dépassant des seuils élevés. Dans ce contexte, la thèse introduit des méthodes de réduction de la dimension, adaptées aux situations où les données sont insuffisantes pour appliquer des méthodes traditionnelles. Ces approches permettent de simplifier l'analyse des données multidimensionnelles tout en assurant la fiabilité des prédictions.

La deuxième partie de la thèse aborde la validation croisée dans des régions où les données sont rares. Elle démontre que les schémas classiques de validation croisée, tels que la méthode K-fold, peuvent induire des biais significatifs dans les scénarios de pénurie de données. Pour pallier cette limite, des alternatives sont proposées, avec l'introduction de bornes de probabilité polynomiales et exponentielles, afin de mieux évaluer le risque de généralisation des modèles. Ces nouvelles bornes permettent de contrôler plus efficacement les erreurs de prédiction et d'améliorer la robustesse des algorithmes dans les régions où les données sont rares. Des ajustements spécifiques à la validation croisée sont également explorés pour minimiser le risque et garantir des estimations fiables, en particulier lorsque les algorithmes sont soumis à des contraintes de stabilité.

La thèse aborde également le défi de la classification déséquilibrée, où une classe minoritaire d'intérêt est souvent sous-représentée. Ce problème est particulièrement présent dans des domaines tels que la détection de fraudes, le diagnostic médical ou la surveillance des infrastructures critiques. La classification sensible aux coûts est privilégiée, car elle modifie l'algorithme d'apprentissage en attribuant des coûts de classification différents pour chaque classe. Cela permet au modèle de porter une attention accrue aux classes minoritaires et de limiter les biais en faveur des classes majoritaires. En inté-

grant des bornes de déviation adaptées aux scénarios de rareté, les résultats démontrent que ces techniques garantissent une performance robuste, même lorsque la probabilité des événements rares est très faible.

Un aspect novateur de cette recherche réside dans la démonstration empirique de la validité des bornes proposées. Les expériences réalisées montrent que ces bornes permettent de certifier l'efficacité des algorithmes, même dans des contextes où les données disponibles sont limitées et où l'incertitude statistique est plus importante. La thèse souligne l'importance de développer des approches théoriques qui intègrent la réalité de la rareté des données, en mettant en avant des solutions adaptées qui s'écartent des hypothèses classiques de disponibilité abondante des données.

En conclusion, cette thèse propose un cadre théorique et méthodologique innovant, spécifiquement conçu pour les environnements où la rareté des données rend les approches classiques inefficaces. Les bornes probabilistes et les ajustements méthodologiques développés sont validés par des études empiriques, démontrant leur capacité à certifier la fiabilité des algorithmes dans des environnements critiques. Cette contribution ouvre des perspectives nouvelles pour l'application de l'apprentissage automatique et de l'analyse statistique dans des domaines sensibles, tels que la finance, la médecine et la gestion des risques, où la précision et la robustesse des prédictions sont indispensables. Ces travaux posent également les bases pour des recherches futures visant à affiner encore davantage les garanties théoriques et à élargir leur application à d'autres types de données et de contextes rares.

Thesis outline and reading guide

We now present the outline of this manuscript :

- Chapter 1 introduces the working frameworks of the manuscript and establishes their connection with the main subject, which is data scarcity. It's essential for setting the stage for the subsequent chapters.
- Chapter 2 develops probability bounds for cross-validation procedures dedicated to the evaluation of algorithms from extreme value analysis.
- Chapter 3 extend the methodology of Sliced Inverse Regression to the field of extreme value analysis.
- Chapter 4 develops sharp bounds for imbalanced classification problems.
- Chapter 5 focuses on studying transfer learning from the point of view of *algorithmic stability*.
- Chapter 6 Serves as a conclusion and opening for further research, this chapter study cross-validation within an algorithmic stability framework, exploring its limitations and setting the stage for future studies on stability in data-scarce environments.

Remark to Readers

This thesis is structured to facilitate focused reading. Each section is designed to be self-contained, complete with its own set of notations and contextual background. Readers with interest in a specific topic or methodology can directly navigate to the relevant section without the necessity of reading preceding sections. This approach is intended to accommodate both comprehensive readers and those seeking insights into particular areas of our study.

Contents

Notation

$:=$	Equal by definition
$\mathbb{N}, \mathbb{R}$	Sets of natural and real numbers
$\mathbb{R}^d$	Set of d-dimensional real-valued vectors
$\langle x, y \rangle$	Inner product of vectors $x, y \in \mathbb{R}^d$
n	number of examples in the full data sample
n_T	number of examples in the training set
n_V	number of examples in the validation set
$\mathcal{X}$	The input space where the examples of a given task belong
$\mathcal{Y}$	The output space of a task
$\mathcal{Z}$	The joint space between inputs and outputs with $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$
$\|x\|_p$	ℓ_p-norm of vector $x \in \mathbb{R}^d$
$\mathbb{R}^{n \times d}$	Set of real matrices of size $n \times d$
$g \in \mathcal{G}$	A predictor g from a hypothesis class $\mathcal{G}$
I_d	Identity matrix of size $d \times d$
$A^\top$	Transpose of matrix A
$\mathrm{Tr}(A), det(A)$	Trace and Determinant of matrix A
$\|A\|_F$	Frobenius norm $\|A\|_F = \mathrm{Tr}(A.A^\top)$
$\mathrm{supp}(\cdot)$	Support of a function or a vector
$\mathbb{1}_E$	Characteristic function of set E
A^c	Complementary set of set A
$\mathbb{P}(\cdot)$	Probability of an event
$\mathbb{E}[\cdot]$	Expectation of a random variable
$\overset{\text{i.i.d}}{\sim}$	Independent and Identically Distributed
$L_2(\pi)$	Set of square integrable functions with respect to measure π
$X \sim P$	Random variable X has distribution P
$\mathcal{N}(\mu, \Sigma)$	Gaussian distribution with mean μ and covariance matrix Σ
$\ell(g, Z)$	A user defined positive loss function and where $Z \in \mathcal{Z}$

Chapter 1

General introduction, motivations and contributions

In the modern context, data is often likened to "new oil" for its considerable value and significance. It's commonly perceived that data is abundant and easily accessible across various domains. Yet, this isn't always the case. In several crucial fields like as medicine (Ahsan and Siddique, 2022), environmental science (Liu et al., 2022), economics (Pandeya et al., 2016), and social sciences (Laurer et al., 2023), the shortage of data presents significant challenges for decision-making and analytical precision due to insufficient data. This data scarcity negatively impacts data-centric methods such as machine learning, which rely on large datasets for precise model training and validation.

Contrary to being a temporary problem, data scarcity is a persistent issue in areas such as studying rare diseases, detecting financial fraud, or conserving endangered species, where limited data is a fundamental characteristic. This necessitates a shift in the mathematical techniques employed for analysis and prediction in these sectors.

Traditional statistical and machine learning techniques often assume the availability of extensive data, an assumption not valid in data-limited situations. This discrepancy between methodological expectations and real-world conditions necessitates a reevaluation of theoretical models. Approaches like cost-sensitive learning, transfer learning, and extreme value analysis, while promising, are theoretically underdeveloped for data-poor environments. The standard theoretical assurances linked with these methods typically assume ample data, creating a gap in their effectiveness for data-scarce scenarios.

To clarify the learning problems addressed, we focus on i.i.d. data $Z_i = (X_i, Y_i)_{i \leq n}$, with common distribution P, lying in a sample space $\mathcal{X} \times \mathcal{Y} = \mathcal{Z}$, and a low probability region $\mathbb{A} \subset \mathcal{Z}$ *i.e.* $P(Z \in \mathbb{A}) = \alpha \ll 1$. This thesis investigates the statistical behavior of algorithms and procedures dedicated to these low-probability regions.

The primary aim of this thesis is to fill this gap by deriving new theoretical assurances that adapt to the challenges of limited data availability. Our goal is to enhance the effectiveness and reliability of statistical guarantees in data-limited contexts. This thesis concentrates on developing statistical convergence rates for:

1. Extreme value analysis.

2. Highly imbalanced classification.

3. Transfer learning.

1.1 Data scarcity in Extreme value analysis

1.1.1 Introduction to Extreme Value Theory (EVT)

Extreme Value Theory (EVT) (de Haan, 1970; Beirlant et al., 2006; Resnick, 2007) specializes in modeling the extraordinary, not the ordinary. It plays a crucial role in industries such as finance, insurance, telecommunications, and environmental sciences for risk management purposes. EVT's core function is to accurately assess the probability of rare occurrences.

The theory offers a method to predict the likelihood of future events that exceed previous extreme records. It aims to understand how random variables behave when they surpass high thresholds, a focus particularly pertinent for heavy-tailed distributions where the likelihood of extreme events is significantly higher. For univariate data, EVT mainly examines the process's highest possible values. The asymptotic nature of these maximums is well-understood and described by the generalized extreme value distribution, as noted by various studies. In multivariate contexts, defining extremes is more intricate due to the absence of a natural order. However, extremes can be identified through various methods such as threshold exceedances, analyzing component-wise maxima, or examining a user-defined norm of vectors. Extreme value theory faces challenges as data dimensions increase. Common dimension-reduction methods like principal component analysis (PCA), which focus on covariance matrices, may not be viable with heavy-tailed distributions where covariance might not be defined. Recent works by Cooley and Thibaud (2019b); Drees and Sabourin (2021) have proposed adaptations to these traditional methods, addressing these limitations.

From a statistical perspective, one of the primary challenges of EVT is the inherently infrequent nature of extreme events, which means that only a small fraction of data is available for analysis. This limited sample size exacerbates the curse of dimensionality in EVT, making it challenging to draw reliable inferences from such sparse data ($k \ll n$).

Moreover, working in extreme regions poses a significant challenge for most, if not all, ML algorithms. Considering extreme data points Z—those where the norm $||Z||$ goes beyond a substantial threshold $t > 0$ their rarity makes them underrepresented in the training set $\mathcal{D}$. Thus, errors in such input space areas may insignificantly affect the holistic prediction error of a predictor $g \in \mathcal{G}$. Leveraging the total probability formula, we can write the statistical risk with respect to a loss function $\ell : \mathcal{G} \times \mathcal{Z} \mapsto \mathbb{R}$ as follows:

$$\begin{aligned}\mathcal{R}[g] &:= \mathbb{E}\left[\ell(g, Z)\right] \\ &= P(\|Z\| < t)\mathbb{E}\left[\ell(g, Z) \mid \|Z\| < t\right] + P(\|Z\| \geq t)\mathbb{E}\left[\ell(g, Z) \mid \|Z\| \geq t\right].\end{aligned}$$

Given the minuscule magnitude of $P(\|Z\| > t)$ and its empirical version, there is no assurance that standard ML methods yield an optimal classifier for the rare region

$$\mathbb{A} = \{z \ : \|z\| > t\}.$$

This means that the measure $\mathbb{E}\left[\ell(g, Z) \mid \|Z\| \geq t\right]$ might not be close to optimal. Nonetheless, in areas like finance, insurance, and aeronautical safety, precise predictions in extreme regions are paramount.

To address this challenge, numerous algorithms specially designed for extreme regions have been introduced in recent times, driven by critical factors like dimensionality reduction and/or anomaly detection (Goix et al. (2016, 2017); Thomas et al. (2017); Chiapino and Sabourin (2016); Drees and Sabourin (2021); Jalalzai and Leluc (2021), see also the review papers Engelke and Ivanovs (2021); Suboh and Aziz (2020)), data augmentation (Jalalzai et al., 2020), adversarial simulation (Bhatia et al., 2021), graphical models (Engelke et al., 2021; Engelke and Volgushev, 2022) and classification in extreme regions (Jalalzai et al., 2018, 2020). A large number of these methods have tuning parameters, and apart from k, their optimal selection is challenging. In the context of EVT, estimating the (conditional) generalization risk is even more critical due to the limited available training data. For such scenarios, cross-validation (CV) presents a fitting solution. Further in this section, we shall present our results regarding cross validation in extreme regions.

Recently, there have been developments from a statistical learning view on EVT, offering non-asymptotic assurances concerning the statistical errors of specific estimators or algorithms. The concentration inequalities for order statistics defined in Boucheron and Thomas (2012) are employed in Boucheron and Thomas (2015) to adaptatively select k for tail index estimation. In Goix et al. (2015), a consistent bound is derived using a Vapnik-Chervonenkis (VC) class of sets in terms of the deviations of the conditional empirical metric in a low occurrence zone $\mathbb{A}$, formulated as $P_{n,\alpha} = \frac{1}{n\alpha}\sum_{i=1}^{n} \mathbb{1}\Big\{Z_i \in (\cdot \cap \mathbb{A})\Big\}$, which proportionally scales as $\mathcal{O}(1/\sqrt{n\alpha}) = \mathcal{O}(1/\sqrt{k})$ with respect to the sample size. Alternative probability bounds with explicit constants are provided in Lhaut et al. (2021) and Clémençon et al. (2022).

In the rest of this section, we discuss challenges highlighted earlier by introducing a new method to reduce the curse of dimensionality. Our approach leverages Sliced Inverse Regression techniques to effectively manage high-dimensional data in extreme regions, simplifying and improving the analysis process. Additionally, we explore risk estimation using cross-validation in rare regions.

1.1.2 Dimensionality reduction in EVT

The fundamental objective of statistical regression is to predict a *dependent variable* $Y \in \mathbb{R}$ using a *covariate vector* $X \in \mathbb{R}^p$, which is commonly known as the *explanatory variables*. When faced with an abundance of explanatory variables, the challenge becomes reducing their dimensionality. Chapter 3 centers on the extreme values of the target variable, defined as $Y\mathbb{1}\{Y > y\}$, where "y" represents a high threshold—typically chosen based on a quantile of Y at a probability level of $1-\alpha$ (with α being exceptionally low). Here, $\mathbb{1}\{A\}$ denotes the indicator function for an event A. In this work, we propose a novel approach to reduce the dimensionality of X specifically to predict these extreme Y values.

Our work considers a conditional extremes model wherein the extreme values of Y are driven by the covariates vector X. The dimension of this vector, p, is significantly larger than the sample size. The curse of dimensionality is even more pronounced in extreme value analysis. In such analyses, only a minuscule portion of the data, represented by the small α, contributes to statistical inference. A real-world instance of this is detecting outliers when there are some given covariates. Typically defined as the extreme values of the dependent variable, outliers correspond only to a minor fraction of the total data.

The aim is to predict the tail distribution of the dependent variable using the covariates. When the dimension of the vector of covariates p is of considerable size, making such predictions intricate.

Before delving deeper, it is worth noting that, while the methodology introduced in this research is framed within the EVT perspective, it is essentially a localized technique relevant to any restricted range of Y. This approach can be suitably modified to address the challenge of reducing dimensionality when predicting Y within specific low-probability regions of various configurations. To put it differently, selecting the tail region above the $1-\alpha$ quantile is not fixed and can be interchanged with any region with probability α. Nonetheless, considering the critical relevance of risk management applications, our emphasis remains on this distinct tail region.

The topic of dimensionality reduction for extremes has garnered significant attention in recent literature. Predominantly, these works cater to the unsupervised setting, which involves examining the extremes of a multidimensional random vector. Broadly, these investigations can be categorized into: clustering techniques (Chautru, 2015; Chiapino et al., 2019a; Janßen and Wan, 2020a), methods for support identification (Goix et al., 2016, 2017; Chiapino and Sabourin, 2016; Chiapino et al., 2019b; Simpson et al., 2020; Meyer and Wintenberger, 2019), Principal Component Analysis targeting the angular aspect of extremes (Cooley and Thibaud, 2019a; Jiang et al., 2020; Drees and Sabourin, 2021), and graphical models tailored for extremes (Hitz and Evans, 2016; Engelke and Hitz, 2020; Asenova et al., 2021; Engelke et al., 2021). Further details can be found in Engelke and Ivanovs (2020) and the references therein. Contrary to the aforementioned studies, our method operates within the supervised realm, focusing on reducing the dimensionality of X to fulfill a specific objective—predicting the large values of the output variable Y.

The underlying assumption of a sufficient linear projection has historical roots in statistics, frequently discussed within the context of sufficient dimension reduction (SDR) spaces (Cook, 2009). Many traditional techniques in supervised dimension reduction, such as Principal Component Regression (Hotelling, 1957), Partial Least Squares (Wold, 1966), Canonical Correlation Analysis (Thompson, 1984), and sparse regularization methods like the Lasso (Jenatton et al., 2011), often pivot around a linear regression relationship between X and Y.

In contrast, SDR operates on the idea of linear dimension reduction: it posits that predicting the dependent variable only requires a handful of linear combinations of covariates. This means that there is a linear subspace E with a moderate dimension $d \leq p$ such that:

$$\mathbb{P}\left(Y \leq t \mid X\right) = \mathbb{P}(Y \leq t | QX), \qquad \forall t \in \mathbb{R}, \quad \text{almost surely}, \tag{1.1}$$

Here, Q symbolizes the orthogonal projection onto E. In essence, Y is dependent on X solely through QX in $\mathbb{R}^d$. This approach anchors itself in the conditional independence concept Dawid (1979); Constantinou and Dawid (2017): the aforementioned condition encapsulates the idea that Y is conditionally independent of X once QX is given. One significant advantage of this approach is its ability to strike a balance between interpretability, which is based on linear operations, and the flexibility provided by its generative model – it does not enforce any specific assumption about the relationship between QX and Y.

Assuming the existence of a significant subspace E satisfying the condition, an intuitive step is to first approximate such a space and subsequently utilize only the variable QX for predicting Y, effectively simplifying the regression dimensionality.

SDR-based estimation can also be perceived as a distinct variant of semi-parametric M-estimation (Delecroix et al., 2006). Another avenue is derivative-based methods, grounded on the idea that the gradient of the regression curve is a constituent of E (Härdle and Stoker, 1989; Hristache et al., 2001; Xia et al., 2007; Dalalyan et al., 2008). More recently, the Reproducing Kernel Hilbert Spaces (RKHS) framework has been leveraged to infer SDR spaces using covariance operators (Fukumizu et al., 2004, 2009).

The methodology most closely associated with our research stems from the inverse regression framework introduced by Li (1991). This includes the Sliced Inverse Regression (SIR) approach and its more advanced counterpart, the Sliced Average Variance Estimate (SAVE) (Cook and Weisberg, 1991). The central concept behind these techniques is that, given certain conditions, the inverse regression curve $\mathbb{E}\left[X|Y\right]$ and its higher moment counterpart — the columns of the conditional covariance matrix $\mathrm{Var}(X|Y)$ — are almost certainly part of the minimal SDR. The Cumulative slicing estimation (CUME) method, introduced in Zhu et al. (2010) and further explored in Portier (2016) through an empirical process perspective, seeks to identify the most expansive subspace of the minimal SDR. This is accomplished by estimating the conditional expectation of X (and its advanced moment version) based on 'slices' of the target variable Y, represented as $\mathbb{1}\{Y < y\}$, followed by an integration of such conditional expectations against y.

A recognized limitation of the SIR approach is its dependence on a specific "linearity condition" (LC) concerning the covariates :

$$\mathbb{E}\left[X \mid QX\right] = QX.$$

The validity of this assumption is further discussed in Hall and Li (1993). This condition is notably met when the covariates are either part of an elliptical random vector or are mutually independent (Cook, 2009; Eaton, 1986). Several enhancements of the SIR method have been proposed to address this limitation. For instance, by employing RKHS, transformations have been suggested to ensure the LC is nearly fulfilled (Wu, 2008; Yeh et al., 2008). Another alternative, especially when deviating from elliptical covariates, involves applying the SIR approach and its more advanced versions to the score functions of the predictor variables (Babichev et al., 2018). Furthermore, in situations where dimensionality (p) exceeds sample size (n), regularization techniques have been put forward to enable feature selection (Li and Yin, 2008). However, these advanced adaptations are beyond the purview of our current study. Here, we primarily focus on the foundational SIR and SAVE approaches. This suggests opportunities for further improvement in subsequent researches. For estimation, we examine a variant of the CUME technique.

Contributions Our contributions are twofold:

1. We extend the principles and methodologies of inverse regression to address the challenges presented by extreme values. In particular, we introduce the concept of *Tail Conditional Independence* which is similar to the SDR but for extreme

regions.

Definition (Tail Conditional Independence (TCI))**.** *Let Y, V, W be random variables defined on $(\Omega, \mathcal{F}, \mathbb{P})$. We assume that Y is real-valued, Borel measurable, while V and W take their values in arbitrary measure spaces. We say that Y is tail conditionally independent from V given W and write $Y \perp\!\!\!\perp_\infty V|W$, if*

$$\frac{\mathbb{E}\left|\,\mathbb{P}\left(Y > y \mid V, W\right) - \mathbb{P}\left(Y > y \mid W\right)\,\right|}{\mathbb{P}\left(Y > y\right)} \xrightarrow[y \to y^+]{} 0, \tag{1.2}$$

*where $y^+ \in \mathbb{R}_+ \cup \{\infty\}$ is the right endpoint (*i.e. *the supremum) of the support of the random variable Y.*

An extreme SDR space for the pair (X, Y) is then defined as the subspace E_e of $\mathbb{R}^p$ such that $Y \perp\!\!\!\perp_\infty X|P_e X$, where P_e is the orthogonal projection on E_e. In other words, E_e is called an extreme SDR space whenever

$$\mathbb{E}\left|\frac{\mathbb{P}(Y > y|Z) - \mathbb{P}(Y > y|P_e Z)}{\mathbb{P}\left(Y > y\right)}\right| \xrightarrow[y \to y^+]{} 0.$$

This TCI setting is a modified version of Gardes (2018)'s probabilistic setting regarding tail conditional independence. In particular, we explain in Remark 3.5 the relevance of our definition for the purpose of predicting tail events and their connections to the statistical learning framework of imbalanced classification.

The TIREX principle is similar to the SIR principle. Indeed, under suitable conditions, in particular, if the extreme conditional expectation converges to some $\ell \in \mathbb{R}^p$ *i.e.*

$$\mathbb{E}\left[Z \mid Y > y\right] \xrightarrow[y \to y^+]{} \ell,$$

then $\ell \in E_e$. This result can be seen as an extension of the standard SIR principle that states that

$$\mathbb{E}\left[Z \mid Y\right] \in E,$$

to extreme regions.

2. We conduct an asymptotic analysis of our proposed estimation strategy, TIREX, which originates from inverse regression, employing specialized tools derived from the theory of empirical processes. More formally, we use as an estimate an empirical version of the quantity $\mathbb{E}[Z \,|\, Y > y\,]$ for a high threshold y growing with the sample size n. A typical choice of such a threshold is the quantile of Y at a probability level $1 - k/n$, where $k = k(n)$ is an intermediate sequence such that $k(n) \to \infty$ and $k(n)/n \to 0$ as $n \to \infty$. Here we propose a refinement of this strategy integrating out the latter quantities over varying quantiles at probability levels $1 - uk/n$ for $u \in (0, 1)$. Such a refinement follows the proven approaches based on the CUME matrice.

   Based on the latter approach, the statistical quantity that we will seek to estimate is the following :

$$C_n(u) = \frac{n}{k}\mathbb{E}\left[Z\mathbb{1}\left\{\tilde{Y} < F^-(uk/n)\right\}\right].$$

where $\tilde{Y}$ is the negative target *i.e.* $\tilde{Y} = -Y$, F is the *c.d.f.* of $\tilde{Y}$ and F^- is the left-continuous inverse of F, $F^-(u) = \inf\{x \in \mathbb{R} : F(x) \geq u\}$.

In Section 3.5 we propose an empirical estimate of C_n and we derive using proper empirical process tools (control of classes of functions changing with n) asymptotic intervals for $\hat{C}_n - C_n$.

Similar results based on the variance matrix ($\mathbb{E}[ZZ^\top \,|\, Y > y\,]$) are also developed in the same section.

However, all these results come with the tuning of parameter d (the dimension of the central space), which is generally selected using cross-validation. In the next section, we shall study the consistency of cross-validation in extreme regions

1.1.3 Cross Validation for EVT

Cross Validation: Context and Challenges

Cross-validation (CV) represents a critical and widely adopted technique for risk estimation and model algorithm selection and is especially relevant to the central focus of this thesis. The primary concept of CV involves dividing data either once or multiple times to calculate the risk associated with different algorithms. Specifically, a portion of the data (referred to as the training sample) is allocated to train each algorithm, while the remaining data (known as the validation sample) is utilized to estimate the risk associated with each algorithm. An example of this can be seen in the 5-fold cross-validation scheme, as illustrated in Figure 1.1.
Within the broader context of risk estimation, CV stands as an improvement over training error (resubstitution error), largely due to its inherent resistance to overfitting. This robustness arises from the independence of the training and validation samples, an assumption that holds when the data is independently and identically distributed (i.i.d.). One of the major reasons for CV's wide acceptance stems from its "universality" with regard to data splitting heuristics.

In a practical setting, given a training sample $\mathcal{D}_T = \{Z_i \mid i \in T\}$, $T \subset [n]$ and a family of candidate predictors $\mathcal{G}$ the predictor g is selected by an algorithm (or learning rule) $\mathcal{A}$, so that the final predictor is given by $\mathcal{A}(T) \in \mathcal{G}$.

To define the cross-validation estimate, we recall the definition of the holdout estimate:

$$\widehat{\mathcal{R}}\Big[\mathcal{A}(T), V\Big] = \frac{1}{n_V} \sum_{i \in V} \ell\left(\mathcal{A}(T), Z_i\right),$$

Where the V is a validation set within $[n] := \big\{1, 2, \ldots, n\big\}$ with cardinal n_V and $T = [n] \setminus V$ denotes the training set. Since T and V contain independent observations, this estimate is often less optimistic than the training error (resubstitution error) $\widehat{\mathcal{R}}\left[\mathcal{A}([n]), [n]\right]$.

Given a family of validation sets in $[n]$, $V_{1:K} = (V_j)_{j=1,\ldots,K}$, the CV estimator of the generalization risk of $\mathcal{A}([n])$ is

$$\widehat{\mathcal{R}}_{\mathrm{CV}}\left[\mathcal{A}, V_{1:K}\right] = \frac{1}{K} \sum_{j=1}^{K} \widehat{\mathcal{R}}\Big[\mathcal{A}(T_j), V_j\Big], \tag{1.3}$$

where $T_j = [n] \backslash V_j$. This estimate is highly advantageous compared to a single holdout as it can significantly reduce the variance Blum et al. (1999); Kumar et al. (2013). However, akin to the holdout, it is evident that these quantities are biased estimators of the statistical risk (generalization risk) $\mathcal{R}\left[\mathcal{A}\left([n]\right)\right]$. This bias becomes pronounced when the training size $card(T)$ is reduced.

We now discuss classical cross-validation techniques: leave-one-out (*l.o.o.*), leave-p-out (*l.p.o.*), and K-fold cross-validation procedures.

Leave-one-out (Stone, 1974) The *l.o.o.* estimate, also known as the deleted estimate, is a foundational exhaustive CV procedure. In this method, each data point is iteratively excluded from the sample and utilized for validation. Specifically, *l.o.o.* is defined by (6.1) with $K = n$, $T_j = [n] \backslash j$, $V_j = j$ for $j \in \{1, 2, \ldots, n\}$.

Leave-p-out (Shao, 1993) with $p \leq n - 1$ is another exhaustive CV where each subset of p data points is sequentially excluded from the sample for validation purposes. Thus, *l.p.o.* is defined by (6.1) with $K = \binom{n}{p}$, and $(V_j)_{1 \leq j \leq B}$ are all subsets of $[n]$ with size $n - p$. Note that *l.p.o.* with $p = 1$ is *l.o.o.*.

Considering $\binom{n}{p}$ training sets can be computationally intractable, even when p is small. K-fold scheme has been proposed as an alternative.

K-fold CV (Geisser, 1975) with $K \in [n]$ was introduced as an alternative to the computationally taxing *l.p.o.*, K-fold CV involves initially dividing the data into K subsamples, each with approximately equal size n/K. Each of these subsamples then successively serves as the validation set. K-fold CV relies on a preliminary partitioning of data into K subsamples of approximately equal to the cardinality n/K . Each subsample successively plays the role of validation sample. Formally, let $A_1, A_2, \ldots, A_K$ be some partition of $[n]$ with $\forall j, card(A_j) = n/K$. Then, the K-fold CV estimator of the risk of $\mathcal{A}\left([n]\right)$ is given by (6.1) with $V_j = A_j$ and $T_j = [n] \backslash A_j$.

Due to its simplicity, K-fold CV is the most used cross-validation scheme in practice. However, because of the discrepancy between the training size and the full sample size, the latter estimate can have a pronounced bias which causes K-fold cross-validation to fail in many contexts (Shao, 1997; Yang, 2006; Arlot, 2008b; Arlot and Lerasle, 2016).

To address the limitations of K-fold, Burman (1989); Fushiki (2011) introduced debiasing correction terms to the K-fold CV estimate to improve its convergence rate. Formally, the bias-corrected CV estimates write as :

$$\widehat{\mathcal{R}}_{CV}^{corr}\left[\mathcal{A}, V_{1:K}\right] = \widehat{\mathcal{R}}_{\mathrm{CV}}\left[\mathcal{A}, V_{1:K}\right] + \widehat{\mathcal{R}}\left[\mathcal{A}([n]), [n]\right] - \frac{1}{K}\sum_{j=1}^{K} \widehat{\mathcal{R}}\left[\mathcal{A}(T_j), [n]\right]. \quad (1.4)$$

The behavior of CV, especially *l.p.o.*, in terms of risk assessment and model identification has been extensively researched from an asymptotic perspective. The outcome usually hinges on the splitting ratio, which is the proportion between the validation and training sets sizes—specifically, $p/(n-p)$ for *l.p.o.* and $1/(K-1)$ for K-fold cross-validation. This was demonstrated by Shao (1993, 1997) for regression, by Yang (2006)

for classification, by Bayle et al. (2020); Austern and Zhou (2020b) for *stable* algorithms as per (Bousquet and Elisseeff, 2002) and by van der Laan et al. (2004) in density estimation contexts. Asymptotic optimality is achieved when this ratio approaches zero as n tends to infinity. Additional asymptotic insights regarding CV in regression are detailed in the book by Gyorfi et al. (2010).

From a non-asymptotic standpoint, determining the theoretical properties of CV estimates is challenging because there is no independence between the terms of the average used in a CV scheme. Commonly used concentration inequalities typically assume independence among the terms used to create the estimator. As far as we know, the limited non-asymptotic results can be divided into two categories: On one hand, we have so-called *insanity check bounds*, indicating that the obtained guarantees show that the CV estimate has a better convergence rate than the hold-out estimate. On the other hand, there is a concept termed *sanity check bounds*. These bounds do not demonstrate that the CV risk estimate surpasses either the hold-out or the training risk estimates. However, they do confirm the consistency of using CV for risk estimation. Subsequently, we review the literature on each type of result.

Variance insanity check bounds The cross-validation estimator offers an advantage in reducing variance compared to just one training-test split. In the keystone paper by Blum et al. (1999), they established insanity check bounds and demonstrated that cross-validation consistently aids in variance reduction, even though they didn't specify the extent of this phenomenon. Subsequently, Kale et al. (2011); Kumar et al. (2013) quantified the extent of variance reduction by CV when a specific type of stable learner is used. However, the latter works do not discuss the bias induced by CV schemes, and in some instances, this bias can result in inconsistency (Shao, 1997; Arlot and Lerasle, 2016).

Sanity check bounds Let us review the existing non-asymptotic analysis that incorporates the CV bias. References such as Devroye and Wagner (1979); Anthony and Holden (1998); Kearns and Ron (1999) have established polynomial upper bounds for the *leave-one-out* (*l.o.o.*) error, assuming a specific weak stability criteria, further utilized in Cornec (2009, 2017). Moreover, Kearns and Ron (1999) (Lemma 4.2) has illustrated that ERM over a VC-class possesses error stability. It is paramount to highlight that exponential bounds for the *l.o.o.* can be formulated under the more robust assumption of *uniform stability*, as demonstrated in Bousquet and Elisseeff (2002). Regarding the K-fold scheme, literature is notably less prolific compared to the *l.o.o.*, especially when delineating upper bounds for risk estimation. To the best of our knowledge, the only non-asymptotic bounds in this domain are provided in Cornec (2009, 2017). The latter references furnish some upper bounds spanning a range of CV strategies. These specific bounds entail a minimum of either exponential or polynomial terms, each reliant on the dimensions of the validation and training datasets. Consequently, the K-fold method yields an exponential bound, given its validation set size is proportionate with the full sample size, contrasting with *l.o.o.* and *l.p.o.*.

The exponential and polynomial upper bounds mentioned earlier serve as sanity check guarantees. This isn't always the case for model selection, as highlighted in works like *e.g.* Wager (2020). Expanding beyond these sanity check limits for CV risk estimators is still an unsolved issue in the general case.

To conclude this section we recall the meaning of polynomial probability upper bound that is an upper bound of the following form :

$$P(\text{Error} > t) \leq C/t^a$$

for some $C > 0$ and $a \geq 1$. While an exponential upper bound write as :

$$P(\text{Error} > t) \leq C_1 \exp\left\{-C_2 t\right\},$$

for some $C_1, C_2 > 0$.

Contributions and Problem Formulation

This thesis is the first of its kind to envision CV for algorithms related to rare regions from a theoretical perspective. The theoretical guarantees are derived for the extreme region $\mathbb{A} = \{z = (x, y) \in \mathcal{Z} : \|x\| \geq t_\alpha\}$ for some (semi)-norm $\|\cdot\|$ and a large threshold t_α chosen as the $(1-\alpha)$-quantile of $\|X\|$ where $\alpha = k/n$ and $k = k(n) = o(n)$. However, these results can be extended easily to any low-probability region.

In such a context, it is natural to measure the performance of a predictor g in terms of an expected loss, conditional to the rare event $\|X\| \geq t_\alpha$. In other words, the quantity of interest is the statistical risk $\mathcal{R}_\alpha[g] = \mathbb{E}\left[\ell(g, Z) \mid \|X\| \geq t_\alpha\right]$.

The hold-out estimate of the statistical error $\mathcal{R}_\alpha\left[\mathcal{A}([n])\right]$ involves a validation index set V disjoint from T and writes as

$$\widehat{\mathcal{R}}_\alpha\Big[\mathcal{A}(T), V\Big] = \frac{1}{n_V \alpha} \sum_{i \in V} \ell\left(\mathcal{A}(T), Z_i\right) \mathbb{1}\Big\{\|X_i\| > \|X_{(\lfloor \alpha n \rfloor)}\|\Big\},$$

where $\left\|X_{(1)}\right\| \geq \ldots \geq \left\|X_{(n)}\right\|$ are the (reverse) order statistics of the sample $(\left\|X_i\right\|)_{i=1,\ldots,n}$.
The CV estimator of the generalization risk of $\mathcal{A}([n])$ in extreme regions is

$$\widehat{\mathcal{R}}_{\text{CV},\alpha}\left[\mathcal{A}, V_{1:K}\right] = \frac{1}{K} \sum_{j=1}^{K} \widehat{\mathcal{R}}_\alpha\Big[\mathcal{A}(T_j), V_j\Big]. \tag{1.5}$$

For clarity reasons, **we suppose further that n is divisible by K so that n/K is an integer**. This condition guarantees, that all validation sets have the same cardinal $n_V = n/K$.

It is essential to note that simply applying the same normalization to the risk and its associated existing upper bounds (Kearns and Ron, 1999; Cornec, 2009, 2017; Arlot and Lerasle, 2016) on the CV error produces an uninformative bound. Namely, dividing by α an upper bound of order $\mathcal{O}(1/\sqrt{n})$ in order to analyze the case of rare classes yields an order $\mathcal{O}(1/(\alpha\sqrt{n}))$ which may not even converge to zero, *e.g.* if $\alpha = \mathcal{O}(\frac{1}{\sqrt{n}})$. This pitfall is a distinctive feature of statistical learning in low probability regions already discussed in (Goix et al., 2015; Lhaut et al., 2021) regarding the deviations of the empirical risk. One purpose of this thesis is to derive consistent non-asymptotic upper bounds for algorithms dedicated to rare regions.

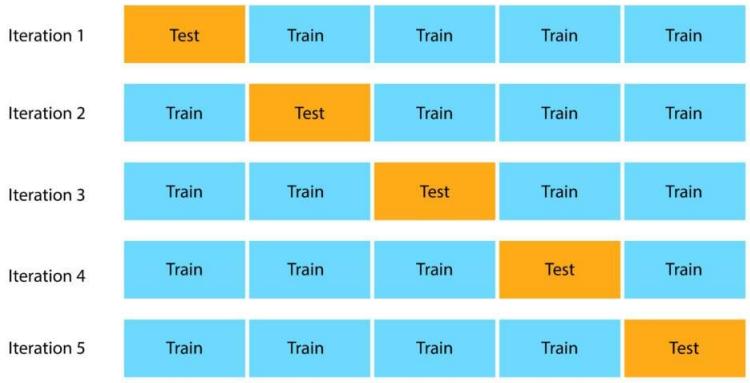


Figure 1.1 – 5-folds cross-validation scheme.

Indeed, previous works on cross-validation (Kearns and Ron, 1999; Cornec, 2017) use Hoeffding's type inequalities. However, this type of inequality doesn't take into account the low variance of $\mathbb{1}\Big\{\|X\| \geq t_\alpha\Big\}$. For the sake of completeness, we recall Hoeffding's inequality

Theorem (Hoeffding's inequality)**.** *Let $Z_1, \ldots, Z_n$ be independent random variables such that $Z_i \in \big[a_i, b_i\big]$ almost surely. For any $\epsilon > 0$, the following statements hold:*

$$P\left(\sum_{i=1}^{n} X_i - \sum_{i=1}^{n} \mathbb{E}\left[X_i\right] > \epsilon\right) \leq \exp\left(-\frac{2\epsilon^2}{\sum_i \left(b_i - a_i\right)^2}\right),$$

$$P\left(\sum_{i=1}^{n} X_i - \sum_{i=1}^{n} \mathbb{E}\left[X_i\right] < -\epsilon\right) \leq \exp\left(-\frac{2\epsilon^2}{\sum_i \left(b_i - a_i\right)^2}\right),$$

$$\textit{and} \quad P\left(\left|\sum_{i=1}^{n} X_i - \sum_{i=1}^{n} \mathbb{E}\left[X_i\right]\right| > \epsilon\right) \leq 2\exp\left(-\frac{2\epsilon^2}{\sum_i \left(b_i - a_i\right)^2}\right).$$

Applying Hoeffding's inequality to control the errors $\left|\frac{1}{n}\sum_{i=1}^{n} \ell(g, Z_i) - \mathbb{E}\left[\ell(g, Z)\right]\right|$ and $\left|\frac{1}{n}\sum_{i=1}^{n} \ell(g, Z_i)\mathbb{1}\{X_i \geq t_\alpha\} - \mathbb{E}\left[\ell(g, Z)\mathbb{1}\{X \geq t_\alpha\}\right]\right|$ yields the same probability upper bound. *i.e.* an upper bound of order $1/\sqrt{n}$. However, using Bernstein's type yields upper bounds of order $1/\sqrt{n}$ and $\sqrt{\alpha/n}$ respectively. Thus, existing works in EVT (Goix et al., 2015; Lhaut et al., 2021) use Bernstein's type inequalities to control the deviations in extreme regions. Following the line of the latter references, we will use the following Bernstein variant to control the deviations of $\widehat{\mathcal{R}}_{\mathrm{CV},\alpha}$.

Theorem. *For a sequence of observations $(Z_1, Z_2, \ldots, Z_n) \in \mathcal{Z}^n$ and some fixed values $z_{1:l} = (z_1, z_2, \ldots, z_l)$ and for some measurable function $f : \mathcal{Z}^n \to \mathbb{R}$, let $W = f(Z_1, Z_2, \ldots, Z_n)$ and define for $l \in [\![1, n]\!]$:*

1. $f_l(z_1, z_2, \ldots, z_l) = \mathbb{E}\Big(W \mid Z_1 = z_1, Z_2 = z_2, \ldots, Z_l = z_l\Big)$,

2. $\Delta_l(z_1, z_2, \ldots, z_{l-1}, z_l) = f_l(z_1, z_2, \ldots, z_{l-1}, z_l) - f_{l-1}(z_1, z_2 \ldots, z_{l-1})$, *(the positive deviations)*

3. $D := \max_{l=1,\ldots,n} \sup_{z_1,\ldots,z_{i-1} \in \mathcal{Z}} \sup_{z \in \mathcal{Z}} \Delta_l \left(z_1, \ldots, z_{l-1}, z\right)$, *(the maximum positive deviation)*

4. $\sigma_l^2(z_{1:l-1}) = \mathrm{Var}\left[\Delta_l(Z_1, Z_2, \ldots, Z_{l-1}, Z') \mid Z_1 = z_1, Z_2 = z_2, \ldots, Z_{l-1} = z_{l-1}\right]$, *where Z' is an independent copy of Z_l,*

5. $\sigma^2 = \sup_{z_{1:l-1} \in \mathcal{Z}^{l-1}} \sum_{l=1}^{n} \sigma_l^2(z_{1:l-1})$ *(the maximum sum of variances).*

Then we have

$$\mathbb{P}(W - \mathbb{E}\left[W\right] > t) \leq \exp\left(\frac{-t^2}{2(\sigma^2 + Dt/3)}\right).$$

The main difficulty compared to existing works that derive non-asymptotic bounds for EVT estimates is the lack of independence between the different terms of the average involved in a CV scheme.

We are now poised to detail our contributions concerning the cross-validation estimate $\widehat{\mathcal{R}}_{\mathrm{CV},\alpha}$.

Contribution we present two novel results for conditional risk estimation by cross-validation $\widehat{\mathcal{R}}_{\mathrm{CV},\alpha}$ when $\mathcal{A}$ is an **empirical risk minimizer** and $\alpha \ll 1$:

(i) We derive an exponential probability bound, contingent on the size of the validation set. This bound acts as a sanity check within the realm of rare events specifically for the K-fold CV scheme. It is not applicable to the *l.p.o.* scheme because, in this instance, the size of the validation set remains fixed at p. More precisely we derive an upper bound, that confirms that, by probability at least $1 - \delta$, one has

$$\left|\widehat{\mathcal{R}}_{\mathrm{CV},\alpha}\left[\mathcal{A}\left([n]\right)\right] - \mathcal{R}_\alpha\left[\mathcal{A}\left([n]\right)\right]\right| = \mathcal{O}\left(\frac{\log(1/\delta)}{\sqrt{n_V \alpha}}\right)$$

(ii) Our second result is a polynomial upper bound, which outperforms the exponential bound in the context of *l.p.o.*. This is because it is exclusively related to the size of the training set. The obtained upper bound yields, with probability at least $1 - \delta$,

$$\left|\widehat{\mathcal{R}}_{\mathrm{CV},\alpha}\left[\mathcal{A}\left([n]\right)\right] - \mathcal{R}_\alpha\left[\mathcal{A}\left([n]\right)\right]\right| = \mathcal{O}\left(\frac{1}{\delta\sqrt{n_T \alpha}}\right).$$

For cases where $\alpha = 1$, our contributions match the state-of-the-art probability upper bounds, apart from certain multiplicative constants and negligible terms. Specifically, for $\alpha = 1$, our exponential (respectively, polynomial) upper bound aligns with existing bounds in Cornec (2017) (risk-fold CV respectively, Kearns and Ron (1999); Cornec (2017)). Addressing the situation where $\alpha \ll 1$ demands distinct proof methodologies due to the low variance (influenced by α) of the random variables at stake. In our approach, we employ a Bernstein-type variant of the bounded difference inequality (McDiarmid, 1998). This mirrors earlier efforts in statistical learning pertaining to EVT, consistent with the works of Goix et al. (2015); Lhaut et al. (2021). A unique

aspect of our research lies in addressing the intricate construction of the cross-validation risk. This risk encompasses a summation of dependent terms, differentiating it from the empirical risk addressed in prior references.

The aforementioned results establish the consistency of K-fold cross-validation (CV) for empirical risk minimizers in rare regions, specifically as $\alpha \to 0$ and $\alpha n \to \infty$. The subsequent section will delve into the performance of CV in conjunction with stable learners. We will particularly concentrate on scenarios where $\alpha = 1$, setting the stage for an in-depth examination of CV applied to stable learners within the context of rare events.

1.2 Bias corrected K-fold: opening the road for analysing stable learners in rare regions

A promising avenue is to derive K-fold probability bounds for rare events in an algorithmic stability framework (see Section 1.4 below for the exact definition). However, to the best of our knowledge all the non-asymptotic guarantees in the standard setting ($\alpha = 1$) concern only the *l.o.o.* and *l.p.o.* schemes or analyses only the variance of K-fold (Kale et al., 2011; Kumar et al., 2013; Abou-Moustafa and Szepesvári, 2017; Bayle et al., 2020; Austern and Zhou, 2020a). None of the latter works imply a universal upper bound regarding the K-fold neither for risk estimation nor for model selection. Indeed their focus is on the variance term of the K-fold error, while they do not take into account the high bias generally induced by this CV scheme (see Shao (1997); Arlot and Lerasle (2016) for instance).

One may wonder whether the absence of a universal bound for the K-fold CV error in an algorithmic stability framework is just a coincidence. We answer in the negative by deriving a lower bound on the K-fold error (Section 6.4) in two different contexts, specifically, regularized empirical risk minimization and stochastic gradient optimization. The latter bound shows that under the uniform stability assumption alone, K-fold CV is inefficient in so far as it can fail in estimating the generalization risk of a uniformly stable algorithm.

Furthermore, we analyze the corrected K-fold (cf. Equation (1.4)) procedure and prove a PAC generalization upper bound covering the general case of uniformly stable learners. As a consequence, the corrected version of the K-fold is shown to be efficient in contrast to the standard version. The corrected K-fold scheme has been investigated in Burman (1989, 1990); Fushiki (2011); Arlot and Lerasle (2016) in the particular frameworks of ordinary linear regression and density estimation. Furthermore, the analysis in the latter references relies on strong regularity assumptions (further details are given in Section 6.5) which aren't satisfied by many modern learning rules like Support Vector Machine (SVM), stochastic gradient descent methods, bagging, etc. Instead our upper bound covers the general case of uniformly stable learners.

Let us briefly recall important notions of stability that will be used in the manuscript. The notion of *stability* was first introduced in Devroye and Wagner (1979) to derive non-asymptotic guarantees for the leave-one-out estimate. Let denote by $[n]$ the set of indices $\{1, \dots, n\}$. The algorithm $\mathcal{A}$ is called stable if removing a training point Z_i, $i \in [n]$, from the $\mathcal{D}_T$ or replacing Z_i with an independent observation Z' drawn from the same distribution does not alter the risk of the output. Later, Bousquet and

Elisseeff (2002) introduced the strongest notion of stability, namely *uniform stability*, an assumption used to derive probability upper bounds for the training error and the *l.o.o.* estimate Bousquet and Elisseeff (2002); Elisseeff et al. (2005); Hardt et al. (2016b); Bousquet et al. (2020); Klochkov and Zhivotovskiy (2021b). Equipped with the above notations, *uniform* stability, also called *leave-one-out* stability, can be defined as follows.

Definition 1.1. *The algorithm $\mathcal{A}$ is said to be* $\beta(n)$-uniformly *stable with respect to a loss function ℓ if, for any $i \in [n]$ and $Z \in \mathcal{Z}_T$, it holds:*

$$\left| \ell\left(\mathcal{A}(\mathcal{D}_T), Z\right) - \ell\left(\mathcal{A}(\mathcal{D}_T^{\setminus i}), Z\right) \right| \leq \beta(n).$$

In practice, uniform stability may be too restrictive since the bound above must hold for all Z, irrespective of its marginal distribution. While weaker, the following notion of stability is still enough to control the leave-one-out deviations Devroye and Wagner (1979); Bousquet and Elisseeff (2002); Elisseeff et al. (2005); Kuzborskij and Orabona (2013).

Definition 1.2. *The algorithm $\mathcal{A}$ has a* hypothesis *stability $\beta(n)$ with respect to a loss function ℓ if, for any $i \in [n]$, it holds:*

$$\left\| \ell\left(\mathcal{A}(\mathcal{D}_T), Z\right) - \ell\left(\mathcal{A}(\mathcal{D}_T^{\setminus i}), Z\right) \right\|_1 \leq \beta(n),$$

where $\left\|X\right\|_q = \left(\mathbb{E}\left[\left|X\right|^q\right]\right)^{1/q}$ *is the L_q norm of X.*

We now recall a direct analog of hypothesis stability: the *pointwise hypothesis stability*. The latter property is used to derive PAC learning bounds for the training error Bousquet and Elisseeff (2002); Elisseeff et al. (2005); Charles and Papailiopoulos (2018).

Definition 1.3. *The algorithm $\mathcal{A}$ has a* pointwise hypothesis *stability $\gamma(n)$ with respect to a loss function ℓ if, for any $i \in [n]$, it holds:*

$$\left\| \ell\left(\mathcal{A}(\mathcal{D}_T), Z_i\right) - \ell\left(\mathcal{A}(\mathcal{D}_T^{\setminus i}), Z_i\right) \right\|_1 \leq \gamma(n).$$

Note that the approach based on stability does not refer to a complexity measure like the VC dimension or the Rademacher complexity. There is no need to prove uniform convergence, and the generalization error depends directly on the stability parameter.

In this thesis, we study the K-fold within a standard stability setting, we answered the question that one cannot construct a vanishing upper bound for the latter estimate.

In other words, we construct a problem where K-fold CV fails to estimate the risk of a uniformly stable learner. Namely, we show that there exists an input space $\mathcal{Z}$, a probability distribution P, and uniformly stable algorithms (regularized empirical risk minimizers and stochastic gradient descent) such as

$$\mathbb{E}\left[\left|\widehat{\mathcal{R}}_{\mathrm{CV}}\left[\mathcal{A}\left([n]\right)\right] - \mathcal{R}\left[\mathcal{A}\left([n]\right)\right]\right|\right] \geq C,$$

for some constant $C > 0$. We also show that, contrarily to the standard K-fold, the bias-corrected K-fold (cf. Equation (1.4)) is consistent, so that for any uniformly stable learner, one has, with probability at least $1 - \delta$,

$$\left| \widehat{\mathcal{R}}_{CV}^{corr} \left[\mathcal{A}\left([n] \right) \right] - \mathcal{R}\left[\mathcal{A}\left([n] \right) \right] \right| = \mathcal{O}\left(\frac{\log(1/\delta)}{\sqrt{n}} \right).$$

This opens the door to the following intriguing future research avenues:

- What type of algorithmic stability is most suitable for analyzing rare events?
- Can we derive non-asymptotic guarantees for the bias-corrected K-fold of order $1/\sqrt{n\alpha}$ within a rare event framework? If not, can we define a new type of stability that enables the latter?

Subsequently, we will explore another type of algorithm that is particularly effective for rare events. Specifically, our focus will be on cost-sensitive learners, which are commonly utilized in imbalanced classification scenarios.

1.3 Cost sensitive learning

Real-world scenarios frequently present us with highly imbalanced datasets, where one class significantly outnumbers the others. This is especially true in fields like fraud detection, medical diagnosis, and rare event prediction. In such cases, the scarcity of data for the minority class poses a unique challenge termed "data scarcity". Training a model on such skewed datasets can lead to a model that is heavily biased towards the majority class, often overlooking the minority instances which, in many applications, are of paramount importance. This imbalance and subsequent data scarcity can lead to models that are superficially accurate but lack depth and practical utility. Addressing this issue requires specialized techniques, algorithms, and a deep understanding of the domain in which the problem resides.

To face the challenges posed by data scarcity in imbalanced classification tasks, various strategies have been proposed, among which oversampling and cost-sensitive learning are prominent. Oversampling involves artificially augmenting the minority class by replicating instances or generating synthetic data, thus equalizing the class distribution. Techniques such as the Synthetic Minority Over-sampling Technique *a.k.a* SMOTE (Chawla et al., 2002) and GAN (Mariani et al., 2018) have gained popularity for their ability to create synthetic samples that lie in the feature space of the minority class, thereby enhancing its representation. While oversampling can indeed enhance the representation of the minority class, it also inherently introduces some level of noise into the data. This is because the synthetic samples, though they lie within the feature space of existing samples, may not necessarily represent genuine instances that would occur in real-world scenarios. In essence, these are "imaginary" data points based on the existing minority samples. Furthermore, there is another dimension to this. If oversampling is applied aggressively (for example as in Figure 1.2), it can lead to the creation of synthetic samples that lie close to the decision boundary between classes. This can make the decision boundary more ambiguous, potentially leading to over-generalization and reduced model performance.

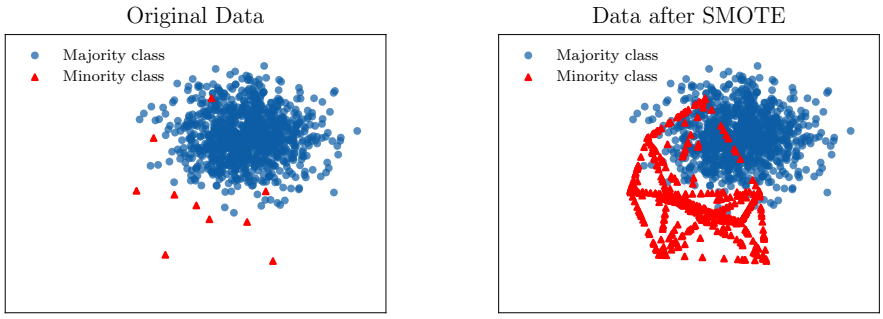


Figure 1.2 – Illustration of the impact of SMOTE oversampling on the noise in a highly imbalanced classification problem.

A simpler strategy is cost-sensitive learning (Elkan, 2001a). It modifies the learning algorithm itself by assigning different misclassification costs to different classes. This ensures that the model pays a higher penalty when misclassifying the minority class, making it more attentive to the nuances of the under-represented class. For instance, a fraudulent credit card transaction might lead to a loss of several hundred dollars, or in e-commerce, not displaying the right item to a customer might mean missing out on that item's sale revenue. Consequently, it is vital for classifiers to detect potential fraud with precision and for online stores to showcase profitable items. Nevertheless, a key challenge in cost-weighting is determining the appropriate costs. A selected cost might hinder the classifier's performance when another cost could have been more appropriate.

In this thesis, we examine the scenario of highly imbalanced **binary** classification. This setting is represented using the same notations as before. Specifically, the sample space in this context is given by $\mathcal{Z} = \mathcal{X} \times \{-1, 1\}$. The positive class, which occurs with a lower probability, is designated as the *minority* class *i.e.* $\mathbb{A} = \Big\{ Z = (X, Y) \in \mathcal{Z} \mid Y = 1 \Big\}$. It is also important to highlight that all proof techniques employed in this thesis remain valid for a **multiclassification** setting. Our focus on binary classification is primarily for notational simplicity.

1.3.1 Formulation of the Problems and Contributions

We consider **empirical risk minimizers** (ERM). The algorithm, in this context, is :

$$\begin{aligned} \mathcal{A}\Big([n]\Big) &= \arg\min_{g \in \mathcal{G}} \frac{1}{n} \sum_{i=1}^{n} \Big(c_+ \ell(g, Z) \mathbb{1}\{Y = 1\} + c_- \ell(g, Z) \mathbb{1}\{Y = -1\} \Big), \\ &:= \arg\min_{g \in \mathcal{G}} \frac{1}{n} \sum_{i=1}^{n} \ell^{\text{bal}}(g, Z). \end{aligned}$$

Where $c_+ > 0$ and $c_- > 0$ denote the costs associated to the positive and negative classes, respectively. By judiciously selecting c_+ and c_-, various imbalanced classification metrics can be covered (see *e.g.* Menon et al. (2013a); Koyejo et al. (2014)). For instance, by setting the loss to the hamming loss *i.e.* $\ell(g, Z) = \mathbb{1}\Big\{ g(X) \neq Y \Big\}$, $c_+ = 1 - \alpha$ and $c_- = \alpha$ where $\alpha = P(Z \in \mathbb{A}) = P(Y = 1)$, the latter algorithm minimizes the AM risk, a metric commonly used in imbalanced classification.

Several theoretical works shed light (Menon et al., 2013a; Koyejo et al., 2014; Xu et al., 2020a) on the consistency of cost-sensitive ERM. In other words, these papers study the excess risk of the aforementioned algorithm, given by

$$\mathcal{R}^{(\mathrm{bal})}\left[\mathcal{A}\left([n]\right)\right] - \mathcal{R}^{\star}$$

Where $\mathcal{R}^{(\mathrm{bal})}\left[\mathcal{A}\left([n]\right)\right] = \mathbb{E}\left[\ell^{(\mathrm{bal})}\left(\mathcal{A}\left([n]\right), Z\right) \mid \mathcal{D}\right]$ is the true *balanced* risk of $\mathcal{A}\left([n]\right)$. While $\mathcal{R}^{\star} = \mathcal{R}^{(\mathrm{bal})}\left[g^{\star}\right]$ and

$$g^{\star} \in \arg\min_{g \in \mathcal{G}} \mathcal{R}^{(\mathrm{bal})}[g].$$

In the latter references, they exhibit a rate of convergence of order $\mathcal{Z}\left(\frac{1}{\alpha\sqrt{n}}\right)$ in **the best case scenario**. However, in practical situations, especially in instances of *absolute rarity* or *relative rarity* (Al-Stouhi and Reddy, 2015) the probability α can often satisfy $\alpha \leq 1/\sqrt{n}$. This leads to a vacuous bound, causing the consistency results to become invalid.

A topic closely associated, yet distinct from the problem at hand, is weighted ERM. Its aim is to learn from data that has inherent biases (refer to *e.g.* Vogel et al. (2020); Bertail et al. (2021) and the sources cited within). This essentially means there is a disparity between the training and target distributions. One can perceive the imbalanced classification challenge as a specific case of this transfer learning scenario, where the training set is skewed, and the objective is a balanced counterpart with evenly distributed class weights. A critical precondition in Bertail et al. (2021) postulates that the target density, in relation to the source, must be finite. In our context, this translates to assuming that p remains at a minimum limit. This is expressly stated in Vogel et al. (2020), where key findings mandate that $p > \epsilon$ where ϵ is a predefined positive value.

The primary aim of this research is to address the aforementioned limitation, striving for generalization guarantees for the balanced risk to remain acute even when p is exceptionally small. Our goal is to determine upper limits for the discrepancies in the empirical risk (and by extension, on the empirical risk minimizer) that align with contemporary standards, substituting the sample size n with np, representing the average size of the infrequent class. As per our understanding, the theoretical findings most aligned with this objective include the normalized Vapnik-style inequalities (Theorem 1.11 in Lugosi (2002)) and relative deviations (Section 5.1 in Boucheron et al. (2005)). Nonetheless, the latter is limited to binary-valued functions and doesn't seamlessly expand to the general real-valued loss functions discussed in our study. Furthermore, they don't offer fast rates for *imbalanced* classification challenges, even though relative deviations are instrumental for fast rates in *standard* classification, as outlined in Section 5 from Boucheron et al. (2005). Also, the upper limits in the referenced materials incorporate a $\log(n)$ component, seemingly conflicting with our intention to uniformly swap n with np. Conclusively, we've yet to discover any theoretical evidence concerning imbalanced classification that employs these bounds to achieve guarantees where dominant terms are reliant on np rather than just n.

More formally, our two major contributions may be summarized as follows:

Standard learning rate We derive an error estimation probability bound for the balanced risk applicable to VC function classes. This error scales at $1/\sqrt{np}$, in contrast to the usual rate of $1/\sqrt{n}$ seen in well-balanced scenarios or the $1/(p\sqrt{n})$ observed in prior imbalanced case studies (for instance, as cited in Xu et al. (2020c)). Practically, our approach caters to scenarios where p is significantly smaller than 1 (indicating stark class imbalance). Our upper bound showcases a pivotal enhancement by a factor of $\sqrt{p}$ when set against existing imbalanced classification works. By applying this boundary to the k-nearest neighbor classification, we deduce a fresh consistency finding: if the value of kp tends to $+\infty$, then the nearest neighbor classifier remains consistent in a relative rarity scenario.

The important improvement of $\sqrt{p}$ is obtained by using a proper **uniform** concentration inequality accounting for the low variance of $\ell(g, Z)\mathbb{1}\{Y = 1\}$, more precisely we use the following result.from Plassier et al. (2023)

Theorem. *Let $(Z, Z_1, \ldots, Z_n)$ be an independent and identically distributed collection of random variables in $(S, \mathcal{S})$. Let $\mathcal{G}$ be a VC class of functions with parameters $v \geq 1$, $A \geq 1$ and uniform envelope $U \geq \sup_{g\in\mathcal{G}, x\in S} |g(x)|$. Let σ be such that $\sigma^2 \geq \sup_{g\in\mathcal{G}} var(g(Z))$ and $\sigma \leq 2U$. For any $n \geq 1$ and $\delta \in (0, 1)$, it holds, with probability $1 - \delta$,*

$$\sup_{g\in\mathcal{G}} \left| \sum_{i=1}^{n} \{g(Z_i) - \mathbb{E}[g(Z)]\} \right| \leq K' \left(\sigma \sqrt{vn \log\left(K'\theta/\delta\right)} + Uv \log\left(K'\theta/\delta\right) \right), \quad (1.6)$$

with $\theta = AU/\sigma$ and $K' > 0$ a universal constant.

The latter theorem allows to show the consistency of the balanced k-nn classifier when $kp \to \infty$ and allows to obtain the following convergence rate for empirical risk minimizers :

$$\mathcal{R}^{(\mathrm{bal})}\left[\mathcal{A}\left([n]\right)\right] - \mathcal{R}^{\star} = \mathcal{O}\left(\frac{\log(1/\delta)}{\sqrt{np}}\right).$$

Fast learning rates We provide fast rates for empirical risk reduction methods when supplemented with a traditional assumption known as the Bernstein condition. Specifically, we show a probability upper bound on the excess risk scaling at $1/(np)$. Formally, we obtain the following convergence rate, with probability at least $1 - \delta$

$$\mathcal{R}^{(\mathrm{bal})}\left[\mathcal{A}\left([n]\right)\right] - \mathcal{R}^{\star} = \mathcal{O}\left(\frac{\log(1/\delta)}{np}\right)$$

This mirrors the rapid rate results in conventional balanced scenarios but with the substitution of the complete sample size n by the number of elements in the minority class np. To the best of our understanding, such fast rates are pioneering in the domain of imbalanced classification studies.

As an example of application, we show that the Bernstein condition is verified for constrained empirical risk minimization with any strongly convex function. This is indeed not surprising since it is a well-known fact that regularization (or constrained optimization) achieves better generalization (Koren and Levy, 2015; van Erven et al.,

2015). Before concluding this section we recall the main ingredient for obtaining fast rates (Bartlett et al., 2005) :

Theorem. *Let $\mathcal{F}$ be a class of functions with ranges in $[a,b]$ and assume that there are some functional $T : \mathcal{F} \to \mathbb{R}^+$ and some constant B such that for every $f \in \mathcal{F}, \mathrm{Var}[f] \leq T(f) \leq BPf$. Let ψ be a sub-root function and let $r^\star$ be the fixed point of ψ,* i.e. *$\psi(r^\star) = r^\star$. Assume that ψ satisfies, for any $r \geq r^\star$,*

$$\psi(r) \geq BP\left(R_n\{f \in \mathcal{F} : T(f) \leq r\}\right)$$

Then, with $c_1 = 704$ and $c_2 = 26$, for any $K > 1$ and every $x > 0$, with probability at least $1 - e^{-x}$,

$$\forall f \in \mathcal{F} \quad Pf \leq \frac{K}{K-1}P_n f + \frac{c_1 K}{B}r^\star + \frac{x\left(11(b-a) + c_2 BK\right)}{n}.$$

Also, with probability at least $1 - e^{-x}$,

$$\forall f \in \mathcal{F} \quad P_n f \leq \frac{K+1}{K}Pf + \frac{c_1 K}{B}r^\star + \frac{x\left(11(b-a) + c_2 BK\right)}{n}.$$

Furthermore, if the functional T verifies $T(\alpha f) \leq \alpha^2 T(f)$ then the same inequalities hold with $c_2 = 6$ and $c_1 = 5$.

1.4 Hypothesis transfer learning

Transfer learning has emerged as a key solution to the problem of data scarcity, especially in test scenarios. It works by using knowledge from a data-rich source domain to boost learning in another related target domain. When there is not enough data in certain environments, models can use transfer learning to pull insights from tasks where data is abundant. This approach not only reduces the need for large datasets in the target domain but also helps avoid some of the computational challenges often found in traditional machine-learning methods. With transfer learning, we can navigate around the usual problems caused by limited data.

From a theoretical standpoint, supervised machine learning operates on a foundational assumption: the samples used for training and testing should come from the same probability distribution. However, in real-world applications, this assumption is frequently challenged. Sometimes, the distributions for training and testing are different, though they have some relation to one another. This specific scenario is termed as Domain Adaptation (DA). Effective DA approaches often rely on using large amounts of unlabeled data from both the original (source) and new (target) domains to adapt the learning model. Previous research has deeply delved into the methods and theory behind DA, emphasizing the importance of specific weighting parameters. But a practical challenge emerges: to estimate these parameters accurately, one needs access to a substantial amount of unlabeled data from both involved domains. This becomes especially complex and computationally demanding when multiple domains are in play or when domains keep changing or expanding. In such contexts, it is crucial to continuously gather and reassess unlabeled data across all domains.

In response to the challenges presented by DA, the methodology of hypothesis transfer learning (HTL) has been developed (Li and Bilmes, 2007; Orabona et al., 2009; Kuzborskij and Orabona, 2013; Perrot and Habrard, 2015; Kuzborskij and Orabona, 2017; Du et al., 2017). The distinctive feature of HTL is its applicability when there is limited or no direct access to the original source domain, or when the relationship between source and target domains is complex. Importantly, HTL operates without presupposing similarity between the source and target distributions, eliminating the requirement to retain extensive source data.

In this thesis, we provide an exploration of HTL. Using the framework of Regularized Empirical Risk Minimization (RERM), our analysis concentrates on binary classification. An integral component of our study pertains to the examination of several surrogate losses, pivotal in machine learning applications. Among these, the exponential loss is notably employed in methodologies such as AdaBoost (Freund and Schapire, 1997b). Furthermore, our focus encompasses logistic, soft-plus, and mean squared error (MSE) losses, among others. It is imperative to emphasize the classification calibration of these surrogate losses (Zhang, 2004a; Bartlett et al., 2006b), which posits them as convex upper bounds for classification error.

A limited number of studies offer theoretical assurances for RERM within the HTL framework, predominantly focusing on the regression context. There has been a stability analysis of the HTL algorithm in the context of RLS for regression by Kuzborskij and Orabona (2013), but it is restricted to the least-squares loss. Subsequently, Kuzborskij and Orabona (2017) explored smooth loss classes and achieved empirical risk statistical rates, which aligns with certain stability guarantees. Nonetheless, the assumption of smoothness is seen as stringent; it doesn't hold for hypotheses derived from the exponential loss or is trivially true for those from the soft plus loss. Additionally, Du et al. (2017) introduced a fresh algorithm for tailoring the source hypothesis for the target domain. However, the theoretical assurances they highlighted come with several robust assumptions, which are difficult to validate in real-world scenarios. The outlined guarantees are influenced by numerous undisclosed parameters (more details are available in Section 5.3, where these assumptions are thoroughly reviewed). There are other theoretical findings on HTL that aren't framed within RERM, as referenced in Li and Bilmes (2007); Morvant et al. (2012); Perrot and Habrard (2015); Dhouib and Redko (2018). Yet, many of these results either hinge on a measure of complexity/distance or are framed differently than classification. For instance, Perrot and Habrard (2015) delves into algorithmic stability in metric learning, using Lipschitz loss functions to examine the excess risk of certain algorithms. The bounds they derived, based on the Lipschitz constant, are not straightforward and are not readily adaptable to many standard classification losses.

In this work, we investigate the statistical risk of some transfer learning procedures dedicated to the binary classification task. To that end, we adopt the angle of algorithmic stability that offers an appealing theoretical framework to analyze such a method. This is the first work exploring algorithmic stability for HTL with the usual classification loss functions.

Our work aims to use the notion of algorithmic stability to derive sharper bounds for the HTL problem.

1.4.1 Mathematical Formulation and Contributions

Hypothesis Transfer Learning (HTL) is formally described as the act of applying a hypothesis, acquired from a source dataset, to a target domain without the requirement of raw source data or any interlink between the two domains. We denote the source domain as $\mathcal{G}_S$ and the target domain $\mathcal{G}_T$. From the target domain, we have n i.i.d. observations, where $n \in \mathbb{N}$ and $n \geq 1$. These are given as $\mathcal{D}_T = \{Z_1, \ldots, Z_n\}$ that belong to $\mathcal{G}_T$ with an underlying distribution P_T. From the source domain, the hypothesis g_S is derived from m observations $\mathcal{D}_S = Z_1^S, \ldots, Z_m^S \in \mathcal{G}_S$, where $m \in \mathbb{N}$ and $m \geq 1$. These observations are defined as under the distribution P_S. The key aspect of HTL is that we do not have access to the source domain's raw observations; only the resulting hypothesis is used. Often, n is significantly smaller than m. The focus of this paper is on binary classification. Our domains combine a source/target variable space, denoted as $\mathcal{X}_S/\mathcal{X}_T$, with the set $\{-1, 1\}$. Therefore, $\mathcal{Z}_S = \mathcal{X}_S \times \{-1, 1\}$ and $\mathcal{Z}_T = \mathcal{X}_T \times \{-1, 1\}$. The aim is to use the hypothesis g_S from the source domain $\mathcal{G}_S$ to enhance the performance of a classification algorithm on $\mathcal{Z}_T$. Specifically, the algorithm is defined as:

$$\begin{aligned} \mathcal{A} : \left(\mathcal{Z}_T\right)^n \times \mathcal{G}_S &\to \mathcal{G}_T \\ \left(\mathcal{D}_T, g_S\right) &\mapsto g_T. \end{aligned}$$

We assume that $\ell(g, Z) = \phi\Big(g(X)Y\Big)$ for some non-negative convex function ϕ and that G_T is a reproducing kernel Hilbert space (RKHS) endowed with a kernel k. This thesis analyses hypothesis transfer learning through RERM. In other words, we consider the following algorithm $\mathcal{A}$:

$$\mathcal{A}(\mathcal{D}_T, h_S) = \hat{g}(\cdot\,; \mathcal{D}_T) + g_S(\cdot), \tag{1.7}$$

where the function $\hat{g} : \mathbb{R}^d \to \mathbb{R}$ is obtained from the target set of data via the minimization problem:

$$\begin{aligned} \hat{g} &= \underset{g \in \mathcal{G}_T}{\arg\min} \frac{1}{n} \sum_{i=1}^{n} \phi\left(\Big(g\left(X_i\right) + g_S\left(X_i\right)\Big) Y_i\right) + \lambda \|g\|_k^2 \\ &= \underset{g \in \mathcal{G}_T}{\arg\min}\, \widehat{\mathcal{R}}\left[g + g_S, [n]\right] + \lambda \|g\|_k^2, \end{aligned} \tag{1.8}$$

In Chapter 5, we examine the statistical risk associated with hypothesis transfer learning (HTL) in binary classification tasks, focusing on algorithmic stability as a theoretical framework for analysis. This is the inaugural study to apply algorithmic stability to HTL with standard classification loss functions. We present a detailed hypothesis stability analysis of HTL in the classification context, applicable to any losses that meet basic conditions. We confirm that our primary assumptions hold true for widely-used classification losses and identify their specific constants. Utilizing these stability insights, we explore the statistical dynamics of the generalization gap and excess risk in HTL. We offer a clear, finite-sample analysis of these factors, emphasizing the statistical characteristics of prevalent losses. More precisely, the magnitude of the obtained bounds is directly related to the quality of g_S on the target domain $\Big($represented by $\mathcal{R}[g_S]\Big)$ instead of the complexity of the hypothesis class Ben-David et al. (2010); Zhang et al. (2012); Cortes et al. (2015); Zhang et al. (2019). In particular, we show that, the RERM HTL algorithm is hypothesis stable with stability parameter β satisfying

$$\beta \leq \Psi_\ell(\mathcal{R}\left[g_S\right])/n,$$

where Ψ_ℓ is a non decreasing function verifying $\Psi_\ell(0) = 0$ and depends solely on ℓ. Determining the function Ψ_ℓ for various loss functions ℓ provides insights into the behavior of HTL when coupled with different cost functions. We compare these loss functions in two distinct transfer scenarios: :

- Positive learning: where the source and target domains are sufficiently similar so that $\mathcal{R}[g_S] \approx 0$.
- Negative learning: where the source hypothesis might adversely affect the target task ($\mathcal{R}[g_S] \to \infty$).

1.5 Publications

The contributions introduced in this dissertation have resulted in the following publications and preprints :

- Chapter 2: **A. Aghbalou**, P. Bertail, F. Portier and A. Sabourin. Cross-validation for Extreme Value Analysis. *arXiv preprint 2202.00488*, 2022. (submitted to a peer reviewed journal)
- Chapter 3: **A. Aghbalou**, F. Portier, A. Sabourin, C. Zhou. Tail inverse regression for dimension reduction with extreme response. In *Bernoulli*, 2024.
- Chapter 4: **A. Aghbalou** and F. Portier, A.sabourin. Sharp error bounds for imbalanced classification: how many examples in the minority class? *arXiv preprint*, 2024. (submitted to a peer reviewed conference)
- Chapter 5: **A. Aghbalou** G. Staerman. Hypothesis Transfer Learning with Surrogate Classification Losses: Generalization Bounds through Algorithmic Stability. In *International Conference on Machine Learning (ICML)*, pages 280-303, 2023.
- Chapter 6: **A. Aghbalou**, F. Portier, A. Sabourin. On the bias of K-fold cross validation with stable learners. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2023.

Chapter 2

Cross-validation for Extreme Value Analysis

Contents

2.1 Introduction

Cross-validation (CV) is a most popular statistical learning tool for estimating the generalization risk of an algorithm and for hyper-parameter or model selection. Despite its widespread usage, the performance of CV depends to a large extent on the specific task considered (risk estimation or model selection with various purposes), on the nature of the statistical problem (regression, density estimation, ...) and on the specific CV scheme (K-fold, leave-one-out, leave-p-out, ...) as discussed below in this introduction.

Extreme Value Analysis (EVA) makes no exception regarding the potential usefulness of CV, particularly in statistical procedures relying on the principle of *Empirical Risk Minimization* (ERM), equally referred to as empirical contrast minimization or M-estimation, depending on the context. This includes in particular parametric inference of multivariate tail dependence (Einmahl et al., 2012, 2018, 2016; Kiriliouk et al., 2019), where CV could be naturally envisioned for evaluating goodness-of-fit or for choosing between competing models. In the recent line of works concerned with dimension reduction in Multivariate Extremes (see the review by Engelke and Ivanovs (2021) and the references therein) the question of hyper-parameters and sparsity level selection cannot be avoided in practice. As an example, in Goix et al. (2016, 2017), the number of subcones of $\mathbb{R}^d$ supporting the limit tail measure, or equivalently the cut-off level below which the empirical mass is deemed negligible, must be chosen by the user. As noted in Remark 5 in Goix et al. (2017) this can be recast as a penalized risk minimization problem. In dimension reduction methods based on Principal Component Analysis (Cooley and Thibaud, 2019b; Jiang et al., 2020; Drees and Sabourin, 2021) the dimension reduction space minimizes an empirical reconstruction risk and the dimension of the output must again be chosen by the user. Here CV could be used to evaluate

the reconstruction error as a natural alternative to the (optimistic) empirical risk computed on the training set. Clustering approaches to dimension reduction constitute still another example of ERM frameworks that have successfully been generalized to the context of EVA (Janßen and Wan, 2020b; Jalalzai and Leluc, 2021) and where, among others, the number of clusters should be chosen.

Beside dimension reduction methods, several statistical learning algorithms incorporating Extreme Value Theory (EVT) have been proposed over the past few years motivated by essential issues such as risk management (Longin, 2000; Gkillas and Katsiampa, 2018), anomaly detection (Chiapino and Sabourin (2016); Thomas et al. (2017); Siffer et al. (2017); Vignotto and Engelke (2020); Chiapino et al. (2020), see also Suboh and Aziz (2020) and the references therein), labeling new classes in a *open set* problem (Rudd et al., 2018), adversarial simulation (Bhatia et al., 2021), extreme quantile regression based on Gradient Boosting (Velthoen et al., 2021), Regression Trees (Farkas et al., 2021) or Generalized Random Forests (Gnecco et al., 2022). Most of these approaches also come with tuning parameters, in addition to k which choice is known to be difficult. Finally, in the supervised framework of classification in extreme regions (Jalalzai et al., 2018, 2020), CV comes as a natural candidate for estimating the generalization risk of the output or, in a high dimensional setting, for feature selection.

Despite the numerous potential applications of CV listed above in a far from exhaustive manner, to our best knowledge the literature is silent about theoretical guarantees enjoyed by CV in an EVA setting. From a mathematical point of view, how to generalize existing theoretical results regarding CV in such a way that the obtained guarantees depend on the number k of extremes, not the full sample size n? Recent works (Boucheron and Thomas, 2012, 2015; Carpentier and Kim, 2015; Goix et al., 2015; Lhaut et al., 2021; Clémençon et al., 2022) focus on finite-sample controls of the deviations of the empirical measure in rare regions, with non-asymptotic upper bounds of the desired order $1/\sqrt{k}$, thus matching the typical asymptotic rates available from the Extreme Value literature (see *e.g.* De Haan and Ferreira (2006), chap. 3,4). However the theoretical properties of CV estimates are notoriously difficult to establish due to the lack of independence between the different terms of the average involved in a CV scheme.

Purpose of this work. Our goal is to open the road to a finite-sample understanding of the guarantees enjoyed by CV in algorithms dedicated to extreme values. To fix ideas, the learning problems we have in mind involve i.i.d. data $Z_i, i \leq n$ in a sample space $\mathcal{Z} \subset \mathbb{R}^d$, and a low probability region $\mathbb{A} \subset \mathcal{Z}$, typically $\mathbb{A} = \{z \in \mathcal{Z} : \|z\| > t_\alpha\}$ for some (semi)-norm $\|\cdot\|$ and a large threshold t_α chosen as the $(1-\alpha)$-quantile of $\|Z\|$ where $\alpha = k/n$ throughout this chapter. In such a context, it is natural to measure the performance of an algorithm in terms of an expected loss, *conditional* to the rare event $\|Z\| > t_\alpha$. This generic setting encompasses in particular the problem of classification in extreme regions (Jalalzai et al., 2018) which we take as our leading example, since classification by means of ERM is a particularly illustrative statistical learning task.

We thus consider the problem of estimating the generalization risk of an ERM classifier, *conditional to a rare event* (see Eq. 2.1 below). Our aim is to obtain *sanity check bounds* (Kearns and Ron, 1999; Cornec, 2009, 2017) regarding the deviations of the CV estimate, that is, bounds that are of the same order of magnitude as the ones regarding the empirical risk itself, of order $\mathcal{O}(1/\sqrt{n\alpha})$ in our case, with multiplicative constants depending on the complexity of the problem. It is worth mentioning at this point that the naive method consisting in applying the same normalization to the risk

and to the associated existing upper bounds on the CV error yields a vacuous bound. Namely, dividing by α an upper bound of order $\mathcal{O}(1/\sqrt{n})$ in order to analyse the case of rare classes yields an order $\mathcal{O}(1/(\alpha\sqrt{n})) = \mathcal{O}(\sqrt{n}/k)$ which may not even converge to zero, *e.g.* if $k = \mathcal{O}(\sqrt{n})$. This pitfall is a distinctive feature of statistical learning in low probability regions already discussed in Goix et al. (2015); Lhaut et al. (2021) regarding the deviations of the empirical risk.

In order to illustrate the significance of our findings we consider a logistic-LASSO regression algorithm trained on extremes observations $\|Z\| \geq t_\alpha$, and we propose to choose the level of the ℓ^1-norm constraint on the parameter by a standard K-fold CV procedure. Such a sparsity-inducing classification algorithm is particularly attractive with high dimensional covariates, as the curse of dimensionality is particularly problematic in EVA due to to the reduced effective sample size $k \ll n$. Our results show that K-fold permits to select a model within a finite collection in a risk consistent manner.

Related works regarding Cross-Validation. As mentioned above CV may serve as a tool for (i) risk estimation, (ii) model selection. Sharp guarantees regarding the former task are typically needed as an intermediate step in order to derive guarantees for the latter task, as discussed *e.g.* in van der Vaart et al. (2006). The model selection task itself may be envisioned from the perspective of $(ii-a)$ estimation, where the goal is to minimize the risk attached to the final output, and $(ii-b)$ model identification, where the goal is to select the 'smallest' possible 'true' model. In the present work we consider mainly task (i). As a by-product, our results allow to derive minimal guarantees regarding task $(ii-a)$. For an in depth review of CV for model selection and risk estimation we refer the reader to Arlot and Celisse (2010) and the reference therein or Wager (2020); Bates et al. (2021) for recent discussions.

A popular working assumption in the statistical learning literature is *algorithmic stability* (Rogers and Wagner (1978); Devroye and Wagner (1979); Anthony and Holden (1998); Kearns and Ron (1999); Bousquet and Elisseeff (2002)). In this chapter we adopt instead the framework of ERM over a class of predictors with finite VC dimension in order to stay close to existing statistical learning viewpoints on EVT mentioned above, leaving the question of stable algorithms to future research.

Our main concern here is risk estimation in a non asymptotic setting. In this context Devroye and Wagner (1979); Anthony and Holden (1998); Kearns and Ron (1999) show polynomial upper bounds on the *leave-one-out* (*l.o.o.*) error under various weak stability assumptions, see also Cornec (2009, 2017). In addition, Kearns and Ron (1999) (Lemma 4.2) show that ERM over a VC-class is in particular error stable. Notice that exponential bounds for the *l.o.o.* can be derived under the stronger assumption of *uniform stability* as *e.g.* in Bousquet and Elisseeff (2002). In view of these facts it is reasonable to expect no more than a polynomial upper bound in our ERM framework on extreme regions for the *l.o.o.* and the *leave-p-out* (*l.p.o.*) without further assumptions. Concerning the K-fold scheme the literature is scarcer than for the *l.o.o.* regarding upper bouds for risk estimation. To our best knowledge the only existing non asymptotic bounds in this respect are derived in Cornec (2009, 2017). In the latter reference an upper bound is obtained for a wide range of CV schemes. The bound incorporates a minimum between an exponential and a polynomial terms, involving respectively the size of the validation and the training sets. This yields an exponential bound for the K-fold since in the latter scheme the validation size is of the same order as the full sample size, contrarily to the *l.o.o.* and *l.p.o.* . Both the exponential and polynomial

upper bounds in the above references are *sanity check* guarantees, which do not prove that the CV risk estimate outperforms neither the hold-out nor the training risk estimate. However they prove that CV is a consistent approach for risk estimation, and as a by-product, for model selection with an estimation purpose (task $(ii-a)$). This is not necessarily the case for model selection with an identification purpose as discussed in several papers such as *e.g.* Wager (2020).

Going beyond sanity check bounds for CV risk estimators remains an open challenge in the general case. One natural question to ask is whether using several training/testing folds improves upon the hold-out method (a single split) or upon using the empirical risk on the training set itself. Although the dependence between the different folds complicates the analysis, partial answers have been brought in various specific settings such as density estimation (Arlot, 2008b; Arlot and Lerasle, 2016) or LASSO regression (Homrighausen and McDonald, 2013; Xu et al., 2020a). Restricting the analysis to the variance of the estimator, Blum et al. (1999) prove that the K-fold reduces the variance and the amount of the latter is quantified in Kale et al. (2011); Kumar et al. (2013) under stability assumptions. Such improved guarantees in the context of rare events are left to future research.

Contributions. We provide three new results for CV-based risk estimation and model selection in a rare region of probability $\alpha \ll 1$:

(i) An exponential probability bound involving the size of the validation set, which yields a sanity check bound in the context of rare events for the K-fold CV scheme but not the *l.p.o.* scheme as the size of the validation set in this case remains constant, equal to p.

(ii) A polynomial upper bound, which outperforms the exponential one in the case of the *l.p.o.* because it only involves the size of the training set.
(iii) For the sake of illustration, we apply our exponential upper bound to the purpose of model selection within a finite family of models in logistic-LASSO regression. In particular we obtain an upper bound on the excess risk scaling as $\mathcal{O}(1/\sqrt{n\alpha})$ w.r.t the sample size with a multiplicative factor depending logarithmically on the number of candidate models.

Both our contributions *(i)* and *(ii)* achieve state-of-the-art guarantees for $\alpha = 1$, up to multiplicative constants and negligible terms. More precisely for $\alpha = 1$ our exponential (*resp.* polynomial) upper bound is of the same nature as the ones in Cornec (2017) (*resp.* Kearns and Ron (1999); Cornec (2017)). However covering the case $\alpha \ll 1$ requires different proof techniques accounting for the low variance (driven by α) of the random variables at stake. In particular we use a Bernstein-type version of the bounded difference inequality due to McDiarmid (1998), following in the footsteps of previous statistical learning works devoted to EVT mentioned above in the spirit of Goix et al. (2015); Lhaut et al. (2021). A distinctive challenge in the present work though is the complicated nature of the variable of interest, that is the cross-validation risk which involves a sum of dependent terms differently from the empirical risk studied in the latter references.

Outline. The statistical framework envisioned in this chapter is introduced in Section 2.2. Our main results theorems 2.9 and 2.12 are presented respectively in Section 2.3 and Section 2.4. Guarantees regarding K-fold for logistic-LASSO regression are derived in Section 2.5. We illustrate the tightness of our bounds in numerical experiments reported in Section 2.6. The supplementary material includes generic statistical tools used in our proofs (Section 2.A), as well as intermediate technical results and detailed proofs of our main results (Section 2.B).

2.2 Extreme values, extreme risk and cross-validation: framework

2.2.1 Conditional risk in an extreme region

Consider a random element O valued in a sample space $\mathcal{Z}$ and a low probability region $\mathbb{A} \subset \mathcal{Z}$ such that $\mathbb{P}\left(Z \in \mathbb{A}\right) = \alpha$ with $0 < \alpha \ll 1$. The probabilistic behavior of Z given that $Z \in \mathbb{A}$ is a main concern in Extreme Value Analysis. Conditioning upon $Z \in \mathbb{A}$, or alternatively rescaling the probability distribution by an appropriate sequence, is a central idea in the asymptotic EVT literature related to the tail empirical processes, see *e.g.* the review paper from Einmahl (1992) or the recent work from Bobbia et al. (2021) and the references therein in relation with local empirical processes. Uniform controls of the deviations of the empirical measure based on an i.i.d. sample $Z_i, i \leq n$ over such a region and, by extension, deviations of an empirical risk conditional to $Z \in \mathbb{A}$, have been analyzed in a non asymptotic setting in several works over the past few years, in various statistical contexts, such as empirical estimation of the stable tails dependence function (Goix et al., 2015) and of the angular measure (Clémençon et al., 2022), classification in extreme regions (Jalalzai et al., 2018), construction of Mass-Volume sets for anomaly detection (Thomas et al., 2017), support recovery (Goix et al., 2017), principal component analysis (Drees and Sabourin, 2021), graphical models (Engelke et al., 2021; Engelke and Volgushev, 2022). Recently Lhaut et al. (2021) compute universal constants involved in the upper bounds. They also discuss various conditioning and combinatoric arguments and concentration tools for such non-asymptotic control.

Here we take as a leading example the problem of ERM classification in extreme regions, following in the footsteps of Jalalzai et al. (2018); Clémençon et al. (2022). In such a context the sample space is $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ with $\mathcal{X} \subset \mathbb{R}^d$ and $\mathcal{Y} = \{-1, +1\}$. The low probability region of interest is then $\mathbb{A} = \{(x, y) \in \mathcal{Z} : \|x\| \geq t_\alpha\}$ where t_α is the $1 - \alpha$ quantile of a norm $\|\cdot\|$ on $\mathcal{X}$. Notice that in our setting, the probability α is known (chosen by the user) whereas the threshold t_α – thus also $\mathbb{A}$ – is unknown because the law of X is unknown. Given a class $\mathcal{G}$ of discrimination rules $g : \mathcal{X} \to \mathbb{R}$ and a loss function $c : \mathcal{G} \times \mathcal{Z} \to \mathbb{R}$, the conditional risk of $g \in \mathcal{G}$ over the rare region $\mathbb{A}$ is

$$\mathcal{R}_\alpha(g) = \mathbb{E}\left[c(g, Z) \mid Z \in \mathbb{A}\right]. \tag{2.1}$$

Remark 2.1 (Relevance of $\mathcal{R}_\alpha$ for Extreme Value Analysis, existing works)**.** *As shown in Jalalzai et al. (2018) for the* $0-1$ *loss associated with binary classifiers, under appropriate regular variation assumptions regarding the class distributions* $\mathcal{L}(X|Y = \sigma 1), \sigma \in \{+, -\}$*, the conditional risk* $\mathcal{R}_\alpha$ *of an* angular *classifier of the kind* $g(x) = \tilde{g}(\theta(x))$*, with* $\theta(x) = \|x\|^{-1}x$*, converges as* $\alpha \to 0$ *to an asymptotic risk* $\mathcal{R}_\infty$*. The latter is the out-of-sample risk of* g *in the extreme region.*

In practice the quantity of interest is not the risk of a fixed discrimination function g, but the risk of the specific $\hat{g}$ issued by an algorithm, also called *learning rule*, given training data $\mathcal{D}_n = (Z_1, \ldots, Z_n)$. Here and throughout, we assume that $\mathcal{D}_n$ is a collection of independent and identically distributed random vectors with common distribution P and $S_n = \{1, , \ldots n\}$ refers to the full index set. Formally, a learning rule can be viewed as a function $\Psi : \sqcup_{m \leq n} \mathcal{S}_m \to \mathcal{G}$, where $\mathcal{S}_m$ is the family of subsets of indices in $\{1, \ldots, n\}$, and $\Psi(S) = \hat{g}$ is the output of the rule trained on $\{Z_i, i \in S\}$. The quantity that we would like to estimate is the generalization risk of the learning rule trained on $\mathcal{D}_n$, $\mathcal{R}_\alpha\Big(\Psi(S_n)\Big)$.

Given a subsample $S \subset \{1, \ldots, n\}$, for some fixed g, an empirical version of $\mathcal{R}_\alpha(g)$ based on S is

$$\widehat{\mathcal{R}}_\alpha(g, S) = \frac{1}{\alpha n_S} \sum_{i \in S} c(g, Z_i) \mathbb{1}\Big\{ \|X_i\| > \|X_{(\lfloor \alpha n \rfloor)}\| \Big\}, \tag{2.2}$$

where $n_S = card(S)$ and $\Big\|X_{(1)}\Big\| \geq \ldots \geq \Big\|X_{(n)}\Big\|$ are the (reverse) order statistics of the sample $\big(\big\|X_i\big\|\big)_{i=1,\ldots,n}$.

Remark 2.2 (Threshold choice)**.** *The random threshold $\|X_{(\lfloor \alpha n \rfloor)}\|$ used for selecting extreme observations in the risk estimate (2.2) is defined using the full index set S_n, not the particular subsample S. An alternative strategy would be to let the random threshold depend on the particular subsample S, using* e.g. *the $\lfloor \alpha n_S \rfloor^{th}$ order statistic within S. In the present work we limit ourselves to the analysis of* CV *estimates of the risk based on the common threshold $\|X_{(\lfloor \alpha n \rfloor)}\|$ which turns out to be convenient in our proofs, see* e.g. *the argument leading to (2.27) in the Appendix. Whether it is possible to obtain similar or better guarantees for the alternative strategy based on a variable threshold remains an open question and would require in any case a substantial modification of our proof techniques.*

Equipped with a definition of an empirical risk $\widehat{\mathcal{R}}_\alpha(g, S)$ (Eq. 2.2) the *hold-out* estimator of the risk $\mathcal{R}_\alpha(\Psi(S_n))$ based on a validation set $V \subset \{1, \ldots, n\}$ and a training set $T = \{1, \ldots, n\} \backslash V$ takes the simple form $\widehat{\mathcal{R}}_\alpha(\Psi(T), V)$. The CV strategy for estimating $\mathcal{R}_\alpha(\Psi(S_n))$ consists in averaging such hold-out estimates over a family of validation sets $V_{1:K} = (V_j)_{j=1,\ldots,K}$, where $V_j \subset \{1, \ldots, n\}$. Namely the CV estimator is

$$\widehat{\mathcal{R}}_{CV,\alpha}(\Psi, V_{1:K}) = \frac{1}{K} \sum_{j=1}^{K} \widehat{\mathcal{R}}_\alpha(\Psi(T_j), V_j), \tag{2.3}$$

where $T_j = \{1, \ldots, n\} \backslash V_j$.

Remark 2.3 (Focus on the estimation error at fixed level α and bias term)**.** *The* ERM *strategy proposed in Jalalzai et al. (2018); Clémençon et al. (2022) in order to choose an appropriate classifier $\hat{g}$ regarding $\mathcal{R}_\infty$ consists in minimizing the empirical version of the subasymptotic risk $\mathcal{R}_\alpha$, $\widehat{\mathcal{R}}_\alpha(g, S_n)$, where S_n is the full index set $\{1, \ldots, n\}$. The statistical guarantees obtained in these papers concern the uniform deviations of $\widehat{\mathcal{R}}_\alpha$ and as a consequence the excess $\mathcal{R}_\alpha$-risk of the empirical risk minimizer $\hat{g}$. The bias term $\mathcal{R}_\infty - \mathcal{R}_\alpha$ is left aside from their statistical analysis. However it is shown in Clémençon et al. (2022) (Remark 3.3 and Appendix D) that for typical multivariate heavy-tailed vectors such that multivariate Cauchy random variables, the biais is of order $O(\alpha)$ and is thus negligible compared with deviation terms.*

In the present work we take a similar approach in that our main focus is on CV*-based estimation of* $\mathcal{R}_\alpha$*. We thus leave the bias term outside our scope. One benefit from this strategy is that our work does not rely on any regular variation assumptions, leaving open the possibility of applications to other contexts outside* EVA *where rare events play a major role, such as anomaly detection or imbalanced classification.*

Remark 2.4 (Binary versus continuous outputs, surrogate loss functions)**.** *In Jalalzai et al. (2018); Clémençon et al. (2022) the analysis is limited to binary classifiers* $g(x) \in \{-1, 1\}$ *and the* $0-1$ *loss* $c(g,(x,y)) = \mathbb{1}\Big\{g(x) \neq y\Big\}$*. Extending their results to more general cost functions such as convex surrogate losses would be an interesting avenue for further research, leveraging ideas summarized in the review paper from Boucheron et al. (2005), Section 4, and the references therein, in particular Zhang (2004b). This would permit to cover the case of computationally realistic algorithms such as Support Vector Machines or logistic regression. In the present work we take a step towards this end and consider general real valued discrimination functions* g *with a bounded loss function* c*, as made precise in our working assumptions* $1-4$ *listed in Section 2.2.3. In practice, in our illustrative example developed in Section 2.2.2 and Section 2.5, we consider a constrained logistic regression problem of LASSO type. We analyse the deviations of the* CV *estimate with respect to the (constrained) logistic expected loss itself. However we do not relate the latter convex surrogate loss with the deviations of the* $0-1$ *error, a task which could be the subject of further work.*

2.2.2 Motivating example: high-dimensional classification with Logistic-LASSO loss

As a motivating example for our work, consider the typical problem of hyper-parameter selection for high-dimensional classification. When the dimension d of the covariate variable X is large, a well documented way to reduce the dimension of the predictor is to add a non differentiable penalty term to a (convex) ERM problem, or equivalently to solve a convex minimization problem under sparsity inducing constraints. We thus consider a LASSO-type logistic regression problem with discrimination functions g_β indexed by a d-dimensional parameter β. The findings of Jalalzai et al. (2018) suggest restricting the attention to angular discrimination functions. In this context $g_\beta(x) = \beta^\top \theta(x)$. Recall that $\theta(x) = \|x\|^{-1}x$ for some norm $\|\cdot\|$. In our experiments we shall choose the sup-norm. The logistic loss function is then, for $z = (x,y) \in \mathbb{R}^d \times \{-1,1\}$ as

$$\begin{aligned} \widehat{\beta}_t = \Psi_{\alpha,t}(S) &= \underset{g \in \mathcal{G}_t}{\arg\min}\, \widehat{\mathcal{R}}_\alpha(g,S), \\ \text{where } \mathcal{G}_t &= \{g_\beta, \beta \in \mathbb{R}^d, \|\beta\|_1 \leq t\} \\ \text{and } \widehat{\mathcal{R}}_\alpha(g_\beta,S) &= \frac{1}{\alpha n_S} \sum_{i \in S} \log\left(1 + \exp\left(-\beta^\top \theta(X_i) Y_i\right)\right) \mathbb{1}\Big\{\|X_i\| > \|X_{(\lfloor \alpha n \rfloor)}\|\Big\}. \end{aligned} \tag{2.4}$$

The learning rule $\Psi_{\alpha,t}$ thus defined is a particular instance of the general setting that we consider under Assumptions 1– 4 presented in the next subsection.

Remark 2.5 (Extensions)**.** *The logistic loss and the* ℓ_1 *constraint (or penalty) are one of many pairs (convex surrogate loss - penalty) commonly considered in statistical learning. As an example, soft margin Support Vector Machines rely on the pair (hinge loss+* ℓ_2

norm). We only consider in Section 2.5 the particular example of logistic regression under ℓ_1 constraint, for the sake of concreteness and simplicity.

2.2.3 Working assumptions

Our main results hold under Assumptions 1 to 4 introduced below.

Assumption 1 (ERM algorithm)**.** *The learning rule denoted by Ψ_α, is an empirical conditional risk minimizer for the probability level α,*

$$\Psi_\alpha(S) = \underset{g\in\mathcal{G}}{\arg\min}\, \widehat{\mathcal{R}}_\alpha(g, S). \tag{2.5}$$

For clarity reasons, we suppose further that n is divisible by K so that n/K is an integer. This condition guarantees, in the case of K-fold cross validation, that all validation sets have the same cardinal $n_V = n/K$. We also need the sequence of validation/training sets to satisfy a certain balance condition which is expressed below.

Assumption 2 (CV scheme balance condition)**.** *The sequence of validation sets $V_1, V_2, \dots V_K$ satisfies*

$$card(V_j) = n_V \quad \forall j \in [\![1, K]\!], \tag{2.6}$$

for some $n_V \in [\![1, n]\!]$. Moreover it holds that

$$\frac{1}{K}\sum_{j=1}^{K} \mathbb{1}\Big\{l \in V_j\Big\} = \frac{n_V}{n} \quad \forall l \in [\![1, n]\!]. \tag{2.7}$$

The next lemma ensures that Assumption 2 holds true for the standard CV procedures (K-fold and *l.p.o.*) and that an identity similar to (6.4) is also valid for the training sets T_j. The proof is provided in Appendix 2.B.1.

Lemma 2.6. *If K divides n, for the leave-one-out, the leave-p-out, and the K-fold procedures, the validation sets $V_{1:K}$ satisfy Assumption 2. Also the the training sets $T_{1:K}$ satisfy*

$$\frac{1}{K}\sum_{j=1}^{K} \frac{\mathbb{1}\Big\{l \in T_j\Big\}}{n_T} = \frac{1}{n} \quad \forall l \in [\![1, n]\!].$$

Remark 2.7. *The condition that K divides n is required for the K-fold* CV *only, in order to ensure that $card(V_j) = n_V$ for all j. However straightforward extensions of our results can be obtained in the case where K does not divide n at the price of some notational complexity.*

We now introduce two assumptions relative to the function class $\mathcal{G}$ and the cost function c. They shall be useful to control the fluctuation of the underlying empirical process. First we require the following standard complexity restriction on the family of functions $(x, y) \mapsto c(g(x), y)$ when g lies in $\mathcal{G}$.

Assumption 3 (finite VC dimension)**.** *The family $\mathcal{G}$ of classifiers and the cost function c are such that the class of functions $z \mapsto c(g, z) = c(g(x), y)$ on $\mathcal{Z}$ has a finite VC-dimension $\mathcal{V}_\mathcal{G}$,* i.e. *the family of subgraphs $\Big\{\{(x, y, t) : t < c(g(x), y)\} : t \in \mathbb{R}, (x, y) \in \mathcal{Z}, g \in \mathcal{G}\Big\}$ has Vapnik dimension $\mathcal{V}_\mathcal{G}$.*

This complexity assumption mainly allows us to use a uniform concentration inequality from Giné and Guillou (2001), which requires that the covering number for the L_2 norm of this family of functions decrease polynomially (with exponent $\mathcal{V}_{\mathcal{G}}$) (see their condition (2.1)). We may thus as well assume the latter weaker condition, which is sometimes easier to check in practice, instead of Assumption 3, without altering our results.

For simplicity we limit ourselves to a cost function bounded by 1. Our result may be extended to any bounded cost function at the price of a multiplicative scaling factor.

Assumption 4 (Normalized cost function)**.** *The cost function c is non-negative and bounded by* 1*,*

$$0 \leq c(g, Z) \leq 1 \qquad \forall (g, Z) \in \mathcal{G} \times \mathcal{Z}.$$

This hypothesis is clearly satisfied for the Hamming loss $c(g, Z) = \mathbb{1}\Big\{g(X) \neq Y\Big\}$.

Remark 2.8 (Boundedness assumption)**.** *The logistic loss considered in our motivating example is not bounded in general.*

However, in the context of classification in extreme regions, we consider angular classifiers, with a ℓ_1 constraint on the parameter β, which amounts to a boundedness assumption on the loss. Indeed $|\beta^\top \theta| \leq \max_{j \leq d} |\theta_j| \sum_{j \leq d} |\beta_j|$, whence, for any $t > 0$, $\sup_{\|\beta\|_1 \leq t, \|\theta\|_\infty = 1} |\beta^\top \theta| \leq t$ for any $t > 0$.

2.3 Exponential bounds for K-fold CV estimates in rare regions

Our first main result Theorem 2.9 below holds true for any CV procedure under assumptions 1 – 4. The leading term of the provided upper bound is $\mathcal{O}(\sqrt{\mathcal{V}_{\mathcal{G}}/(n_V \alpha)})$. In the case of the K-fold $1/n_V = \mathcal{O}(1/n)$. Thus,the bound of Theorem 2.9 becomes $\mathcal{O}(\sqrt{\mathcal{V}_{\mathcal{G}} \log(1/\delta)/(n\alpha)})$. The latter bound is indeed a sanity check bound as it matches (up to unknown multiplicative constants) the one relative to the empirical risk conditional to a rare event established in Jalalzai et al. (2018), Th. 2, where $k = n\alpha$.

Theorem 2.9 (Exponential CV bound for rare events)**.** *Under assumptions 1, 2, 3, 4, we have, with probability $1 - 15\delta$,*

$$\left| \widehat{\mathcal{R}}_{CV,\alpha}(\Psi_\alpha, V_{1:K}) - \mathcal{R}_\alpha\Big(\Psi_\alpha(S_n)\Big) \right| \leq E_{CV}(n_T, n_V, \alpha) + \frac{20}{3n\alpha} \log\left(\frac{1}{\delta}\right) + 20\sqrt{\frac{2}{n\alpha} \log\left(\frac{1}{\delta}\right)},$$

where

$$E_{CV}(n_T, n_V, \alpha) = M\sqrt{\mathcal{V}_{\mathcal{G}}}\left(\frac{1}{\sqrt{n_V \alpha}} + \frac{4}{\sqrt{n_T \alpha}}\right) + \frac{5}{n_T \alpha},$$

and where $M > 0$ is a universal constant.

Proof [Sketch of the proof]

Introduce the pseudo-empirical risk

$$\tilde{\mathcal{R}}_\alpha(g,S) = \frac{1}{\alpha n_S}\sum_{i\in S} c(g,Z_i)\mathbb{1}\Big\{\|X_i\| > t_\alpha\Big\}. \tag{2.8}$$

Notice that when the distribution of $\|X\|$ is unknown, $\tilde{\mathcal{R}}_\alpha$ is not observable and only $\widehat{\mathcal{R}}_\alpha$ is a genuine statistic. However $\tilde{\mathcal{R}}_\alpha$ will serve as an intermediate quantity in the proofs. Define the average pseudo-empirical risk of the family $\Big(\Psi_\alpha(T_j)\Big)_{0\le j\le K}$ by

$$\tilde{\mathcal{R}}_{CV,\alpha}(\Psi_\alpha, V_{1:K}) = \frac{1}{K}\sum_{j=1}^{K}\tilde{\mathcal{R}}_\alpha(\Psi_\alpha(T_j), V_j) \tag{2.9}$$

and the average 'true' risk by

$$\mathcal{R}_{CV,\alpha}(\Psi_\alpha, V_{1:K}) = \frac{1}{K}\sum_{j=1}^{K}\mathcal{R}_\alpha(\Psi_\alpha(T_j)). \tag{2.10}$$

Using the previous quantities, write the following decomposition

$$\left|\widehat{\mathcal{R}}_{CV,\alpha}(\Psi_\alpha, V_{1:K}) - \mathcal{R}_\alpha\Big(\Psi_\alpha(S_n)\Big)\right| \le \mathrm{D}_{\mathrm{t}_\alpha} + \mathrm{D}_{\mathrm{cv}} + \mathrm{Bias}, \tag{2.11}$$

with

$$\mathrm{D}_{\mathrm{t}_\alpha} = |\widehat{\mathcal{R}}_{CV,\alpha}(\Psi_\alpha, V_{1:K}) - \tilde{\mathcal{R}}_{CV,\alpha}(\Psi_\alpha, V_{1:K})|, \tag{2.12}$$

$$\mathrm{D}_{\mathrm{cv}} = |\tilde{\mathcal{R}}_{CV,\alpha}(\Psi_\alpha, V_{1:K}) - \mathcal{R}_{CV,\alpha}(\Psi_\alpha, V_{1:K})|, \tag{2.13}$$

$$\mathrm{Bias} = |\mathcal{R}_{CV,\alpha}(\Psi_\alpha, V_{1:K}) - \mathcal{R}_\alpha\Big(\Psi_\alpha(S_n)\Big)|. \tag{2.14}$$

The remainder of the proof (see Section 2.B.3) consists in deriving upper bounds for each terms of the error decomposition (2.11), from which the result follows.

The term $\mathrm{D}_{\mathrm{t}_\alpha}$ measures the deviation between the cross validation estimator when using the order statistics and the cross validation estimator when using the 'true' level t_α, which can be bounded using Bernstein inequality, taking advantage of the small variance of the random indicator function $\mathbb{1}\Big\{\|X\| > t_\alpha\Big\}$. The term D_{cv} measures the deviations of $\tilde{\mathcal{R}}_{CV,\alpha}(\Psi_\alpha, V_{1:K})$ from its mean. It is controlled by a uniform bound (over the class $\mathcal{G}$) on the deviations of the empirical risk evaluated on the validation sample. To do so we leverage recent arguments leading to a bound on such deviations on low probability regions (as *e.g.* in Goix et al. (2015); Jalalzai et al. (2018)). Finally the term Bias is the bias of the cross validation procedure, the control of which relies on the specific nature (ERM) of the considered learning algorithm. Indeed in this context the bias may be upper bounded in terms of the supremum deviations of the empirical risk evaluated on the training sets T_j. ■

Theorem 2.9 can be used to obtain exponential bounds for the K-fold CV estimate. From Lemma 2.6, Assumption 2 regarding the sequence of masks $V_{1:K}$ holds true for the K-fold CV procedure. Consequently Theorem 2.9 applies with $n_V = n/K$ and $n_T = n - n_V = \frac{K-1}{K}n$. In the following corollary, $V_{1:K}^{\text{K-fold}}$ denote the sequence of validation sets associated to K-fold.

Corollary 2.10. *Under the assumptions of Theorem 2.9, the K-fold* CV *estimate (with $K \geq 2$) for the conditional risk (2.1) satisfies with probability $1 - 15\delta$,*

$$\left| \widehat{\mathcal{R}}_{CV,\alpha}(\Psi_\alpha, V_{1:K}^{K\text{-}fold}) - \mathcal{R}_\alpha\Big(\Psi_\alpha(S_n)\Big) \right| \leq E_{K\text{-}fold}(n,K,\alpha) + \frac{20}{3n\alpha}\log\left(\frac{1}{\delta}\right) + 20\sqrt{\frac{2}{n\alpha}\log\left(\frac{1}{\delta}\right)},$$

with,

$$E_{K\text{-}fold}(n,K,\alpha) = 5M\sqrt{\frac{\mathcal{V}_{\mathcal{G}}K}{n\alpha}} + \frac{5K}{(K-1)n\alpha}.$$

Discussion. *As mentioned in the introduction, for $\alpha = 1$ the upper bound in Theorem 2.9 and its application to the K-fold in Corollary 2.10 are of the same nature as in Cornec (2017), Proposition 4.1, which apply to the same context as ours,* i.e. ERM *over a hypothesis class of finite VC-dimension. Covering the case $\alpha \ll 1$ requires special proof techniques with a Bernstein-type inequality due to McDiarmid (1998) and recalled in Proposition 2.17 in the supplement. Doing so improves by a factor $\sqrt{\alpha}$ over the naive method consisting in diving both sides of the existing bounds by α. As discussed in the introduction this naive method yields a potentially diverging bound as $\alpha = \alpha(n) = \frac{k}{n} \to 0$. Also the organization of our proof is different, in particular a key simplifying step is the balance condition of the* CV *schemes which applies to the K-fold (Lemma 2.6), a fact which (to our best knowledge) is not mentioned in the existing literature. Finally, though Kumar et al. (2013) quantify the amount of variance reduction brought by the K-fold, the bias term is left outside the analysis in this reference. We haven't found any comparable finite sample upper bound in the literature devoted to stable algorithms, a natural question to ask since* ERM *over a VC class is error stable (Kearns and Ron, 1999).*

Remark 2.11 (On the universal constants)**.** *A drawback of our results in the present section and in the following one (Section 2.4) is the presence of universal constants in our upper bounds. These unknown constants derive from our control of the Rademacher averages using Giné and Guillou (2001), who themselves resort to chaining arguments. This is a standard issue in statistical learning. In most cases these constants may be replaced with additional logarithmic terms with respect to the sample size or may be computed explicitly. For the empirical training risk in low probability regions these improvements are respectively achieved in Lhaut et al. (2021) and in Clémençon et al. (2022), Theorem A.1. We leave this question for further research regarding the* CV *risk.*

Despite the satisfactory sanity check bound obtained thus far for the K-fold (Corollary 2.10), note that the term $\mathcal{O}\left(\sqrt{\mathcal{V}_{\mathcal{G}}/(n_V\alpha)}\right)$ in the upper bound of Theorem 2.9 does not even converge to 0 in the *l.p.o.* setting because the size n_V of the validation set remains constant, equal to p. Thus, Theorem 2.9 is not adapted to the latter type of CV schemes. In the next section we obtain (Theorem 2.12) an alternative upper bound involving only the size n_T of the training set which allows to cover the *l.p.o.* case.

2.4 Polynomial bounds for *l.p.o.* CV estimates in rare regions

Theorem 2.9 provides trivial bounds for CV schemes with small test size. In contrast our second main result (Theorem 2.12 below)

yields a sanity-check bound for a wider class of CV procedures, including leave-one-out and leave-p-out. In particular, we show that, with high probability, the error is at most $\mathcal{O}(\sqrt{\mathcal{V}_{\mathcal{G}}/(n_T\alpha)})$. Most –if not all– CV procedures satisfy $1/n_T = \mathcal{O}(1/n)$ and the latter bound is thus of order $\mathcal{O}(\sqrt{\mathcal{V}_{\mathcal{G}}/(n\alpha)})$.

Theorem 2.12 (Polynomial cross-validation bounds for rare events)**.** *Under assumptions 1, 2, 3, 4 one has with probability* $1 - 17\delta$*,*

$$\left|\widehat{\mathcal{R}}_{CV,\alpha}(\Psi_\alpha, V_{1:K}) - \mathcal{R}_\alpha\Big(\Psi_\alpha(S_n)\Big)\right| \leq E'_{CV}(n_T, \alpha) + \frac{1}{\delta\sqrt{n_T\alpha}}(5M\sqrt{\mathcal{V}_{\mathcal{G}}} + M_5),$$

where $M, M_5 > 0$ *are universal constants,* M *is the same as in Theorem 2.9 and*

$$E'_{CV}(n_T, \alpha) = \frac{9M\sqrt{\mathcal{V}_{\mathcal{G}}}}{\sqrt{\alpha n_T}} + \frac{9}{n_T\alpha}.$$

Proof [Sketch of the proof] First write

$$\left|\widehat{\mathcal{R}}_{CV,\alpha}\ (\Psi_\alpha, V_{1:K}) - \mathcal{R}_\alpha\Big(\Psi_\alpha(S_n)\Big)\right| \leq \text{Bias} + |\widehat{\mathcal{R}}_{CV,\alpha}(\Psi_\alpha, V_{1:K}) - \mathcal{R}_{CV,\alpha}(\Psi_\alpha, V_{1:K})|\,, \tag{2.15}$$

where Bias is defined by (2.14).

The upper bound for the term Bias obtained in the proof of Theorem 2.9 is of order $O(1/\sqrt{n_T\alpha})$, see (2.37) in the supplement for details. Since $1/n_T = O(1/n)$ in the CV schemes that we consider, the latter bound is sufficient to obtain a sanity check bound. However, in that proof, the term $|\widehat{\mathcal{R}}_{CV,\alpha}(\Psi_\alpha, V_{1:K}) - \mathcal{R}_{CV,\alpha}(\Psi_\alpha, V_{1:K})|$ is upper bounded by the sum $\mathrm{D}_{t_\alpha} + \mathrm{D}_{\mathrm{cv}}$ defined in (2.12) and (2.13). The probability upper bound for the latter term involves a term of order $\mathcal{O}(1/\sqrt{n_V\alpha})$, see (2.28) in the supplement, which is not satisfactory for small n_V. Therefore one needs an alternative control for $|\widehat{\mathcal{R}}_{CV,\alpha}(\Psi_\alpha, V_{1:K}) - \mathcal{R}_{CV,\alpha}(\Psi_\alpha, V_{1:K})|$. The main ingredient to proceed is the following Markov-type inequality

$$\mathbb{P}(\widehat{\mathcal{R}}_{\mathrm{CV},\alpha}(\Psi_\alpha, V_{1:K}) - \mathcal{R}_{CV,\alpha}(\Psi_\alpha, V_{1:K}) \geq t) \leq \frac{\mathbb{E}\Big(|\hat{\mathcal{R}}_\alpha\Big(\Psi_\alpha(S_n), S_n\Big) - \mathcal{R}_\alpha(\Psi_\alpha(S_n))|\Big)}{t} + \frac{\mathbb{E}(\mathrm{D}_{t_\alpha} + \text{Bias})}{t}, \tag{2.16}$$

which holds true under the stipulated assumptions. The proof is deferred to the supplement (Lemma 2.24).

It is shown in Section 2.B.5 from the supplement that $\mathbb{E}(\text{Bias})$ and $\mathbb{E}(\mathrm{D}_{t_\alpha})$ are both upper bounded by $O(1/\sqrt{n_T\alpha})$ (inequalities (2.44,2.45)). In addition the probability upper bound on the supremum deviations on the rare region (Lemma 2.22 also used in the proof of Theorem 2.9) shows that the latter quantity is sub-Gaussian, which yields (Vershynin

(2018), Proposition 2.5.2) an upper bound for $\mathbb{E}\Big(|\hat{\mathcal{R}}_\alpha(\Psi_\alpha(S_n), S_n) - \mathcal{R}_\alpha(\Psi_\alpha(S_n)|\Big)$ of the same order of magnitude as the other terms in the r.h.s. of (2.16).

The final step of the proof is to derive a probability upper bound for the opposite of the l.h.s. of (2.16), that is $\mathcal{R}_{CV,\alpha}(\Psi_\alpha, V_{1:K}) - \widehat{\mathcal{R}}_{\mathrm{CV},\alpha}(\Psi_\alpha, V_{1:K})$. We use the fact (proved in Lemma 2.23) that the CV risk estimate $\widehat{\mathcal{R}}_{\mathrm{CV},\alpha}(\Psi_\alpha, V_{1:k})$ is always larger than the empirical risk $\widehat{\mathcal{R}}_\alpha$ evaluated on its minimizer $\Psi_\alpha(S_n)$, thus

$$\begin{aligned}\mathcal{R}_{CV,\alpha}(\Psi_\alpha, V_{1:K}) - \widehat{\mathcal{R}}_{\mathrm{CV},\alpha}(\Psi_\alpha, V_{1:K}) &\leq \mathcal{R}_{CV,\alpha}(\Psi_\alpha, V_{1:K}) - \widehat{\mathcal{R}}_\alpha(\Psi_\alpha(S_n), S_n) \\ &\leq \mathrm{Bias} + |\mathcal{R}_\alpha(\Psi(S_n)) - \widehat{\mathcal{R}}_\alpha(\Psi_\alpha(S_n), S_n)|, \quad (2.17)\end{aligned}$$

where the last inequality follows from the definition of Bias in (2.14) and the triangle inequality. From the proof of Theorem 2.9, Bias admits a probability upper bound involving only n and n_T (see (2.37) and (2.34). The second term in the r.h.s. of (2.17) is less than the supremum deviations of the empirical risk $\widehat{\mathcal{R}}_\alpha$, which shares the same property (Lemma 2.22). Adding up the upper bounds for each term of the r.h.s. of (2.16) and (2.17) concludes the proof, see Section 2.B.5 in the supplement for details. ■

Using Theorem 2.12 and following the same steps as in the proof of Corollary 2.10, we obtain a sanity-check guarantee regarding leave-p-out estimates.

Corollary 2.13 (leave-p-out sanity check for rare events)**.** *Under the assumptions of Theorem 2.12, the* l.p.o. CV *estimate for the conditional risk (2.1) satisfies with probability* $1 - 17\delta$*,*

$$|\widehat{\mathcal{R}}_{\mathrm{CV},\alpha}(\Psi_\alpha, V_{1:K}^{\mathrm{lpo}}) - \mathcal{R}_\alpha\Big(\Psi_\alpha(S_n)\Big)| \leq E_{lpo}(n,p,\alpha) + \frac{1}{\delta\sqrt{(n-p)\alpha}}(5M\sqrt{\mathcal{V}_\mathcal{G}} + M_5),$$

with

$$E_{lpo}(n,p,\alpha) = 9M\sqrt{\frac{\mathcal{V}_\mathcal{G}}{(n-p)\alpha}} + \frac{9}{(n-p)\alpha}\,.$$

Discussion. *As it is the case in Section 2.3, our polynomial bounds from Theorem 2.12 and Corollary 2.13 are of the same nature as the state-of-the art for* $\alpha = 1$*, that is Kearns and Ron (1999), Theorem 4.2, and Cornec (2017), Proposition 4.3. Again we improve by a factor* $\sqrt{\alpha}$ *upon the naive method dividing existing bounds by a factor* α*. Concerning the presence of unknown constants, see Remark 2.11. In addition to covering the case of rare events, our results extend those of the latter reference in several directions, namely they encompass the* l.p.o. *scheme whereas Kearns and Ron (1999) only consider the* l.o.o.*, and they apply to any bounded cost function, not only the Hamming loss. Also the organisation of our proof is different, for example the risk decomposition (2.15) is new.*

Remark 2.14 (Tightness of the polynomial bound)**.** *A natural question to ask is whether or not the polynomial rate (*w.r.t. *the probability* δ*) is tight concerning the* l.p.o. CV *scheme. The answer is yes, in the* ERM *context, in the general case (that is with a classical risk function and* $\alpha = 1$*). Indeed Kearns and Ron (1999) show that, without further assumptions on the algorithm* Ψ *and the cost function* c*, the bound* $1/\delta$ *can be attained. We conjecture that the same is true for* $\alpha < 1$*, leaving this question for further work.*

Remark 2.15 (Comparison between the bounds from theorems 2.9 and 2.12)**.** *Although Theorem 2.12 also applies to the K-fold, the bound provided by Theorem 2.9 is sharper for this particular* CV *scheme for small values of δ due to its exponential nature. In other words Theorem 2.12 has a greater level of generality than Theorem 2.9 because the upper bound in the latter involved n_V, contrarily to the former. The price to pay is a slower tail decay (polynomial versus exponential).*

2.5 Application to logistic-LASSO regression

We now turn to an application of our results to high dimensional classification as introduced in Section 2.2.2. Recall that the learning rule for fixed constraint level $t > 0$ takes the form (see Eq. 2.4)

$$\Psi_{\alpha,t}(S) = \underset{\beta \in \mathcal{B}_t}{\arg\min} \frac{1}{\alpha n} \sum_{i \in S} c\left(g_\beta, (X_i, Y_i)\right) \mathbb{1}\Big\{\|X_i\| > \|X_{(\lfloor n\alpha \rfloor)}\|\Big\}.$$

Recall also the logistic loss with angular discrimination function, $c(g_\beta, (x, y)) = \log(1 + \exp(-\beta^\top \theta(x) y))$. In practice the aim of the parameter selection procedure is to choose the 'best' parameter t^* within a finite grid $\mathcal{T} \subset \mathbb{R}^+$, regarding the risk of the associated learning rule $\Psi_{\alpha,t}$, that is

$$t^* = \underset{t \in \mathcal{T}}{\arg\min}\, \mathcal{R}_\alpha\left(\Psi_{\alpha,t}\left(S_n\right)\right).$$

In view of the exponential nature of the upper bound for K-fold CV obtained in Section 2.3 compared to the polynomial bound for *l.p.o.* CV (Section 2.4) and because K-fold is computationally faster than *l.p.o.* we consider a selection procedure based on a K-fold CV estimate of $\mathcal{R}_\alpha\left(\Psi_{\alpha,t}\left(S_n\right)\right)$,

$$\hat{t} = \underset{t \in \mathcal{T}}{\arg\min}\, \widehat{\mathcal{R}}_{\mathrm{CV}}\left(\Psi_{\alpha,t}, V_{1:K}^{\text{K-fold}}\right).$$

We obtain in Lemma 2.16 an upper bound in probability for the excess risk $\mathcal{R}\left(\Psi_{\alpha,\hat{t}}(S_n)\right) - \mathcal{R}\left(\Psi_{\alpha,t^*}(S_n)\right)$. Since this upper bound converges to 0 as $\alpha \to 0$ with $\alpha n \to \infty$, our result ensures in particular the consistency of the selection procedure in extreme regions.

Lemma 2.16. *Suppose that the sample space $\mathcal{Z}$ is bounded so that Assumption 4 holds. Then, the excess risk $\mathcal{R}_\alpha(\Psi_{\alpha,\hat{t}}(S_n)) - \mathcal{R}_\alpha(\Psi_{\alpha,t^*}(S_n))$ verifies, with probability $1 - 15\delta$,*

$$\mathcal{R}_\alpha\Big(\Psi_{\alpha,\hat{t}}\left(S_n\right)\Big) - \mathcal{R}_\alpha\left(\Psi_{\alpha,t^*}\left(S_n\right)\right) \leq \max(\mathcal{T})\left[2E^{\mathcal{T}}_{\textit{K-fold}}(n, K, \alpha) + \frac{40}{3n\alpha}\log\left(\frac{|\mathcal{T}|}{\delta}\right) + 40\sqrt{\frac{2}{n\alpha}\log\left(\frac{|\mathcal{T}|}{\delta}\right)}\right],$$

with

$$E^{\mathcal{T}}_{\textit{K-fold}}(n, K, \alpha) = 5M_{\mathcal{T}}\sqrt{\frac{(d+1)K}{n\alpha}} + \frac{5K}{(K-1)n\alpha},$$

for some universal constant $M_{\mathcal{T}} > 0$ depending only on $\mathcal{T}$.

Proof By definition of $\hat{t}$, one has

$$\widehat{\mathcal{R}}_{\mathrm{CV}}\left(\Psi_{\alpha,\hat{t}}, V_{1:K}^{\text{K-fold}}\right) \leq \widehat{\mathcal{R}}_{\mathrm{CV}}\left(\Psi_{\alpha,t^*}, V_{1:K}^{\text{K-fold}}\right).$$

It follows that,

$$\begin{aligned}\mathcal{R}\left(\Psi_{\alpha,\hat{t}}\left(S_n\right)\right) - \mathcal{R}\left(\Psi_{\alpha,t^*}\left(S_n\right)\right) &\leq \mathcal{R}\left(\Psi_{\alpha,\hat{t}}\left(S_n\right)\right) - \widehat{\mathcal{R}}_{\mathrm{CV}}\left(\Psi_{\alpha,\hat{t}}, V_{1:K}^{\text{K-fold}}\right) \\ &\quad + \widehat{\mathcal{R}}_{\mathrm{CV}}\left(\Psi_{\alpha,t^*}, V_{1:K}^{\text{K-fold}}\right) - \mathcal{R}\left(\Psi_{\alpha,t^*}\left(S_n\right)\right) \\ &\leq 2\sup_{t\in\mathcal{T}}\left|\widehat{\mathcal{R}}_{\mathrm{CV}}\left(\Psi_{\alpha,t}\left(S_n\right), V_{1:K}^{\text{K-fold}}\right) - \mathcal{R}\left(\Psi_{\alpha,t}\left(S_n\right)\right)\right|. \end{aligned} \tag{2.18}$$

For fixed $t \in \mathcal{T}$, all the required assumptions of Theorem 2.9 are met, except that the cost function is not bounded by 1 but rather by t, thus also by $\max(\mathcal{T})$. As explained in Remark 2.8 our results still apply, up to multiplication of all upper bounds by a factor $\max(\mathcal{T})$. We may thus use Corollary 2.10 with $\mathcal{V}_{\mathcal{G}} = d + 1$, so that with probability $1 - 15\delta$,

$$\left|\widehat{\mathcal{R}}_{CV,\alpha}(\Psi_{\alpha,t}, V_{1:K}^{\text{K-fold}}) - \mathcal{R}_\alpha\Big(\Psi_{\alpha,t}(S_n)\Big)\right| \leq$$

$$\max(\mathcal{T})\left[E_{\text{K-fold}}^t(n,K,\alpha) + \frac{20}{3n\alpha}\log\left(\frac{1}{\delta}\right) + 20\sqrt{\frac{2}{n\alpha}\log\left(\frac{1}{\delta}\right)}\right],$$

where $M_t > 0$ is a universal constant and

$$E_{\text{K-fold}}^t(n,K,\alpha) = 5M_t\sqrt{\frac{(d+1)K}{n\alpha}} + \frac{5K}{(K-1)n\alpha}.$$

By setting $M_{\mathcal{T}} = \max_{t\in\mathcal{T}} M_t$, we obtain by union bound, with probability $1 - 15\delta$,

$$\sup_{t\in\mathcal{T}}\left|\widehat{\mathcal{R}}_{CV,\alpha}(\Psi_{\alpha,t}, V_{1:K}^{\text{K-fold}}) - \mathcal{R}_\alpha\Big(\Psi_{\alpha,t}(S_n)\Big)\right| \leq$$

$$\max(\mathcal{T})\left[E_{\text{K-fold}}^{\mathcal{T}}(n,K,\alpha) + \frac{20}{3n\alpha}\log\left(\frac{|\mathcal{T}|}{\delta}\right) + 20\sqrt{\frac{2}{n\alpha}\log\left(\frac{|\mathcal{T}|}{\delta}\right)}\right]. \tag{2.19}$$

Combining inequalities (2.18) and (2.19) yields the desired result. ■

2.6 Numerical Experiments

The aim of our experiments is to illustrate the tightness of our bounds. The question we ask is whether the error (*resp.* excess risk) upper bound of order $O(1/\sqrt{n\alpha})$ describes accurately the behaviour of the CV error (*resp.* excess risk). Note that the problem of obtaining lower bounds for the generalization risk of classification algorithms in extreme regions remains to this date an open question in the statistical learning literature dedicated to extremes. For simplicity we limit our experiments to the K-fold scheme with $K = 10$.

2.6.1 CV error for risk estimation

Experimental setting. We consider the simple setting of a one dimensional threshold-based classifier ($\mathcal{G} = \{\text{sign}(X-\delta) \mid \delta \in \mathbb{R}\}$) minimizing the Hamming loss $l(g,(X,Y)) = \mathbb{1}\Big\{g(X) \neq Y\Big\}$. We investigate the risk estimation error of the CV estimator $\widehat{\mathcal{R}}_{\text{CV},\alpha}(\Psi_\alpha, V_{1:K})$ defined in (2.3) for several values of α within the range $\Big[1\%, 20\%\Big]$. In practice we compute $\widehat{\mathcal{R}}_{\text{CV},\alpha}$ using a dataset $\mathcal{D}_n$, of size $n = 2.10^4$ and evaluate the generalization risk of the trained rule $\Psi_\alpha(S_n)$ on a test set ($\mathcal{D}_{Test}$, of size $n_{test} = 2.10^6$). We perform $n_{simu} = 10^4$ experiments and we report the average and the upper 0.90 quantile of the abolute error obtained over the n_{simu} experiments. In other words we monitor the absolute generalization gap $\Big|\widehat{\mathcal{R}}_{\text{CV},\alpha}(\Psi_\alpha, V_{1:K}) - \mathcal{R}_\alpha(\Psi_\alpha(S_n))\Big|$ approximated by the quantity $\Big|\widehat{\mathcal{R}}_{\text{CV},\alpha}(\Psi_\alpha, V_{1:K}) - \widehat{\mathcal{R}}_\alpha(\Psi_\alpha(\mathcal{D}_n), \mathcal{D}_{Test})\Big|$ and we report a Monte-Carlo approximation of its expected value and its quantile of order 0.90 for different value of α.

Datasets. we generate a balanced binary classification dataset $Z_i = (X_i, Y_i) \in \mathcal{Z} = \mathbb{R} \times \{0,1\}$ with $\mathbb{P}(Y=0) = \mathbb{P}(Y=1) = 1/2$. Both classes are sampled from a t-student distribution, with respective parameters $(\mu_i, \sigma_i, \nu_i), i = 0, 1$. We set $\mu_0 = -\mu_1 = 1$, $\sigma_0 = \frac{3}{5}$, $\sigma_1 = 3$, and $\nu_1 = \nu_2 = 1.5$.

2.6.2 CV excess risk for model selection

We now describe the empirical analysis of the model selection upper bound presented in Lemma 2.16.

Experimental setting. We consider the problem of tuning the penalty parameter of a Lasso logistic regression model. Note that, instead of using the constrained formulation of the Lasso (cf. Equation (2.4)) , we consider in our experiments the Lagrangian formulation:

$$\Psi_{\alpha,\lambda}(S) = \underset{\beta \in \mathbb{R}^d}{\arg\min} \frac{1}{\alpha n_S} \sum_{i \in S} \left(c\left(\beta^T \Theta\left(X_i\right), Y_i\right) + \lambda \|\beta\|_1 \right) \mathbb{1}\Big\{\|X_i\| > \|X_{(\lfloor \alpha n \rfloor)}\|\Big\},$$

with penality parameter λ ranging in a finite logarithmic grid

$$\Delta = \{10^{i/30} - 1 \mid i \in [\![1, 30]\!]\}.$$

The reason for using the penalized formulation in practice is mainly a computational one: the latter version can be solved by many standard optimization algorithms (stochastic gradient descent for instance) contrarily to the constrained one that requires special and time consuming optimization routines (see *e.g.* Lee et al. (2006); Homrighausen and McDonald (2017)). Notice that we leave a gap between theory and practice to be filled in further work. Indeed, analyzing the penalized Lasso requires different proof techniques and more assumptions. For example, Homrighausen and McDonald (2017) work under a realizability assumption while Chetverikov et al. (2021b) make some moment assumptions.

In the sequel, we study the excess risk of the model selected by K-fold cross validation $\mathcal{R}_\alpha(\Psi_{\alpha,\hat{\lambda}}(S_n)) - \mathcal{R}_\alpha(\Psi_{\alpha,\lambda^*}(S_n))$ for several values of α within the range $\Big[1\%, 10\%\Big]$. Similarly to the previous experiment we select $\hat{\lambda}$ using a dataset $\mathcal{D}_n$ of size $n = 10^4$, then we use a test set $\mathcal{D}_{test}$ of size $n_{test} = 10^6$ to estimate $\mathcal{R}_\alpha$ and choose λ^* accordingly

Figure 2.1 – K-fold CV risk estimation absolute error as a function of $n\alpha$ (logarithmic scale): mean and upper quantile at level 0.90

. Finally, we report the average model selection excess risk and its corresponding 0.9 quantile for different values of α over $n_{simu} = 10^4$ Monte Carlo simulations.
Datasets. We generate a balanced binary classification dataset $Z_i = (X_i, Y_i) \in \mathcal{Z} = \mathbb{R}^{20} \times \{0, 1\}$ with $\mathbb{P}(Y = 0) = \mathbb{P}(Y = 1) = 1/2$. Both classes are sampled from a t-multivariate-student distribution, with respective sparse parameters $(\mu_i, \sigma_i, \nu_i), i = 0, 1$. We set $\mu_0 = -\mu_1 = (e_5, 0, \ldots, 0)$, $\sigma_0 = \sigma_1 = 10 I_{20}$, $\nu_1 = \nu_2 = 1.5$ and $e_5 = (1, \ldots, 1)$ is a 5 dimensional unit vector.

2.6.3 Results

Figure 2.1 displays the risk estimation error of the cross validation estimator $\widehat{\mathcal{R}}_{\mathrm{CV},\alpha}(\Psi_\alpha, V_{1:K})$ as a function of α on the logarithmic scale. As suggested by our theoretical findings, the average error and its quantile indeed decrease at rate $O(1/\sqrt{n\alpha})$ as a function of α. This confirms that our bounds may be sharp up to multiplicative constants.
One must note that in the model selection case (Figure 2.2) the rate of convergence appears to be faster than $1/\sqrt{n\alpha}$ for values of $n\alpha$ ranging between 500 and 1000. This is not surprising insofar as it corroborates the findings of many recent works where it is established that the Lasso algorithm enjoys an *algorithmic stability* (Bousquet and Elisseeff, 2002) property which induces fast rates for CV estimates (Celisse and Guedj, 2016; Abou-Moustafa and Szepesvári, 2019). For the smallest values of $n\alpha$ (less than 500) a slower rate is observed. This might be explained by the findings of Homrighausen and McDonald (2013); Chetverikov et al. (2021a), who show that, outside the context of extreme values, the rate of convergence for cross-validation estimates using k training samples deteriorates as the value of $c = \frac{\ln(d)}{\ln(k)}$ increases.

2.A Generic technical tools

We recall the following McDiarmid's extension of Bernstein inequality (Theorem 3.8 in McDiarmid (1998)).

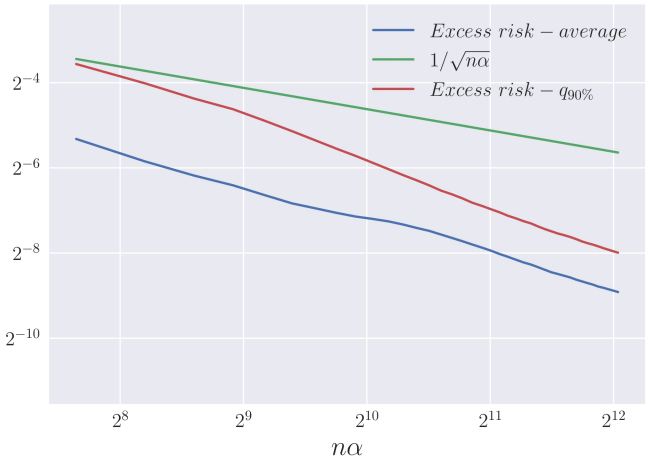


Figure 2.2 – K-fold CV excess risk as a function of $n\alpha$ (logarithmic scale): mean and upper quantile at level 0.90

Proposition 2.17. *For a sequence of observations $(Z_1, Z_2, \ldots, Z_n) \in \mathcal{Z}^n$ and some fixed values $z_{1:l} = (z_1, z_2, \ldots, z_l)$ and for some measurable function $f : \mathcal{Z}^n \to \mathbb{R}$, let $W = f(Z_1, Z_2, \ldots, Z_n)$ and define for $l \in [\![1, n]\!]$:*

1. $f_l(z_1, z_2, \ldots, z_l) = \mathbb{E}\left(W \mid Z_1 = z_1, Z_2 = z_2, \ldots, Z_l = z_l\right)$,
2. $\Delta_l(z_1, z_2, \ldots, z_{l-1}, z_l) = f_l(z_1, z_2, \ldots, z_{l-1}, z_l) - f_{l-1}(z_1, z_2 \ldots, z_{l-1})$, *(the positive deviations)*
3. $D := \max_{l=1,\ldots,n} \sup_{z_1,\ldots,z_{i-1} \in \mathcal{Z}} \sup_{z \in \mathcal{Z}} \Delta_l\left(z_1, \ldots, z_{l-1}, z\right)$, *(the maximum positive deviation)*
4. $\sigma_l^2(z_{1:l-1}) = \mathrm{Var}\left[\Delta_l(Z_1, Z_2, \ldots, Z_{l-1}, Z') \mid Z_1 = z_1, Z_2 = z_2, \ldots, Z_{l-1} = z_{l-1}\right]$, *where Z' is an independent copy of Z_l,*
5. $\sigma^2 = \sup_{z_{1:l-1} \in \mathcal{Z}^{l-1}} \sum_{l=1}^{n} \sigma_l^2(z_{1:l-1})$ *(the maximum sum of variances).*

Then we have

$$\mathbb{P}(W - \mathbb{E}\left[W\right] > t) \leq \exp\left(\frac{-t^2}{2(\sigma^2 + Dt/3)}\right).$$

We also recall Proposition 2.5.2 from Vershynin (2018), which provides an upper bound for the expectation of sub-Gaussian random variables.

Proposition 2.18. *Let X be a real valued random variable and suppose that*

$$\mathbb{P}(X \geq t) \leq C_1.exp(-t^2/C_2^2),$$

for some $C_1, C_2 > 0$. then it holds that

$$\mathbb{E}(X) \leq M_2 C_2,$$

where $M_2 > 0$ is a universal constant depending only on C_1.

2.B Intermediate results and detailed proofs

2.B.1 Proof of Lemma 2.6

Since the leave-one-out is a special case of K-fold with $K = n$ (or leave-p-out with $p = 1$) it suffices to prove the statement concerning the cases of the leave-p-out and the K-fold.

K-Fold. For this procedure, the validation sets is a partition of $[\![1, n]\!]$:

$$\bigcup_{j=1}^{K} V_j = [\![1, n]\!] \; and \; V_j \bigcap V_k = \varnothing \; , \; \forall j \neq k \in [\![1, K]\!]. \tag{2.20}$$

Under the assumption that K divides n, the condition $card(V_j) = n/K := n_V$ for all the validation sets V_j holds, as stipulated in (6.3). Thus we have

$$n = \sum_{j=1}^{K} card(V_j) = K n_V. \tag{2.21}$$

Furthermore, under (2.20), any index $l \in [\![1, n]\!]$ belongs to a unique validation test $V_{j'}$ and to all the train sets $T_j = V_j^c$ with $j \neq j'$. Hence, we both have

$$\begin{cases} \sum_{j=1}^{K} \mathbb{1}\Big\{ l \in T_j \Big\} = K - 1, & \text{and} \\ \sum_{j=1}^{K} \mathbb{1}\Big\{ l \in V_j \Big\} = 1. \end{cases}$$

Using (2.21) and the fact that $n_T = n - n_V = (K - 1)n_V$ yields the desired result.

Leave-p-out. In the leave-p-out procedure, the sequence of validation sets is the family of all subsamples V_j of $\mathcal{D}_n$ of size $card(V_j) = p$, thus $K = \binom{n}{p}$. On the other hand, any index $l \in [\![1, n]\!]$ belongs to $\binom{n-1}{p-1}$ validation sets. Indeed constructing a V_j such as $l \in V_j$ is equivalent to first picking l and then choosing $p - 1$ elements from $[\![1, n]\!] \setminus \{l\}$. Hence we have

$$\sum_{j=1}^{K} \mathbb{1}\Big\{ l \in V_j \Big\} = \binom{n-1}{p-1} \; , \; \forall l \in [\![1, n]\!].$$

Using the identity $n\binom{n-1}{p-1} = p\binom{n}{p}$ we obtain

$$\frac{1}{K n_V} \sum_{j=1}^{K} \mathbb{1}\Big\{ l \in V_j \Big\} = \frac{1}{p\binom{n}{p}} \sum_{j=1}^{K} \mathbb{1}\Big\{ l \in V_j \Big\} = 1/n.$$

A similar argument applies to the sequence $T_{1:K}$, which completes the proof.

2.B.2 Intermediate results for the proofs of theorems 2.9 and 2.12

In this section we gather the main intermediate results involved in the proofs of our main results theorems 2.9 and 2.12, which are of interest in their own.

A key tool to our proofs is a Bernstein-type inequality relative to the deviation of a generic random variable $W = f(Z_1, \dots, Z_n)$ from its mean (McDiarmid (1998)) that is recalled for convenience in section 2.A of this supplement (Proposition 2.17). The control of the deviations involves both a maximum deviation term and a variance term. We leverage this result to control the deviations of the pseudo-empirical risk $\tilde{\mathcal{R}}_\alpha$ defined in (2.8) averaged over the K validation sets $V_{1:K}$. These deviations are embodied by the random variable W defined in Lemma 2.19, Equation (2.22), which is a key quantity when analysing the deviations of any CV risk estimate. Controlling the deviations of W is the main purpose of this section.

Lemma 2.19. *Let $\mathcal{D}_n = (Z_1, Z_2, \dots, Z_n) \in \mathcal{Z}^n$ be a sequence of random variables, and let $V_1, V_2, \dots V_K$ be validation sets that verify Assumption 2 with size n_V. Moreover, suppose that assumptions 3 and 4 regarding the class $\mathcal{G}$ and c hold. Define*

$$W = \frac{1}{K} \sum_{j=1}^{K} \sup_{g \in \mathcal{G}} |\tilde{\mathcal{R}}_\alpha(g, V_j) - \mathcal{R}_\alpha(g)|, \tag{2.22}$$

where $\tilde{\mathcal{R}}_\alpha$ is defined by Equation 2.8. Then the random variable W satisfies the Bernstein-type inequality

$$P\left(W - E(W) \geq t\right) \leq \exp\left(\frac{-n\alpha t^2}{2(4 + t/3)}\right).$$

Proof We introduce for convenience the rescaled variable $W_\alpha = \alpha W$. Then W_α may be written as

$$W_\alpha = \frac{1}{K} \sum_{j=1}^{K} \left[\frac{1}{n_V} \sup_{g \in \mathcal{G}} \left| \sum_{i \in V_j} c(g, Z_i) \mathbb{1}_\alpha(X_i) - \mathbb{E}\left[c(g, Z) \mathbb{1}_\alpha(X)\right] \right| \right],$$

where we use the shorthand notation $\mathbb{1}_\alpha(X) = \mathbb{1}\Big\{\|X\| \geq t_\alpha\Big\}$. We derive an upper bound on $\mathbb{P}\left(W_\alpha - \mathbb{E}(W_\alpha) > t\right)$ using Proposition 2.17. Namely we show that the maximum deviations term D and σ^2 from the latter statement are respectively bounded by $D \leq 1/n$ and $\sigma^2 \leq 4\alpha/n$. To do so we compute explicitly the five quantities defined in the satement of Proposition 2.17 in our particular context.

1. The conditional expectations f_l from Proposition 2.17 are (recall that $z_i = (x_i, y_i)$),

$$\begin{aligned}
&f_l(z_1, z_2, \ldots, z_l) \\
&= \mathbb{E}\left(W_\alpha \mid Z_1 = z_1, Z_2 = z_2, \ldots, Z_l = z_l\right) \\
&= \frac{1}{Kn_V} \sum_{j=1}^{K} \mathbb{E}\left[\sup_{g \in \mathcal{G}} \left| \sum_{i \in V_{j,l}} c(g, z_i) \mathbb{1}_\alpha(x_i) + \sum_{i \in V_j \setminus V_{j,l}} c(g, Z_i) \mathbb{1}_\alpha(X_i) \right.\right. \\
&\qquad\qquad \left.\left. - n_V \mathbb{E}\left[c(g, Z) \mathbb{1}_\alpha(X)\right] \right| \right] \\
&= \frac{1}{Kn_V} \sum_{j=1}^{K} \mathbb{E}\left[\sup_{g \in \mathcal{G}} \left| h_{j,g}\left(z_{1:l}, Z_{l+1:n}\right) \right| \right],
\end{aligned}$$

where $V_{j,l} = V_j \cap [\![1, l]\!]$ are the validation indices which belong to the interval $[\![1, l]\!]$, and

$$h_{j,g}\left(z_{1:l}, Z_{l+1:n}\right) = \sum_{i \in V_{j,l}} c(g, z_i) \mathbb{1}_\alpha(x_i) + \sum_{i \in V_j \setminus V_{j,l}} c(g, Z_i) \mathbb{1}_\alpha(X_i) - n_V \mathbb{E}\left[c(g, Z) \mathbb{1}_\alpha(X)\right].$$

2. Recall the definition of the positive deviations Δ_l,

$$\Delta_l(z_1, z_2, \ldots, z_{l-1}, z_l) = f_l(z_1, z_2, \ldots, z_{l-1}, z_l) - f_{l-1}(z_1, z_2 \ldots, z_{l-1}).$$

In view of the expression for f_l from step 1, we may thus write

$$\Delta_l(z_{1:l}) = \frac{1}{Kn_V} \sum_{j=1}^{K} \mathbb{E}\left[\sup_{g \in \mathcal{G}} \left| h_{j,g}\left(z_{1:l}, Z_{l+1:n}\right) \right| - \sup_{g \in \mathcal{G}} \left| h_{j,g}\left(z_{1:l-1}, Z_{l:n}\right) \right| \right].$$

Now, notice that $V_{j,l} = V_{j,l-1}$ if $l \notin V_j$ and $V_{j,l} = V_{j,l-1} \cup \{l\}$ otherwise. Hence

$$h_{j,g}\left(z_{1:l}, Z_{l+1:n}\right) - h_{j,g}\left(z_{1:l-1}, Z_{l:n}\right) = \mathbb{1}\{l \in V_j\} \left(c(g, z_l) \mathbb{1}_\alpha(x_l) - c(g, Z_l) \mathbb{1}_\alpha(X_l) \right).$$

Using the fact that, for any functions f, g, it holds that $\left| \sup|f| - \sup|g| \right| \leq \sup|f - g|$, we obtain

$$|\Delta_l(z_{1:l})| \leq \frac{1}{Kn_V} \sum_{j=1}^{K} \left[\mathbb{1}\{l \in V_j\} \mathbb{E} \underbrace{\sup_{g \in \mathcal{G}} \left| c(g, z_l) \mathbb{1}_\alpha(x_l) - c(g, Z_l) \mathbb{1}_\alpha(X_l) \right|}_{\text{(Assumption 4)} \leq 1} \right] \tag{2.23}$$

and deduce that

$$\begin{aligned}
|\Delta_l(z_{1:l})| &\leq \frac{1}{Kn_V} \sum_{j=1}^{K} \mathbb{1}\{l \in V_j\} \\
\text{(by Assumption 2)} &\leq \frac{1}{n}.
\end{aligned}$$

3. The maximum positive deviation is defined by

$$D := \max_{l=1,\ldots,n} \sup_{z_1,\ldots,z_{i-1}\in\mathcal{Z}} \sup_{z\in\mathcal{Z}} \Delta_l\left(z_1,\ldots,z_{l-1},z\right).$$

From the previous step, we immediately obtain

$$D \leq 1/n.$$

4. Let $Z' = (X', Y')$ be an independent copy of $Z = (X, Y)$ and let $z_{1:l} = (z_1, z_2, \ldots, z_l)$. Recall the conditional variance term from the statement of Proposition 2.17, $\sigma_l^2(z_{1:l-1}) := \mathrm{Var}\left[\Delta_l(Z_1, Z_2, \ldots, Z_{l-1}, Z') \mid Z_1 = z_1, Z_2 = z_2, \ldots, Z_{l-1} = z_{l-1}\right]$. Then σ_l^2 may be upper bounded as follows,

$$\begin{aligned}\sigma_l^2(z_{1:l-1}) &\leq \mathbb{E}\left[\Delta_l(Z_1, Z_2, \ldots, Z_{l-1}, Z')^2 \mid Z_1 = z_1, Z_2 = z_2, \ldots, Z_{l-1} = z_{l-1}\right] \\ &= \mathbb{E}\left[\Delta_l(z_1, z_2, \ldots, z_{l-1}, Z')^2\right].\end{aligned}$$

Now using (2.23), write

$$\begin{aligned}\sigma_l^2 &\leq \frac{1}{(Kn_V)^2}\mathbb{E}\left[\left(\sum_{j=1}^{K} \mathbb{1}\{l \in V_j\}\mathbb{E}\Big[\sup_{g\in\mathcal{G}} |c(g, Z')\mathbb{1}_\alpha(X') - c(g, Z_l)\mathbb{1}_\alpha(X_l)| \mid Z'\Big]\right)^2\right] \\ (|c| \leq 1) &\leq (\frac{1}{Kn_V})^2\mathbb{E}\left[\left(\sum_{j=1}^{K} \mathbb{1}\{l \in V_j\}\mathbb{E}\Big[\sup_{g\in\mathcal{G}} \underbrace{\mathbb{1}_\alpha(X') + \mathbb{1}_\alpha(X_l)}_{\text{independent of } g} \mid Z'\Big]\right)^2\right] \\ &= (\frac{1}{Kn_V})^2\mathbb{E}\left[\left(\sum_{j=1}^{K} \mathbb{1}\{l \in V_j\}\underbrace{(\mathbb{1}_\alpha(X') + \alpha)}_{\text{independent of } j}\right)^2\right] \\ &\leq \mathbb{E}\left[(\mathbb{1}_\alpha(X') + \alpha)^2\right]\left(\frac{1}{Kn_V}\sum_{j=1}^{K} \mathbb{1}\{l \in V_j\}\right)^2 \\ &= (\alpha + 3\alpha^2)\left(\frac{1}{Kn_V}\sum_{j=1}^{K} \mathbb{1}\{l \in V_j\}\right)^2 \\ \text{(by (6.4))} &= \frac{\alpha + 3\alpha^2}{n^2} \\ (\alpha \leq 1) &\leq \frac{4\alpha}{n^2}.\end{aligned}$$

5. Finally we get $\sigma^2 = \sum_{l=1}^{n} \sup_{z_{1:l-1}} \sigma_l^2(z_{1:l-1}) \leq \frac{4\alpha}{n}$.

At this stage, applying proposition 2.17 gives

$$\mathbb{P}(W_\alpha - \mathbb{E}\left[W_\alpha\right] > t) \leq \exp\left\{\frac{-nt^2}{2(4\alpha + t/3)}\right\}.$$

Therefore for $W = W_\alpha/\alpha$ one obtains

$$\mathbb{P}(W - \mathbb{E}\left[W\right] > t) \leq \exp\left\{\frac{-n\alpha t^2}{2(4 + t/3)}\right\}.$$

■

To obtain a genuine probability bound on Z *via* the latter lemma, one also needs to control the term $E(Z)$. This is the purpose of the next lemma, the spirit of which is similar to Lemma 14 in Goix et al. (2015). The main difference *w.r.t.* to the latter reference is that we handle any bounded cost function (not only the Hamming loss), using a bound for Rademacher averages from Giné and Guillou (2001) which applies in this broader setting.

Lemma 2.20. *In the setting of Lemma 2.19, W satisfies*

$$\mathbb{E}(W) \leq \frac{M\sqrt{\mathcal{V}_\mathcal{G}}}{\sqrt{\alpha n_V}}.$$

where $M > 0$ is a universal constant.

Proof

Notice first that, since the observations are i.i.d. ,

$$\mathbb{E}\left[\frac{1}{K}\sum_{j=1}^{K}\sup_{g\in\mathcal{G}}\left|\tilde{\mathcal{R}}_\alpha(g, V_j) - \mathcal{R}_\alpha(g)\right|\right] = \mathbb{E}\left[\sup_{g\in\mathcal{G}}\left|\tilde{\mathcal{R}}_\alpha(g, V_1) - \mathcal{R}_\alpha(g)\right|\right].$$

That is, $\mathbb{E}(W) = \mathbb{E}(W_{V_1})$, where for a subset of indices $S = \{1, \ldots, n_S\}$, we denote

$$W_S = \sup_{g\in\mathcal{G}}\left|\tilde{\mathcal{R}}_\alpha(g, S) - \mathcal{R}_\alpha(g)\right|.$$

In order to bound $\mathbb{E}\left[W_S\right]$ defined above, we follow the same steps as in the proof of Lemma 14 in Goix et al. (2015), where most arguments also hold true for a bounded VC class of cost functions.

In particular, we use the symmetrization technique. Consider Rademacher random variables $\mathcal{E} = (\epsilon_1, \epsilon_2, \ldots, \epsilon_{n_S})$ taking values in $\{-1, 1\}$ and introduce the randomized process

$$W_\mathcal{E} = \sup_{g\in\mathcal{G}}\left|\frac{1}{\alpha n_S}\sum_{i=1}^{n_S}\epsilon_i c(g, Z_i)\mathbb{1}\left\{\|X_i\| \geq t_\alpha\right\}\right|.$$

It can be shown using the same classical steps as in the proof of Lemma 13 in Goix et al. (2015) that

$$\mathbb{E}(W_S) \leq 2\mathbb{E}(W_{\mathcal{E}}) \,.$$

The key argument to proceed is to condition the above expectation upon the number of indices i such that $\|X_i\| \geq t_\alpha$. This conditioning trick is a standard technique for deriving non asymptotic bounds in the EVT framework (Goix et al. (2015); Lhaut et al. (2021)). Introduce a random variable $Z_{i,\alpha}$, which has the same distribution as $(Z \mid \|X\| \geq t_\alpha)$ and notice that

$$\sum_{i=1}^{n_S} \epsilon_i c(g, Z_i) \mathbb{1}\Big\{\|X_i\| \geq t_\alpha\Big\} \sim \sum_{i=1}^{\mathcal{N}} \epsilon_i c(g, Z_{i,\alpha}) \,,$$

where $\mathcal{N}$ has a Binomial distribution $\mathcal{B}(n_S, \alpha)$. Equipped with these notations

$$\mathbb{E}(W_{\mathcal{E}}) = \mathbb{E}(\phi(\mathcal{N})) \,,$$

where

$$\phi(N) = \mathbb{E} \sup_{g \in \mathcal{G}} \left| \frac{1}{\alpha n_S} \sum_{i=1}^{N} \epsilon_i c(g, Z_{i,\alpha}) \right| = \frac{N}{\alpha n_S} \mathbb{E} \sup_{g \in \mathcal{G}} \frac{1}{N} \left| \sum_{i=1}^{N} \epsilon_i c(g, Z_{i,\alpha}) \right| .$$

Then by a classical Rademacher complexity arguments for finite VC-classes (see Giné and Guillou (2001), Proposition 2.1),

$$\begin{aligned} \mathbb{E} \sup_{g \in \mathcal{G}} \frac{1}{N} \left| \sum_{i=1}^{N} \epsilon_i c(g, Z_{i,\alpha}) \right| &\leq \frac{M_1' \sqrt{\mathcal{V}_{\mathcal{G}}}}{\sqrt{N}} + \frac{M_2' \mathcal{V}_{\mathcal{G}}}{N} \\ &\leq M' \sqrt{\frac{\mathcal{V}_{\mathcal{G}}}{N}}, \end{aligned}$$

for some universal constant $M' > 0$, whence

$$\phi(N) \leq \frac{N}{\alpha n_S} \frac{M' \sqrt{\mathcal{V}_{\mathcal{G}}}}{\sqrt{N}} = \frac{M' \sqrt{\mathcal{V}_{\mathcal{G}} N}}{\alpha n_S}$$

By concavity we have $\mathbb{E}(\sqrt{N}) \leq \sqrt{\mathbb{E}(N)} = \sqrt{\alpha n_S}$, and we obtain

$$\mathbb{E}(W_S) \leq 2\mathbb{E}(W_{\mathcal{E}}) \leq \frac{2M' \sqrt{\mathcal{V}_{\mathcal{G}}}}{\sqrt{n_S \alpha}} \,.$$

The result follows. ■

The following probability upper bound for Z follows immediately by combining Lemma 2.19 and Lemma 2.20.

Corollary 2.21. *In the setting of Lemma 2.19,*

we have

$$P\left(W - \frac{M \sqrt{\mathcal{V}_{\mathcal{G}}}}{\sqrt{\alpha n_V}} \geq t \right) \leq \exp\left(\frac{-n\alpha t^2}{2(4 + t/3)} \right) ,$$

where W is defined in (2.22).

To conclude this section, we extend Theorem 10 in Goix et al. (2015) bounding the supremum deviations of the empirical measure on low probability regions, to handle the case of any cost function c absolutely bounded by one.

Lemma 2.22. *Recall the definitions of the risk $\mathcal{R}_\alpha$ and its empirical version $\widehat{\mathcal{R}}_\alpha$ given in Section 2.2 and introduce the (random) supremum deviations*

$$W' = \sup_{g\in\mathcal{G}} |\widehat{\mathcal{R}}_\alpha(g, S_n) - \mathcal{R}_\alpha(g)|.$$

If $\mathcal{G}$ is a family of classifiers with finite VC-dimension and c a bounded cost function with $\sup_{g,z} |c(g, z)| \leq 1$, then, the following Bernstein-type inequality holds,

$$\mathbb{P}(W' - Q(n, \alpha) \geq t) \leq 3 \exp\left(\frac{-n\alpha t^2}{2(4 + t/3)}\right),$$

where $Q(n, \alpha) = B(n, \alpha) + \frac{1}{n\alpha}$ and B is defined by

$$B(n, \alpha) = \frac{M\sqrt{\mathcal{V}_\mathcal{G}}}{\sqrt{\alpha n}} \tag{2.24}$$

and M is a universal constant.

Proof Write $W' \leq W_1 + W_2$ with :

$$W_1 = \sup_{g\in\mathcal{G}} |\widehat{\mathcal{R}}_\alpha(g, S_n) - \tilde{\mathcal{R}}_\alpha(g, S_n)|,$$

$$W_2 = \sup_{g\in\mathcal{G}} |\tilde{\mathcal{R}}_\alpha(g, S_n) - \mathcal{R}_\alpha(g)|.$$

Concerning W_2, applying Corollary 2.21 with $K = 1$, $V_1 = S_n$, yields

$$\mathbb{P}(W_2 - B(n, \alpha) \geq t) \leq \exp\left(\frac{-n\alpha t^2}{2(4 + t/3)}\right). \tag{2.25}$$

We now focus on W_1. Define

$$u_i = \left| \mathbb{1}\Big\{ \|X_i\| > \|X_{(\lfloor \alpha n \rfloor)}\| \Big\} - \mathbb{1}\Big\{ \|X_i\| \geq t_\alpha \Big\} \right|$$

and notice that $W_1 \leq \frac{1}{n\alpha} \sum_{i=1}^n u_i$. It is known (see for instance the bound for the term A in Jalalzai et al. (2018), page 12) that

$$\frac{1}{n\alpha} \sum_{i=1}^n u_i \leq \frac{1}{\alpha} \left| \frac{1}{n} \sum_{i=1}^n \mathbb{1}\Big\{ \|X\| \geq t_\alpha \Big\} - \alpha \right| + \frac{1}{n\alpha}.$$

Now, by noticing that $\mathrm{Var}(\mathbb{1}\Big\{ \|X\| \geq t_\alpha \Big\}) \leq \alpha$ and using Bernstein's inequality, we get

$$\begin{aligned} \mathbb{P}\left(\left| \frac{1}{n} \sum_{i=1}^n \mathbb{1}\Big\{ \|X\| \geq t_\alpha \Big\} - \alpha \right| \geq t \right) &\leq 2 \exp\left(\frac{-nt^2}{2(\alpha + t/3)}\right) \\ &\leq 2 \exp\left(\frac{-nt^2}{2(4\alpha + t/3)}\right). \end{aligned}$$

Finally, dividing by α, we get

$$\mathbb{P}\left(\frac{1}{n\alpha}\left|\sum_{i=1}^{n}\mathbb{1}\Big\{\|X\|\geq t_\alpha\Big\}-\alpha\right|\geq t\right)\leq 2\exp\left(\frac{-n\alpha t^2}{2(4+t/3)}\right).$$

Therefore we finally obtain

$$\mathbb{P}\left(\frac{1}{n\alpha}\sum_{i=1}^{n}u_i-1/n\alpha\geq t\right)\leq 2\exp\left(\frac{-n\alpha t^2}{2(4+t/3)}\right). \tag{2.26}$$

The result follows using $W'\leq W_1+W_2$ and $W_1\leq\frac{1}{n\alpha}\sum_{i=1}^{n}u_i$. ■

2.B.3 Detailed proof of Theorem 2.9

In view of the argument following the statement, we derive probability upper bounds for the three terms $\mathrm{D}_{\mathrm{t}_\alpha}$, D_{cv} and Bias defined in equations (2.12, 2.13, 2.14).

Probability bound for $\mathrm{D}_{\mathrm{t}_\alpha}$ (see (2.12)). Using the fact that the cost function verifies $0\leq c\leq 1$, write

$$\mathrm{D}_{\mathrm{t}_\alpha}=|\widehat{\mathcal{R}}_{CV,\alpha}(\Psi_\alpha,V_{1:K})-\tilde{\mathcal{R}}_{CV,\alpha}(\Psi_\alpha,V_{1:K})|\leq U,$$

where $U=\frac{1}{Kn_V\alpha}\sum_{j=1}^{K}\sum_{i\in V_j}u_i$ and $u_i=\left|\mathbb{1}\Big\{\|X_i\|>\|X_{(\lfloor\alpha n\rfloor)}\|\Big\}-\mathbb{1}\Big\{\|X_i\|\geq t_\alpha\Big\}\right|$ as the proof of Lemma 2.22. Now notice that

$$\begin{aligned}U&=\frac{1}{Kn_V\alpha}\sum_{j=1}^{K}\sum_{i=1}^{n}u_i\mathbb{1}\Big\{i\in V_j\Big\}\\&=\frac{1}{Kn_V\alpha}\sum_{i=1}^{n}u_i\sum_{j=1}^{K}\mathbb{1}\Big\{i\in V_j\Big\}\\&=\frac{1}{n\alpha}\sum_{i=1}^{n}u_i.\end{aligned}$$

The last line follows from Assumption 2. Hence, using Inequality (2.26) from the proof of Lemma 2.22, we obtain

$$\mathbb{P}(\mathrm{D}_{\mathrm{t}_\alpha}-1/n\alpha\geq t)\leq 2\exp\left(\frac{-n\alpha t^2}{2(4+t/3)}\right). \tag{2.27}$$

Probability bound for D_{cv} (see 2.13). First, notice that

$$\begin{aligned}\mathrm{D}_{\mathrm{cv}}&=|\frac{1}{K}\sum_{j=1}^{K}\Big[\tilde{\mathcal{R}}_\alpha\Big(\Psi_\alpha(T_j),V_j\Big)-\mathcal{R}_\alpha\Big(\Psi_\alpha(T_j)\Big)\Big]|\\&\leq\frac{1}{K}\sum_{j=1}^{K}\sup_{g\in\mathcal{G}}|\tilde{\mathcal{R}}_\alpha(g,V_j)-\mathcal{R}_\alpha(g)|.\end{aligned}$$

The r.h.s of the latter display is the quantity Z defined in Lemma 2.19, Equation 2.22. As a consequence of this lemma,

$$\mathbb{P}(\mathrm{D_{cv}} - B(n_V, \alpha) \geq t) \leq \exp\left(\frac{-n\alpha t^2}{2(4 + t/3)}\right), \tag{2.28}$$

with

$$B(n_V, \alpha) = \frac{M\sqrt{\mathcal{V}_\mathcal{G}}}{\sqrt{\alpha n_V}},$$

for some universal constant $M > 0$.

Probability bounds for Bias **(see 2.14).** We write

$$\begin{aligned} \text{Bias} &= \frac{1}{K}\left|\sum_{j=1}^{K}\left(\mathcal{R}_\alpha(\Psi_\alpha(T_j)) - \mathcal{R}_\alpha\Big(\Psi_\alpha(S_n)\Big)\right)\right| \\ &\leq C_1 + C_2, \end{aligned} \tag{2.29}$$

with

$$C_1 = \frac{1}{K}\left|\sum_{j=1}^{K}\Big(\mathcal{R}_\alpha(\Psi_\alpha(T_j)) - \widehat{\mathcal{R}}_\alpha(\Psi_\alpha(T_j), T_j) + \widehat{\mathcal{R}}_\alpha(\Psi_\alpha(T_j), T_j) - \mathcal{R}^*_\alpha\Big)\right|,$$

$$C_2 = \left|\mathcal{R}^*_\alpha - \widehat{\mathcal{R}}_\alpha(\Psi_\alpha(S_n), S_n) + \widehat{\mathcal{R}}_\alpha(\Psi_\alpha(S_n), S_n) - \mathcal{R}_\alpha\Big(\Psi_\alpha(S_n)\Big)\right|$$

and

$$\mathcal{R}^*_\alpha = \inf_{g\in\mathcal{G}} \mathcal{R}_\alpha(g).$$

Using the fact that $\widehat{\mathcal{R}}_\alpha(\Psi_\alpha(T_j), T_j) = \inf_{g\in\mathcal{G}} \widehat{\mathcal{R}}_\alpha(g, T_j)$ (Assumption 1) and for any real functions h and f, $|\inf h - \inf f| \leq \sup|h - f|$, write

$$\begin{aligned} \frac{1}{K}\left|\sum_{j=1}^{K}\left(\widehat{\mathcal{R}}_\alpha(\Psi_\alpha(T_j), T_j) - \mathcal{R}^*_\alpha\right)\right| &= \frac{1}{K}\left|\sum_{j=1}^{K}\inf_{g\in\mathcal{G}}\widehat{\mathcal{R}}_\alpha(g, T_j) - \inf_{g\in\mathcal{G}}\mathcal{R}_\alpha(g)\right| \\ &\leq \frac{1}{K}\sum_{j=1}^{K}\left|\inf_{g\in\mathcal{G}}\widehat{\mathcal{R}}_\alpha(g, T_j) - \inf_{g\in\mathcal{G}}\mathcal{R}_\alpha(g)\right| \\ &\leq \frac{1}{K}\sum_{j=1}^{K}\sup_{g\in\mathcal{G}}|\widehat{\mathcal{R}}_\alpha(g, T_j) - \mathcal{R}_\alpha(g)|. \end{aligned}$$

Then, by using the triangle inequality, deduce that

$$\begin{aligned} C_1 &\leq \frac{2}{K}\sum_{j=1}^{K}\sup_{g\in\mathcal{G}}|\widehat{\mathcal{R}}_\alpha(g, T_j) - \mathcal{R}_\alpha(g)| \\ &\leq 2(W_1 + W_2), \end{aligned} \tag{2.30}$$

with

$$W_1 = \frac{1}{K}\sum_{j=1}^{K} \sup_{g\in\mathcal{G}} |\widehat{\mathcal{R}}_\alpha(g, T_j) - \tilde{\mathcal{R}}_\alpha(g, T_j)|,$$

$$W_2 = \frac{1}{K}\sum_{j=1}^{K} \sup_{g\in\mathcal{G}} |\tilde{\mathcal{R}}_\alpha(g, T_j) - \mathcal{R}_\alpha(g)|.$$

In order to bound Z_2 we use the fact that the statements of Lemmas 2.19, 2.20 and Corollary 2.21 still hold true if one substitutes the training sets T_j for the validation sets V_j, and the training size n_T for n_V, as revealed by an inspection of the proofs. The only difference between the two arguments is in steps 2 and 4 from the proof of Lemma 2.19 where we use the identity $\frac{1}{K}\sum_{j=1}^{K} \frac{\mathbb{1}\left\{l \in T_j\right\}}{n_T} = \frac{1}{n}$ from Lemma 2.6 instead of Identity (6.4). Thus we obtain, from the twin statement of Corollary 2.21,

$$\mathbb{P}(W_2 - 2B(n_T, \alpha) \geq 2t) \leq 2\exp\left(\frac{-n\alpha t^2}{2(4 + t/3)}\right), \tag{2.31}$$

where $B(n, \alpha)$ is defined in (2.24). Similarly the term W_1 can be bounded following the same argument as in the first paragraph (Probability bound for $\mathrm{D}_{\mathrm{t}_\alpha}$) up to replacing V_j with T_j and n_V with n_T :

$$\begin{aligned} W_1 &\leq \frac{1}{Kn_T\alpha}\sum_{j=1}^{K}\sum_{i=1}^{n} u_i \mathbb{1}\left\{i \in T_j\right\} \\ &= \frac{1}{Kn_T\alpha}\sum_{i=1}^{n} u_i \sum_{j=1}^{K} \mathbb{1}\left\{i \in T_j\right\} \\ &= \frac{1}{n\alpha}\sum_{i=1}^{n} u_i, \end{aligned}$$

where the last line follow from the claim after Assumption 2 ($\frac{1}{K}\sum_{j=1}^{K} \mathbb{1}\left\{l \in T_j\right\} = \frac{n_T}{n}$). This yields using similar arguments as before

$$\mathbb{P}(W_1 - \frac{1}{n\alpha} \geq 2t) \leq 4\exp\left(\frac{-n\alpha t^2}{2(4 + t/3)}\right).$$

Using the fact that $n_T \leq n$ it follows

$$\mathbb{P}(W_1 - \frac{1}{n_T\alpha} \geq 2t) \leq 4\exp\left(\frac{-n\alpha t^2}{2(4 + t/3)}\right). \tag{2.32}$$

Decomposition (2.30) combined with inequalities (2.31), (2.32) leads to

$$\mathbb{P}\left(C_1 - 2Q(n_T, \alpha) \geq 4t\right) \leq 6\exp\left(\frac{-n\alpha t^2}{2(4 + t/3)}\right), \tag{2.33}$$

with

$$\begin{aligned} Q(n, \alpha) &= B(n, \alpha) + \frac{1}{n\alpha} \\ &= \frac{M\sqrt{\mathcal{V}_\mathcal{G}}}{\sqrt{\alpha n}} + \frac{1}{n\alpha}. \end{aligned} \tag{2.34}$$

Using the same technique, one has

$$C_2 \leq 2 \sup_{g \in \mathcal{G}} |\widehat{\mathcal{R}}_\alpha(g, S_n) - \mathcal{R}_\alpha(g)|.$$

Then by Lemma 2.22, we obtain

$$\mathbb{P}\Big(C_2 - 2Q(n,\alpha) \geq 4t\Big) \leq 6 \exp\left(\frac{-n\alpha t^2}{2(4+t/3)}\right). \tag{2.35}$$

Moreover, notice that we have

$$\begin{cases} 4B(n_T,\alpha) \geq 2B(n,\alpha) + 2B(n_T,\alpha), \\ \frac{4}{n_T\alpha} \geq \frac{2}{n_T\alpha} + \frac{2}{n\alpha}. \end{cases}$$

Therefore we get

$$4Q(n_T,\alpha) \geq 2Q(n,\alpha) + 2Q(n_T,\alpha). \tag{2.36}$$

Combining equations (2.33), (2.35), (2.36) and decomposition (2.29) yields

$$\mathbb{P}\Big(\text{Bias} - 4Q(n_T,\alpha) \geq 8t\Big) \leq 12 \exp\left(\frac{-n\alpha t^2}{2(4+t/3)}\right). \tag{2.37}$$

Assembling terms. Using equations (2.27), (2.28), (2.37) and the decomposition (2.11), deduce the inequality

$$\mathbb{P}\left(\left|\widehat{\mathcal{R}}_{CV,\alpha}(\Psi_\alpha, V_{1:K}) - \mathcal{R}_\alpha\Big(\Psi_\alpha(S_n)\Big)\right| - E_{CV}(n_T, n_V, \alpha) \geq 10t\right) \leq 15 \exp\left(\frac{-n\alpha t^2}{2(4+t/3)}\right), \tag{2.38}$$

with

$$\begin{aligned} E_{CV}(n_T, n_V, \alpha) &= B(n_V,\alpha) + 4Q(n_T,\alpha) + \frac{1}{n_T\alpha} \\ &= M\sqrt{\mathcal{V}_\mathcal{G}}\left(\frac{1}{\sqrt{n_V\alpha}} + \frac{4}{\sqrt{n_T\alpha}}\right) + \frac{5}{n_T\alpha}. \end{aligned}$$

The last line follows, using the definitions of B (eq. 2.24) and Q (eq. 2.34).

By inverting inequality (2.38), one has, with probability $1 - 15\delta$,

$$\left|\widehat{\mathcal{R}}_{CV,\alpha}(\Psi_\alpha, V_{1:K}) - \mathcal{R}_\alpha\Big(\Psi_\alpha(S_n)\Big)\right| \leq E_{CV}(n_T, n_V, \alpha) + \frac{20}{3n\alpha}\log\left(\frac{1}{\delta}\right) + 20\sqrt{\frac{2}{n\alpha}\log\left(\frac{1}{\delta}\right)},$$

which is the desired result.

2.B.4 Intermediate results for the proof of Theorem 2.12

Lemma 2.23. *Let Ψ_α be the ERM rule on the tail region of level $1-\alpha$ defined in (2.5). Given a dataset $\mathcal{D}_n = (Z_1, Z_2, \ldots, Z_n) \in \mathcal{Z}^n$ it holds that*

$$\widehat{\mathcal{R}}_\alpha(\Psi_\alpha(S_n), S_n) \leq \widehat{\mathcal{R}}_{CV,\alpha}(\Psi_\alpha, V_{1:K}).$$

In other words the CV *risk estimate of the ERM rule cannot be less than the empirical risk evaluated on the full dataset.*

Proof The argument for $\alpha \neq 1$ is the same as the one for $\alpha = 1$ (standard ERM) which may be found in Kearns and Ron (1999). We reproduce it for the sake of completeness. By definition of $\Psi_\alpha(S_n)$, one has

$$\forall j \in [\![1, n]\!] \, , \, \widehat{\mathcal{R}}_\alpha(\Psi_\alpha(T_j), S_n) \geq \widehat{\mathcal{R}}_\alpha(\Psi_\alpha(S_n), S_n),$$

since $\widehat{\mathcal{R}}_\alpha(g, S_n) = \frac{1}{n}\left(n_V \widehat{\mathcal{R}}_\alpha(g, V_j) + n_T \widehat{\mathcal{R}}_\alpha(g, T_j)\right)$, for $g \in \mathcal{G}$. It follows that

$$\frac{1}{n}\left(n_V \underbrace{\widehat{\mathcal{R}}_\alpha(\Psi_\alpha(T_j), V_j)}_{\text{validation error}} + n_T \underbrace{\widehat{\mathcal{R}}_\alpha(\Psi_\alpha(T_j), T_j)}_{\text{training error}} \right) \geq \frac{1}{n}\Big(n_V \widehat{\mathcal{R}}_\alpha(\Psi_\alpha(S_n), V_j) + n_T \widehat{\mathcal{R}}_\alpha(\Psi_\alpha(S_n), T_j)\Big).$$

Since $\Psi_\alpha(T_j)$ minimizes the training error on the j'th training set T_j, in particular we have

$$\widehat{\mathcal{R}}_\alpha(\Psi_\alpha(S_n), T_j) \geq \widehat{\mathcal{R}}_\alpha(\Psi_\alpha(T_j), T_j),$$

hence

$$\widehat{\mathcal{R}}_\alpha(\Psi_\alpha(S_n), V_j) \leq \widehat{\mathcal{R}}_\alpha(\Psi_\alpha(T_j), V_j) \, , \, \forall j \in [\![1, K]\!]. \tag{2.39}$$

In addition the average empirical risk of $\Psi_\alpha(S_n)$ is equal to the empirical risk on the full dataset, indeed

$$\begin{aligned}
\frac{1}{K}\left(\sum_{j=1}^{K} \widehat{\mathcal{R}}_\alpha(\Psi_\alpha(S_n), V_j) \right) &= \frac{1}{K n_V}\left(\sum_{j=1}^{K} \sum_{i \in V_j} c\Big(\Psi_\alpha(S_n), O_i\Big) \mathbb{1}\left\{ \|X_i\| > \|X_{\left(\lfloor \alpha n \rfloor\right)}\| \right\} \right) \\
&= \frac{1}{K n_V}\left(\sum_{j=1}^{K} \sum_{i=1}^{n} c\Big(\Psi_\alpha(S_n), O_i\Big) \mathbb{1}\Big\{ i \in V_j \Big\} \mathbb{1}\Big\{ \|X_i\| > \|X_{(\lfloor \alpha n \rfloor)}\| \Big\} \right) \\
&= \frac{1}{K n_V}\left(\sum_{i=1}^{n} c\Big(\Psi_\alpha(S_n), O_i\Big) \mathbb{1}\Big\{ \|X_i\| > \|X_{(\lfloor \alpha n \rfloor)}\| \Big\} \right) \sum_{j=1}^{K} \mathbb{1}\Big\{ i \in V_j \Big\} \\
\text{(By Assumption 2)} &= \widehat{\mathcal{R}}_\alpha\left(\Psi_\alpha(S_n), S_n\right).
\end{aligned}$$

Thus by averaging Inequality (2.39), we get

$$\widehat{\mathcal{R}}_\alpha(\Psi_\alpha(S_n), S_n) \leq \widehat{\mathcal{R}}_{CV,\alpha}(\Psi_\alpha, V_{1:K}).$$

■

The following lemma is used in the proof of our second main result concerning *l.p.o.* risk estimation, see Inequality (2.16). It is a generalization of Markov inequality that is particularly useful for cross-validation estimates. Our proof shares similarities with the proof of Theorem 4.1 in Kearns and Ron (1999) formulated under general algorithmic stability assumptions.

Lemma 2.24. *In the setting of Theorem 2.9, we have*

$$\mathbb{P}(\widehat{\mathcal{R}}_{\mathrm{CV},\alpha}(\Psi_\alpha, V_{1:K}) - \mathcal{R}_{CV,\alpha}(\Psi_\alpha, V_{1:K}) \geq t) \leq \frac{\mathbb{E}(\mathrm{D}_{\mathrm{t}_\alpha} + \mathrm{Bias} + |\hat{\mathcal{R}}_\alpha(\Psi_\alpha(S_n), S_n) - \mathcal{R}_\alpha(\Psi_\alpha(S_n))|)}{t},$$

where Bias *(resp* $\mathrm{D}_{\mathrm{t}_\alpha}$*) is defined by equation* (2.14) *(resp* (2.12)*).*

Proof Set $\widehat{\mathcal{R}}_{\mathrm{CV},\alpha} = \widehat{\mathcal{R}}_{\mathrm{CV},\alpha}(\Psi_\alpha, V_{1:K})$, $\mathcal{R}_{CV,\alpha} = \mathcal{R}_{CV,\alpha}(\Psi_\alpha, V_{1:K})$, $\hat{\mathcal{R}}_\alpha = \hat{\mathcal{R}}_\alpha(\Psi_\alpha(S_n), S_n)$, $\mathcal{R}_\alpha = \mathcal{R}_\alpha(\Psi_\alpha(S_n))$.
For any integrable real valued random variable L, and any $t > 0$ write

$$\mathbb{E}\left[L\right] = \mathbb{P}\left(L \geq t\right)\mathbb{E}\left[L \,|\, L \geq t\right] + \mathbb{E}\left[L\mathbb{1}\left\{L < t\right\}\right].$$

Reorganising, we obtain the following generalized Markov inequality,

$$\mathbb{P}\left(L \geq t\right) = \frac{\mathbb{E}\left[L\right] - \mathbb{E}\left[L\mathbb{1}\left\{L < t\right\}\right]}{\mathbb{E}\left[L \,|\, L \geq t\right]} \leq \frac{\mathbb{E}(L) - \mathbb{E}(L\mathbb{1}\left\{L < t\right\})}{t}.$$

Letting $L = \widehat{\mathcal{R}}_{\mathrm{CV},\alpha} - \mathcal{R}_{CV,\alpha}$ we obtain

$$\begin{aligned}\mathbb{P}(\widehat{\mathcal{R}}_{\mathrm{CV},\alpha} - \mathcal{R}_{CV,\alpha} \geq t) = \frac{\mathbb{E}(\widehat{\mathcal{R}}_{\mathrm{CV},\alpha} - \mathcal{R}_{CV,\alpha})}{t} \\ \frac{-\mathbb{E}\left[(\widehat{\mathcal{R}}_{\mathrm{CV},\alpha} - \mathcal{R}_{CV,\alpha})\mathbb{1}\left\{\widehat{\mathcal{R}}_{\mathrm{CV},\alpha} - \mathcal{R}_{CV,\alpha} \leq t\right\}\right]}{t}. \quad (2.40)\end{aligned}$$

Using the fact that $\mathbb{E}(\tilde{\mathcal{R}}_{CV,\alpha} - \mathcal{R}_{CV,\alpha}) = 0$ and that $\mathrm{D}_{\mathrm{t}_\alpha} = |\widehat{\mathcal{R}}_{\mathrm{CV},\alpha} - \tilde{\mathcal{R}}_{CV,\alpha}|$, one gets

$$\begin{aligned}\mathbb{E}(\widehat{\mathcal{R}}_{\mathrm{CV},\alpha} - \mathcal{R}_{CV,\alpha}) &= \mathbb{E}(\widehat{\mathcal{R}}_{\mathrm{CV},\alpha} - \tilde{\mathcal{R}}_{CV,\alpha}) \\ &\leq \mathbb{E}(\mathrm{D}_{\mathrm{t}_\alpha}). \quad (2.41)\end{aligned}$$

Now using lemma 2.23 write

$$\begin{aligned}\mathbb{E}\left[(\mathcal{R}_{CV,\alpha} - \widehat{\mathcal{R}}_{\mathrm{CV},\alpha})\mathbb{1}_{\widehat{\mathcal{R}}_{\mathrm{CV},\alpha} - \mathcal{R}_{CV,\alpha} \leq t}\right] &\leq \mathbb{E}\left[(\mathcal{R}_{CV,\alpha} - \widehat{\mathcal{R}}_\alpha)\mathbb{1}_{\widehat{\mathcal{R}}_{\mathrm{CV},\alpha} - \mathcal{R}_{CV,\alpha} \leq t}\right] \\ &\leq \mathbb{E}\left[|\mathcal{R}_{CV,\alpha} - \widehat{\mathcal{R}}_\alpha|\mathbb{1}_{\widehat{\mathcal{R}}_{\mathrm{CV},\alpha} - \mathcal{R}_{CV,\alpha} \leq t}\right] \\ &\leq \mathbb{E}\left[|\mathcal{R}_{CV,\alpha} - \widehat{\mathcal{R}}_\alpha|\right] \\ &\leq \mathbb{E}\left[|\mathcal{R}_{CV,\alpha} - \mathcal{R}_\alpha|\right] + \mathbb{E}\left[|\mathcal{R}_\alpha - \widehat{\mathcal{R}}_\alpha|\right] \\ &= \mathbb{E}\left[\mathrm{Bias}\right] + \mathbb{E}\left[|\mathcal{R}_\alpha - \widehat{\mathcal{R}}_\alpha|\right]. \quad (2.42)\end{aligned}$$

Where the Bias term in the last line is defined in (2.14). Combining inequality (2.40) with equations (2.41) and (2.42) yields

$$\mathbb{P}(\widehat{\mathcal{R}}_{\mathrm{CV},\alpha} - \mathcal{R}_{CV,\alpha} \geq t) \leq \frac{\mathbb{E}(\mathrm{D}_{\mathrm{t}_\alpha} + \mathrm{Bias} + |\hat{\mathcal{R}}_\alpha - \mathcal{R}_\alpha|)}{t},$$

which concludes the proof. ■

2.B.5 Proof of Theorem 2.12

In view of the discussion following the statement of the theorem (namely the risk decomposition (2.15) and the Markov-type inequality (2.16)) and the bound for the term Bias obtained in (2.37), we only need to obtain bounds for the expectations $\mathbb{E}\Big(|\hat{\mathcal{R}}_\alpha(\Psi_\alpha(S_n), S_n) - \mathcal{R}_\alpha(\Psi_\alpha(S_n))|\Big)$, $\mathbb{E}(\mathrm{D}_{\mathrm{t}_\alpha})$, and $\mathbb{E}(\mathrm{Bias})$. The proof will then be completed by combining together the different terms.

Bounding $\mathbb{E}(\mathrm{D}_{\mathrm{t}_\alpha})$. By Equation (2.27), one has

$$\mathbb{P}(\mathrm{D}_{\mathrm{t}_\alpha} - \frac{1}{n\alpha} \geq t) \leq \exp\left(\frac{-n\alpha t^2}{2(4 + t/3)}\right) .$$

On the one hand, under Assumption 4 one has $\mathbb{P}(\mathrm{D}_{\mathrm{t}_\alpha} - \frac{1}{n\alpha} \geq t) = 0$ for $t \geq 2$. On the other hand, the following inequality holds,

$$\forall t \leq 2\,,\ 2(4 + t/3) \leq 10\,.$$

Hence we may write, for $t \geq 0$,

$$\mathbb{P}(\mathrm{D}_{\mathrm{t}_\alpha} - \frac{1}{n\alpha} \geq t) \leq \exp\left(\frac{-n\alpha t^2}{10}\right) . \tag{2.43}$$

Therefore by Proposition 2.18, we get

$$\begin{aligned}\mathbb{E}(\mathrm{D}_{\mathrm{t}_\alpha}) &\leq \frac{1}{n\alpha} + \frac{M_1}{\sqrt{n\alpha}} \\ &\leq \frac{1}{n_T\alpha} + \frac{M_1}{\sqrt{n_T\alpha}}\,,\end{aligned} \tag{2.44}$$

for some universal constant $M_1 > 0$.

Bounding $\mathbb{E}(\mathrm{Bias})$. Using Eéquation (2.37) and reasoning as in the previous paragraph leads to

$$\mathbb{E}(\mathrm{Bias}) \leq 4Q(n_T, \alpha) + \frac{M_2}{\sqrt{n_T\alpha}}, \tag{2.45}$$

where $Q(n, \alpha)$ is defined by (2.34) and $M_2 > 0$ is a universal constant, independent of $\mathcal{G}, n$ and α.

Bounding $\mathbb{E}(|\hat{\mathcal{R}}_\alpha(\Psi_\alpha(S_n), S_n) - \mathcal{R}_\alpha(\Psi_\alpha(S_n))|)$. By Lemma 2.22 we obtain

$$\mathbb{P}(\Big|\hat{\mathcal{R}}_\alpha(\Psi_\alpha(S_n), S_n) - \mathcal{R}_\alpha(\Psi_\alpha(S_n))\Big| - Q(n, \alpha) \geq t) \leq 3\exp\left(\frac{-n\alpha t^2}{2(4 + t/3)}\right) . \tag{2.46}$$

Then we get

$$\begin{aligned}\mathbb{E}(\Big|\hat{\mathcal{R}}_\alpha(\Psi_\alpha(S_n),&S_n) - \mathcal{R}_\alpha(\Psi_\alpha(S_n))\Big|) \\ &\leq Q(n, \alpha) + \frac{M_3}{\sqrt{n\alpha}} \\ &\leq Q(n_T, \alpha) + \frac{M_3}{\sqrt{n_T\alpha}}\,,\end{aligned} \tag{2.47}$$

for some universal constant $M_3 > 0$.

Combining equations (2.44), (2.45), (2.47) with Lemma 2.24 gives

$$\mathbb{P}(\widehat{\mathcal{R}}_{CV,\alpha}(\Psi_\alpha, V_{1:K}) - \mathcal{R}_{CV,\alpha}(\Psi_\alpha, V_{1:K}) \geq t) \leq \frac{5Q(n_T,\alpha) + \frac{M_4}{\sqrt{n_T\alpha}}}{t} + \frac{1/(n_T\alpha)}{t}\,, \quad (2.48)$$

where $M_4 = M_1 + M_2 + M_3$. The next step is to derive a probability bound for $\mathcal{R}_{CV,\alpha}(\Psi_\alpha, V_{1:K}) - \widehat{\mathcal{R}}_{CV,\alpha}(\Psi_\alpha, V_{1:K})$. We have

$$\begin{aligned}
&\mathbb{P}(\mathcal{R}_{CV,\alpha}(\Psi_\alpha, V_{1:K}) - \widehat{\mathcal{R}}_{CV,\alpha}(\Psi_\alpha, V_{1:K}) - 5Q(n_T,\alpha) \geq 9t) \\
&\qquad \leq \mathbb{P}(\mathcal{R}_{CV,\alpha}(\Psi_\alpha, V_{1:K}) - \widehat{\mathcal{R}}_{\alpha}(\Psi_\alpha(S_n), S_n) - 5Q(n_T,\alpha) \geq 9t) \\
&\qquad \leq \mathbb{P}(\mathcal{R}_{CV,\alpha}(\Psi_\alpha, V_{1:K}) - \mathcal{R}_{\alpha}(\Psi_\alpha(S_n)) - 4Q(n_T,\alpha) \geq 8t) \\
&\qquad\qquad + \mathbb{P}(\mathcal{R}_{\alpha}(\Psi_\alpha(S_n)) - \widehat{\mathcal{R}}_{\alpha}(\Psi_\alpha(S_n), S_n) - Q(n_T,\alpha) \geq t) \\
&\qquad \leq \mathbb{P}(\left|\widehat{\mathcal{R}}_{\alpha}(\Psi_\alpha(S_n), S_n)) - \mathcal{R}_{\alpha}(\Psi_\alpha(S_n))\right| - Q(n_T,\alpha) \geq t) \\
&\qquad\qquad + \mathbb{P}(\text{Bias} - 4Q(n_T,\alpha) \geq 8t) \\
&\text{(By (2.37) + (2.46))} \leq 15\exp\left(\frac{-n\alpha t^2}{2(4+t/3)}\right). \qquad (2.49)
\end{aligned}$$

The first inequality follows from the fact that $\widehat{\mathcal{R}}_{CV,\alpha}(\Psi_\alpha, V_{1:K}) \geq \widehat{\mathcal{R}}_{\alpha}(\Psi_\alpha(S_n), S_n)$ (lemma 2.23). The second inequality is obtained by a union bound. The third inequality follows from the definition of Bias (eq. 2.14). Combining (2.48), (2.49) and that

$$\begin{aligned}
\mathbb{P}(|X| - 5Q(n_T,\alpha) \geq 9t) &\leq \mathbb{P}(X - 5Q(n_T,\alpha) \geq 9t) + \mathbb{P}(-X - 5Q(n_T,\alpha) \geq 9t) \\
&\leq \mathbb{P}(X \geq 9t) + \mathbb{P}(-X - 5Q(n_T,\alpha) \geq 9t),
\end{aligned}$$

leads to

$$\begin{aligned}
&\mathbb{P}\left(\left|\widehat{\mathcal{R}}_{CV,\alpha}(\Psi_\alpha, V_{1:K}) - \mathcal{R}_{CV,\alpha}(\Psi_\alpha, V_{1:K})\right| - 5Q(n_T,\alpha) \geq 9t\right) \\
&\leq \frac{5Q(n_T,\alpha)}{t} + \frac{M_4/\sqrt{n_T\alpha}}{t} + \frac{(1/n_T\alpha)}{t} + 15\exp\left(\frac{-n\alpha t^2}{2(4+t/3)}\right). \qquad (2.50)
\end{aligned}$$

Finally, using (2.15), we get

$$\begin{aligned}
&\mathbb{P}\left(\left|\widehat{\mathcal{R}}_{CV,\alpha}(\Psi_\alpha, V_{1:K}) - \mathcal{R}_{\alpha}\Big(\Psi_\alpha(S_n)\Big)\right| - 9Q(n_T,\alpha) \geq 17t\right) \\
&\leq \mathbb{P}(\text{Bias} - 4Q(n_T,\alpha) \geq 8t) \\
&\quad + \mathbb{P}\left(\left|\widehat{\mathcal{R}}_{CV,\alpha}(\Psi_\alpha, V_{1:K}) - \mathcal{R}_{CV,\alpha}(\Psi_\alpha, V_{1:K})\right| - 5Q(n_T,\alpha) \geq 9t\right) \qquad (2.51) \\
&\leq \frac{5Q(n_T,\alpha) + (M_4/\sqrt{n_T\alpha})}{t} + \frac{1/(n_T\alpha)}{t} + 27\exp\left(\frac{-n\alpha t^2}{2(4+t/3)}\right). \qquad (2.52)
\end{aligned}$$

The last line follows from (2.37) and (2.50). Since for any $t \geq 2$,

$$\mathbb{P}\left(\left|\widehat{\mathcal{R}}_{CV,\alpha}(\Psi_\alpha, V_{1:K}) - \mathcal{R}_{\alpha}\Big(\Psi_\alpha(S_n)\Big)\right| - 9Q(n_T,\alpha) \geq 17t\right) = 0$$

we can restrict our attention to the case $t \leq 2$, for which we have

$$27 \exp\left(\frac{-n\alpha t^2}{2(4+t/3)}\right) \leq 27 \exp\left(\frac{-n\alpha t^2}{10}\right).$$

Using that $\exp(-x) \leq \frac{1}{x}$ for $x \geq 0$, we deduce that

$$27 \exp\left(\frac{-n\alpha t^2}{2(4+t/3)}\right) \leq \frac{270}{n\alpha t^2}.$$

Using the latter inequality and inverting (2.51), we get that with probability $1 - 17\delta$,

$$\left|\widehat{\mathcal{R}}_{CV,\alpha}(\Psi_\alpha, V_{1:K}) - \mathcal{R}_\alpha\Big(\Psi_\alpha(S_n)\Big)\right| \leq 9Q(n_T,\alpha) + \frac{5Q(n_T,\alpha) + (M_4/\sqrt{n_T\alpha}) + (1/n_T\alpha)}{\delta} + \sqrt{\frac{270}{n_T\alpha\delta}},$$

Using the fact that $\sqrt{\frac{1}{\delta}} \leq \frac{1}{\delta}$ (since $\delta \leq 1$), the latter inequality becomes:

$$|\widehat{\mathcal{R}}_{CV,\alpha}(\Psi_\alpha, V_{1:K}) - \mathcal{R}_\alpha\Big(\Psi_\alpha(S_n)\Big)| \leq 9Q(n_T,\alpha) + \frac{1}{\delta\sqrt{n_T\alpha}}(5Q(n_T,\alpha) + M_5) + \frac{1}{\delta n_T\alpha},$$

with $M_5 = M_4 + \sqrt{270}$. Replacing Q with its expression (Equation 2.34) gives the desired result.

2.C Optimal classifier in extreme regions

The aim of this section is to extend the result in Jalalzai et al. (2018) (See Remark 2.1) to a wider class of losses. For the sake of brevity, we only introduce necessary notations, more insights can be found in the original work. Suppose that the conditional distribution of X given $Y = 1$ (*resp.* $Y = -1$) is regularly varying with index 1 and exponent measure μ_+ (*resp.* μ_-) that is : for $A \subset [0,\infty]^d \backslash \{0\}$ a measurable set such that $0 \notin \partial A$ and $\mu_+(\partial A) \neq 0$,

$$b(t)\mathbb{P}\left\{t^{-1}X \in A \mid Y = 1\right\} \underset{t\to\infty}{\longrightarrow} \mu_+(A),$$

for some function $b(t)$ satisfying

$$\forall x \in \mathbb{R}\,, \ \frac{b(tx)}{t} \underset{t\to+\infty}{\longrightarrow} x^{-1}.$$

We denote by $Z_\infty = (X_\infty, Y_\infty)$ an *extreme* observation drawn from the limit measures : $\mathbb{P}(X_\infty \mid Y = 1) = \mu_+$, $\mathbb{P}(X_\infty \mid Y = -1) = \mu_-$ and

$$\mathbb{P}\left(Y_\infty = 1\right) = p_\infty := \lim_{t\to\infty} \mathbb{P}\left(Y = 1 \mid \|X\| \geq t\right).$$

Furthermore, let $\eta_\infty(x) = \mathbb{P}(Y_\infty = 1 \mid X_\infty = x)$, $\mathcal{R}_t(g) = \mathbb{E}\left[c(g,O) \mid \|X\| \geq t\right]$ and $\mathcal{R}_\infty\left(g\right) = \limsup_{t\to\infty} \mathcal{R}_t\left(g\right)$. The function η_∞ is constant along rays (see Jalalzai et al. (2018)) that is $\eta_\infty(tx) = \eta_\infty\left(x\right)$ for all $t \geq 0$, so that $\eta_\infty(x) = \eta_\infty\Big(\Theta\left(x\right)\Big)$. In the

following, we will use the latter property to show that the optimal predictor with respect to $\mathcal{R}_\infty$ is angular (depends only on the angle Θ). Before stating the main result of this section, let's recall a standard assumption in the EVT framework (Jalalzai et al., 2018; Cai et al., 2011).

Assumption 5. *The limiting regression function η_∞ is continuous on $S = \left\{x \in \mathcal{X} \mid \|x\| = 1\right\}$ and*

$$\sup_{\theta \in S} \left|\eta(\Theta(t\theta)) - \eta_\infty(\theta)\right| \underset{t\to\infty}{\longrightarrow} 0,$$

where $\eta(x) = \mathbb{P}(Y = 1 \mid X = x)$ is the standard regression function.

We further assume that :

$$\forall z = (x, y) \in \mathcal{Z} \quad , \quad c(g, z) = \phi(g(x)y),$$

for some function ϕ so that

$$\mathcal{R}_t(g) = \mathbb{E}\left[\eta\left(X\right)\phi\left(g(X)\right) + \left(1 - \eta\left(X\right)\right)\phi\left(-g(X)\right)\right]. \tag{2.53}$$

The next theorem allows to deduce that the optimal classifier g^*_∞ is constant along rays.

Theorem 2.25. *Suppose that $g^* = \arg\min_{g\in\mathbb{R}^{\mathcal{X}}} \mathcal{R}_t(g) = a \circ \eta$ for some function a. Furthermore assume that Assumption 5 holds. Then, if*

$$f(x) = x\phi(a(x)) + (1 - x)\phi(-a(x))$$

is uniformly continuous on Range(η)*, one has,*

$$g^*_\infty = \arg\min_{g\in\mathbb{R}^{\mathcal{X}}} \mathcal{R}_\infty(g) = a \circ \eta_\infty.$$

In other words, the exists a discrimination function g^*_∞ that minimizes the asymptotic risk and depends on the angle of the input only.

Proof First notice that $\mathcal{R}_t(g^*) \leq \mathcal{R}_t(g)$ for all $g \in \mathbb{R}^{\mathcal{X}}$. Thus by taking the limit superior on both sides, it is clear that $g^* = \arg\min_{g\in\mathbb{R}^{\mathcal{X}}} \mathcal{R}_\infty(g)$. It remains to show that

$$\lim_{t\to\infty} \mathcal{R}_t(g^*_\infty) - \mathcal{R}_t(g^*) = 0$$

.

Now using Equation (2.53) write,

$$\begin{aligned}\mathcal{R}_t(g^*_\infty) - \mathcal{R}_t(g^*) &= \frac{\int_{\{\|x\|>t\}} f \circ \eta(x) - f \circ \eta_\infty(x) \mathrm{d}P(x)}{\mathbb{P}\{\|X\| > t\}} \\ &\leq A_t \end{aligned} \tag{2.54}$$

where $A_t = \sup_{\{\|x\|\geq t\}} \left|f \circ \eta(x) - f \circ \eta_\infty(x)\right|$. On the other hand, since η_∞ is constant along rays, Assumption 5 is equivalent to

$$\sup_{\{\|x\|\geq t\}} \left|\eta(x) - \eta_\infty(x)\right| \underset{t\to\infty}{\longrightarrow} 0.$$

Therefore, by using the uniform continuity of f, one obtains

$$A_t = \sup_{\{\|x\| \geq t\}} \left| f \circ \eta(x) - f \circ \eta_\infty(x) \right| \underset{t \to \infty}{\longrightarrow} 0.$$

Combining the last fact with Inequality (2.54) yields the desired result. ■

2.C.1 Examples

We present the main applications of Theorem 2.25, formally we focus on Logistic regression ($c(g, z) = \phi(g(x)y) = \log(1 + \exp\Big(-g(x)y\Big))$) and SVM ($c(g, z) = \phi(g(x)y) = \max\Big(0, 1 - yg(x)\Big)$) losses.

Corollary 2.26 (Logistic Regression)**.** *Suppose that* $0 < \eta_1 \leq \eta(x) \leq \eta_2 < 1$ *for all* $x \in \mathcal{X}$*. Then, under Assumption 5, for the logistic regression loss, one has,*

$$g^*_\infty = \arg\min_{g \in \mathbb{R}^{\mathcal{X}}} \mathcal{R}_\infty(g) = \ln\left(\frac{\eta_\infty}{1 - \eta_\infty}\right).$$

Proof The proof consists on verifying the assumptions of Theorem 2.25 and using the well know fact (see *e.g.* Zhang (2004b)) that, for logistic regression, the bayes predictor is given by $g^* = a \circ \eta$ with

$$a(x) = \ln\left(\frac{x}{1 - x}\right).$$

Furthermore, the function f defined in Theorem 2.25 is given by,

$$\begin{aligned} f(x) &= x\phi(a(x)) + (1 - x)\phi(-a(x)) \\ &= -x \ln(x) - (1 - x) \ln(1 - x) \end{aligned}$$

which is uniformly continuous on $\big[\eta_1, \eta_2\big]$. Thus all conditions of Theorem 2.25 are fulfilled and the proof is complete.

■

Mimicking the latter proof and using results from Bartlett et al. (2006a) yields the following corollary

Corollary 2.27 (Support Vector Machine)**.** *Suppose that Assumption 5 holds, then for the hinge loss (*SVM*), one has,*

$$g^*_\infty = \arg\min_{g \in \mathbb{R}^{\mathcal{X}}} \mathcal{R}_\infty(g) = \operatorname{sign}(2\eta - 1).$$

Proof For SVM, the function f defined in Theorem 2.25 writes as $f(x) = 1 - |2x - 1|$ (see *e.g.* Zhang (2004b)), which is uniformly continuous. The rest of the proof is identical to the previous proof and thus omitted. ■

Chapter 3

Tail Inverse Regression: dimension reduction for prediction of extremes

Contents

3.1 Introduction

Dimension reduction is a crucial matter in supervised learning problems where the goal is to predict a *dependent variable* $Y \in \mathbb{R}$ or summaries of it, when the dimension p of the *covariate vector* $X \in \mathbb{R}^p$ is large. In this chapter we consider dimension reduction for prediction of tail events, by which we mean events of the kind $\{Y > y\}$, for arbitrarily large values of y. This stylized statistical problem relates to a wide range of practical applications such as supervised anomaly detection, system monitoring with a large number of sensors, prediction of extreme weather conditions or financial risk management. For instance, in financial risk management, a typical concern is to identify risk factors, which will be further used to explain extreme events such as financial market crashes, see *e.g.* Fama and French (1993, 2015). Risk factors are often lower dimensional functionals based on a large number of stock returns. Identifying such risk factors that can predict financial market crashes is therefore an example of dimension reduction for the problem of predicting tail events.

Our focus on extreme values connects our work with the field of Extreme Value Theory (EVT) which has been successfully applied to model tail events with potentially catastrophic impact. Statistical inference in this framework is performed using the most extreme realizations of the random variable under consideration. We refer the interested reader to the monographs Beirlant et al. (2006); De Haan and Ferreira (2007); Resnick (2013, 2007). Notice that the curse of dimensionality is particularly troublesome in extreme value analysis where only a small fraction of the data, reflected by the low probability $\mathbb{P}\left(Y > y\right)$, is used for inference. Before proceeding further we remark

that the method proposed in this study, although motivated by and formulated in an EVT framework, does not rely on the minimal assumptions typically required in EVT such as a power law decay. It is in fact a local method related to any small range of Y and as such, it could be easily adapted to tackle the problem of dimension reduction for prediction of Y within low probability regions of other shapes. However in view of the importance of applications towards risk management, we concentrate on this specific tail region.

Dimension reduction in EVT. The subject of dimension reduction for extremes has inspired numerous recent works. The vast majority of them are devoted to the unsupervised setting, *i.e.* analyzing the extremes of a high dimensional random vector. Such studies can be divided into the following categories: clustering methods (Chautru (2015); Chiapino et al. (2019a); Janßen and Wan (2020a)), support identification, (Goix et al. (2016, 2017); Chiapino and Sabourin (2016); Chiapino et al. (2019b); Simpson et al. (2020); Meyer and Wintenberger (2019)), Principal Component Analysis of the angular component of extremes (Cooley and Thibaud (2019a); Jiang et al. (2020); Drees and Sabourin (2021)), and graphical models for extremes (Hitz and Evans (2016); Engelke and Hitz (2020); Asenova et al. (2021)); see also Engelke and Ivanovs (2020) and the references therein.

By contrast, our approach takes place in the supervised setting. Our main informal assumption is that *a low dimensional orthogonal projection PX is sufficient for predicting extreme values of Y*. In other words the extreme values of Y can be entirely explained by a limited number of linear combinations of the components of X. In this setting, the only existing works are, to our best knowledge, Gardes (2018) and Bousebata et al. (2021). In Gardes (2018), the informal assumption emphasized above is made precise by a specific notion of *tail conditional independence*, reported in Equation (3.6) below. Dimension reduction is considered under this condition. Gardes (2018) demonstrates the usefulness of such a reduction for statistical estimation of large conditional quantiles. Even though we follow in the footsteps of Gardes (2018) in terms of informal goal, our framework differs significantly from Gardes (2018)'s on several key aspects. First, the specific definition of tail conditional independence that we propose (See Definition 3.4 in Section 3.3) is not equivalent to Gardes (2018)'s condition (3.6). We carry out an in-depth comparison of both conditions and we show that neither one of them implies the other, in Section B from the supplementary material Aghbalou et al. (2021). Second, our assumption is motivated by a downstream task (predicting the occurrence of a tail event) which is different from, although related to the one motivating Gardes (2018) (estimation of extreme conditional quantiles). Third, the statistical guarantees brought by Gardes (2018) are obtained under the assumption that the dimension reduction space is already known. In the cited reference an estimation method is indeed proposed for the dimension reduction space, however its statistical properties are only analyzed *via* simulations. Instead, we bring statistical guarantees regarding the estimation of a sufficient projection subspace itself. We discuss qualitatively the positive impact it may have for prediction of tail events in Remark 3.5. Lastly, the computational cost of TIREX depends only polynomially on the ambient dimension p, which is not the case with the current estimation method in Gardes (2018), as discussed in Section 3.6.

Another study related to our work is the recently published paper Bousebata et al. (2021), where the authors adopt a partial least square strategy to uncover the relation between linear combinations of covariates and the extreme values of the target. Their

model assumptions differ from ours substantially: the inverse regression model assumed in Bousebata et al. (2021) implies a single-index relationship between extreme values of the response and the covariates. In addition, the model requires regular variation of the dependent variable Y and of the link function. Lastly, the model relies on finite variance of Y. In contrast, our approach is somewhat 'free' from most restrictions on the distribution of (X, Y) except from the well-known linearity condition and constant variance condition, typically needed for SIR. Such conditions concern only the distribution of the covariates. Since we do not impose regular variation, we can handle not only thin-tailed but also extremely heavy-tailed dependent variables with no finite variance or even mean.

Sufficient Dimension Reduction and inverse methods. The underlying assumption of a sufficient linear projection subspace has been formalized under the notion of Sufficient Dimension Reduction (SDR) space (Cook (2009)). Many classical approaches to supervised dimension reduction rely on a linear regression model between X and Y. This is the case *e.g.* for Principal component regression (Hotelling (1957)), Partial least squares (Wold (1966)), Canonical correlation analysis (Thompson (1984)) or penalized methods with sparsity inducing regularization such as the Lasso (Jenatton et al. (2011)). Differently, *SDR* builds upon a *linear dimension reduction* assumption: only a small number of *linear* combinations of covariates is useful for predicting the dependent variable. In other words, there exists a linear subspace E (an SDR space) of a moderate dimension $d \leq p$, such that

$$\mathbb{P}\left(Y \leq t \,\middle|\, X\right) = \mathbb{P}(Y \leq t | PX), \qquad \forall t \in \mathbb{R}, \quad \text{almost surely}, \tag{3.1}$$

where P is the orthogonal projector on E, *i.e.* Y depends on X only through $PX \in \mathbb{R}^d$. This framework relies heavily on the notion of conditional independence Dawid (1979); Constantinou and Dawid (2017): Condition (3.1) characterizes the fact that Y is conditionally independent from X given PX. One major advantage of this approach is that it strikes a balance between interpretability of the dimension reduction based on linear operations and flexibility of the generative model – no assumption is made regarding the dependence structure between PX and Y.

Under the assumption that there exists a non trivial subspace E such that (3.1) holds, a natural idea is to estimate such a subspace first, and then use only the variable PX to predict Y, thus reducing the dimensionality of the regression problem. The estimation problem based on SDR can also be viewed as a specific case of semi-parametric M-estimation (Delecroix et al. (2006)). Alternatively, one may consider derivative based methods, relying on the fact that the gradient of the regression curve belongs to E (Härdle and Stoker (1989); Hristache et al. (2001); Xia et al. (2007); Dalalyan et al. (2008)). Recently, the framework of Reproducing Kernel Hilbert Spaces (RKHS) has been employed to estimate SDR spaces by means of covariance operators (Fukumizu et al. (2004, 2009)).

The family of methods to which our work relates most is the inverse regression paradigm initiated by Li (1991), including the Sliced Inverse Regression (SIR) strategy and its second order variant Sliced Average Variance Estimate (SAVE) (Cook and Weisberg (1991)). The main idea underlying these methods is that under appropriate assumptions the inverse regression curve $\mathbb{E}\left[X|Y\right]$ and its second moment variant – the columns of the conditional covariance matrix $\mathrm{Var}X|Y$ – almost surely belong to the minimal SDR. Cumulative slicing estimation (CUME), proposed in Zhu et al. (2010) and further

analyzed in Portier (2016), aims at recovering the largest possible subspace of the minimal SDR. It is achieved by estimating the conditional expectation and variance of X, conditioning on 'slices' of the target Y, in the form of $\mathbb{1}\{Y < y\}$, and then aggregating such conditional expectations and variances by integration with respect to y.

A well-known restriction of the SIR strategy is that it relies on a so-called *linearity condition* (LC) regarding the covariates, namely equation (3.2) in the next section, see Hall and Li (1993) for a justification. The required condition is satisfied in particular if the covariates form an elliptical random vector or are independent (Cook (2009); Eaton (1986)). There are various extensions of SIR permitting to overcome this restriction. Using RKHS, it has been proposed to transform the data in a way that LC is approximately satisfied (Wu (2008); Yeh et al. (2008)). Another possibility allowing to depart from elliptical covariates is to apply the SIR methodology and its higher order variants to score functions of the explanatory variables (Babichev et al. (2018)). Finally, the high dimensional case $p > n$ calls for regularization methods which permit in addition to perform feature selection (Li and Yin (2008)). All these extensions are out of the scope of the present work, in which we restrict ourselves to the original SIR and SAVE methods, thus leaving room for several improvement in further works. For estimation purposes we consider a variant of CUME.

Contributions and outline. Our contributions are twofold. First, we develop in Section 3.3 a modified version of Gardes (2018)'s probabilistic setting regarding tail conditional independence. In particular we explain in Remark 3.5 the relevance of our definition for the purpose of predicting tail events and its connections to the statistical learning framework of imbalanced classification. We discuss thoroughly the distinctions between the two alternative definitions for tail conditional independence in Section 3.3.2, where we also provide examples of models satisfying one or the other. Second, we show in Section 3.4 that our definition permits to extend inverse regression principles and methods to this extreme values setting (theorems 3.11, 3.16). We derive an asymptotic analysis for our proposed estimation strategy TIREX stemming from inverse regression, using specific tools from the theory of empirical processes (Section 3.5). We illustrate the finite sample performance of TIREX with simulated and real world data sets in Section 3.6, in particular we demonstrate empirically the usefulness of TIREX for tail events prediction. The code developed for TIREX is available online [1] and some technical proofs and additional comments are deferred to the supplementary material Aghbalou et al. (2021).

We start-off in Section 3.2 by recalling the necessary background regarding conditional independence of random variables, SDR spaces, and inverse regression.

3.2 Background: dimension reduction space and Sliced Inverse Regression

Conditional independence of random variables Y and V given W is defined *e.g.* in Constantinou and Dawid (2017) as follows: the conditional distribution of Y given (V, W) is the same as the conditional distribution of Y given W, almost surely. Several char-

[1] https://github.com/anassag/TIREX

acterizations are recalled below, the equivalence of which are proved in Constantinou and Dawid (2017), Proposition 2.3.

Definition 3.1 (conditional independence)**.** *Let Y, V, W be random variables defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and taking values in arbitrary measure spaces. The variables Y and V are called conditionally independent given W, a property denoted by $Y \perp\!\!\!\perp V \mid W$, if the equivalent conditions below are satisfied.*

(CI-1) For all $A_Y \in \sigma(Y)$, $\mathbb{P}\left(Y \in A \,\middle|\, V, W\right) = \mathbb{P}\left(Y \in A \,\middle|\, W\right)$, almost surely.

(CI-2) For all real-valued functions f and g, measurable and bounded,

$$\mathbb{E}\left[f(Y)g(V) \,\middle|\, W\right] = \mathbb{E}\left[f(Y) \,\middle|\, W\right] \mathbb{E}\left[g(V) \,\middle|\, W\right], \quad a.s.$$

(CI-3) For any real-valued function g, measurable and bounded,

$$\mathbb{E}\left[g(V) \,\middle|\, Y, W\right] = \mathbb{E}\left[g(V) \,\middle|\, W\right], \quad a.s.$$

Notice that the existence of regular versions of conditional probability distributions is not required in Definition 3.1. However in this paper, Y is real valued, thus the existence of such a regular version for the conditional distribution of Y given (V, W) is granted. As a consequence we may write, without additional precautions, expressions of the kind '$\mathbb{P}\left(Y \in A \,\middle|\, V = v, W = w\right)$'. The latter quantity is defined as the value of the conditional probability kernel at point $((v, w), A)$.

In the context of supervised dimension reduction, we consider $V = X$ and search for a projection $W = PX$ of X on a lower dimensional subspace E satisfying the above conditions. We assume for simplicity that the covariance matrix $\Sigma = \mathrm{Cov} X$ is invertible and for ease of presentation we introduce a standardized covariate vector $Z = \Sigma^{-1/2}(X - m)$ where $m = \mathbb{E}\left[X\right]$. We consider in the remaining of this paper the problem of regressing Y on Z, which amounts to assuming that both m and Σ are known, so that the vector Z is observed. Thus $\mathrm{Var}\left[Z\right] = I_p$ and $\mathbb{E}\left[Z\right] = 0$. An SDR space (Cook (2009); Cook and Ni (2005)) is a subspace E of $\mathbb{R}^p$ such that $Y \perp\!\!\!\perp Z \,\middle|\, PZ$ where P is the orthogonal projection on E, which is equivalent to condition (3.1) in the introduction. Our results easily extend to general covariates X (see *e.g.* Cook and Weisberg (1991)) at the price of an additional notational burden. Notice already that in terms of non-standardized covariates X, a subspace $\tilde{E}$ of $\mathbb{R}^p$ with associated orthogonal projector $\tilde{P}$ is an SDR space for the pair (X, Y) if and only if $\tilde{E} = \Sigma^{-1/2}E$ where E is an SDR space for Z.

A central space is an SDR subspace E_c for the pair (Z, Y) of minimal dimension. In our context of finite dimensional covariates a central space always exists since the ambient space $\mathbb{R}^p$ itself is an SDR space. Uniqueness is not guaranteed in general but holds true under mild assumptions ensuring that an intersection of SDR spaces is an SDR space (see *e.g.* Portier and Delyon (2013), Theorem 1). In such a case one may refer without ambiguity to *the* central space.

First and second order inverse methods, respectively named SIR (Li (1991)) and SAVE (Cook and Weisberg (1991)) are two of many methods to estimate SDR spaces. Both rely on the fact that under appropriate assumptions detailed below, first and second moments of the covariate vector, conditioning upon the target, belong to an SDR space. In the sequel, E is an SDR space, and P denotes the orthogonal projection on E. Then

$Q = I_p - P$ is the orthogonal projection on $E^\perp$, the orthogonal complement of E. The required conditions are the Linearity Condition (LC):

$$\mathbb{E}\left[Z \mid PZ\right] = PZ \quad \text{a.s.} \tag{3.2}$$

and the additional Constant Conditional Variance (CCV),

$$\text{Var}\left[Z \mid PZ\right] \text{ is constant} \quad \text{a.s.} \tag{3.3}$$

Under both LC and CCV, we have that $\mathbb{E}\left[\text{Var}\left[Z|PZ\right]\right] = \mathbb{E}\left[ZZ^T\right] - \mathbb{E}\left[PZ(PZ)^T\right] = I_p - P$ and therefore the constant matrix in (3.3) is necessarily the projection $Q = I_p - P$ on the orthogonal complement of E.

Notice that LC and CCV depend on an unknown SDR space. Assuming that LC holds for all orthogonal projectors is in fact equivalent to assuming that the covariate vector Z is spherically symmetric, *i.e.* $Z = \rho U$ where $\rho \perp\!\!\!\perp U$, ρ is a non negative random variable and U is uniformly distributed over the unit sphere of $\mathbb{R}^p$, as proved in Eaton (1986). Among spherical variables with finite second moment, CCV is equivalent to being Gaussian ((Bryc, 2012, Theorem 4.1.4)).

The following proposition in Li (1991) encapsulates the main idea of SIR. We give below the (classical) proof for the sake of completeness.

Proposition 3.2 (SIR principle)**.** *If E is an SDR space for which LC (3.2) is satisfied, then $Q(\mathbb{E}\left[Z|Y\right]) = 0$ a.s., that is, $\mathbb{E}\left[Z|Y\right] \in E$ a.s.*

Proof By the tower rule from conditional expectation,

$$\begin{aligned}\mathbb{E}\left[Z \mid Y\right] &= \mathbb{E}\left[\mathbb{E}\Big(Z \,|\, Y, PZ\Big) \,\Big|\, Y\right] = \mathbb{E}\left[\mathbb{E}(Z \,|\, PA) \,\Big|\, Y)\right] \\ &= \mathbb{E}\left[PZ \mid Y\right] = P\,\mathbb{E}\left[Z \mid Y\right]\end{aligned}$$

where the second equality comes from conditional independence and the third one follows from the linearity condition (3.2). Thus $Q\mathbb{E}\left[Z|Y\right] = 0$. ■

The SIR method advocated first by Li (1991) consists in estimating first conditional expectations $C_h = \mathbb{E}\left[Z \,\middle|\, Y \in I(h)\right]$, $h = 1, \ldots, H$, where $I(h), h = 1, \ldots, H$ are called slices and form a partition of the sample range of Y (or the support of the density function if Y is continuous). From Proposition 3.2, those estimates lie in the vicinity of the SDR space. Next, performing a Principal Component Analysis (PCA) on the C_h's, one obtains a good approximation of E. More precisely, the SIR estimate of E is given by the largest eigenvectors associated to the SIR matrix,

$$M_{\text{SIR}} = \sum_{h=1}^{H} p_h^{-1} C_h C_h^T,$$

where $p_h = \mathbb{P}\left(Y \in I(h)\right)$; see Li (1991). Various estimation procedures of SDR spaces are proposed in Cook and Ni (2005); Zhu et al. (2010). In the latter reference, the matrix

$$M_{\text{CUME}} = \mathbb{E}\Big[m(Y)m(Y)^T\Big], \tag{3.4}$$

with $m(y) = \mathbb{E}\left[Z\mathbb{1}\{Y \leq y\}\right]$, is introduced as an alternative to the SIR matrix. One advantage of this approach is that the slicing parameter h is no longer needed. In addition the estimate of the matrix M_{CUME} benefits from the aggregating effect of the expectation sign which is typically associated with variance reduction.

A pitfall of SIR is that it is not guaranteed that the C_h's span the entire space E, so that SIR may be inconsistent. This may happen in particular when the regression function $\mathbb{E}\left[Y|Z\right]$ admits some symmetry properties (Li, 1991, Remark 4.5), a phenomenon referred to as the SIR pathology. In this case, Li (1991) and Cook and Weisberg (1991) recommend to use higher order moments such as the conditional variance of Z given Y to obtain a second order matrix with wider range. This second order method requires that CCV (3.3) is satisfied in addition to LC, in which case the following result holds. Here and throughout, span(M) stands for the column space of matrix M.

Proposition 3.3 (SAVE principle)**.** *If E is an SDR space for which LC (3.2) and CCV (3.3) are satisfied, then*

$$Q\Big(\mathbb{E}\Big[ZZ^\top \,|\, Y\Big] - I_p\Big) = 0 \qquad a.s.,$$

in other words $\text{span}\Big(\mathbb{E}[ZZ^\top \,|\, Y] - I_p) \subset E \quad a.s.$

Proof We reformulate here the arguments of Cook and Weisberg (1991) in our notational framework for convenience. An immediate consequence of assumptions (3.2) and (3.3) is that $\mathbb{E}\left[ZZ^T|PZ\right] = Q + PZZ^TP$. From a conditioning argument and the conditional independence assumption, $\mathbb{E}\left[ZZ^T|Y\right] = Q + P\mathbb{E}\left[ZZ^T|Y\right]P$. Rearranging gives $\mathbb{E}\left[ZZ^T|Y\right] - I_p = P(\mathbb{E}\left[ZZ^T|Y\right] - I_p)P$, thus $Q(\mathbb{E}\left[ZZ^T|Y\right] - I_p) = 0$. ■

Notice that Propositions 3.2 and 3.3 together imply that $Q(\text{Var}\left[Z \,\middle|\, Y\right] - I_p) = 0$. Finally for estimation purpose the extension of the CUME method to the second order framework is termed CUVE (cumulative variance estimation) by Zhu et al. (2010). In the case of standardized covariates, it consists in estimating the matrix $M_{\text{CUVE}} = \mathbb{E}\Big[W(Y)W(Y)^\top\Big]$, where $W(y) = \text{Var}\left[Z\mathbb{1}\{Y \leq y\}\right] - F_Y(y)I_p$ is a second order moments matrix which column space is included in $\tilde{E}$. The latter fact is obtained by a slight modification of the argument leading to the SAVE principle.

3.3 Tail conditional independence, Extreme SDR space

3.3.1 Definition for Tail Conditional Independence

The focus on the largest values of the target variable Y suggests to weaken the classical definition of conditional independence, so that the equivalent conditions (CI-1)-(CI-3) hold only for Y exceeding a high threshold tending to its right endpoint. Namely, in a similar (but different) manner as in Gardes (2018) we define tail conditional independence as a variant of condition (CI-1) from Definition 3.1. In the sequel the right endpoint (*i.e.* the supremum) of the support of the random variable Y is denoted by

y^+. The limits as $y \to y^+$ as understood as the limits as $y \to y^+, y < y^+$. We assume that $\mathbb{P}\big(Y > y\big) \to 0$ as $y \to y^+$, in particular we exclude the case of point masses at y^+.

Definition 3.4 (Tail Conditional Independence (TCI))**.** *Let Y, V, W be random variables defined on $(\Omega, \mathcal{F}, \mathbb{P})$. We assume that Y is real valued, Borel measurable, while V and W take their values in arbitrary measure spaces. We say that Y is tail conditionally independent from V given W and write $Y_\infty \perp\!\!\!\perp V|W$, if*

$$\frac{\mathbb{E}\Big|\mathbb{P}\big(Y > y \;\big|\; V, W\big) - \mathbb{P}\big(Y > y \;\big|\; W\big)\Big|}{\mathbb{P}\big(Y > y\big)} \xrightarrow[y \to y^+]{} 0. \tag{3.5}$$

Contrary to conditional independence, tail conditional independence is not symmetric: $Y_\infty \perp\!\!\!\perp V \;\big|\; W$ does not imply that $V_\infty \perp\!\!\!\perp Y \;\big|\; W$.

In Gardes (2018)'s work, tail conditional independence is defined in a somewhat more technical manner, see Definition 1 from the cited reference. However a necessary condition (see Equation (2) in that paper) is the almost sure convergence of the $\sigma(V, W)$-measurable ratio,

$$\frac{\mathbb{P}\big(Y > y \;\big|\; V, W\big) - \mathbb{P}\big(Y > y \;\big|\; W\big)}{\mathbb{P}\big(Y > y \;\big|\; W\big)} \xrightarrow[y \to y^+]{} 0, \quad \text{a.s.} \tag{3.6}$$

In the sequel we refer to our notion of tail conditional independence defined in (3.5) as TCI, while we write TCI-G to refer to L. Gardes' condition (3.6). Both definitions are motivated by similar but different downstream tasks, namely prediction of extreme values for TCI in connection to the AM risk criterion (see Remark 3.5 below), versus estimation of large conditional quantiles (see Section 3.1 in Gardes (2018)).

In Subsection 3.3.2 below we work out a generic example where TCI holds and on this occasion, we discuss briefly the differences between TCI and TCI-G. In order not to interrupt the flow of ideas a more thorough comparison between the two definitions is relegated to the supplementary (Section B).

In practice TCI allows for an extension of the SIR framework to handle extreme values (Section 3.4). Whether it is possible to obtain a similar extension with TCI-G is an open question. We conjecture a negative answer because our Tail Inverse Regression principles theorems 3.11, 3.16 rely on a specific consequence of TCI, namely Property (iii) from Proposition 3.7 below. In spirit our definition for TCI and the subsequent Tail inverse regression framework developed in Section 3.4 below is compatible with the main notions underlying graphical models for extremes (Engelke and Hitz (2020)) and One component regular variation (Hitz and Evans (2016)). These connections are further detailed in Remarks 3.12 and 3.13 from Section 3.4.

Meanwhile the next remark sheds light on the relevance of the proposed definition of TCI for statistical learning applications.

Remark 3.5 (TCI and Imbalanced Classification)**.** *Predicting exceedances over arbitrarily high thresholds y may be viewed as a family of binary classification problems indexed by y. Indeed for fixed y, consider the binary target $T = \mathbb{1}\{Y > y\}$ with marginal class probability $\pi = \pi_y = \mathbb{P}\big(Y > y\big)$. The goal is thus to predict T, by means of the covariate vector $X = (V, W)$ where $V \in \mathbb{R}^{p-d}, W \in \mathbb{R}^d$. As $y \to y^+$, $\pi_y \to 0$. This is a typical instance of* class imbalance*, a well documented potential issue in binary classification which has been the subject of several works in the statistical learning literature,*

see e.g. *the recent papers Menon et al. (2013b) or Xu et al. (2020b) and the references therein. A classifier is a binary function* h *defined on* $\mathbb{R}^p$. *Given a family of candidate classifiers* $h \in \mathcal{H}$ *the goal is to select a 'good' candidate based on a training set and an appropriate notion of a theoretical risk and its empirical counterpart. When* π *is so close to zero that the probability of a classification error* $\mathbb{P}\Big(h(X) \neq T, T = 1\Big)$ *is negligible compared with* $\mathbb{P}\Big(h(X) \neq T, T = 0\Big)$, *the traditional* $0-1$ *risk* $R(h) = \mathbb{P}\Big(h(X) \neq T\Big)$ *is driven by the latter term and tends to favor the trivial classifier* $h \equiv 0$. *One standard approach aiming at granting more importance to the minority class when required by the application context (*e.g. *if the event* $\{T = 1\}$, *although rare, has an overwhelming impact) is to consider the* Arithmetic Mean Risk *(AM risk in short), see* e.g. *Menon et al. (2013b),*

$$\mathcal{R}_{AM}(h) = \frac{1}{2}\left[\mathbb{P}\left(h(X) = 1 \,\middle|\, T = 0\right) + \mathbb{P}\left(h(X) = 0 \,\middle|\, T = 1\right)\right]. \tag{3.7}$$

Generalizations to arbitrary weight vectors are considered in Xu et al. (2020b). In a dimension reduction context consider the classes

$$\mathcal{H} = \{h : \mathbb{R}^p \to \{0,1\},\ \textit{measurable w.r.t. } \mathcal{B}(\mathbb{R}^p)\}\ ,$$
$$\mathcal{H}_W = \{h \in \mathcal{H} :\ \forall (v,w) \in \mathbb{R}^{p-d} \times \mathbb{R}^d,\ h(v,w) = \tilde{h}(w),\ \tilde{h}\ \textit{is measurable w.r.t. } \mathcal{B}(\mathbb{R}^d)\}.$$

Let us refer to the classification problem attached respectively to $\mathcal{H}$ *and* $\mathcal{H}_W$ *as the* full problem *and the* reduced problem. *The Bayes classifier for each problem are respectively minimizers of the AM risk over the full family* $\mathcal{H}$ *and the reduced one* $\mathcal{H}_W$,

$$h^* \in \arg\min_{h \in \mathcal{H}} \mathcal{R}_{AM}(h)\ ; \qquad h^*_W \in \arg\min_{h \in \mathcal{H}_W} \mathcal{R}_{AM}(h).$$

The main ingredient of the subsequent analysis are the regression functions $\eta(x) = \mathbb{P}\left(T = 1 \,\middle|\, X = x\right)$ *and* $\eta_W(w) = \mathbb{P}\left(T = 1 \,\middle|\, W = w\right)$. *A modification of standard arguments (see the supplementary material, Section A) yields explicit expressions for the Bayes classifiers* $h^*(x) = \mathbb{1}\Big\{\eta(x) > \pi\Big\}$, $h^*_W(x) = \mathbb{1}\Big\{\eta_W(w) > \pi\Big\}$. *In addition the Bayes risks are*

$$\begin{aligned} \mathcal{R}_{AM}(h^*) &= \mathbb{E}\left[\min\left(\frac{\eta(X)}{\pi}, \frac{1-\eta(X)}{1-\pi}\right)\right]\ ; \\ \mathcal{R}_{AM}(h^*_W) &= \mathbb{E}\left[\min\left(\frac{\eta_W(W)}{\pi}, \frac{1-\eta_W(W)}{1-\pi}\right)\right]. \end{aligned} \tag{3.8}$$

Because $\mathcal{H}_W \subset \mathcal{H}$ *we must have* $\mathcal{R}_{AM}(h^*_W) \geq \mathcal{R}_{AM}(h^*)$. *The difference between the two may be seen as a bias term: the price to pay for dimension reduction. Indeed for any random choices* $\widehat{h} \in \mathcal{H}$, $\widehat{h}_W \in \mathcal{H}_W$, *which are typically the outputs of a statistical learning algorithm applied respectively to the full covariate space and the reduced one, the excess risk for the reduced problem decomposes as*

$$\mathcal{R}_{AM}(\widehat{h}_W) - \mathcal{R}_{AM}(h^*) = \underbrace{\mathcal{R}_{AM}(\widehat{h}_W) - \mathcal{R}_{AM}(h^*_W)}_{A} + \underbrace{\mathcal{R}_{AM}(h^*_W) - \mathcal{R}_{AM}(h^*)}_{B}\ .$$

The first term (A) in the right-hand side is the excess risk stemming from the particular choice of the learning algorithm, which typically increases with the dimension of the input W. *In particular when* $p-d$ *is large, the excess risk term A will be typically less than its*

counterpart in the full problem $\mathcal{R}_{AM}(\widehat{h})-\mathcal{R}_{AM}(h^*)$. *The second term (B) is the bias term above mentioned. The bias-variance compromise is in favour of dimensionality reduction* via *projection on the second variable* W *whenever* $A+B \leq \mathcal{R}_{AM}(\widehat{h})-\mathcal{R}_{AM}(h^*)$.

We now derive an upper bound on the bias term B *which is closely connected to our definition of* TCI. *Notice that for any finite set* $\mathcal{X}$ *and any pair of real functions* (f,g) *it holds that* $|\min_{x\in\mathcal{X}} f(x) - \min_{x\in\mathcal{X}} g(x)| \leq \max_{x\in\mathcal{X}} |f(x)-g(x)|$. *This, combined with* (3.8) *above and Jensen inequality, implies that*

$$\begin{aligned} B = \mathcal{R}_{AM}(h^*_W) - \mathcal{R}_{AM}(h^*) &\leq \mathbb{E}\left| \min\left(\frac{\eta_W(W)}{\pi}, \frac{1-\eta_W(W)}{1-\pi}\right) - \min\left(\frac{\eta(X)}{\pi}, \frac{1-\eta(X)}{1-\pi}\right)\right| \\ &\leq \mathbb{E}\left\{ \max\left(\frac{\eta(X)-\eta_W(W)}{\pi}, \frac{(1-\eta(X))-(1-\eta_W(W))}{1-\pi}\right)\right\} \\ &= \mathbb{E}\left|\frac{\eta(X)-\eta_W(W)}{\pi}\right|, \end{aligned} \tag{3.9}$$

where the latter identity holds whenever $\pi \leq 1/2$. *Now, with* $T = \mathbb{1}\{Y > y\}$,

$$\mathbb{E}\left|\frac{\eta(X)-\eta_W(W)}{\pi}\right| = \frac{\mathbb{E}\left|\mathbb{P}\left(Y>y \mid V,W\right) - \mathbb{P}\left(Y>y \mid W\right)\right|}{\mathbb{P}\left(Y>y\right)}.$$

One recognizes the TCI *criterion in the latter expression. Thus* TCI *means that the bias term* B *vanishes as* $y \to y^+$, *so that* projection on W is relevant for the problem of predicting the rare event $\{Y > y\}$, *for large values of* y. *The cut-off value* y *above which* $\mathcal{R}_{AM}(\hat{h}_W) \leq \mathcal{R}_{AM}(\hat{h})$, *that is* $A+B \leq \mathcal{R}_{AM}(\hat{h})-\mathcal{R}_{AM}(h^*)$ *(in expectation or with high probability), depends on two main factors: (i) the rate of convergence of* B_y *to zero and (ii) the sensitivity of the learning algorithm to the curse of dimensionality for a given sample size. Indeed both excess risks* $\mathcal{R}_{AM}(\hat{h})-\mathcal{R}_{AM}(h^*)$ *and* $\mathcal{R}_{AM}(\hat{h}_W)-\mathcal{R}_{AM}(h^*_W)$ *typically converge to zero (in expectation or in probability) with the sample size, at a different rate which depends on the respective dimensions* p, d. *Precise quantification of this cut-off point for specific learning algorithms and finite sample sizes is outside the scope of the present work and left for future research.*

3.3.2 Examples and discussion

In this section we provide a generic example based on a mixture model where the TCI condition (3.5) is satisfied under mild assumptions. We discuss an alternative additive model in Remark 3.6. We consider several particular instances of the generic mixture model and on this occasion we discuss the similarities and differences between TCI and the TCI-G condition (3.6) proposed in Gardes (2018). Some technical proofs are deferred to Section B in the supplementary material, as well as additional comments, examples and counter-examples allowing for a better understanding of the differences between the two definitions.

Our leading example is constructed as follows: Let the target Y be distributed according to a mixture

$$Y = BY_1 + (1-B)Y_2\,,$$

where B is a Bernoulli variable with parameter $\theta \in (0,1)$, and Y_1, Y_2 are real variables defined through their conditional survival functions

$$S_1(y,V) = \mathbb{P}\left(Y_1 > y \mid V\right)\,, \qquad S_2(y,W) = \mathbb{P}\left(Y_2 > y \mid W\right).$$

Here, the covariate variables V, W are respectively valued in $\mathbb{R}^{p-d}$ and $\mathbb{R}^d$ with marginal distributions that we denote by P_V and P_W. The full covariate vector is $X = (V, W) \in \mathbb{R}^p$. We assume that the variables (B, V, W) are independent. Notice that independence between V and W ensures that the Linearity Condition and Constant Conditional Variance condition are automatically satisfied. In this context, straightforward calculations (as detailed in the supplementary material, Section B) show that

$$\frac{\left|\mathbb{P}\left(Y > y \mid V, W\right) - \mathbb{P}\left(Y > y \mid W\right)\right|}{\mathbb{P}\left(Y > y\right)} = \frac{\theta(S_1(y, V) - S_1(y))}{\theta S_1(y) + (1-\theta)S_2(y)}.$$

The TCI condition is that the expectancy of the above ratio vanishes as $y \to y^+$ and it is not difficult to imagine several models for (Y_1, V) and (Y_2, W) for which it is the case, as exemplified below.

Remark 3.6 (Variant: additive model)**.** *The mixture model described here is by no means the only option to construct examples of variables*(Y, V, W) *satisfying the* TCI *assumption. Another natural example is an additive model* $Y = Y_1 + Y_2$*, where* Y_1 *and* Y_2 *are respectively driven by* V *and* W*, while* Y_1 *has lighter tails than* Y_2*. The mathematical derivations are somewhat more intricate because convolutions are involved instead of sums of distribution functions. However special cases can be worked out. In the supplementary material we consider* $Y_1 = V \in \mathbb{R}$*,* $Y_2 = W\zeta \in \mathbb{R}$ *where* ζ *is heavy-tailed and* V, W *have a compact support which is bounded away from* 0 *and we show that* TCI *holds. More general statements might be obtained using results regarding sums of regularly varying random variables (Jessen and Mikosch (2006)). We leave this question to further works.*

As an example in the generic mixture model described above, consider the case where Y_1 and Y_2 are themselves defined as multiplicative mixtures

$$Y_1 = \sum_{i=1}^{p-d} M_i^{(1)} V_i \epsilon_i \,, \qquad Y_2 = \sum_{j=1}^{d} M_j^{(2)} W_j \zeta_j \,, \tag{3.10}$$

where $M^1 = (M_1^1, \ldots, M_{p-d}^1)$ is a multinomial vector with weight parameter $\pi^1 = (\pi_1^1, \ldots, \pi_{p-d}^1)$, that is $\sum_{i=1}^{p-d} M_i^1 = 1$ and $\mathbb{P}(M_i^1 = 1) = \pi_i^1$; M_2 is as well a multinomial variable with parameter $\pi^2 = (\pi_1^2, \ldots, \pi_d^2)$; and the variables ϵ_i, $i \leq p-d$ and ζ_j, $j \leq d$ are multiplicative noises, with different tail behaviour. Assume for simplicity that all ϵ_j's (*resp.* ζ_j's) share the same survival function S_ϵ (*resp.* S_ζ) and that for all $s, t > 0$,

$$\lim_{y \to \infty} S_\epsilon(s^{-1}y)/S_\zeta(t^{-1}y) = 0. \tag{3.11}$$

Condition (3.11) is satisfied *e.g.* with Pareto noises, $S_\epsilon(y) = y^{-\alpha_1}, S_\zeta(y) = y^{-\alpha_2}$ with $\alpha_1 > \alpha_2 > 0$, or with Exponential versus Pareto noises, $S_\epsilon(y) = e^{-\alpha_1 y}, S_\zeta(y) = y^{-\alpha_2}$, $\alpha_1, \alpha_2 > 0$. The random vectors $M^1, M^2, \epsilon, \zeta, V, W$ are independent. Finally the covariate vectors V and W are made of independent components V_i, W_j, with nonnegative, bounded support included in an interval $[a, b]$ with $0 \leq a < b < \infty$.

In this generic example, Y_1 has a lighter tail than Y_2, so that it is the main risk factor regarding large values of Y, and it is intuitively desirable for a formal definition of tail conditional independence to be such that Y is tail conditionally independent from V given W here.

We now consider two special cases regarding the marginal distributions of the covariates V_j, W_j. recall that $[a, b]$ contains the support of each V_i and each W_j.

(i) As a first go assume that $a > 0$. Then both TCI and TCI-G hold. The proof is deferred to the supplementary material, Section B.4.

(ii) Assume now that $a = 0$, more specifically that each variable V_j, W_j follows a binary Bernoulli distribution with parameter $\tau \in (0, 1)$ (the choice of a common τ merely simplifies the notations). In Section B.5 from the supplementary material we show that TCI-G does not hold, while TCI does.

Notice that the difference between the two cases concerns only the marginal distribution of the covariate, namely whether $\mathbb{P}\Big(W_j = 0\Big) > 0$ is key. This seemingly minor variation results in fact in potential failure of TCI-G, while TCI remains true. The main conclusions of our comparison between the two definitions (TCI and TCI-G) in the supplementary material, Section B, may be summarized as follows.

1. Neither condition implies the other in general, except for discrete covariates where TCI-G implies TCI.

2. TCI-G criterion concerns the additional information brought by V regarding the probability of the event $Y > y$, *after* conditioning on W. The criterion is satisfied if the additional information is negligible, for *all* possible values $W = w$, even those values such that the conditional distribution of Y given $W = w$ is shorter tailed than the marginal distribution of Y. Indeed TCI-G is primarily designed for quantile regression, and the focus is not on the tail of Y's distribution, but instead on the tails of the conditional distributions of Y given W. This is the informal reason why TCI-G is not satisfied in the example above, Case (ii).

3. In constrast TCI is designed for prediction of extreme values of Y. It is an integrated version of TCI-G with respect to the variable (V, W), with a weight function granting more importance to w's such that the ratio $\mathbb{P}\left(Y > y \,\middle|\, W = w\right) / \mathbb{P}\left(Y > y\right)$ is large as $y \to y^+$. In words, TCI is comparatively more sensitive to values w such that the conditional probability given $W = w$ of an exceedance $Y > y$ is large.

3.3.3 Technical consequences of TCI, parallel with traditional conditional independence

Definition 3.4 implies equivalent weak formulations of the traditional conditions (CI-1,CI-2,CI-3) reviewed in the background section.

Proposition 3.7. *If $Y_\infty \perp\!\!\!\perp V \,\middle|\, W$ in the sense of Definition 3.4, then the following equivalent conditions* (i), (ii), (iii) *hold.*

(i) For any real-valued functions g and h, measurable and bounded, we have

$$\frac{\mathbb{E}\Big[g(V)h(W)\Big(\mathbb{P}\left(Y > y \,\middle|\, V, W\right) - \mathbb{P}\left(Y > y \,\middle|\, W\right)\Big)\Big]}{\mathbb{P}\left(Y > y\right)} \xrightarrow[y \to y^+]{} 0.$$

(ii) *For any real-valued functions g and h, measurable and bounded, we have*

$$\frac{\mathbb{E}\Big[\, h(W)\Big(\, \mathbb{E}\left[\mathbb{1}\{Y>y\}g(V) \;\middle|\; W\right] - \mathbb{E}\left[\mathbb{1}\{Y>y\} \;\middle|\; W\right]\mathbb{E}\left[g(V) \;\middle|\; W\right] \,\Big)\Big]}{\mathbb{P}\left(Y>y\right)} \xrightarrow[y\to y^+]{} 0.$$

(iii) *For any real-valued functions g and h, measurable and bounded, we have*

$$\frac{\mathbb{E}\Big[\, h(W)\mathbb{1}\{Y>y\}\,\Big(\mathbb{E}\left[g(V) \;\middle|\; Y,W\right] - \mathbb{E}\left[g(V) \;\middle|\; W\right]\Big)\Big]}{\mathbb{P}\left(Y>y\right)} \xrightarrow[y\to y^+]{} 0.$$

Remark 3.8 (Relevance of Proposition 3.7 for our purpose)**.** *From a technical perspective, Property (iii) in Proposition 3.7 is key to obtain the tail analogues of the SIR and SAVE principles (Theorems 3.11, 3.16 in Section 3.4). This is not surprising insofar as the traditional condition (CI-3) for conditional independence in Definition 3.1 is central to prove the SIR/SAVE principles.*

Whether the converse implication from Proposition 3.7 holds true in general, i.e. *whether Conditions (i), (ii), (iii) imply* TCI *remains an open question which is not directly relevant for our purposes and thus left for future works.*

Proof [Proof of Proposition 3.7]

We first show the equivalence (ii)⇔(iii) by proving that the left-hand sides of the two conditions are identical. Indeed if g and h are bounded and measurable, then

$$\begin{aligned}\mathbb{E}\Big[h(W)\mathbb{1}\{Y>y\}\mathbb{E}\left[g(V) \;\middle|\; Y,W\right]\Big] =& \mathbb{E}\Big[h(W)\mathbb{1}\{Y>y\}g(V)\Big]\\ =& \mathbb{E}\Big[h(W)\mathbb{E}\left[\mathbb{1}\{Y>y\}g(V) \;\middle|\; W\right]\Big],\end{aligned}$$

while

$$\mathbb{E}\Big[h(W)\mathbb{1}\{Y>y\}(\mathbb{E}\left[g(V) \;\middle|\; W\right])\Big] = \mathbb{E}\Big[h(W)\mathbb{E}\left[\mathbb{1}\{Y>y\} \;\middle|\; W\right]\mathbb{E}\left[g(V) \;\middle|\; W\right]\Big].$$

To show that (ii)⇒(i), note that

$$\begin{aligned}\mathbb{E}\Big[g(V)h(W)\mathbb{E}\left[\mathbb{1}\{Y>y\} \;\middle|\; V,W\right]\Big] &= \mathbb{E}\Big[g(V)h(W)\mathbb{1}\{Y>y\}\Big]\\ &= \mathbb{E}\Big[h(W)\mathbb{E}\left[g(V)\mathbb{1}\{Y>y\} \;\middle|\; W\right]\Big]\\ &= \mathbb{E}\Big[h(W)\mathbb{E}\left[g(V) \;\middle|\; W\right]\mathbb{E}\left[\mathbb{1}\{Y>y\} \;\middle|\; W\right]\Big] + r_1(y)\\ &= \mathbb{E}\Big[g(V)h(W)\mathbb{E}\left[\mathbb{1}\{Y>y\} \;\middle|\; W\right]\Big] + r_1(y),\end{aligned}$$

where $\lim_{y\to y^+} r_1(y)/\mathbb{P}\left(Y>y\right) = 0$ by Condition (ii).

The argument for the converse implication (ii)⇐(i) is similar:

$$\begin{aligned}\mathbb{E}\left[h(W)\mathbb{1}\{Y>y\}g(V)\right] &= \mathbb{E}\Big[h(W)\mathbb{E}\left[\mathbb{1}\{Y>y\} \;\middle|\; V,W\right]g(V)\Big]\\ &= \mathbb{E}\Big[h(W)\mathbb{E}\left[\mathbb{1}\{Y>y\} \;\middle|\; W\right]g(V)\Big] + r_2(y),\end{aligned}$$

where $\lim_{y\to y^+} r_2(y)/\mathbb{P}\left(Y > y\right) = 0$ under condition (i).

Finally we show that Property (i) from Proposition 3.7 is satisfied under the TCI assumption from Definition 3.4. Let g, h be bounded, measurable functions defined on $\mathcal{V}, \mathcal{W}$ respectively and let $\|g\|_\infty$ and $\|h\|_\infty$ denote their supremum norm. By Jensen's inequality,

$$\begin{aligned}
&\mathbb{P}\left(Y > y\right)^{-1} \left|\mathbb{E}\left[g(V)h(W)\left(\mathbb{P}\left(Y > y \mid V, W\right) - \mathbb{P}\left(Y > y \mid W\right)\right)\right]\right| \\
&\leq \|g\|_\infty \|h\|_\infty \mathbb{P}\left(Y > y\right)^{-1} \mathbb{E}\left|\mathbb{P}\left(Y > y \mid V, W\right) - \mathbb{P}\left(Y > y \mid W\right)\right|,
\end{aligned}$$

where the right hand side tends to zero under Condition (3.5) from Definition 3.4.

■

3.3.4 Extreme dimension reduction spaces

In the context of statistical regression, we now define extreme sufficient dimension reduction subspaces in a similar fashion to the usual SDR spaces.

Definition 3.9 (Extreme SDR space and extreme central space)**.**

- *An extreme SDR space for the pair (Z, Y) is a subspace E_e of $\mathbb{R}^p$ such that $Y_\infty \perp\!\!\!\perp Z \mid P_e Z$, where P_e is the orthogonal projection on E_e. In other words E_e is called an extreme SDR space whenever*

$$\mathbb{E}\left|\frac{\mathbb{P}(Y > y|Z) - \mathbb{P}(Y > y|P_e Z)}{\mathbb{P}\left(Y > y\right)}\right| \xrightarrow[y\to y^+]{} 0. \tag{3.12}$$

- *An extreme central space $E_{e,c}$ for the pair (Z, Y) is an extreme SDR space of minimum dimension.*

Investigating sufficient conditions ensuring uniqueness of an extreme central space is left for future studies. Instead, in the present manuscript we shall consider *an* extreme SDR space E_e and we shall show that under appropriate assumptions, inverse extreme regression objects, namely limits of conditional expectations $\mathbb{E}\left[Z \mid Y > y\right]$ (Theorem 3.11) and second order variants (Theorem 3.16) belong to E_e. In particular they belong to any extreme *central* space.

Remark 3.10 (Relationship between the central space and its extreme counterpart)**.** *Because Equation (3.12) holds true for any $y \in \mathbb{R}$ when E_e is chosen as a (non extreme) SDR space for the pair (Z, Y), any SDR space for (Z, Y) is an extreme SDR space. Thus, upon uniqueness of the central space E_c and the extreme central space $E_{e,c}$, it holds that $E_{e,c} \subset E_c$. Examples of other dimension reduction subspaces more specific than E_c but not related to the extreme value of Y include the central mean subspace (Cook and Li, 2002) and the central quantile subspace Christou (2020).*

3.4 Tail Inverse Regression

In the sequel, we consider an extreme SDR space $E_e \subset \mathbb{R}^p$ for the pair (Z, Y) in the sense of Definition 3.9. That is, we assume that $Y_\infty \perp\!\!\!\perp Z \mid P_e Z$ as in Definition 3.4, where P_e is the orthogonal projection on E_e. Also we define $Q_e = I_p - P_e$. In order to adapt the SIR strategy to this tail conditional independence framework, we show the following result which is a 'tail version' of the SIR principle (Proposition 3.2). In the remainder of this chapter let $\|\cdot\|$ denote any norm on a finite dimensional vector space.

Theorem 3.11 (TIREX1 principle)**.** *Assume the following conditions regarding the pair (Z, Y) and the extreme SDR space E_e.*

1. *(Uniform integrability):*
   The random variables $g_{1,A}(Z) = \|Z\| \mathbb{1}\Big\{\|Z\| > A\Big\}$, $g_{2,A}(Z) = \mathbb{E}\left[\|Z\| \mathbb{1}\Big\{\|Z\| > A\Big\} \;\middle|\; P_e Z\right]$ indexed by $A \in \mathbb{R}$ satisfy

$$\lim_{A \to \infty} \limsup_{y \to y^+} \mathbb{E}\left[g_{k,A}(Z) \;\middle|\; Y > y\right] = 0, \qquad k = 1, 2\,; \tag{3.13}$$

2. *(LC) The standardized vector Z satisfies the linearity condition (3.2) relative to P_e;*

3. *(Convergence of conditional expectations) For some $\ell \in \mathbb{R}^p$,*

$$\mathbb{E}\left[Z \mid Y > y\right] \xrightarrow[y \to y^+]{} \ell. \tag{3.14}$$

Then $\ell \in E_e$.

Proof We need to show that $Q_e \ell = 0$. By continuity of the projection operator Q_e it is enough to show that $Q_e \mathbb{E}\left[Z \mid Y > y\right] \to 0$ as $y \to y^+$. On the other hand the linearity condition (LC) (3.2) ensures that $Q_e \mathbb{E}\left[Z \mid P_e Z\right] = Q_e P_e Z = 0$ almost surely. Thus letting $p_y = \mathbb{P}(Y > y)$ one may write

$$\begin{aligned} Q_e \mathbb{E}\left[Z \mid Y > y\right] &= p_y^{-1} Q_e \mathbb{E}\Big[Z \mathbb{1}\{Y > y\}\Big] \\ &= p_y^{-1} \mathbb{E}\Big[\Big(Q_e \mathbb{E}\Big[Z | P_e Z, Y\Big] - Q_e \mathbb{E}\left[Z \mid P_e Z\right]\Big) \mathbb{1}\{Y > y\}\Big], \end{aligned}$$

because the second term of the difference inside the expectation of the second line is zero.

Let $A > 0$ and consider separately the case when $Z \le A$ and $Z > A$, so that

$$\begin{aligned} & Q_e \mathbb{E}\Big[Z \mathbb{1}\{Y > y\}\Big] \\ &= Q_e \mathbb{E}\Big[\Big(\mathbb{E}\Big[Z \mathbb{1}\Big\{\|Z\| \le A\Big\} \;\Big|\; P_e Z, Y\Big] - \mathbb{E}\Big[Z \mathbb{1}\Big\{\|Z\| \le A\Big\} \;\Big|\; P_e Z\Big]\Big) \mathbb{1}\{Y > y\}\Big] \\ &\quad + Q_e \mathbb{E}\Big[\Big(\mathbb{E}\Big[Z \mathbb{1}\Big\{\|Z\| > A\Big\} \;\Big|\; P_e Z, Y\Big] - \mathbb{E}\Big[Z \mathbb{1}\Big\{\|Z\| > A\Big\} \;\Big|\; P_e Z\Big]\Big) \mathbb{1}\{Y > y\}\Big]. \end{aligned}$$

For the first term of the above display, using Condition (ii) of Proposition 3.7 with $h = 1$ and $g(Z) = Z\mathbb{1}\Big\{\|Z\| < A\Big\}$, we obtain that

$$p_y^{-1}\, Q_e \mathbb{E}\Bigg[\Big(\mathbb{E}\left[Z\mathbb{1}\Big\{\|Z\| \le A\Big\} \;\middle|\; P_e Z, Y\right] - \mathbb{E}\left[Z\mathbb{1}\Big\{\|Z\| \le A\Big\} \;\middle|\; P_e Z\right]\Big)\mathbb{1}\{Y > y\}\Bigg] \xrightarrow[y\to y^+]{} 0. \tag{3.15}$$

For the second term corresponding to $Z > A$, we use that $\|Q_e z\| \le \|z\|$, the Jensen inequality and the triangular inequality, which yields

$$\begin{aligned}
&\Bigg\|Q_e \mathbb{E}\Bigg[\Big(\mathbb{E}\left[Z\mathbb{1}\Big\{\|Z\| > A\Big\} \;\middle|\; P_e Z, Y\right] - \mathbb{E}\left[Z\mathbb{1}\Big\{\|Z\| > A\Big\} \;\middle|\; P_e Z\right]\Big)\mathbb{1}\{Y > y\}\Bigg]\Bigg\| \\
&\le \mathbb{E}\Big(\{\mathbb{E}\left[\|Z\|\mathbb{1}\Big\{\|Z\| > A\Big\} \;\middle|\; P_e Z, Y\right] + \mathbb{E}\left[\|Z\|\mathbb{1}\Big\{\|Z\| > A\Big\} \;\middle|\; P_e Z\right]\}\mathbb{1}\{Y > y\}\Big) \\
&= \mathbb{E}\left[g_{1,A}(Z)\mathbb{1}\{Y > y\}\right] + \mathbb{E}\left[g_{2,A}(Z)\mathbb{1}\{Y > y\}\right]
\end{aligned}$$

By (3.15) and the previous decomposition, we have shown that

$$\limsup_{y\to y^+} \|Q_e \mathbb{E}\left[Z|Y > y\right]\| \le \limsup_{y\to y^+} \mathbb{E}\left[g_{1,A}(Z) \;\middle|\; Y > y\right] + \limsup_{y\to y^+} \mathbb{E}\left[g_{2,A}(Z)|Y > y\right].$$

By further letting $A \to \infty$, by Assumption (3.13), the right-hand side is arbitrarily small. This shows that $\lim_{y\to y^+} Q_e \mathbb{E}\left[Z \;\middle|\; Y > y\right] = 0$ and the proof is complete.

■

Remark 3.12 (special case: Tail conditional distribution)**.** *A particular framework justifying the existence of the limit ℓ (Condition (3.69) in the statement of Theorem 3.11) is the following. Assume that the covariate Z admits a* tail conditional distribution *given Y, in the sense that the distribution of Z conditional to $Y > y$ converges as $y \to y^+$. In other words assume that there is a probability distribution μ, that we may call the* tail conditional distribution *of Z given Y, such that for all bounded, continuous function g defined on $\mathbb{R}^p$,*

$$\mathbb{E}\left[g(Z) \;\middle|\; Y > y\right] \xrightarrow[y\to y^+]{} \mu(g) := \int_{\mathbb{R}^p} g \,\mathrm{d}\mu.$$

By virtue of Proposition 2.20 in Van der Vaart (1998), if the uniform integrability condition (3.13) is satisfied regarding the functions $g_{1,A}$ and if Z admits a tail conditional distribution μ relative to Y, then it holds that

$$\mathbb{E}\left[Z \;\middle|\; Y > y\right] \xrightarrow[y\to y^+]{} m_\mu := \int z d\mu(z),$$

so that condition (3.69) automatically holds with $\ell = m_\mu$.

Remark 3.13 (relationships with graphical models for extremes)**.** *The above notion of tail conditional distribution reveals a connection between the present work and graphical modeling approaches in EVT. Namely, assuming a tail conditional distribution of Z given Y, and requiring in addition that the random variable Y is regularly varying, is equivalent to assuming* one-component regular variation *of the pair (Y, Z), a concept first introduced by Hitz and Evans (2016). See in particular their Theorem 1.4, where the pair (X, Y) plays the role of the pair (Y, Z) in the present work.*

The notion of conditional independence at extreme levels also plays an important role in Engelke and Hitz (2020). However our work departs significantly from the latter, in so far as the general context in the cited reference is that of unsupervised learning. All considered variables play a symmetric role –there is no target variable nor covariate –, and they rely on an assumption of joint multivariate regular variation of the considered random vector which is by no means necessary in our context.

Remark 3.14 (Special case: extreme central space)**.** *Upon uniqueness of the extreme central space $E_{e,c}$, under the assumptions of Theorem 3.11 we obtain that $\ell \in E_{e,c}$.*

Remark 3.15 (Sufficient condition for uniform integrability)**.** *Using the fact that for any $\varepsilon > 0$, $\mathbb{1}\Big\{\|Z\| > A\Big\} \leq \|Z\|^{\varepsilon}/A^{\varepsilon}$, a sufficient condition for the uniform integrability condition (3.13) is that*

$$\limsup_{y \to y^+} \frac{\mathbb{E}\left[\|Z\|^{1+\varepsilon}\, \mathbb{1}\{Y > y\}\right]}{\mathbb{P}\left(Y > y\right)} < \infty,$$

for some $\varepsilon > 0$.

A natural strategy in view of Theorem 3.11 is to consider empirical counterparts of the conditional expectations $\mathbb{E}\left[Z \,\middle|\, Y > y\right]$ for large values of y so as to estimate the limit value ℓ, which belongs to any extreme SDR space. Asymptotic statistical guarantees for this approach are derived in Section 3.5. However an obvious limitation of Theorem 3.11 is that it recovers a single direction within an extreme SDR space, namely the line $\{t\ell, t \in \mathbb{R}\}$ in the case where $\ell \neq 0$. If a unique extreme central space exists and if this subspace is one dimensional, then indeed the generated line and the extreme central space coincide. To consider situations where the minimum dimension of an extreme SDR space is greater than one, we develop an extreme analogue of the SAVE framework by considering conditional second order moments. The main result justifying this approach is encapsulated in Theorem 3.16 below.

Theorem 3.16 (TIREX2 principle)**.** *Assume (Z, Y) and the extreme SDR space E_e satisfy the assumptions of Theorem 3.11 and that in addition,*

1. *(second order uniform integrability):*

$$\lim_{A \to \infty} \limsup_{y \to y^+} \mathbb{E}\left[h_{k,A}(Z) \,\middle|\, Y > y\right] = 0, \qquad k = 1, 2\,, \tag{3.16}$$

   where $h_{1,A}(Z) = \|Z\|^2 \mathbb{1}\Big\{\|Z\| > A\Big\}$ and $h_{2,A}(Z) = \mathbb{E}\left[\|Z\|^2 \mathbb{1}\Big\{\|Z\| > A\Big\} \,\middle|\, P_e Z\right]$ for $A \in \mathbb{R}$;

2. *(CCV) The standardized vector Z satisfies the constant variance condition (3.3) relative to P_e;*

3. *(Convergence of conditional expectations) For some $S \in \mathbb{R}^{p \times p}$,*

$$\mathbb{E}\Big[ZZ^\top \,|\, Y > y\Big] \xrightarrow[y \to y^+]{} S + \ell\ell^\top. \tag{3.17}$$

Then $\operatorname{span}(S - I_p) \subset E_e$, i.e. $Q_e(S - I_p) = 0$.

Notice that the existence of $\ell = \lim \mathbb{E}\left[Z \,\middle|\, Y > y\right]$ is part of the assumptions of Theorem 3.11 and that in the latter framework, $Q_e \ell \ell^\top = 0$. Thus condition (3.71) is equivalent to requiring that $\mathrm{Var}\left[Z \,\middle|\, Y > y\right]$ converges to some limit variance S as $y \to y^+$ and the conclusion can be rephrased as $Q_e(\,\mathbb{E}[ZZ^\top \,|\, Y > y\,] - I_p\,) \to 0$ as $y \to y^+$, or equivalently $Q_e(\mathrm{Var}\left[Z \mid Y > y\right] - I_p) \to 0$. The technique of the proof is similar to that for Theorem 3.11. The key is to observe that the Constant Conditional Variance assumption allows to introduce a difference $(\,\mathbb{E}[ZZ^\top \,|\, P_e Z, Y] - \mathbb{E}[ZZ^\top \,|\, P_e Z]\,)\mathbb{1}\{Y > y\}$ which is asymptotically negligible because of the TCI assumption. The details are gathered in the supplement, Section C.

3.5 Estimation

This section is devoted to the statistical implementation of our main results from Section 3.4. Theorems 3.11 and 3.16 show that the quantities ℓ and S in the limits of the two statements are key to estimate the extreme SDR space, because $\ell \in E_e$ and $\mathrm{span}(S - I_p) \subset E_e$. A natural first idea would be to use as an estimate an empirical version of the quantities $\mathbb{E}[Z \,|\, Y > y\,]$ or $\mathbb{E}[ZZ^\top \,|\, Y > y\,]$ for a high threshold y growing with the sample size n. A typical choice of such a threshold is the quantile of Y at a probability level $1 - k/n$, where $k = k(n)$ is an intermediate sequence such that $k(n) \to \infty$ and $k(n)/n \to 0$ as $n \to \infty$. Here we propose a refinement of this strategy integrating out the latter quantities over varying quantiles at probability levels $1 - uk/n$ for $u \in (0, 1)$. Such a refinement follows the proven approaches based on the CUME and CUVE matrices described in the background section 3.2. For this purpose we introduce and prove the asymptotic normality of the empirical processes associated with the first and second order method, that are respectively the specialisation of the SIR/CUME and the SAVE/CUVE processes to extreme regions of the target Y.

Even though the first order method is potentially less fruitful than the second order one since the limit ℓ in Theorem 3.11 is a single vector, it helps build the intuition about the statistical theory for both the first order and second order methods. In addition, the first order method turns out to be more stable in some of our experiments.

Some notations are introduced in Section 3.5.1. We provide asymptotic theory for the first and second order empirical processes in Section 3.5.2. Section 3.5.3 summarizes the methods we suggest for estimating E_e.

3.5.1 Framework and notations

For any right-continuous cumulative distribution function H (be it empirical or not), we shall denote by H^- the left-continuous inverse of H, $H^-(u) = \inf\{x \in \mathbb{R} \,:\, H(x) \geq u\}$. Recall that with these conventions, for $u \in [0, 1]$ and $x \in \mathbb{R}$, we have

$$H(x) \geq u \iff x \geq H^-(u). \tag{3.18}$$

For any i.i.d. sample $(T_i)_{i \leq n}$ associated with a real random variable T with cumulative distribution H, we use the standard definition of the empirical distribution function,

$$\hat{H}(x) = n^{-1} \sum_{i=1}^{n} \mathbb{1}\{T_i \leq x\}. \tag{3.19}$$

For notational and mathematical convenience we shall work with the negative target $\tilde{Y} = -Y$ and denote the *c.d.f.* of $\tilde{Y}$ as F , which we assume to be continuous in the remainder of this chapter. For simplicity we write k instead of $k(n)$ for the intermediate sequence defined at the beginning of this section, as is customary in extreme value statistics. Consider the first order and second order inverse regression functions $C_n(u), B_n(u)$,

$$C_n(u) = \frac{n}{k}\mathbb{E}\left[Z\mathbb{1}\left\{\tilde{Y} < F^{-}(uk/n)\right\}\right] , \tag{3.20}$$

$$B_n(u) = \frac{n}{k}\mathbb{E}\left[(ZZ^\top - I_p)\mathbb{1}\left\{\tilde{Y} < F^{-}(uk/n)\right\}\right] . \tag{3.21}$$

The empirical versions of (3.20) and (3.21) based on an independent sample (Z_i, Y_i) identically distributed as the pair (Z, Y) are

$$\hat{C}_n(u) = \frac{1}{k}\sum_{i=1}^{n} Z_i\mathbb{1}\left\{\tilde{Y}_i \le \hat{F}^{-}(uk/n)\right\}, \tag{3.22}$$

$$\hat{B}_n(u) = \frac{1}{k}\sum_{i=1}^{n}(Z_iZ_i^\top - I_p)\mathbb{1}\left\{\tilde{Y}_i \le \hat{F}^{-}(uk/n)\right\}. \tag{3.23}$$

Extensions to the more realistic situation where the pair (X, Y) is observed with the mean and covariance of X unknown are gathered in Section E from the supplementary .

3.5.2 Main result

The remainder of this section aims at establishing the weak convergence of the (tail) empirical processes associated with TIREX, respectively $\sqrt{k}(\hat{C}_n(u)-C_n(u))$ and $\sqrt{k}(\hat{B}_n(u)-B_n(u))$ in the space of bounded functions $\ell^\infty([0,1])$. This is achieved in Corollary 3.18.

A key point of our analysis, which follows from the continuity of F, is that the functions $C_n(u), B_n(u)$ and their estimates $\hat{C}_n(u), \hat{B}_n(u)$ are invariant under the transformation $U = F(\tilde{Y})$. More precisely, with the latter notation, we have the following identities

$$C_n(u) = \frac{n}{k}\mathbb{E}\Big[\, Z\mathbb{1}\Big\{U < uk/n\Big\}\,\Big] \,, B_n(u) = \frac{n}{k}\mathbb{E}\Big[\,(ZZ^\top - I_p)\mathbb{1}\Big\{U < uk/n\Big\}\,],$$

and the processes $\hat{C}_n(u), \hat{B}_n(u)$ remain the same when constructed from the initial sample $(X_i, \tilde{Y}_i)$ or when constructed from the uniformized sample (X_i, U_i). Indeed for $u \in [0,1]$, it holds that

$$\mathbb{1}\left\{\tilde{Y}_i \le \hat{F}^{-}(uk/n)\right\} = \mathbb{1}\left\{U_i \le \hat{F}_U^{-}(uk/n)\right\}, \qquad \text{a.s.},$$

where $\hat{F}_U$ is the empirical distribution function associated with the uniform sample $U_1, \ldots, U_n$, see Fact D.1 in the supplementary material for a short proof. These facts are a known feature of rank based estimators; see for instance Fermanian et al. (2004) in the copula estimation context and Portier (2016) in the standard SIR context.

We now state our main result which is formulated in terms of a generic random pair (V, Y), an i.i.d. sample thereof $(V_i, Y_i), i \le n$, and a measurable function $h : \mathbb{R}^r \to \mathbb{R}^q$, where Y is the response variable as above, the covariate V is a random vector of size

$r \in \mathbb{N}^*$ and h is such that the random vector $h(V)$ has finite second moments. Define

$$D_n(u) = \frac{n}{k}\mathbb{E}\left[h(V)\mathbb{1}\left\{\tilde{Y} < F^-(uk/n)\right\}\right],$$
$$\hat{D}_n(u) = \frac{1}{k}\sum_{i=1}^{n} h(V_i)\mathbb{1}\left\{\tilde{Y}_i \leq \hat{F}^-(uk/n)\right\}.$$

Weak convergence of the TIREX1 and TIREX2 processes (Corollary 3.18) is obtained upon setting $V = Z$ and respectively $h_C(z) = z$ and $h_B(z) = \mathrm{vec}(zz^\top - I_p)$, where for any matrix $M \in \mathbb{R}^{r\times s}$, $\mathrm{vec}(M)$ denotes the vector of size $r \times s$ obtain by concatenating the columns of M.

Theorem 3.17 (Tail empirical process for a generic pair (V, Y))**.** *Suppose that the distribution function F of $\tilde{Y} = -Y$ is continuous and that, letting $U = F(\tilde{Y})$, it holds that*

1. *for any $j \in \{1, \ldots, q\}$, the functions $u \mapsto \mathbb{E}\left[h(V)_j\mathbb{1}\{U \leq u\}\right]$ and $u \mapsto \mathbb{E}\left[h(V)_j^2\mathbb{1}\{U \leq u\}\right]$ are differentiable on $(0, 1)$ with a continuous derivative at 0,*
2. *for all $M \geq 0$, $S(M) := \lim_{\delta\to 0}\mathbb{E}\left[h(V)h(V)^\top\mathbb{1}\left\{\|V\| \geq M\right\} \,\middle|\, U \leq \delta\right]$ exists and is such that $\lim_{M\to\infty} S(M) = 0$,*
3. *as $\delta \to 0$, $\mathbb{E}\left[h(V) \,\middle|\, U \leq \delta\right]$ converges to a limit $\nu \in \mathbb{R}^q$.*

Then we have as $n \to \infty, k \to \infty, k/n \to 0$,

$$\left\{\sqrt{k}(\widehat{D}_n(u) - D_n(u))\right\}_{u\in[0,1]} \rightsquigarrow \left\{W(u)\right\}_{u\in[0,1]},$$

where W is a Gaussian process with mean zero and covariance function

$$(s, t) \mapsto s \wedge t\left(\Xi - \nu\nu^\top\right), \tag{3.24}$$

with ν as in the 3^{textrd} Condition of the statement and

$$\Xi = S(0) = \lim_{\delta\to 0}\mathbb{E}\left[h(V)h(V)^\top \,\middle|\, U \leq \delta\right] \in \mathbb{R}^{q\times q}. \tag{3.25}$$

Corollary 3.18 (Weak convergence of the TIREX1 and TIREX2 processes)**.** *By choosing the pair $(V, Y) = (Z, Y)$ and assuming that the function $h_C(z) = z$ (*resp. *$h_B(z) = \mathrm{vec}(zz^\top - I_p)$) satisfies the assumptions of Theorem 3.17, the TIREX1 process $\sqrt{k}(\hat{C}_n(u) - C_n(u))$ (*resp. *the TIREX2 process $\sqrt{k}(\hat{B}_n(u) - B_n(u))$ converges weakly in $\ell^\infty(0, 1)$ to a tight Gaussian process W_C (*resp. *W_B) with covariance function given by (3.24) with $V = Z$ and $h = h_C$ (*resp. *$h = h_B$)*

Proof [Proof of Theorem 3.17]

Consider the pseudo-empirical version of $D_n(u)$,

$$\widetilde{D}_n(u) = k^{-1}\sum_{i=1}^{n} h(V_i)\mathbb{1}\left\{U_i \leq uk/n\right\} = k^{-1}\sum_{i=1}^{n} h(V_i)\mathbb{1}\left\{\tilde{Y}_i \leq F^-(uk/n)\right\}. \tag{3.26}$$

Notice that $\widetilde{D}_n$ is not observed but serves as an intermediate quantity through the following key identity:

$$\hat{D}_n(u) = \widetilde{D}_n\left(\frac{n}{k}\hat{F}_U^-(uk/n)\right),$$

where $\hat{F}_U$ is the empirical *c.d.f.* associated with the sample $(U_i, i \leq n)$. Introducing the process

$$\widetilde{\Gamma}(u) = \sqrt{k}\left(\widetilde{D}_n(u) - D_n(u)\right), \quad u \in [0,1], \tag{3.27}$$

we have the following decomposition

$$\sqrt{k}(\hat{D}_n(u) - D_n(u)) = \widetilde{\Gamma}\left(\frac{n}{k}\,\hat{F}^-\left(uk/n\right)\right) + \sqrt{k}\left(D_n\left(\frac{n}{k}\,\hat{F}_U^-\left(uk/n\right)\right) - D_n(u)\right). \tag{3.28}$$

In the remainder of the proof, we show that the first term can be replaced by $\widetilde{\Gamma}(u)$, while the second term can be replaced by $-\nu\hat{\gamma}_1(u)$ where $\hat{\gamma}_1$ is the tail empirical process for uniform random variables,

$$\hat{\gamma}_1(u) = \sqrt{k}\left(\frac{n}{k}\hat{F}_U(uk/n) - u\right). \tag{3.29}$$

Finally we show that the process $(\hat{\gamma}_1(u), \widetilde{\Gamma}(u))_{u\in[0,1]}$ converges jointly to a Gaussian process.

Intermediate results, uniform tail processes

The main tools that we use in our proof of Theorem 3.17 concern the weak convergence of the tail empirical (quantile) process associated with a uniform response variables. Many approaches have been considered to handle the behavior of such processes, see Csorgo et al. (1986) for general empirical processes and Einmahl and Mason (1988) for the tail version. For the sake of completeness we provide in the supplementary (Section D.3) a different, direct proof of Lemma 3.19 below, relying on 'classes of function changing with n' (Van Der Vaart and Wellner (1996))

Lemma 3.19. *Under the assumptions of Theorem 3.17, the process $\widetilde{\Gamma}$ defined in (3.27) converges weakly in $\ell^\infty(0,1)$ to a tight Gaussian process $\widetilde{W}$ with covariance function*

$$(u_1, u_2) \mapsto (u_1 \wedge u_2)\Xi\,,$$

where Ξ is defined in (3.25)

An immediate consequence of Lemma 3.19, obtained upon setting $V = Z$ and $h(V) = 1$, is the weak convergence of the tail empirical process for uniform random variables introduced in (3.29).

Corollary 3.20. *As $n \to \infty$, $k \to \infty$, and $k/n \to 0$, the uniform tail empirical process (3.29) weakly converges to a standard Brownian motion W_1.*

Combining Corollary 3.20 and an appropriate variant of Vervaat's lemma (see Section D.2 from the supplementary material) we obtain in Section D.4 from the same supplement, the following result.

Lemma 3.21. *As $n \to \infty$, $k \to \infty$,*

$$\sup_{u\in(0,1]} \left| \sqrt{k} \left(\frac{n}{k} \hat{F}_U^-(uk/n) - u \right) + \hat{\gamma}_1(u) \right| = o_{\mathbb{P}}(1).$$

Separate and joint convergence in Decomposition (3.28)

We now show the following three relations: as $n \to \infty$,

$$\sup_{u\in(0,1]} \left| \widetilde{\Gamma}\left(\frac{n}{k}\, \hat{F}_U^-\Big(uk/n\Big) \right) - \widetilde{\Gamma}(u) \right| = o_{\mathbb{P}}(1), \tag{3.30}$$

$$\sup_{u\in(0,1]} \left| \sqrt{k} \left(D_n\left(\frac{n}{k}\, \hat{F}_U^-\Big(uk/n\Big) \right) - D_n(u) \right) + \nu\, \hat{\gamma}_1(u) \right| = o_{\mathbb{P}}(1), \tag{3.31}$$

$$\begin{pmatrix} \hat{\gamma}_1(u) \\ \widetilde{\Gamma}(u) \end{pmatrix} \rightsquigarrow W'(u), \tag{3.32}$$

where W' is a centered Gaussian process on $(0,1]$ with covariance function $(s,t) \mapsto s \wedge t\, \Xi'$. Here $\Xi' = S'(0)$ is the limit second moment matrix from Lemma 3.19 applied to $h'(V) = (1, h(V))$. More specifically, with this choice of h', we have

$$\Xi' = \begin{pmatrix} 1 & \nu^\top \\ \nu & \Xi \end{pmatrix} \in \mathbb{R}^{(q+1)\times(q+1)}$$

where $\Xi = \lim_{\delta\to 0} \mathbb{E}\left[h(V)h(V)^\top \,|\, U \le \delta\right]$.

We first prove (3.30). From Lemma 3.19, the process $\widetilde{\Gamma}$ is tight, whence asymptotically equi-continuous, meaning that

$$\lim_{\delta\downarrow 0} \limsup_n \mathbb{P}\left(\sup_{|s-t|\le\delta} |\widetilde{\Gamma}(s) - \widetilde{\Gamma}(t)| \ > \epsilon \right) = 0.$$

Also, from Lemma 3.21 and Corollary 3.20, $\sup_{u\in(0,1]} |(n/k)\hat{F}_U^-(uk/n) - u| = o_{\mathbb{P}}(1)$. Combining the two yields (3.30).

To prove (3.31), we apply the mean value theorem to get that

$$\begin{aligned}
&\sqrt{k}\left\{ D_n\left(\frac{n}{k}\hat{F}_U^-\Big(uk/n\Big) \right) - D_n(u) \right\} \\
&= \frac{n}{\sqrt{k}} \left\{ \mathbb{E}\Big[h(V) \mathbb{1}\Big\{ \Big\{ U \le u_n \Big\} \Big]_{u_n = \hat{F}_U^-(uk/n)} - \mathbb{E}\Big[h(V) \mathbb{1}\Big\{ \Big\{ U \le uk/n \Big\} \Big] \right\} \\
&= \frac{n}{\sqrt{k}} \tilde{g}(\tilde{U}_{u,n}) \left\{ \hat{F}_U^-\Big(uk/n\Big) - uk/n \right\} \\
&= \sqrt{k}\, \tilde{g}(\tilde{U}_{u,n}) \left\{ \frac{n}{k}\, \hat{F}_U^-\Big(uk/n\Big) - u \right\},
\end{aligned}$$

where $\tilde{g}(x)$ is the derivative of $x \mapsto \mathbb{E}\left[h(V)\mathbb{1}\{U \leq x\}\right]$ at point x and $\tilde{U}_{u,n}$ lies on the line segment between $\hat{F}_U^-(uk/n)$ and uk/n. Lemma 3.21 and Corollary 3.20 imply that $\tilde{U}_{u,n} \to 0$ in probability uniformly over $u \in [0,1]$, thus by continuity of $\tilde{g}$ at 0, $g(\tilde{U}_{u,k}) = \tilde{g}(0) + o_{\mathbb{P}}(1)$. We can further calculate $\tilde{g}(0)$ based on Assumption 3 in Theorem 3.17 as follows,

$$\tilde{g}(0) = \lim_{u\to 0} \mathbb{E}\left[h(V)\mathbb{1}\{U \leq u\}\right]/u = \lim_{u\to 0} \mathbb{E}\left[h(V) \,\middle|\, U \leq u\right] = \nu.$$

Therefore, the relation (3.31) is proved by applying Lemma 3.21, and the Slutsky's lemma. Finally, (3.32) follows from applying Lemma 3.19 to the function $h'(V) = (1, h(V))$.

Conclusion

By combining the decomposition in (3.28) with the relations (3.30)-(3.32), we obtain that, as $n \to \infty$,

$$\left\{\sqrt{k}\left(\widehat{D}_n(u) - D_n(u)\right)\right\}_{u\in[0,1]} \rightsquigarrow W := (-\nu, I_q)\, W'$$

which is a Gaussian process with covariance function

$$\begin{aligned}\Sigma(s,t) &= s \wedge t\ (-\nu, I_q) \begin{pmatrix} 1 & \nu^\top \\ \nu & \Xi \end{pmatrix} \begin{pmatrix} -\nu^\top \\ I_q \end{pmatrix} \\ &= s \wedge t \left(\Xi - \nu\nu^\top\right).\end{aligned}$$

■

3.5.3 Proposed estimation method

This section summarizes the main steps of the first and second order methods that we propose based on the processes $\hat{C}_n$ and $\hat{B}_n$. We first introduce TIREX1 and TIREX2 matrices in parallel with the matrix M_{CUME} defined in (3.4) in our framework, following the integral based methods proposed by Zhu et al. (2010), see also Portier (2016). In line with the CUME (3.4) matrix, we define

$$\begin{aligned} M_{\text{TIREX1}} &= \int_0^1 C_n(u) C_n(u)^\top \,\mathrm{d}u\,, \\ M_{\text{TIREX2}} &= \int_0^1 B_n(u) B_n(u)^\top \,\mathrm{d}u\,, \end{aligned} \tag{3.33}$$

where C_n and B_n are defined in (3.20) and (3.21) respectively. We omit the dependency of the matrices on n, k for convenience. An easy but important observation which underlies our strategy for estimating an extreme SDR space is the following lemma.

Lemma 3.22 (Consistency of the TIREX matrices)**.**
(i) *Under the assumptions of Theorem 3.11,*

$$M_{TIREX1} \longrightarrow \frac{1}{3}\ell\ell^\top \quad \text{as } n \to \infty.$$

(ii) *Under the assumptions of Theorem 3.16,*

$$M_{TIREX2} \longrightarrow \frac{1}{3}(S - I_p + \ell\ell^\top)^2 \quad \text{as } n \to \infty.$$

Proof Under the assumptions of the first statement, for fixed u, $C_n(u)C_n(u)^\top \to u^2\ell\ell^\top$ as $n \to \infty$. The result follows by dominated convergence on $(0,1)$, which applies by virtue of Condition (3.13). Indeed this uniform integrability assumption ensures that for some constant $A > 0$, for n large enough, for all $u \in (0,1)$,

$$\begin{aligned}\|C_n(u)\| &= \left\| u\mathbb{E}\Big[Z \,|\, \tilde{Y} < F^-(uk/n)\Big] \right\| \\ &\leq u(A + \mathbb{E}\left[\|Z\|\mathbb{1}\Big\{\|Z\| > A\Big\} \,|\, \tilde{Y} < F^-(uk/n)\right] \leq u(A+1).\end{aligned}$$

The argument for the second statement is similar, up to a call to Condition (3.16) instead of (3.13). ■

As a consequence of Lemma 3.22, both column spaces of M_{TIREX1} and M_{TIREX2} are asymptotically included in E_e. The column space of M_{TIREX1} has dimension one while that of M_{TIREX2} can be of any dimension not higher than that of E_e. We propose the following estimation procedures based respectively on the processes $\hat{C}_n$ and $\hat{B}_n$.

TIREX1

1. Choose $k \ll n$ and $1 \leq d \leq p$.
2. Compute the estimated TIREX1 matrix, $\widehat{M}_{\text{TIREX1}} = \int_0^1 \hat{C}_n(u)\hat{C}_n^\top(u)\,\mathrm{d}u$ using the identity given in (3.34).
3. Perform an eigen decomposition of $\widehat{M}_{\text{TIREX1}}$ and keep the first d eigenvectors $(e_i, i \leq d)$.
4. ouptut: $\hat{E}_e = \text{span}(\{e_i, i \leq d\})$.

Choosing $d > 1$ is not immediately justified because the limit of M_{TIREX1} is a rank one matrix $\ell\ell^\top/3$ as indicated in Lemma 3.22. However, empirical evidence suggests that allowing $d > 1$ can be useful to recover more components among the extreme central subspace basis. This is why we include this option in the algorithm.

TIREX2

1. Choose $k \ll n$ and $1 \leq d \leq p$.
2. Compute the estimated TIREX2 matrix, $\widehat{M}_{\text{TIREX2}} = \int_0^1 \hat{B}_n(u)\hat{B}_n^\top(u)\,\mathrm{d}u$ using the identity given in (3.35).
3. Perform an eigen decomposition of $\widehat{M}_{\text{TIREX2}}$ and keep the first d eigenvectors $(e_i, i \leq d)$ associated with the highest eigen values.
4. ouptut: $\hat{E}_e = \text{span}(\{e_i, i \leq d\})$

We make the following remarks regarding the relationships between our main theoretical result Corollary 3.18 and the proposed estimation methods TIREX1 and TIREX2.

Remark 3.23 (Asymptotic normality of the TIREX matrices)**.** *The asymptotic normality of the random matrices* $\sqrt{k}(\widehat{M}_{TIREX1} - M_{TIREX1})$ *and* $\sqrt{k}(\widehat{M}_{TIREX2} - M_{TIREX2})$ *could be obtained as a further consequence of Corollary 3.18 with straightforward calculations. This can be achieved by using the Delta-method as in the proof of Portier (2016), Proposition 5. For the sake of conciseness we leave the detailed proof to interested readers.*

Remark 3.24 (Bias term)**.** *Notice that the TIREX matrices* M_{TIREX} *are deterministic but subasymptotic quantities which depend on the the choice of the ratio* k/n*. The ultimate goal in view of Lemma 3.22 would be to obtain the limit distribution of* $\sqrt{k}(\widehat{M}_{TIREX1} - \frac{1}{3}\ell\ell^\top)$ *and* $\sqrt{k}(\widehat{M}_{TIREX2} - \frac{1}{3}(S - I_p + \ell\ell^\top))$*. An obvious way to do so is to assume that the bias terms* $\sqrt{k}(M_{TIREX1} - \frac{1}{3}\ell\ell^\top)$ *and* $\sqrt{k}(M_{TIREX2} - \frac{1}{3}(S - I_p + \ell\ell^\top))$ *converge to zero in probability, and use Slutsky's lemma.*

Remark 3.25 (Principal Component Analysis of the TIREX matrices)**.** *The output of the TIREX methods is the eigen spaces of the estimated TIREX matrices. An important final step is to show that such eigen spaces converges to the space spanned by the limits* $1/3\ell\ell^\top$ *and* $1/3(S - I_p + \ell\ell^\top)$*. A possible starting point would be to use results from perturbation theory, see* e.g. *(Zwald and Blanchard, 2005, Theorem 3) where the Frobenius norm of the error is controlled by the inverse of a spectral gap.Since this problem is left aside even in the traditional inverse regression literature we leave this question to further research while demonstrating the performance of the TIREX algorithms by numerical experiments.*

Remark 3.26 (Choices of d, k)**.** *The choice of the intermediate sequence* k *of extreme order statistics is a standard issue in extreme value statistics. In our experiments (Section 3.6) we propose to choose* k *by cross-validation. Theoretical investigation regarding this strategy is beyond the scope of this work. Similarly, the choice of* d *in the PCA decomposition of the matrix* $\widehat{M}_{TIREX2}$ *is a recurrent question in the PCA literature, which is also left to further research. In practice a natural and widely used strategy is an elbow method applied to the plot of the estimated eigen values. In the supervised learning context, we recommend to choose* d *by cross-validation. More generally (outside the supervised learning context), testing for the rank of the underlying matrix is a convenient method to infer the value of* d*. Such an approach has been succcessfully employed in the SDR literature (Portier and Delyon, 2014) where the test statistics are usually based on the eigenvalues amplitude. Finally recall that the limit of the matrix* M_{TIREX1} *has rank one, so that the default choice of* $d = 1$ *in the first order method is legitimate. Investigating theoretical guarantees for choosing the value of* d *in the TIREX context is beyond the scope of this thesis and left for future work.*

3.6 Experiments

This section focuses on the practical usefulness of TIREX for finite sample sizes based on simulated and real data. We first give some details about the implementation of TIREX (Section 3.6.1) and discuss its computational complexity. We discuss the improvement brought by TIREX over the estimation method proposed by Gardes (2018). Second, with synthetic datasets of various dimensions, we explore the estimation performance

of TIREX1 and TIREX2 for various values of k, as measured by a distance between the estimated and true extreme SDR spaces (Section 3.6.2). On this occasion we compare the estimation performance of TIREX with that of its closest alternatives, namely Gardes (2018)'s method, CUME and CUVE. Finally in Section 3.6.3 we compare TIREX with several existing dimension reduction tools for predicting tail events on several real data sets of relatively high dimension.

3.6.1 TIREX implementation

In a preliminary step common to all our experiments, the covariates are empirically standardized and we set $\hat{Z}_i = \hat{\Sigma}^{-1/2}(X_i - \hat{m})$ with $\hat{m} = n^{-1}\sum_{i=1}^n X_i$ and $\hat{\Sigma} = n^{-1}\sum_{i=1}^n (X_i - \hat{m})(X_i - \hat{m})^T$. Working with empirically standardized covariates to estimate an extreme SDR space E_e is equivalent to working with raw covariates to estimate $\tilde{E}_e = \Sigma^{-1/2}E_e$ up to remainder terms of order $O_{\mathbb{P}}(1/\sqrt{n})$, see Section C in Aghbalou et al. (2021). By abuse of notation we use the same symbols in the present section to denote both the empirical processes constructed with the $\hat{Z}_i$'s and the Z_i's.

We start off by deriving an explicit, computationally efficient formula for the matrices $\widehat{M}_{\text{TIREX1}}$ and $\widehat{M}_{\text{TIREX2}}$. Let $(\hat{Z}_{(1)}, Y_{(1)}), (\hat{Z}_{(2)}, Y_{(2)}), \ldots, (\hat{Z}_{(n)}, Y_{(n)})$ be such that $Y_{(1)} \geq \cdots \geq Y_{(n)}$. From the definition of $\hat{C}_n$, we have $\hat{C}_n(u) = \frac{1}{k}\sum_{i=1}^{\lceil ku \rceil} \hat{Z}_{(i)}$. This implies that $\hat{C}_n$ is piece-wise constant, more precisely for $j \in \{1, \ldots, n\}$, whenever $u \in ((j-1)/k, j/k]$, we have $k\hat{C}_n(u) = \sum_{i=1}^j \hat{Z}_{(i)} := \hat{S}_j$. Since $\widehat{M}_{\text{TIREX1}} = \sum_{j=1}^k \int_{(j-1)k}^{j/k} \hat{C}_n(u)\hat{C}_n(u)^\top du$, it follows that

$$\widehat{M}_{\text{TIREX1}} = \frac{1}{k^3}\sum_{j=1}^k \hat{S}_j \hat{S}_j^\top. \tag{3.34}$$

Evaluating the latter display requires $O(n\log(n))$ operations for sorting the Y's values; kd operations to compute the $\hat{S}_j$, $j = 1, \ldots, k$ (because $\hat{S}_j$ can be deduced from $\hat{S}_{j-1}$ with one operation); and $O(kd^2)$ operations to compute the matrix $\widehat{M}_{\text{TIREX1}}$. The overall cost is then of order $n\log(n) + kd^2$. Similar arguments regarding the second order matrix $\widehat{M}_{\text{TIREX2}}$ lead to the expression

$$\widehat{M}_{\text{TIREX2}} = \frac{1}{k^3}\sum_{j=1}^k \hat{T}_j \hat{T}_j^T, \tag{3.35}$$

with $T_j = \sum_{i=1}^j (\hat{Z}_{(i)}\hat{Z}_{(i)}^T - I_p)$.

The final step is to perform an eigen-decomposition of the estimated matrix $\widehat{M}_{\text{TIREX1}}$ (*resp.* $\widehat{M}_{\text{TIREX2}}$). Given the alleged dimension d of E_e, the vector space generated by the d eigen vectors associated to the d largest eigenvalues of the matrix (with multiplicities, assuming uniqueness of the corresponding eigen space for simplicity) constitutes the TIREX estimate $\widehat{E}_e$. The non standard SDR space can be estimated by multiplying the obtained directions by $\hat{\Sigma}^{-1/2}$.

Computational complexity.

Evaluating (3.34) requires $O(n\log(n))$ operations for sorting the Y's values; kp operations to compute the $\hat{S}_j$, $j = 1, \ldots, k$ (because $\hat{S}_j$ can be deduced from $\hat{S}_{j-1}$ with one operation); and $O(kp^2)$ operations to compute the matrix $\widehat{M}_{\text{TIREX1}}$. The overall

cost is then of order $n\log(n) + kp^2$. Similarly the overall cost for $\widehat{M}_{\text{TIREX2}}$ is of order $n\log(n)+kp^4$. Finally the eigen-decompostion based on SVD requires $O(p^3)$ operations.

In contrast the estimation procedure proposed in Gardes (2018) relies on an optimization strategy over a $p-d$-dimensional grid where d is the reduced dimension, and has an important computational cost when $d > 1$ according to the author (see Sections 3.2 and 4.1 of the cited reference). The existing implementation of Gardes (2018)'s method is restricted to $d = 1$ and the experiments conducted in that paper are limited to $p = 4$. Whether it is possible to bypass the curse of dimensionality in Gardes (2018)'s framework remains an open question. For these reasons we limit our comparison with Gardes (2018)'s method in our experiments to low dimensional examples, Models A and C, introduced below.

3.6.2 Performance for tail SDR estimation, synthetic data

We consider three particular instances of the mixture model presented in Section 3.3.2. The heavy tailed noise variables $\zeta_j, j \leq d$ follow identical Pareto distributions, $\mathbb{P}\left(\zeta_j > t\right) = t^{-\alpha_2}$ with $\alpha_2 = 10$. The short-tailed noise variables $\epsilon_j, j \leq p-d$ are exponentially distributed, $\mathbb{P}\left(\epsilon_j > t\right) = e^{-\alpha_1 t}, t > 0$, with rate parameter $\alpha_1 = 10$. The variables $(\zeta_j, j \leq d; \epsilon_j, j \leq p-d)$ are independent.

Model A. We consider Case (i) from the generic example (continuous covariates) with $\theta = 0.5$, $a = 1, b = 10$. For simplicity we take all covariate variables uniformly distributed over the interval $[a, b] = [1, 10]$. Recall that in this context, both TCI and TCI-G hold. Then according to both definitions the d-dimensional subspace of $\mathbb{R}^p$ generated by the canonical basis vectors $(e_{p-d+1}, \ldots, e_p)$ is an extreme SDR space. We set $p = 2$, $d = 1$.

Model B. Here we set $p = 30, d = 5$, all other setup remains unchanged comparing with Model A.

Model C. We use the distribution described in Case (ii) from Section 3.3.2, where the covariates are Bernoulli variables. In this context, TCI holds but TCI-G does not. We set the Bernoulli parameter to $\tau = 0.5$. To maintain the comparability between TIREX and Gardes (2018) we set $p = 2, d = 1$.

Experimental setting.

The sample size is set to $n = 10^4$ for Models A and C, and to $n = 10^5$ for Model B. The TIREX matrices following (3.34) and (3.35) are computed for 150 different values of k within the range $[\![n/100, n]\!]$. The orthogonal projection on the subspace generated by their first d eigen vectors constitutes our estimates $\hat{P}_e$. In other words we consider for simplicity that d is known by the user, as discussed in Remark 3.26. The quality of the estimator is measured by the squared Frobenius norm of the error, $\|\hat{P}_e - P_e\|_F^2$. We evaluate the squared bias $\|P_e - \mathbb{E}[\hat{P}_e]\|_F^2$, the variance $\mathbb{E}[\|\hat{P}_e - \mathbb{E}[\hat{P}_e]\|_F^2]$, and the MSE $\mathbb{E}[\|\hat{P}_e - P_e\|_F^2]$ using TIREX, based on $N = 200$ repetitions. Thus the maximum relative error of the MSE estimate, *i.e.* the maximum standard deviation of the estimate divided by the estimate itself, over all models and all values of k, is 0.11, which is sufficiently small for a qualitative interpretation of the results.

In addition we compare the relative performances of TIREX1 and Gardes (2018)'s method for Models A and C. We leave TIREX2 outside the comparison because our results (Figure 3.1) show that TIREX1 is a better option in this setting. To alleviate the computational cost we perform only $N = 100$ repetitions and we estimate the projectors for two values of k, namely $k = n^{2/3} \approx 464$ as recommended in Gardes (2018) and $k = 2000$ which is close to the value minimizing the MSE with TIREX1 for both models, considering our results below.

Results. Figure 3.1 displays the squared bias, variance and MSE for TIREX1 and TIREX2 as a function of k. The curves illustrate the typical bias-variance trade-off in Extreme Value Analysis regarding the choice of k, and confirm the findings of Corollary 3.18. Small values of k are associated with large variance, while large values result in a large bias. Notice that choosing $k = n$ with TIREX1 (*resp.* TIREX2) amounts to applying the standard SIR method CUME (*resp.* CUVE). Our results show that the MSE in this case is typically much larger (due to the bias) than with moderate k's, namely with $k \approx 2000$ for $n = 10^4$ and $k \approx 15000$ for $n = 10^5$.

In some cases, comparatively larger variances occur for $k \approx n/2$. We interpret this as an unstable transitional regime between two extremal behaviors: On the one hand, for small values of k, only the very largest values of Y are selected. These are mostly generated by the second component Y_2 of the mixture model, the heavy-tailed one. On the other hand when k is large, both components Y_1, Y_2 are equally involved in the computation of M_{TIREX}.

The variance attached to the second order method TIREX2 tends to be larger than that of the first order method TIREX1. However, when the dimension of the extreme SDR space is greater than one (Model B), TIREX1 fails to recover more than one direction, and TIREX2 is preferable. This fact illustrates the conclusion of Theorem 3.11, see also Lemma 3.22, where a single vector (or a rank-one matrix) is identified in the limit. TIREX2 does not suffer from this flaw since the associated limit in Lemma 3.22 is a matrix offering potentially more than one direction in the SDR space. As a conclusion, one should definitely prefer TIREX1 over TIREX2 when the extreme values of Y are known to be explained by a single linear combination of $Z_1, \ldots, Z_p$. Otherwise it is necessary to resort to TIREX2 to discover additional directions, even though the estimates may have a higher variance.

Table 3.1 displays the results of the comparison with Gardes (2018)'s method in terms of MSE and execution time. In Model A where Gardes (2018)'s assumptions are satisfied, Gardes (2018)'s method performs better than TIREX for the two values of k considered. However its execution time, even in this low dimensional setting is several orders of magnitude higher than that of TIREX. In Model C, as suggested by the theory, Gardes (2018)'s method fails to recover the tail SDR space (in the sense of TCI, not TCI-G). By contrast, TIREX can recover the tail SDR space within very short execution time.

	Model A, TIREX1	Model A, Gardes (2018)	Model C, TIREX1	Model C, Gardes (2018)
$k = 464$	2.10^{-3} (2 *s*)	4.10^{-4} (6 *h*)	4.10^{-3} (2.3 *s*)	1 (4.3 *h*)
$k = 2000$	5.10^{-4} (1.7 *s*)	5.10^{-5} (6.5 *h*)	9.10^{-4} (3.2 *s*)	0.8 (8.5 *h*)

Table 3.1 – MSE for TIREX and Gardes (2018)'s method in Models A and C, 100 replications. Execution times on a standard laptop are in brackets, with *h* and *s* indicating hour and second respectively.

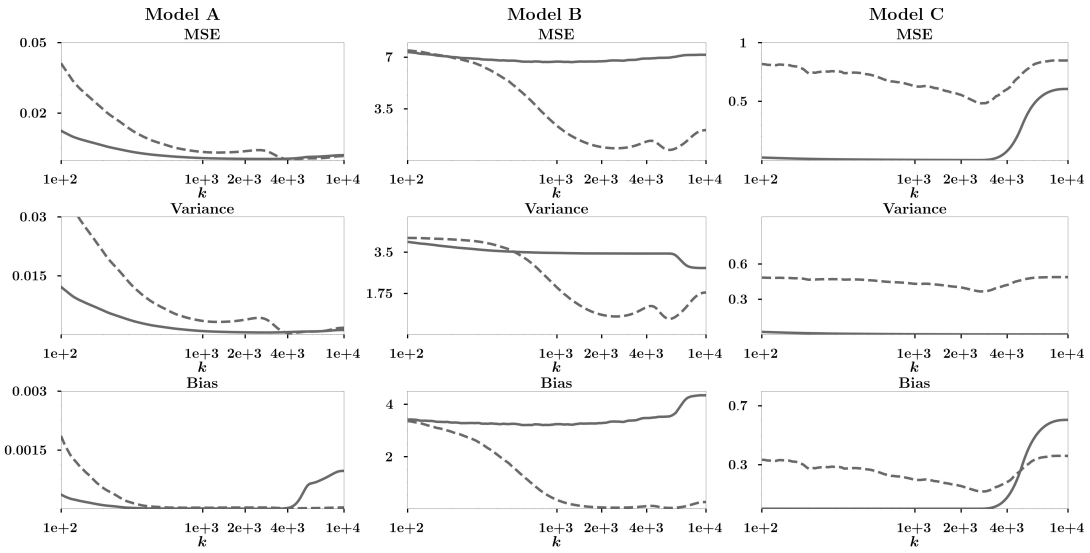


Figure 3.1 – Performance in terms of Frobenius norm of the error, as a function of k, with TIREX1 (solid line) and TIREX2 (dotted line), in Models A,B,C. Mean squared error, bias and variance computed over 100 repetitions.

3.6.3 Predicting tail events with TIREX on real datasets

We now investigate the relevance of TIREX as a dimension reduction tool for predicting unusually large values of Y. As explained in Remark 3.5, this may be viewed as a classification task: predict an exceedance $\{Y > y\}$ with the help of p covariates $X \in \mathbb{R}^p$. Reducing the dimension allows to escape the curse of the dimensionality using the projected covariates, however it generally induces a bias which may influence the (weighted) risk of an error. The most important observation in Remark 3.5 is that, if $Y_\infty \perp\!\!\!\perp X \mid P_e X$, the bias term vanishes in the limit $y \to y^+$. Since TIREX aims precisely at estimating P_e such that $Y_\infty \perp\!\!\!\perp X \mid P_e X$, a reasonable hope is that it would generally perform better than other dimension reduction algorithms targeting different reduction subspaces $P \neq P_e$ that would not enjoy this property.

Experimental setting.

We follow a two-steps procedure: first, run a dimension reduction algorithm (TIREX or another existing method) and project the covariates X_i on the estimated SDR space; second apply a classification algorithm to predict the event $Y_i > y$ with the help of the projected covariates. For all dimension reduction methods entering the comparison, the dimension of the reduced subspace is set to $d = 2$.

Throughout our experiments the second step is fixed: We use a nearest neighbors algorithm with a number of neighbors set to 5. In the end the performance of the competing dimension reduction methods is measured in terms of the AM risk (3.7) and the AUC (Area under the ROC Curve) of the nearest neighbors classifier trained on the reduced covariates. The number of observations k in TIREX is selected based on 5-fold cross-validation with the AUC criterion.

Competitors.

TIREX is compared with several alternative methods using the full dataset for estimation, not only the subset associated with the largest values of the target. Namely we consider in a supervised setting the standard SDR estimates obtained with the CUME and CUVE methods introduced in Section 3.2. In an unsupervised setting we consider routinely used methods available in the Python Scikit-learn package Pedregosa et al. (2011), namely Principal Component Analysis (PCA), Singular Value Decomposition (SVD) which is a non-centered version of PCA, Locally Linear Embedding (LLE), and Isomap (IMP). The latter two methods are non-linear generalizations of of PCA (Roweis and Saul (2000), Tenenbaum et al. (2000), see also Chojnacki and Brooks (2009); Bengio et al. (2003)) which are widely applied in many contexts such as data visualization Elgammal and Lee (2004); Tenenbaum et al. (2000), or classification Vlachos et al. (2002), among others. Considering the dimensions $p \in [\![18, 103]\!]$ of the datasets described below, Gardes (2018)'s method for dimension reduction could not be included in the comparison for the algorithmic complexity reasons described above.

Data sets.

Eight datasets are used. Three of them come from the UCI repository[2]: *Residential* (372 apartment sale prices, with 103 covariates); *crime* (1994 per capita violent crimes with 122 socio-economic covariates); *Parkinsons* (5875 voice recordings along with 25 attributes). Three other datasets come from the Delve repository[3]: *Bank* (8192 rejection rates of different banks, with 32 features each); *CompAct* (8192 CPU's times with 27 covariates); *PUMA32* (8192 angular accelerations of a robot arm, with 32 attributes). Finally, two other data are obtained from the LIACC repository[4]: *Ailerons* (13750 control action on the ailerons of an aircraft with 40 attributes) and *Elevator* (16559 control action on the elevators of an aircraft with 18 attributes).

Results.

For all datasets, y is chosen equal to the 0.98-quantile of the target $(Y_i)_{i=1,\dots,n}$ except for *Residential* where the 0.90-quantile has been used to counterbalance the small sample size. The results in terms of AM risk and AUC are summarized in Tables 3.2 and 3.3 respectively. In the vast majority of cases, TIREX1 or TIREX2 performs better than the other methods. On these examples, TIREX1 is often superior to TIREX2, which indicates that the added flexibility introduced by the second order moments does not compensate for the increased variance.

[2] https://archive.ics.uci.edu
[3] http://www.cs.toronto.edu/~delve/data/datasets.html
[4] https://www.dcc.fc.up.pt/~ltorgo/Regression/DataSets.html

	TIREX1	TIREX2	CUME	CUVE	PCA	SVD	LLE	IMP
Bank	0.434	**0.378**	0.42	0.392	0.418	0.474	0.486	0.432
Crime	**0.412**	0.5	0.471	0.47	0.502	0.469	0.47	0.5
CompAct	**0.208**	0.279	0.287	0.313	0.242	0.243	0.271	0.253
Residential	**0.158**	0.353	0.421	0.447	0.479	0.479	0.49	0.49
Parkinsons	**0.252**	0.346	0.268	0.346	0.469	0.469	0.455	0.47
Puma32	0.492	0.501	0.5	0.5	0.5	0.5	0.501	**0.49**
Elevators	**0.446**	**0.446**	0.471	0.463	0.5	0.5	0.5	0.5
Ailerons	**0.307**	0.329	0.314	0.33	0.498	0.499	0.498	0.501

Table 3.2 – AM risk of the nearest neighbors classifier with reduced covariates obtained with different dimension reduction methods.

	TIREX1	TIREX2	CUME	CUVE	PCA	SVD	LLE	IMP
Bank	**0.771**	0.696	0.698	0.684	0.736	0.689	0.608	0.65
Crime	0.666	0.67	0.616	0.686	0.678	**0.773**	0.672	0.661
CompAct	0.893	0.887	**0.899**	0.871	0.876	0.874	0.868	0.885
Residential	**0.902**	0.827	0.674	0.745	0.667	0.659	0.666	0.694
Parkinsons	**0.901**	0.818	0.852	0.82	0.742	0.753	0.743	0.748
Puma32	**0.711**	0.578	0.616	0.515	0.587	0.577	0.537	0.547
Elevators	0.686	**0.694**	0.615	0.672	0.528	0.537	0.514	0.514
Ailerons	**0.853**	0.834	0.828	0.832	0.502	0.515	0.514	0.525

Table 3.3 – AUC of the nearest neighbors classifier with reduced covariates obtained with different dimension reduction methods.

3.A Proofs for Remark 1

In this section, for the sake of completeness, we prove two facts regarding classification with the AM risk in the *full problem* defined in Remark 3.5 from the current chapter. First the classifier

$$h^*(x) = \mathbb{1}\Big\{\eta(x) > \pi\Big\} \tag{3.36}$$

is a minimizer of the AM risk ; Second, the associated Bayes risk is given by

$$\mathcal{R}_{\mathrm{AM}}(h^*) = \mathbb{E}\left[\min\left(\frac{\eta(X)}{\pi}, \frac{1-\eta(X)}{1-\pi}\right)\right]. \tag{3.37}$$

We introduce the AM loss function

$$\ell_{\mathrm{AM}}(\hat{t}, t) = \frac{1}{1-\pi}\mathbb{1}\Big\{\hat{t} = 1, t = 0\Big\} + \frac{1}{\pi}\mathbb{1}\Big\{\hat{t} = 0, t = 1\Big\}$$

so that for any classifier, $\mathcal{R}_{\mathrm{AM}}(h) = \mathbb{E}\Big[\ell_{\mathrm{AM}}(h(X), T)\Big]$. Consider now the conditional AM risk

$$\widetilde{\mathcal{R}_{\mathrm{AM}}}(h, x) = \mathbb{E}\Big[\ell_{\mathrm{AM}}(h(X), T) \;\Big|\; X = x\Big],$$

thus $\mathcal{R}_{\mathrm{AM}}(h) = \mathbb{E}\left[\widetilde{\mathcal{R}_{\mathrm{AM}}}(h, X)\right]$. We also have

$$\begin{aligned}\widetilde{\mathcal{R}_{\mathrm{AM}}}(h, x) &= \frac{1}{1-\pi}\mathbb{1}\Big\{h(x) = 1\Big\}(1-\eta(x)) + \frac{1}{\pi}\mathbb{1}\Big\{h(x) = 0\Big\}\eta(x) \\ &= \frac{1-\eta(x)}{1-\pi} + \mathbb{1}\Big\{h(x) = 0\Big\}\left[\frac{\eta(x)}{\pi} - \frac{1-\eta(x)}{1-\pi}\right]. \end{aligned} \tag{3.38}$$

Also, the classifier in (3.36) may be written equivalently as $h^*(x) = \mathbb{1}\left\{\frac{\eta(x)}{\pi} > \frac{1-\eta(x)}{1-\pi}\right\}$. Thus for any classifier h, we may write the difference in conditional risks as

$$\begin{aligned}\widetilde{\mathcal{R}_{\mathrm{AM}}}(h, x) - \widetilde{\mathcal{R}_{\mathrm{AM}}}(h^*, x) &= \frac{\eta - \pi}{\pi(1-\pi)}\Big[\mathbb{1}\Big\{h(x) = 0\Big\} - \mathbb{1}\Big\{h^*(x) = 0\Big\}\Big] \\ &= \left|\frac{\eta - \pi}{\pi(1-\pi)}\right| \mathbb{1}\Big\{h(x) \neq h^*(x)\Big\}\end{aligned}$$

The latter display is nonnegative, which shows that h^* defined in (3.36) indeed minimizes the AM risk. Turning to our second claim, notice that we may write, using (3.38),

$$\begin{aligned}\widetilde{\mathcal{R}_{\mathrm{AM}}}(h^*, x) &= \begin{cases} \eta(x)/\pi & \text{if } \eta(x)/\pi > (1-\eta(x))/(1-\pi) \\ (1-\eta(x))/(1-\pi) & \text{otherwise} \end{cases} \\ &= \min\left(\frac{\eta(x)}{\pi}, \frac{1-\eta(X)}{1-\pi}\right).\end{aligned}$$

This proves (3.37).

3.B Proofs for Section 3.2 and additional comments

In this section we provide the full proofs regarding our examples and counter-examples from Section 3.2 regarding the generic mixture model. On this occasion we conduct a thorough comparison between the two definitions of tail conditional independence TCI and TCI-G, see Equations (3.5) and (3.6). For convenience write $S(y) = \mathbb{P}(Y > y); S(y, W) = \mathbb{P}(Y > y|W); S(y, W, V) = \mathbb{P}(Y > y|W, V)$. The relevant quantities are respectively the ratios

$$R(y, V, W) = \frac{S(y, V, W) - S(y, W)}{S(y)}\ , \text{ and } \tilde{R}(y, V, W) = \frac{S(y, V, W) - S(y, W)}{S(y, W)}. \tag{3.39}$$

The TCI condition is that $\mathbb{E}|R(y, V, W)| \to 0$ as $y \to y^+$, whereas TCI-G means that $\tilde{R}(y, V, W) \to 0$ as $y \to y^+$, almost surely. Notice already that our criterion (3.1) is an integrated version of (3.2), with a weight function

$$\rho(y, W) = S(y, W)/S(y), \tag{3.40}$$

such that $\rho(y, W) \geq 0$ and $\mathbb{E}\left[\rho(y, W)\right] = 1$ for all y. Namely, TCI means that

$$\mathbb{E}\left|\tilde{R}(y, V, W)\rho(y, W)\right| \xrightarrow[y \to y^+]{} 0 \tag{3.41}$$

3.B.1 Additional notations regarding the generic mixture model from Section 3.3.2

We introduce in the context of Section 3.3.2 the additional notations

$$S_1(y) = \mathbb{P}(Y_1 > y) = \int S_1(y,v)\,\mathrm{d}P_V(v)\ , \quad S_2(y) = \mathbb{P}\left(Y_2 > y\right) = \int S_2(y,w)\,\mathrm{d}P_W(w).$$

With these notations, using the independence assumption regarding the pair (V,W) we may write

$$S(y,v,w) = \theta S_1(y,v) + (1-\theta)S_2(y,w)\ ; \quad S(y,w) = \theta S_1(y) + (1-\theta)S_2(y,w)\ ;$$
$$S(y) = \theta S_1(y) + (1-\theta)S_2(y).$$

Thus, the ratios R, $\tilde{R}$ defined at the beginning of this section and involved in TCI and TCI-G write respectively

$$R(y,v,w) = \frac{\theta(S_1(y,v) - S_1(y))}{\theta S_1(y) + (1-\theta)S_2(y)}\ , \qquad \tilde{R}(y,v,w) = \frac{\theta(S_1(y,v) - S_1(y))}{\theta S_1(y) + (1-\theta)S_2(y,w)}\ . \tag{3.42}$$

Notice already that

$$|R(y,v,w)| \leq \frac{\theta}{1-\theta}\frac{S_1(y,v) + S_1(y)}{S_2(y)}\ , \tag{3.43}$$

$$|\tilde{R}(y,v,w)| \leq \frac{\theta}{1-\theta}\left(\frac{S_1(y,v)}{S_2(y,w)} + \int \frac{S_1(y,v')}{S_2(y,w)}\,\mathrm{d}P_V(v')\right)\ . \tag{3.44}$$

Finally, specializing to the case where Y_1 and Y_2 follow the mixture model described in the same section of the main chapter, the conditional survival functions for Y_1, Y_2 are, for $y > b$,

$$S_1(y,v) = \sum_{i=1}^{p-d} \mathbb{1}\{v_i > 0\}\pi_i^1 S_\epsilon(y/v_i)\ , \qquad S_2(y,w) = \sum_{j=1}^{d} \mathbb{1}\Big\{w_j > 0\Big\}\pi_j^2 S_\zeta(y/w_j) \tag{3.45}$$

We now discuss the main differences between the two definitions. Natural questions to ask are *(i)* whether one definition is more appropriate than the other depending on the context ; *(ii)* whether one condition is stronger than the other, possibly under additional assumptions.

As for Question *(i)*, in spirit, as reflected by the equivalent condition (3.41), TCI is comparatively more sensitive to values $W = w$ such that the conditional probability of an exceedance $Y > y$ is large, which is an appealing feature for identifying tail risk factors as described in the introduction. On the other hand, one advantage of TCI-G's scaling is that the ratio $\tilde{R}$ introduced at the beginning of this section is a *relative* deviation, which is arguably easily interpretable. However TCI-G's criterion takes into account *all* possible values w, even those such that the conditional distribution of Y given $W = w$ is shorter tailed than the marginal distribution of Y. The focus in TCI-G is not exactly on the tail of Y's distribution, but rather on the tails of the conditional distributions of Y given W.

Before turning to Question *(ii)*, we discuss the differences between the two conditions in terms of mode of convergence.

3.B.2 Convergence almost-surely or in expectation in TCI-G or TCI

Almost sure convergence $\tilde{R}(y,V,W) \to 0$ as $y \to y^+$ implies $\mathbb{E}|\tilde{R}(y,V,W)| \to 0$. Indeed by conditioning on W, we have $\mathbb{E}\,[\tilde{R}(y,V,W)] = 0$ so that, denoting by z_+ (resp. z_-) the negative (resp. positive) part of a real z, it holds that $\mathbb{E}\,[\tilde{R}(y,V,W)_+] = \mathbb{E}\,[\tilde{R}(y,V,W)_-]$. As a consequence

$$\mathbb{E}\,|\tilde{R}(y,V,W)| = 2\mathbb{E}\,[\tilde{R}(y,V,W)_-].$$

However for all y, v, w, $\tilde{R}(y,v,w) \geq -1$ so that $0 \leq \tilde{R}(y,V,W)_- \leq 1$. By dominated convergence, if Condition (3.2) holds, then also $\mathbb{E}\,[\tilde{R}(y,V,W)_-] \to 0$, and the above display implies that $\mathbb{E}\,|\tilde{R}(y,V,W)|$ converges to 0 as well. This argument is not valid regarding the tail behaviour of $R(y,V,W)$ because it is not true that $R(y,V,W) \geq -1$ almost surely.

We are now ready to examine Question *(ii)*, that is, whether one condition (TCI or TCI-G) implies the other, in general or under simplifying assumptions.

3.B.3 Special case: discrete covariates with finite support

In order to build up the intuition, consider the special case where the covariates have a finite support. This is a sensible assumption for real life applications where observations are discretized.

We thus consider here finitely supported covariates $V \in \{v_1, \ldots, v_m\}$, $W \in \{w_1, \ldots, w_n\}$. Denote $p(v_i) = \mathbb{P}(V = v_i), p(w_j) = \mathbb{P}(W = w_j), p(v_i, w_j) = \mathbb{P}(V = v_i, W = w_j)$. Assume for simplicity that for all $(i,j) \in \{1,\ldots,m\} \times \{1,\ldots,n\}$, we have $p(v_i, w_j) > 0$.

First, in this case, almost sure convergence and convergence in expectation are equivalent for both ratios R and $\tilde{R}$ introduced at the beginning of this section. In other words

$$\mathbb{E}|R(y,V,W)| \xrightarrow[y\to y^+]{} 0 \iff |R(y,V,W)| \xrightarrow[y\to y^+]{} 0, \text{ almost surely ;} \tag{3.46}$$

$$\mathbb{E}|\tilde{R}(y,V,W)| \xrightarrow[y\to y^+]{} 0 \iff |\tilde{R}(y,V,W)| \xrightarrow[y\to y^+]{} 0, \text{ almost surely .} \tag{3.47}$$

Indeed

$$\mathbb{E}|R(y,V,W)| = \sum_{i=1}^{m}\sum_{j=1}^{n} p(v_i,w_j)\left|\frac{S(y,v_i,w_j) - S(y,w_j)}{S(y)}\right|.$$

The latter display converges to 0 as $y \to y^+$ if and only if each terms in the finite summation does, that is, if and only if $\forall (i,j), R(y,v_i,w_j) \to 0$ as $y \to y^+$. This proves (3.46), and the argument for (3.47) is similar.

Second, TCI-G implies TCI, meaning that our definition is weaker than Gardes (2018)'s in this discrete setting. To see this, in view of the equivalence between L^1 and almost sure convergences, it is enough to show that the ratio $R(y,v_i,w_j)/\tilde{R}(y,v_i,w_j)$ is uniformly upper bounded when y, i and j vary. However for all (y,i,j),

$$\frac{R(y,v_i,w_j)}{\tilde{R}(y,v_i,w_j)} = \rho(y,w_j) = S(y,w_j)/S(y) = \frac{S(y,w_j)}{\sum_{k=1}^{n} p(w_k)S(y,w_k)} \leq \frac{1}{p(w_j)} \leq 1/\min_{k\leq n} p(w_k) < \infty.$$

As a consequence, if $\tilde{R}(y,V,W) \to 0$ almost surely, then also $R(y,V,W) \to 0$ almost surely as $y \to y^+$ and the result follows.

3.B.4 Example in the mixture model where both TCI and TCI-G hold

We consider the setting of Section 3.3.2 from the main chapter, and in particular the case where the lower bound of the support of each W_j is positive, $a > 0$.

We verify that the upper bounds (3.43) and (3.44) uniformly converge to 0. First, using (3.45), we have

$$\frac{S_1(y) + S_1(y, v)}{S_2(y)} \leq 2\frac{\sup_{v\in[a,b]^{p-d}} S_1(y, v)}{\inf_{w\in[a,b]^d} S_2(y, w)} \leq 2\frac{S_\epsilon(y/b)}{S_\zeta(y/a)},$$

where the right-hand-side converges to 0 as $y \to \infty$ under Condition (3.7). Thus the upper bound in (3.43) uniformly converges to 0 and TCI holds by dominated convergence.

Turning to $\tilde{R}$, we also have

$$\sup_{(v,w)\in[a,b]^p} \frac{S_1(y, v)}{S_2(y, w)} \leq \frac{\sup_{v\in[a,b]^{p-d}} S_1(y, v)}{\inf_{w\in[a,b]^d} S_2(y, w)} \leq \frac{S_\epsilon(y/b)}{S_\zeta(y/a)} \xrightarrow[y\to\infty]{} 0.$$

Thus, by dominated convergence the right-hand-side of (3.44) converges to 0 as $y \to \infty$ so that TCI-G holds as well.

In the general case the situation is much more complex and it turns out that neither condition implies the other, as revealed by the counter-examples constructed in the next two subsections.

3.B.5 Counter-example in the mixture model where TCI holds but TCI-G does not

In contrast to the latter subsection, we now consider the case where the support of the W_j's includes 0, so that $a = 0$. Namely we take each variable V_j, W_j following a binary Bernoulli distribution with parameter $\tau \in (0, 1)$Thus $\mathbb{P}\Big(W = (0, \ldots, 0)\Big) = (1 - \tau)^d > 0$. Notice already that the right-hand side of (3.44) is not bounded because $S_2(y, w = (0, \ldots, 0)) = 0$ for $y > 0$. Also, from (3.42),

$$\tilde{R}(y, v, w = (0, \ldots, 0)) = \frac{S_1(y, v) - S_1(y)}{S_1(y)}.$$

In this specific example Y_1 and Y_2 have point masses at 0 and we have for $y > 0$, $S_1(y) = \sum_j \pi_j^1 \tau S_\epsilon(y) = \tau S_\epsilon(y)$ while for $v_1 = (1, \ldots, 1)$, $S_1(y, v_1) = \sum_j \pi_j^1 S_\epsilon(y) = S_\epsilon(y)$. Thus in the above display, $\tilde{R}(y, v_1, 0) = (1 - \tau)/\tau$ for all $y > 1$ and TCI-G does not hold.

Finally we show that TCI holds by examining the right-hand side of (3.43). The argument above shows that

$$\frac{S_1(y, v) + S_1(y)}{S_2(y)} \leq \frac{(1 + \tau)S_\epsilon(y)}{\tau S_\zeta(y)}.$$

This proves uniform convergence to 0 in (3.43) under Condition (3.7) and concludes the argument.

3.B.6 Counter-example where TCI-G holds but TCI does not

In this example we depart from the mixture model forming the basis of the two latter examples. The idea behind is to build the survival functions in such a way that $\limsup_{y\to\infty} \rho(y, W) = \infty$ (see (3.40) for the definition of ρ), with probability one, while TCI-G holds.

In addition to the notations introduced at the beginning of this section, we introduce the ratio

$$q(y, v, w) = S(y, v, w)/S(y, w).$$

Thus $\tilde{R}(y, v, w) = q(y, v, w) - 1$ and $R(y, v, w) = (q(y, v, w) - 1)\rho(y, w)$. We denote respectively by $P_W, P_{V,W}$ the marginal distribution of W and the joint distribution of (V, W). Here we define V, W as independent uniform variables, $P_W = P_V = \mathcal{U}_{[-1/2,1/2]}$ and $P_{V,W} = P_V \otimes P_W$. We shall build (S, ρ, q) such that $\mathrm{h}|q(y, V, W)-1| \to 0$ as $y \to \infty$, almost surely, (so that TCI-G holds) while $\limsup \mathbb{E}\left[|q(y, V, W) - 1|\rho(y, W)\right] > 0$ as $y \to \infty$ (so that TCI does not hold).

The functions $S(y), q(y, v, w), \rho(y, w)$ define a joint distribution of (Y, V, W) with no mass at the right end point of Y if conditions (3.48) (3.49) and (3.50) below hold.

$$S \text{ is non-increasing }, \quad \lim_{y\to y^+} S(y) = 0, \quad S(y) \geq 0; \tag{3.48}$$

$$\begin{gathered} P_W\text{-almost surely, the function } y \mapsto \rho(y, W)S(y) \text{ is non-increasing, and} \\ \lim_{y\to y^+} \rho(y, W)S(y) = 0, \quad \rho(y, W) \geq 0, \quad \mathbb{E}\left[\rho(y, W)\right] = 1\,, \forall y; \end{gathered} \tag{3.49}$$

$$\begin{gathered} P_{V,W}\text{-almost surely, the function } y \mapsto q(y, V, W)\rho(y, W)S(y) \text{ is non-increasing, and} \\ \lim_{y\to y^+} q(y, V, W)\rho(y, W)S(y) = 0, \quad q(y, V, W) \geq 0, \quad \mathbb{E}\left[q(y, V, W) \;\middle|\; W\right] = 1\,, \forall y. \end{gathered} \tag{3.50}$$

Construction of $S(y), \rho(y, w)$

We let $S(y) = e^{-y}$, $y \geq 0$, and we construct ρ such that $\mathbb{P}\left(\limsup_{y\to\infty} \rho(y, w) = \infty\right) = 1$ while (3.49) is satisfied. To this end define for $n \geq 2$, and $0 \leq j \leq n$,

$$L_n = \sum_{k<n, k\geq 2} k^2\,; \qquad L_{n,j} = L_n + jn. \tag{3.51}$$

Thus $L_2 = 0, L_3 = 4$, $L_n \leq n^3$ for $n \geq 2$ and $L_{n,n} = L_{n+1}$. Also $\mathbb{R}_+ = \sqcup_{n\geq 2} \sqcup_{0\leq j<n} [L_{n,j}, L_{n,j+1})$. Also for $y \geq 0$, we denote by $(n(y), j(y))$ the unique pair of integers such that $y \in [L_{n,j}, L_{n,j+1})$.

For $n \geq 2, 0 \leq j < n$, we define $\rho(y, w)$ for $y \in [L_{n,j}, L_{n,j+1})$ and $w \in [-1/2, 1/2]$ as follows: let $I_{n,j} = [1/2 - (j + 1)/n, 1/2 - j/n]$, then

$$\rho(y, w) = 1 + w + \frac{n}{4\pi} \sin\Big(\pi(y - L_{n,j})/n\Big) \left[\mathbb{1}\Big\{w \in I_{n,j}\Big\} - \mathbb{1}\Big\{w \notin I_{n,j}\Big\}/(n - 1)\right]. \tag{3.52}$$

Notice that for all $w \in [-1/2, 1/2]$, the function $y \mapsto \rho(y, w)$ is continuous. Also for all $w \in [-1/2, 1/2]$ we have $\limsup_y \rho(y, w) = +\infty$. Indeed for any fixed $n \geq 2$, let j such that $w \in I_{n,j}$. Then letting

$$y_n = L_{n,j} + n/2,$$

we have $\rho(y_n, w) = w + n/(4\pi) \geq n/(4\pi) - 1/2$. The sequence y_n converges to ∞ and is such that $\rho(y_n, w) \to \infty$ as $n \to \infty$, which proves the claim.

We now verify that the conditions gathered in (3.49) hold.

1. First for all $y \geq 0$,

$$\mathbb{E}\Big[\rho(y, W)\Big] = 1 + \mathbb{E}\big[W\big] + \frac{n}{4\pi} \sin\Big(\pi(y - L_{n,j})/n\Big)\Big[1/n - (n-1)/(n(n-1))\Big] = 1.$$

2. We show that for all y, $\rho(y, W) \geq 1/3$ almost surely. By construction, $\rho(y, W) \geq 1/2 - \frac{n(y)}{4(n(y)-1)\pi}$. Since $m/(m-1) \leq 2$ for $m \geq 2$, we obtain

$$\rho(y, W) \geq 1/2 - \frac{2}{4\pi} \geq 1/2 - 1/6 = 1/3.$$

3. We now show that $y \mapsto S(y)\rho(y, w)$ is non increasing for all $w \in [-1/2, 1/2]$. Since both S and ρ are continuous functions of y, with derivatives from the right which we denote respectively $S'(y)$ and $\rho'(y, w)$, we need to show that $\rho'(y, w) < -\rho(y, w)S'(y)/S(y)$. Here $S'(y)/S(y) = -1$, and from the above point we obtain $-\rho(y, w)S'(y)/S(y) \geq 1/3$. To conclude we show that

$$\forall y \geq 0, w \in [-1/2, 1/2], \rho'(y, w) \leq 1/4.$$

   Let $y > 0$ and $(n, j) = (n(y), j(y))$ as above. On the one hand if $w \in I_{n(y),j(y)}$ we have $0 \leq \rho'(y, W) \leq 1/4$. On the other hand if $w \notin I_{n(y),j(y)}$, we have $\rho'(y, w) < 0$. In both cases $\rho'(y, w) \leq 1/4 \leq 1/3 \leq -\rho(y, w)S'(y)/S(y)$, which conludes the argument.

4. Finally we verify that $\lim_{y\to\infty} \rho(y, W)S(y) = 0$, almost surely. To see this, notice that for all $y > 0, w \in [-1/2, 1/2], |\rho(y, w)| \leq 3/2 + \frac{n(y)}{4\pi}$. Now since $L_n \geq n^2$, $\{n : L_n \leq y\} \subset \{n : n^2 \leq y\}$, so that $n(y) = \sup\{n : L_n \leq y\} \leq \sup\{n : n^2 \leq y\} \leq \sqrt{y}$. Thus $|\rho(y, w)|e^{-y} \leq (3/2 + \sqrt{y}/(4\pi))e^{-y} \to 0$ as $y \to \infty$.

Construction of $q(y, v, w)$

Recall $n(y)$ from the beginning of the above paragraph. Define

$$q(y, v, w) = 1 + v\left[\mathbb{1}\left\{w > \frac{1}{2} - \frac{1}{n(y)}\right\} + \mathbb{1}\left\{w \leq \frac{1}{2} - \frac{1}{n(y)}\right\} \exp\left(-\frac{y + \left\lceil \frac{1}{1/2-w} \right\rceil}{4}\right)\right] \tag{3.53}$$

We now verify that the function $y \mapsto q(y, v, w)S(y, w) = S(y, v, w)$ is non increasing. Notice already that all the other constraints gathered in (3.50) are satisfied. Since for fixed (v, w), both $y \mapsto q(y, v, w)$ and $y \mapsto S(y, w)$ are continuous, it is enough to verify

that the derivative from the right of $y \mapsto q(y,v,w)S(y,w)$ is negative or null, that is (since $q \geq 1/2$ is positive), we need to ensure that

$$\frac{q'(y,v,w)}{q(y,v,w)} \leq -\frac{S'(y,w)}{S(y,w)}. \tag{3.54}$$

With our definition of $S(y,w)$ from Subsection 3.B.6,

$$-S'(y,w)/S(y,w) = (\rho(y,w) - \rho'(y,w))/\rho(y,w) = 1 - \rho'(y,w)/\rho(y,w) \geq 1 - \frac{1/4}{1/3} = 1/4.$$

If we denote $y(w) = y - \left\lceil \frac{1}{1/2-w} \right\rceil$, we have

$$q'(y,v,w)/q(y,v,w) = -\frac{1}{4}\mathbb{1}\left\{ w \leq \frac{1}{2} - \frac{1}{n(y)} \right\} v\exp(-y(w)/4)/(1+v\exp(-y(w)/4)).$$

The above display is always less than $1/4$ so that (3.54) holds for all $y > 0$ and $v, w \in [-1/2, 1/2]$ and (3.50) is satisfied. This fact combined with the argument in Subsection 3.B.6 implies that the functions (S, ρ, q) define a proper joint distribution for (Y, V, W).

Conclusion

We have constructed a joint distribution for (Y,V,W) in Sections 3.B.6, 3.B.6, such that $P_{V,W}$-almost surely, $q(y,V,W) \to 1$ as $y \to \infty$, as can be seen immediately from the definition of q in (3.53). Thus (Y,V,W) satisfy TCI-G. However, for all $n \geq 0$, let $y_n = L_n + n/2$ (see Subsection 3.B.6), so that by construction $n(y_n) = n$. Notice that $I_{n,0} = [1/2 - 1/n, 1/2]$ and for $w \in I_{n,0}$, we have $\rho(y_n, w) = 1 + w + n/(4\pi) \geq n/16$ and $q(y_n, v, w) = 1 + v$. Thus

$$\begin{aligned}
\mathbb{E}\left[|R(y_n,V,W)|\right] &= \mathbb{E}\left[|q(y_n,V,W) - 1|\rho(y_n,W)\right] \\
&\geq \mathbb{E}\left[|q(y_n,V,W) - 1|\rho(y_n,W)\mathbb{1}\left\{W > 1/2 - 1/n\right\}|\right] \\
&= \mathbb{P}\left(W > 1/2 - 1/n\right)\mathbb{E}\left[|q(y_n,V,W) - 1|\rho(y_n,W) \;\middle|\; W \in I_{n,0}\right] \\
&\geq \frac{1}{n}\mathbb{E}\left[|V|n/16\right] \geq \mathbb{E}\left[|V|\right]/16 = 1/64.
\end{aligned}$$

We have shown that $\limsup_{y\to\infty} \mathbb{E}\left[|R(y,V,W)|\right] > 0$, so that TCI does not hold, which concludes the counter-example.

3.B.7 Additive Mixture Model (Remark 3.6)

We end this section devoted to examples with a full derivation of the additive mixture example mentioned in Remark 3.6. We consider here an additive mixture $Y = Y_1 + Y_2$. The first (light-tailed) component is $Y_1 = V \in [a,b]$ $(-\infty < a < b < \infty)$; and the second (heavy-tailed) one is $Y_2 = W\xi$ where $W \in [c,d]$ with $0 < c < d < \infty$, and ξ has a continuous survival function $S_\xi(y) := 1 - F_\xi(y)$ satisfying $q(y) = y^\alpha S_\xi(y) \to C$ as $y \to \infty$, for some $\alpha, C > 0$. In addition, we assume that V and W are independent.

We show that TCI holds, that is $Y_\infty \perp\!\!\!\perp V \mid W$. Introducing the function

$$g(v,w,y) = w^{-\alpha} y^{\alpha} S_\xi[(y-v)/w],$$

we have that

$$\begin{aligned} &\sup_{[v,w]\in[a,b]\times[c,d]} \Big| g(v,w,y) - C| \\ &= \sup_{v,w\in[a,b]\times[c,d]} \left| \left(1-\frac{v}{y}\right)^{-\alpha} \left(\frac{y-v}{w}\right)^{\alpha} S_\xi\left[\frac{y-v}{w}\right] - c \right| \xrightarrow[y\to\infty]{} 0. \end{aligned} \tag{3.55}$$

The last limit relation follows from $\frac{y-v}{w} \to \infty$ and $1 - v/y \to 1$, uniformly for $v, w \in [a,b]\times[c,d]$ as $y \to \infty$. We have that

$$\begin{aligned} \mathbb{P}(Y > y|V,W) &= S_\xi((y-V)/W) = W^\alpha y^{-\alpha} g(V,W,y) \\ \mathbb{P}(Y > y|W) &= W^\alpha y^{-\alpha} \int_a^b g(v,W,y) f_1(v)\,\mathrm{d}v, \\ \mathbb{P}(Y > y) &= y^{-\alpha} \int_c^d \int_a^b w^\alpha g(v,w,y) f_1(v) f_2(w)\,\mathrm{d}v\,\mathrm{d}w. \end{aligned}$$

Thus

$$\frac{\mathbb{P}\left(Y > y|X\right) - \mathbb{P}\left(Y > y|W\right)}{\mathbb{P}\left(Y > y\right)} = \frac{W^\alpha \left\{ g(v,W,y) - \int_a^b g(v,W,y) f_1(v)\,\mathrm{d}v \right\}}{\int_c^d \int_a^b w^\alpha g(v,w,y) f_1(v) f_2(w)\,\mathrm{d}v\,\mathrm{d}w}$$

By (3.55) and dominated convergence, $\int_c^d \int_a^b g(v,w,y) f_1(v) f_2(w)\,\mathrm{d}v\,\mathrm{d}w \to c\mathbb{E}(W^\alpha)$ as $y \to \infty$. Regarding the numerator, Cauchy's inequality implies that

$$\begin{aligned} &\mathbb{E}\left| W^\alpha \left\{ g(v,W,y) - \int_a^b g(v,W,y) f_1(v)\,\mathrm{d}v \right\} \right| \\ \le &\sqrt{\mathbb{E}W^{2\alpha}} \sqrt{\mathbb{E}\left\{ g(v,W,y) - \int_a^b g(v,W,y) f_1(v)\,\mathrm{d}v \right\}^2}. \end{aligned}$$

The right-hand side tends to zero by noting that $\mathbb{E}W^{2\alpha} < \infty$ and applying the dominated convergence theorem twice to the second term. The proof is complete.

3.C Proof of Theorem 2

We need to show that

$$Q_e \mathbb{E}\left[ZZ^\top - I \;\middle|\; Y > y \right] \xrightarrow[y\to y^+]{} 0. \tag{3.56}$$

Notice first that from (LC) (2.1) and (CCV) (2.2) it holds that $Q_e(\mathrm{Var}Z \,|\, P_e Z - I_p) = -Q_e P_e = 0$. Thus also

$$Q_e \mathbb{E}\left[ZZ^\top - I_p \;\middle|\; P_e Z \right] = Q_e(\mathrm{Var}Z \,|\, P_e Z - I_p) + Q_e E[Z \,|\, P_e Z] E[Z \,|\, P_e Z]^\top = Q_e P_e = 0.$$

As a consequence

$$
\begin{aligned}
Q_e\mathbb{E}\left[(ZZ^\top - I_p)\mathbb{1}\{Y > y\}\right] &= Q_e\mathbb{E}\left[\mathbb{E}\Big((ZZ^\top - I_p)\mathbb{1}\{Y > y\} \,|\, P_eZ, Y\Big)\right] \\
&= Q_e\mathbb{E}\Big[\Big(\mathbb{E}\left[ZZ^\top - I_p \,\Big|\, P_eZ, Y\right] - \mathbb{E}\left[ZZ^\top - I_p \,\Big|\, P_eZ\right]\Big)\mathbb{1}\{Y > y\}\Big] \\
&= Q_e\mathbb{E}\Big[\Big(\mathbb{E}\left[ZZ^\top \,\Big|\, P_eZ, Y\right] - \mathbb{E}\left[ZZ^\top \,\Big|\, P_eZ\right]\Big)\mathbb{1}\{Y > y\}\Big]
\end{aligned}
$$

Thus in order to show (3.56) it is sufficient to show that for all pair $(i,j) \in \{1,\dots,p\}^2$, writing $p_y = \mathbb{P}\left(Y > y\right)$,

$$
p_y^{-1}\mathbb{E}\Big[\Big(\mathbb{E}\left[Z_iZ_j \,\Big|\, P_eZ, Y\right] - \mathbb{E}\left[Z_iZ_j \,\Big|\, P_eZ\right]\Big)\mathbb{1}\{Y > y\}\Big] \xrightarrow[y\to y^+]{} 0 \tag{3.57}
$$

Fixing $i,j \le p$ and following the same path as in Theorem 3.11. We decompose the left-hand side of (3.57) for any $A > 0$ as a sum $C_1(A,y) + C_2(A,y)$ where

$$
\begin{aligned}
C_1(A,y) = p_y^{-1}\mathbb{E}\Big[\Big(&\mathbb{E}\left[Z_iZ_j\mathbb{1}\Big\{\|Z\| \le A\Big\} \,\Big|\, P_eZ, Y\right] - \dots \\
&\dots\mathbb{E}\left[Z_iZ_j\mathbb{1}\Big\{\|Z\| \le A\Big\} \,\Big|\, P_eZ\right]\Big)\mathbb{1}\{Y > y\}\Big], \\
C_2(A,y) = p_y^{-1}\mathbb{E}\Big[\Big(&\mathbb{E}\left[Z_iZ_j\mathbb{1}\Big\{\|Z\| > A\Big\} \,\Big|\, P_eZ, Y\right] - \dots \\
&\dots\mathbb{E}\left[Z_iZ_j\mathbb{1}\Big\{\|Z\| > A\Big\} \,\Big|\, P_eZ\right]\Big)\mathbb{1}\{Y > y\}\Big].
\end{aligned}
$$

Point (iii) of Proposition 3 with $h = 1$ and $g(Z) = Z_iZ_j\mathbb{1}\Big\{\|Z\| \le A\Big\}$ ensures that $C_1(A,y) \to 0$ as $y \to y^+$ for any fixed A. On the other hand, using that $|Z_iZ_j| \le \frac{1}{2}(|Z_i|^2 + |Z_j|^2) \le \frac{1}{2}\|Z\|_2^2 \le c\|Z\|^2$ for some constant c we may bound $|C_2(A,y)|$ as follows,

$$
\begin{aligned}
|C_2(A,y)| &\le p_y^{-1}c\mathbb{E}\left[\mathbb{E}\left[\|Z\|^2\mathbb{1}\Big\{\|Z\| > A\Big\} \,\Big|\, P_eZ, Y\right]\mathbb{1}\{Y > y\}\right] + \dots \\
&\quad \dots p_y^{-1}c\mathbb{E}\left[\mathbb{E}\left[\|Z\|^2\mathbb{1}\Big\{\|Z\| > A\Big\} \,\Big|\, P_eZ\right]\mathbb{1}\{Y > y\}\right] \\
&= p_y^{-1}c\Big(\mathbb{E}\left[\|Z\|^2\mathbb{1}\Big\{\|Z\| > A\Big\}\mathbb{1}\{Y > y\}\right] + \mathbb{E}\Big(\mathbb{E}\left[\|Z\|^2\mathbb{1}\Big\{\|Z\| > A\Big\} \,\Big|\, P_eZ\right]\mathbb{1}\{Y > y\}\Big)\Big) \\
&= c\mathbb{E}\left[h_{1,A}(Z) \,\Big|\, Y > y\right] + \mathbb{E}\left[h_{2,A}(Z) \,\Big|\, Y > y\right].
\end{aligned}
$$

Hence, in view of Condition (4.4) for any $\epsilon > 0$ there exists some $A > 0$ such that

$$
\limsup_{y\to y^+} |C_2(A,y)| \le \epsilon,
$$

whence $\limsup_{y\to y^+} |C_2(A,y)| + |C_1(A,y)| \le \epsilon$, which shows (3.57) and completes the proof.

3.D Proofs and auxiliary results for Section 5

3.D.1 Inverse of empirical *c.d.f.* 's and order statistics

The following general fact is used on several occasions in our proofs:

Fact 3.D.1. *For $u \in (0,1]$, $\hat{H}^-(u) = T_{(\lceil nu \rceil)}$ and for $z \in [0, n-1]$:*

$$(\hat{H}(T_i) < (z+1)/n) \Leftrightarrow (T_i \leq \hat{H}^-(z/n)).$$

Proof The first statement follows from the definition of $\hat{H}^-$. Thus, using (5.1),

$$\begin{aligned} \hat{H}(T_i) < (z+1)/n) &\iff T_i < \hat{H}^-((z+1)/n) = T_{(\lceil z+1 \rceil)} = T_{(\lceil z \rceil + 1)} \\ &\iff T_i \leq T_{(\lceil z \rceil)} = \hat{H}^-(z/n). \end{aligned}$$

■

3.D.2 Vervaat's Lemma

We quote Lemma 4.3 in Segers (2015), which is a variant of "Vervaat's lemma", i.e., the functional delta method for the mapping sending a monotone function to its inverse.

Lemma 3.27. *Let $G : \mathbb{R} \to [0,1]$ be a continuous distribution function. Let $0 < r_n \to \infty$ and let $\hat{G}_n$ be a sequence of random distribution functions such that, in $\ell^\infty(\mathbb{R})$, we have $r_n(\hat{G}_n - G) \rightsquigarrow \beta \circ G$, as $n \to \infty$, where β is a random element of $\ell^\infty([0,1])$ with continuous trajectories. Then $\beta(0) = \beta(1) = 0$ almost surely and as $n \to \infty$,*

$$\sup_{u \in [0,1]} r_n |(G\{\hat{G}_n^-(u)\} - u) + (\hat{G}_n\{G^-(u)\} - u)| = o_{\mathbb{P}}(1).$$

3.D.3 Proof of Lemma 1

Because we only need to show that for any $u \in \mathbb{R}^p$, $v^T \widetilde{\Gamma} \rightsquigarrow v^T \widetilde{W}$, to prove that $\widetilde{\Gamma}$ is asymptotically tight we may consider the case where $q = 1$, i.e., $h(V) \in \mathbb{R}$. Denoting by ψ the derivative of $x \mapsto \mathbb{E}\left[h(V)^2 \mathbb{1}\Big\{\{U \leq x\}\Big\}\right]$, there exist by assumption positive constants (c_0, δ_0) such that for all $\delta \leq \delta_0$, $\psi(\delta) \leq c_0$. Similarly, because we assume the existence of $\Xi = S(0)$ (Assumption 2 in Theorem 3), there exist positive constants (c_1, δ_1) such that for all $\delta \leq \delta_1$, $\mathbb{E}[h(V)^2 \,|\, U < \delta_1] \leq c_1$. We assume in the following argument that $k/n \leq \delta_0 \wedge \delta_1$.

We apply Theorem 2.11.23 in Van Der Vaart and Wellner (1996) (Classes of functions changing with n) with

$$\begin{aligned} f_{n,u}(V,U) &= \sqrt{\frac{n}{k}} h(V) \mathbb{1}\Big\{\{U \leq uk/n\}\Big\}, \\ \mathcal{F}_n &= \{f_{n,u} \,:\, u \in [0,1]\}, \\ F_n(V,U) &= \sqrt{\frac{n}{k}} |h(V)| \mathbb{1}\Big\{\{U \leq k/n\}\Big\}. \end{aligned}$$

We start by verifying equation 2.11.21 in Van Der Vaart and Wellner (1996). First, we have

$$\mathbb{E}\left[F_n(V,U)^2\right] \leq c_1.$$

Second, for any $\eta > 0$ and $M > 0$, it holds that (for n, k large enough)

$$\begin{aligned}
&\mathbb{E}\left[F_n(V,U)^2 \mathbb{1}\Big\{\{F_n(V,U) > \eta\sqrt{n}\}\Big\}\right] \\
&= \left(\frac{n}{k}\right) \mathbb{E}\left[|h(V)|^2 \mathbb{1}\Big\{\{U \leq k/n\}\Big\} \mathbb{1}\Big\{\{|h(V)| \mathbb{1}\Big\{\{U \leq k/n\}\Big\} > \eta\sqrt{k}\}\Big\}\right] \\
&\leq \left(\frac{n}{k}\right) \mathbb{E}\left[|h(V)|^2 \mathbb{1}\Big\{\{U \leq k/n\}\Big\} \mathbb{1}\Big\{\{|h(V)| > \eta\sqrt{k}\}\Big\}\right] \\
&\leq \left(\frac{n}{k}\right) \mathbb{E}\left[|h(V)|^2 \mathbb{1}\Big\{\{U \leq k/n\}\Big\} \mathbb{1}\Big\{\{|h(V)| > M\}\Big\}\right].
\end{aligned}$$

Hence

$$\limsup_{n\to\infty} \mathbb{E}\left[F_n(V,U)^2 \mathbb{1}\Big\{\{F_n(U) > \eta\sqrt{n}\}\Big\}\right] \leq S(M).$$

But M is arbitrary so the latter display is arbitrarily small. Third, by the Mean Value Theorem, whenever $u \leq t$, $\exists \tilde{t} \in (u,t)$ such that

$$\begin{aligned}
\mathbb{E}\left[(f_{n,u}(V,U) - f_{n,t}(V,U))^2\right] &= \left(\frac{n}{k}\right) \mathbb{E}\left[h(V)^2 \mathbb{1}\Big\{\{uk/n \leq U \leq tk/n\}\Big\}\right] \\
&= \psi(\tilde{t}k/n)(t-u) \\
&\leq c_0(t-u).
\end{aligned}$$

This implies that

$$\sup_{|u-t|\leq\delta_n} \mathbb{E}\left[(f_{n,u}(V,U) - f_{n,t}(V,U))^2\right] \to 0, \text{ as } \delta_n \to 0.$$

It remains to check the entropy condition for the class $\mathcal{F}_n$. Let $0 < \epsilon < 1$, and denote by $u_i = i\epsilon$, $i = 0, \ldots, N$ and $u_{N+1} = 1$ with $N = \lfloor 1/\epsilon \rfloor$. Denote respectively by $f^+_{n,u}$ and $f^-_{n,u}$ the positive and negative parts of $f_{n,u}$ and by $\mathcal{F}^+_n, \mathcal{F}^-_n$ the associated classes. The functions (f^+_{n,u_i}) (resp. (f^-_{n,u_i})) forms an (ϵ, L_2)-bracketing of $\mathcal{F}^+_n$ (resp. $\mathcal{F}^-_n$), i.e., for any $u \in [0,1]$, there exists i such that

$$f^+_{n,u_i} \leq f^+_{n,u} \leq f^+_{n,u_{i+1}},$$

and

$$\mathbb{E}\left[(f^+_{n,u_{i+1}}(V,U) - f^+_{n,u_i}(V,U))^2\right] \leq c_0\epsilon.$$

Similar inequalities remain valid for $\mathcal{F}^-_n$. Hence considering the functions $f_{n,i} = f^+_{n,u_i} - f^-_{n,u_{i+1}}$, we have that for $u \in [u_i, u_{i+1}]$, $i = 0, \ldots, N$,

$$f_{n,u}(\,\cdot\,) = f^+_{n,u}(\,\cdot\,) - f^-_{n,u}(\,\cdot\,) \in [f_{n,i}(\,\cdot\,), f_{n,i+1}(\,\cdot\,)],$$

thus there exists $C > 0$ such that

$$\mathcal{N}_{[]}(\epsilon \|F_n\|_{L_2(P)}, \mathcal{F}_n, L_2(P)) \leq C/\epsilon^2.$$

The entropy condition is satisfied as for all $\delta_n \to 0$,

$$\int_0^{\delta_n} \sqrt{\log \mathcal{N}_{[]}(\epsilon \|F_n\|_{L_2(P)}, \mathcal{F}_n, L_2(P))}\, d\epsilon \to 0. \tag{3.58}$$

Consequently, the process $\widetilde{\Gamma}$ is tight. Finally the covariance functions at $s \leq t$ are given by

$$\begin{aligned}
\mathrm{Cov}\widetilde{\Gamma}_h(s), \widetilde{\Gamma}_h(t) &= \mathbb{E}\left[n/k h(V)h(V)^\top \mathbb{1}\Big\{U \leq sk/n\Big\}\right] - \cdots \\
&\qquad\qquad n/k \mathbb{E}\left[h(V)\mathbb{1}\Big\{U \leq sk/n\Big\}\right] \mathbb{E}\left[h(V)\mathbb{1}\Big\{U \leq tk/n\Big\}\right] \\
&= s\mathbb{E}\left[h(V)h(V)^\top \,\middle|\, U \leq sk/n\right] - \cdots \\
&\qquad\qquad k/n\ st \mathbb{E}\left[h(V) \,\middle|\, U \leq sk/n\right] \mathbb{E}\left[h(V) \,\middle|\, U \leq tk/n\right]
\end{aligned}$$

The first term in the right-hand side converges to $s\,\Xi = (s \wedge t)\,\Xi$ while the second term goes to zero from Assumption 3 in Theorem 3's statement. This concludes the proof.

3.D.4 Proof of Lemma 2

We apply Lemma 3.27 (Vervaat) to the distribution functions

$$\hat{G}_n(u) = \begin{cases} 0 & \text{for } u < 0 \\ \hat{F}_U(uk/n)/\hat{F}_U(k/n) & \text{for } 0 \leq u \leq 1 \\ 1 & \text{for } 1 < u \end{cases}, \qquad G(u) = \begin{cases} 0 & \text{for } u < 0 \\ u & \text{for } 0 \leq u \leq 1 \\ 1 & \text{for } 1 < u \end{cases}.$$

The quantile functions of $\hat{G}_n$ and G are respectively, for any $u \in [0,1]$,

$$\hat{G}_n^-(u) = \frac{\hat{F}_U^-(u\hat{F}_U(k/n))}{k/n}, \qquad G^-(u) = u,$$

Now we prove that the conditions of Lemma 3.27 are satisfied with $r_n = \sqrt{k}$ and β a Brownian bridge with covariance function $u_1 \wedge u_2 - u_1 u_2$. Define

$$a_n = \frac{k/n}{\hat{F}_U(k/n)}$$

and write

$$\begin{aligned}
\sqrt{k}(\hat{G}_n(u) - u) &= \left(\frac{\sqrt{k}}{\hat{F}_U(k/n)}\right)\Big(\hat{F}_U(uk/n) - u\hat{F}_U(k/n)\Big) \\
&= a_n\sqrt{k}\left(\frac{n}{k}\hat{F}_U(uk/n) - u\frac{n}{k}\hat{F}_U(k/n)\right) \\
&= a_n\sqrt{k}\left((\frac{n}{k}\hat{F}_U(uk/n) - u) - u(\frac{n}{k}\hat{F}_U(k/n) - 1)\right) \\
&= a_n\Big(\hat{\gamma}_1(u) - u\hat{\gamma}_1(1)\Big) \\
&= a_n\hat{\gamma}_2(u),
\end{aligned}$$

where $\hat{\gamma}_1$ is defined in (5.12) and

$$\hat{\gamma}_2(u) = \hat{\gamma}_1(u) - u\hat{\gamma}_1(1). \tag{3.59}$$

Now use that $a_n \to 1$ in probability and that $\sup_{u\in[0,1]} |\hat{\gamma}_2(u)| = O_{\mathbb{P}}(1)$ (both are consequences of Corollary 2) to to conclude (invoking Slutsky's lemma) that

$$\begin{aligned}\sqrt{k}(\hat{G}_n(u) - u) &= \hat{\gamma}_2(u) + (a_n - 1)\hat{\gamma}_2(u) \\ &= \hat{\gamma}_2(u) + o_{\mathbb{P}}(1),\end{aligned} \tag{3.60}$$

where the stochastic convergence $o_{\mathbb{P}}(1)$ is uniform in $u \in [0,1]$. In particular that $\sqrt{k}(\hat{G}_n(u) - u)$ weakly converges to a Brownian bridge with covariance function $u_1 \wedge u_2 - u_1 u_2$. The conclusion of Lemma 3.27 is that

$$\sup_{u\in(0,1]} \left|\hat{\gamma}_3(u) + \sqrt{k}(\hat{G}_n(u) - u)\right| = o_{\mathbb{P}}(1),$$

with

$$\hat{\gamma}_3(u) = \sqrt{k}(\hat{G}_n^-(u) - G^-(u)) = \sqrt{k}\left(\frac{n}{k}\hat{F}_U^-(u\hat{F}_U(k/n)) - u\right). \tag{3.61}$$

Consequently, using (3.60),

$$\sup_{u\in(0,1]} \left|\hat{\gamma}_3(u) + \hat{\gamma}_2(u)\right| = o_{\mathbb{P}}(1). \tag{3.62}$$

Remark that

$$\sqrt{k}\left((n/k)\hat{F}_U^-(uk/n) - u\right) = \hat{\gamma}_3(ua_n) + u\sqrt{k}(a_n - 1). \tag{3.63}$$

and that, as $\hat{\gamma}_1(1) = \sqrt{k}((n/k)\hat{F}_U(k/n) - 1)$,

$$\sqrt{k}(a_n - 1) = -a_n\hat{\gamma}_1(1) \tag{3.64}$$

Using the triangle inequality and (3.63), we get

$$\begin{aligned}&\left|\sqrt{k}\left((n/k)\hat{F}_U^-(uk/n) - u\right) + \hat{\gamma}_1(u)\right| \\ &= \left|\hat{\gamma}_3(ua_n) + u\sqrt{k}(a_n - 1) + \hat{\gamma}_1(u)\right| \\ &\le \left|\hat{\gamma}_3(ua_n) + \hat{\gamma}_2(ua_n)\right| + \left|u\sqrt{k}(a_n - 1) + \hat{\gamma}_1(u) - \hat{\gamma}_2(ua_n)\right| \\ &= \left|\hat{\gamma}_3(ua_n) + \hat{\gamma}_2(ua_n)\right| + |\hat{\gamma}_1(u) - \hat{\gamma}_1(ua_n)|,\end{aligned}$$

where the last line is deduced from (3.64) and $\hat{\gamma}_2(u) = \hat{\gamma}_1(u) - u\hat{\gamma}_1(1)$. Whenever $u \in [0, 1/2]$, we have, with probability going to 1, that $ua_n \in [0,1]$. Moreover, because $a_n \to 0$ in probability, there exists $\delta_n \to 0$ such that the event $|u - ua_n| \le |a_n| \le \delta_n$ has probability going to 1. On these events, it holds

$$\begin{aligned}\sup_{u\in(0,1/2]} \left|\hat{\gamma}_3(ua_n) + \hat{\gamma}_2(ua_n)\right| &\le \sup_{u\in(0,1]} \left|\hat{\gamma}_3(u) + \hat{\gamma}_2(u)\right| = o_{\mathbb{P}}(1) \\ \sup_{u\in(0,1/2]} |\hat{\gamma}_1(u) - \hat{\gamma}_1(ua_n)| &= \sup_{u\in(0,1],v\in(0,1],|u-v|\le\delta_n} |\hat{\gamma}_1(u) - \hat{\gamma}_1(v)| = o_{\mathbb{P}}(1).\end{aligned}$$

We have used (3.62) and the asymptotic equicontinuity of $\hat{\gamma}_1$. Consequently we have shown that, whenever $n \to \infty$, $k \to \infty$, we have

$$\sup_{u\in(0,1/2]} \left|\sqrt{k}\left((n/k)\hat{F}_U^-(uk/n) - u\right) + \hat{\gamma}_1(u)\right| = o_{\mathbb{P}}(1).$$

To obtain the stated result, apply this with $2k$ in place of k.

3.E Extension to non-standardized covariates

In this section we extend our inverse regression framework to the case of non-standardized covariates X. Section 3.E.1 recalls standard results for that matter. In Section 3.E.2 the extensions of the TIREX1 and TIREX2 principles are presented. The proofs of these results are omitted since they follow from classical arguments from non-standardized covariates combined with our proofs with standardized covariates from Section 3. In Section 3.E.3 we show that estimating the mean vector and covariance matrix for standardization does not change the asymptotic behavior of the latter tail processes.

3.E.1 SIR and SAVE principles with non-standardized covariates

We first recall some necessary background from the theory of inverse regression with non-standardized covariates, as exposed *e.g.* in Cook and Weisberg (1991).

SDR spaces

Recall from Section 2 that in terms of non-standardized covariates $X = m + \Sigma^{1/2}Z$, a subspace $\tilde{E}$ of $\mathbb{R}^p$ is a SDR space for the pair (X, Y) if and only if $\tilde{E} = \Sigma^{-1/2}E$ where E is a SDR space for the pair (Z, Y). We denote in the sequel by $\tilde{P}$ the orthogonal projector onto such a SDR space $\tilde{E}$ and we define $\tilde{Q} = I_p - \tilde{P}$.

Linearity and constant variance conditions

Conditions LC (2.1) and CCV (2.2) regarding the standardized variable Z are respectively equivalent to

$$\mathbb{E}\left[X|\tilde{P}X\right] = b + B\tilde{P}X \tag{3.65}$$

for some $b \in \mathbb{R}^p$ and $B \in \mathbb{R}^{p\times p}$, and

$$\mathrm{Var}X|\tilde{P}X \text{ is constant} \quad \text{a.s.} \tag{3.66}$$

SIR principle and CUME matrix

The extension of the SIR principle (Proposition 1) in terms of non-standardized covariates, is that under condition (3.65), it holds that

$$\Sigma^{-1}(\mathbb{E}\left[X|Y\right] - m) \in \tilde{E}. \tag{3.67}$$

As a consequence the CUME matrix defined in (2.3) must be replaced with the matrix $\tilde{M}_{\mathrm{CUME}} = \mathbb{E}\left[\tilde{m}(Y)\tilde{m}(Y)^T\right]$, with

$$\tilde{m}(y) = \mathbb{E}\left[(X - m)\mathbb{1}\{Y \le y\}\right],$$

in which case it holds that

$$\mathrm{span}(\tilde{M}_{\mathrm{CUME}}) \subset \Sigma\tilde{E} = \Sigma^{1/2}E.$$

SAVE principle

The parallel statement of Proposition 2 is that under conditions (3.65) and (3.66), we have

$$\text{span}(\Sigma^{-1}(\text{Var}\left[X \,\middle|\, Y\right] - \Sigma)) \subset \tilde{E} \quad a.s., \tag{3.68}$$

or equivalently $\text{span}(\Sigma^{-1}(\mathbb{E}\Big[(X-m)(X-m)^\top \,|\, Y\Big] - \Sigma)) \subset \tilde{E}$.

3.E.2 TIREX principles with non-standardized covariates

It follows from Definition 3 that E_e is an extreme SDR space for the pair (Z, Y) if and only if $\tilde{E}_e = \Sigma^{-1/2}E_e$ is an extreme SDR space for the pair (X, Y), in the sense that, denoting by $\tilde{P}_e$ the orthogonal projection on $\tilde{E}_e$, $Y_\infty \perp\!\!\!\perp X \,|\, \tilde{P}_e X$.

We now state the analogue statement to Theorem 1 in terms of the non-standardized covariate X.

Proposition 3.28 (non-standardized TIREX1 principle)**.** *The assumptions of Theorem 1 are equivalent to*

1. $\lim_{A\to\infty} \limsup_{y\to y^+} \mathbb{E}\left[\tilde{g}_{k,A}(X) \,\middle|\, Y > y\right] = 0$, $k = 1, 2$ *where* $\tilde{g}_{1,A}(X) = \|X\|\mathbb{1}\Big\{\|X\| > A\Big\}$ *and* $\tilde{g}_{2,A}(X) = \mathbb{E}\left[\|X\|\mathbb{1}\Big\{\|X\| > A\Big\} \,\middle|\, \tilde{P}_e X\right]$, *where* $\tilde{P}_e$ *is the orthogonal projector on* $\tilde{E}_e = \Sigma^{-1/2}E_e$.
2. *The covariate vector satisfies the non-standardized Linearity Condition* (3.65)
3. *For some* $\tilde{\ell} \in \mathbb{R}^p$, *with* $m = \mathbb{E}\left[X\right]$

$$\mathbb{E}\left[X \,\middle|\, Y > y\right] - m \xrightarrow[y\to y^+]{} \tilde{\ell}. \tag{3.69}$$

In such a case $\tilde{\ell} = \Sigma^{1/2}\ell$ *where* ℓ *is the limit defined in Theorem 1 and the conclusion is that*

$$\Sigma^{-1}\tilde{\ell} \in \tilde{E}_e.$$

Proposition 3.29 (non-standardized TIREX2 principle)**.** *Assume that* (X, Y) *and the extreme SDR space satisfy the assumptions of Proposition 3.28 (non-standardized TIREX1 principle) and that in addition,*

1. *(second order uniform integrability):*

$$\lim_{A\to\infty} \limsup_{y\to y^+} \mathbb{E}\left[\tilde{h}_{k,A}(X) \,\middle|\, Y > y\right] = 0, \qquad k = 1, 2\,, \tag{3.70}$$

where $\tilde{h}_{1,A}(X) = \|X\|^2\mathbb{1}\Big\{\|X\| > A\Big\}$ *and* $\tilde{h}_{2,A}(X) = \mathbb{E}\left[\|X\|^2\mathbb{1}\Big\{\|X\| > A\Big\} \,\middle|\, \tilde{P}_e X\right]$,

2. *(CCV) The covariate vector* X *satisfies the non-standardized constant variance condition* (3.66) *relative to* $\tilde{P}_e$,

3. *(Convergence of conditional expectations) For some* $\tilde{S} \in \mathbb{R}^{p\times p}$,

$$\mathbb{E}\left[XX^\top \,\middle|\, Y > y\right] \xrightarrow[y\to y^+]{} \tilde{S} + \tilde{\ell}\tilde{\ell}^\top, \tag{3.71}$$

where $\tilde{\ell}$ *is the limit appearing in Proposition 3.28.*

Then

$$\operatorname{span}(\Sigma^{-1}(\tilde{S} - \Sigma)) \subset \tilde{E}_e,$$

i.e. $\tilde{Q}_e\Sigma^{-1}(\tilde{S} - \Sigma) = 0$.

3.E.3 Estimation with non-standardized covariates

Consider the non-standardized versions of the matrices $M_{\text{TIREX1}}, M_{\text{TIREX2}}$ from Section 5 defined as follows:

$$\begin{aligned} \tilde{M}_{\text{TIREX1}} &= \int_0^1 C_n^m(u)C_n^m(u)^\top \,\mathrm{d}u\,, \text{ with} \\ C_n^m(u) &= \frac{n}{k}\mathbb{E}\left[(X-m)\mathbb{1}\left\{\tilde{Y} < F^-(uk/n)\right\}\right]\,, \end{aligned} \tag{3.72}$$

and

$$\begin{aligned} \tilde{M}_{\text{TIREX2}} &= \int_0^1 B_n^{m,\Sigma}(u)B_n^{m,\Sigma}(u)^\top \,\mathrm{d}u\,, \text{ with} \\ B_n^{m,\Sigma}(u) &= \frac{n}{k}\mathbb{E}\left[\Big((X-m)(X-m)^\top - \Sigma\Big)\mathbb{1}\left\{\tilde{Y} < F^-(uk/n)\right\}\right]. \end{aligned} \tag{3.73}$$

In view of Propositions 3.28 and 3.29, under the same assumptions therein, $\operatorname{span}(\tilde{M}_{\text{TIREX1}})$ and $\operatorname{span}(\tilde{M}_{\text{TIREX2}})$ become close to $\Sigma\tilde{E}_e$ as $n \to \infty$, in the sense that

$$\lim_{n\to\infty} \tilde{Q}_e\Sigma^{-1}\tilde{M}_{\text{TIREX1}} = \lim_{n\to\infty} \tilde{Q}_e\Sigma^{-1}\tilde{M}_{\text{TIREX2}} = 0,$$

where $\tilde{Q}_e$ is the orthogonal projector on $\tilde{E}_e^\perp$.

Notice that we can write $C_n^m, B_n^{m,\Sigma}$ in terms of C_n, B_n as follows:

$$\begin{aligned} C_n^m(u) &= \Sigma^{1/2}C_n(u) \\ B_n^{m,\Sigma}(u) &= \Sigma^{1/2}B_n(u)\Sigma^{1/2} \end{aligned} \tag{3.74}$$

Despite the apparent simplicity of (3.74), in the estimation step with unknown covariate's mean and covariance, one must replace m and Σ in Equations (3.72) and (3.73) with some estimates, *e.g.* the empirical ones which we denote by $\hat{m}, \hat{\Sigma}$. Namely we consider the processes

$$\begin{aligned} \hat{C}_n^{\hat{m}}(u) &= \frac{1}{k}\sum_{i=1}^n (X_i - \hat{m})\mathbb{1}\left\{\tilde{Y}_i \le \hat{F}^-(uk/n)\right\}\,, \\ \hat{B}_n^{\hat{m},\hat{\Sigma}}(u) &= \frac{1}{k}\sum_{i=1}^n \left((X_i-\hat{m})(X_i-\hat{m})^\top - \hat{\Sigma}\right)\mathbb{1}\left\{\tilde{Y}_i \le \hat{F}^-(uk/n)\right\}] \end{aligned} \tag{3.75}$$

and define the non-standardized TIREX1 and TIREX2 tail empirical processes respectively as

$$\sqrt{k}\left(\hat{C}_n^{\hat{m}} - C_n^m\right) \text{ and } \sqrt{k}\left(\hat{B}_n^{\hat{m},\hat{\Sigma}} - B_n^{m,\Sigma}\right). \tag{3.76}$$

We assume that the conditions for the Central Limit Theorem regarding the estimators $\hat{m}$ and $\hat{\Sigma}$ are met. For instance, we assume that X admits fourth order moments, an assumption which is needed anyway for the weak convergence of the TIREX2 process, see Corollary 1. Thus we work under the assumption that

$$\hat{m} = m + O_{\mathbb{P}}(1/\sqrt{n})\,; \quad \hat{\Sigma} = \Sigma + O_{\mathbb{P}}(1/\sqrt{n}). \tag{3.77}$$

Proposition 3.30 (Weak convergence of non-standardized TIREX processes)**.** *Under Assumption* (3.77)*,*

1. *The standardized TIREX1 process* $\sqrt{k}(\hat{C}_n - C_n)$ *converges weakly in* $\ell^\infty([0,1])$ *to a tight Gaussian process* W_1 *if and only if its non-standardized version defined in* (3.76) *converges weakly, in the same space, to the Gaussian process* $\Sigma^{1/2}W_1$*.*

2. *If weak convergence of the TIREX1 process holds true, then the standardized TIREX2 process* $\sqrt{k}(\hat{B}_n - B_n)$ *converges weakly in* $\ell^\infty([0,1])$ *to a tight Gaussian process* W_2 *if and only if its non-standardized version defined in* (3.76) *converges weakly, in the same space, to the Gaussian process* $\Sigma^{1/2}W_2\Sigma^{1/2}$*.*

Proof [Proof of Proposition 3.30]

1. Substituting $X - m$ with $\Sigma^{1/2}Z$ we obtain

$$\begin{aligned}\hat{C}^{\hat{m}}(u) &= \frac{1}{k}\sum_{i=1}^{n}\Sigma^{1/2}(Z_i + m - \hat{m})\mathbb{1}\left\{\tilde{Y}_i \le \hat{F}^-(uk/n)\right\} \\ &= \Sigma^{1/2}\left\{\hat{C}_n(u) + \Delta_n(u)(m - \hat{m})\right\}\end{aligned} \tag{3.78}$$

   where $\hat{C}_n$ is defined in (5.5) in terms of Z and

$$\Delta_n(u) := \frac{n}{k}\hat{F}\left(\hat{F}^-(uk/n)\right) \le \frac{n}{k}\hat{F}\left(\hat{F}^-(k/n)\right) = \frac{n}{k}\hat{F}(\tilde{Y}_{(k)}) = 1. \tag{3.79}$$

   Combining the latter upper bound, (3.78) and (3.74) we obtain

$$\sqrt{k}\left(\hat{C}_n^{\hat{m}} - C_n^m\right) = \Sigma^{1/2}\sqrt{k}(\hat{C}_n(u) - C_n(u)) + R_n(u), \tag{3.80}$$

   where $\sup_{u\in[0,1]} R_n(u) = O_{\mathbb{P}}(\sqrt{k/n}) = o_{\mathbb{P}}(1)$ and the main term $\sqrt{k}(\hat{C}_n(u) - C_n(u))$ is the standardized TIREX1 process. The first assertion of the statement follows from the Slutsky's lemma.

2. The argument for the second order method is similar though the computation is more involved. We have

$$\begin{aligned}\hat{B}_n^{\hat{m},\hat{\Sigma}}(u) &= \Sigma^{1/2}\left\{\frac{1}{k}\sum_{i\le n}\left(\left(Z_i+\Sigma^{-1/2}(m-\hat{m})\right)\left(Z_i+\Sigma^{-1/2}(m-\hat{m})\right)^\top - \Sigma^{-1/2}\hat{\Sigma}\Sigma^{-1/2}\right)\times\cdots\right.\\ &\qquad\left.\cdots\mathbb{1}\left\{\tilde{Y}_i\le\hat{F}^-(uk/n)\right\}\right\}\Sigma^{1/2}\\ &= \Sigma^{1/2}\left\{\hat{B}_n(u)+A_{1,n}\Delta_n(u)+A_{2,n}(u)\right\}\Sigma^{1/2}\end{aligned}$$

with $\Delta_n(u)\le 1$ as in (3.79) and

$$\begin{aligned}A_{1,n} &= \left(I_p-\Sigma^{-1/2}\hat{\Sigma}\Sigma^{-1/2}\right)+\Sigma^{-1/2}(m-\hat{m})(m-\hat{m})^\top\Sigma^{-1/2},\\ A_{2,n} &= \Sigma^{-1/2}(m-\hat{m})\hat{C}_n^\top(u)+\hat{C}_n(u)(m-\hat{m})^\top\Sigma^{-1/2}.\end{aligned}$$

Under the assumption that the TIREX1 empirical process converges weakly we have that $\sup_u \hat{C}_n(u)=O_{\mathbb{P}}(1)$, and using (3.77) and (3.74) we obtain

$$\sqrt{k}\left(\hat{B}_n^{\hat{m},\hat{\Sigma}}(u)-B_n^{m,\Sigma}(u)\right)=\Sigma^{1/2}\sqrt{k}\left(\hat{B}_n(u)-B_n(u)\right)\Sigma^{1/2}+R_n'(u)$$

with $\sup_u R_n'(u)=O_{\mathbb{P}}(\sqrt{k/n})=o_{\mathbb{P}}(1)$. The second assertion follows.

■

Chapter 4

Probability bounds for imbalanced classification

Contents

4.1 Introduction

Consider the problem of binary classification with covariate X and target $Y \in \{-1, 1\}$. The flagship approach to this problem in statistical learning is Empirical Risk Minimization (ERM), which produces approximate minimizers of $\mathcal{R}(g) = \mathbb{E}\Big[\ell(g(X), Y)\Big]$, given a loss function ℓ and a family of candidate classifiers $g \in \mathcal{G}$, with the help of observed data. with classifier g, $\ell_g(X, Y) = \ell(g(X), Y)$. However, when the underlying distribution is imbalanced, that is $p = \mathbb{P}(Y = +1)$ is relatively small, minimizing empirical version of $\mathcal{R}$ often leads to trivial classification rules for which the majority class is always predicted, because minimizing $\mathcal{R}(g)$ in that case is similar to minimizing $\mathbb{E}\Big[\ell(g(X), Y) \,|\, Y = -1\Big]$. Indeed by the law of total probabilities, $\mathcal{R}(g) = p\mathbb{E}\Big[\ell(g(X), Y) \,|\, Y = +1\Big] + (1-p)\mathbb{E}\Big[\ell(g(X), Y) \,|\, Y = -1\Big]$ and the former term is negligible with respect to the latter when $p \ll 1$. For this reason, even though standard ERM approaches might enjoy satisfactory generalization properties over imbalanced distributions, with respect to the standard risk $\mathcal{R}$, they may lead to unpleasantly high false negative rates and in general the average error on the minority class has no reason to be small, as its contribution to the overall risk $\mathcal{R}$ is negligible. This is typically what should be avoided in many applications when false negatives are of particular concern, among which medical diagnosis or anomaly detection for aircraft engines, considering the tremendous cost of an error regarding a positive example.

Bypassing the shortcoming described above is the main goal of many works regarding imbalanced classification. The existing literature may be roughly divided into oversampling approaches such as SMOTE and GAN (Chawla et al., 2002; Mariani et al., 2018),

undersampling procedures (Liu et al., 2009; Triguero et al., 2015) and risk balancing procedures also known as *cost-sensitive learning* (Scott, 2012; Xu et al., 2020c). Here we focus on the latter approach which enjoys numerous benefits, including simplicity, improved decision-making (Elkan, 2001a; Viaene and Dedene, 2005) , improved class probability estimation (Wang et al., 2019a; Fu et al., 2022), better resource allocation (Xiong et al., 2015; Ryu et al., 2017) and increased fairness (Menon and Williamson, 2018; Agarwal et al., 2018). By incorporating the varying costs of misclassification into the learning process, it enables models to make more informed and accurate predictions for the minority class, leading to higher-quality predictions. Balancing the risk consists of minimizing risk measures that differ significantly from the standard empirical risk, by means of an appropriate weighting of the negative and positive errors, in order to achieve a balance between the contributions of the positive and negative classes to the overall risk. In the present chapter we consider the balanced-risk, $\mathcal{R}_{\text{bal}}(g) = \mathbb{E}\left[\ell(g(X), Y) \,|\, Y = +1\right] + \mathbb{E}\left[\ell(g(X), Y) \,|\, Y = -1\right]$. Other metrics might be considered as detailed for instance in Table 1 in Menon et al. (2013a) which we do not analyze here for the sake of conciseness, even though our techniques of proof may be straightforwardly extended to handle these variants.

Empirical risk minimization based on the balanced risk is a natural idea, which is widely exploited by practitioners and has demonstrated its practical relevance in several operational contexts (Elkan, 2001b; Sun et al., 2007; Wang et al., 2016; Khan et al., 2018; Pathak et al., 2022). From a theoretical perspective, class imbalance has been the subject of several works. For instance, the consistency of the resulting classifier is investigated in Koyejo et al. (2014). Several different risk measures and loss functions are considered in Menon et al. (2013a) where results of asymptotic nature are established, for fixed $p > 0$, as $n \to \infty$. Also in the recent work by Xu et al. (2020c), generalization bounds are established for the imbalanced multi-class problem for a robust variant of the balanced risk considered here. Their main results from the perspective of class imbalance, is their Theorem 1 where the upper bound on the (robust) risk includes a term scaling as $1/(p\sqrt{n})$. A related subject is weighted ERM where the purpose is to learn from biased data (see *e.g.* Vogel et al. (2020); Bertail et al. (2021) and the references therein), that is, the training distribution and the target distribution differ. The imbalanced classification problem may be seen as a particular instance of this transfer learning problem, where the training distribution is imbalanced and the target is a balanced version of it with equal class weights. A necessary assumption in Bertail et al. (2021) is that the density of the target with respect to the source is bounded, which in our context is equivalent to requiring that p is bounded away from 0, an explicit assumption in Vogel et al. (2020) where the main results impose that $p > \epsilon$ for some fixed $\epsilon > 0$.

The common working assumption in the cited references that p is bounded from below, renders their application disputable in concrete situations where the number of positive examples is negligible with respect to a wealth of negative instances. To our best knowledge the literature is silent regarding such a situation. More precisely, we have not found neither asymptotic results covering the case where p depends on n in such a way that $p \to 0$ as $n \to \infty$; nor finite sample bounds which would remain sharp even in situations where p is much smaller than $1/\sqrt{n}$. Such situations arise in many examples in machine learning (see *e.g.* the motivating examples in the next section). However, existing works assume that the sizes of both classes are of comparable magnitude, which leaves a gap between theory and practice. A possible explanation is that existing works

do not exploit the full potential of the *low variance* of the loss functions on the minority class typically induced by boundedness assumptions combined with a low expected value associated with a small p.

It is the main purpose of this work to overcome this bottleneck and obtain generalization guarantees for the balanced risk which remain sharp even for very small p, that is, under sever class imbalance. Our purpose is to obtain upper bounds on the deviations of the empirical risk (and thus on the empirical risk minimizer) matching the state-of-the art, up to replacing the sample size n with np, the mean size of the rare class. To our best knowledge, the theoretical results which come closest to this goal are normalized Vapnik-type inequalities (Theorem 1.11 in Lugosi (2002)) and relative deviations (Section 5.1 in Boucheron et al. (2005)). However the latter results only apply to binary valued functions and as such do not extend immediately to general real valued loss functions which we consider in this chapter, nor do they yield fast rates for *imbalanced* classification problems, although relative deviations play a key role in establishing fast rates in *standard* classification as reviewed in Section 5 from Boucheron et al. (2005). Also, as explained above, we have not found any theoretical result regarding imbalanced classification which would leverage these bounds in order to obtain guarantees with leading terms depending on np instead of n.

Our main tools are (i) Bernstein-type concentration inequalities (that is, upper bounds including a variance term) for empirical processes that are consequences of Talagrand inequalities such as in Giné and Guillou (2001), (ii) fine controls of the expected deviations of the supremum error in the vicinity of the Bayes classifier, by means of local Rademacher complexities Bartlett et al. (2005); Bartlett and Mendelson (2006). Our contributions are two-fold.

1. We establish an estimation error bound on the balanced risk which holds true for VC classes of functions, which scales as $1/\sqrt{np}$ instead of the typical rate $1/\sqrt{n}$ in well-balanced problem, or $1/(p\sqrt{n})$ in existing works regarding the imbalanced case (*e.g.* as in Xu et al. (2020c)). Thus, in practice, our setting encompasses the case where $p \ll 1$ (severe class imbalanced) and our upper bound constitutes a crucial improvement by a factor $\sqrt{p}$ compared with existing works in imbalanced classification. Applying the previous bound to the k-nearest neighbor classification rule, we obtain the following new consistency result: as soon as kp goes to infinity, the nearest neighbors classification rule is consistent in case of relative rarity.

2. We obtain fast rates for empirical risk minimization procedures under an additional classical assumption called a Bernstein condition. Namely we prove upper bounds on the excess risk scaling as $1/(np)$, which matches fast rate results in the standard, balanced case, up to replacing the full sample size n with the expected minority class size np. To our best knowledge such fast rates are the first of their kind in the imbalanced classification literature.

Outline. Some mathematical background about imbalanced classification and some motivating examples are given in Section 4.2. In Section 4.3, we state our first non-asymptotic bound on the estimation error over VC class of functions and consider application to k-nearest neighbor classification rules. In Section 4.4, fast convergence rates are obtained and an application to ERM is given. Finally, some numerical experiments are provided in Section 4.5 to illustrate the theory developed in the chapter. All proofs of the mathematical statements are in the appendix.

4.2 Definition and notation

Consider a standard binary classification problem where random covariates X, defined over a space $\mathcal{X}$, are employed to distinguish between two classes defined by their labels $Y = 1$ and $Y = -1$. The underlying probability measure is denoted by $\mathbb{P}$ and the associated expectancy, by $\mathbb{E}$. The law of (X, Y) on the sample space $\mathcal{X} \times \mathcal{Y} := \mathcal{X} \times \{-1, 1\}$, is denoted by P. We assume that the label $Y = 1$ corresponds to minority class, i.e., $p = \mathbb{P}(Y = 1) \ll 1$. In the sequel we assume that $p > 0$, even though p may be arbitrarily small.

We adopt notation from empirical process theory. Given a measure μ on $\mathcal{X} \times \mathcal{Y}$ and a real function f defined over $\mathcal{X} \times \mathcal{Y}$, we denote $\mu(f) = \int f d\mu$. Then $f = \mathbb{1}_C$ for a measurable set C, we may write interchangeably $\mu(f) = \mu(\mathbb{1}_C) = \mu(C)$. We denote by P_+ the conditional law of (X, Y) given that $Y = +1$, thus

$$P_+(f) = \frac{\mathbb{E}(f(X,Y)\mathbb{1}\{Y = 1\})}{p} = \mathbb{E}(f(X,Y) \mid Y = 1).$$

In addition, we denote by $\mathrm{Var}_+(f)$ the conditional variance of $f(X, Y)$ given that $Y = +1$. The conditional distribution and variance P_- and Var_- are defined similarly, conditional to $Y = -1$.

In this chapter we consider general discrimination functions (also called *scores*) $g : \mathcal{X} \to \mathbb{R}$ and loss functions $\ell : \mathbb{R} \times \{-1, 1\} \to \mathbb{R}$, and our results will hold under boundedness and Vapnik-type complexity assumptions detailed below in Sections 4.3, 4.4. Given a score function g and a loss ℓ, it is convenient to introduce the function $\ell_g : (x, y) \mapsto \ell(g(x), y)$. With this notation the (unbalanced) risk of the score function g is $\mathcal{R}(g) = \mathbb{E}[\ell_g(X, Y)]$. Notice that the standard $0-1$ misclassification risk, $\mathcal{R}^{0-1}(g) = \mathbb{P}\left(g(X) \neq Y\right)$, is retrieved when g takes values in $\{-1, 1\}$ and $\ell(g(x), y) = \mathbb{1}\{g(x) \neq y\}$, or when g is real valued and $\ell(g(x), y) = \mathrm{sign}(-g(x)y)$. Allowing for more general scores and losses is a standard approach in statistical learning allowing to bypass the NP-hardness of the minimization problem associated with $\mathcal{R}^{0-1}$. Typically (although this is not formally required for our results to hold), the function ℓ_g takes the form $\ell_g(x, y) = \phi\left(-g(x)y\right)$, where ϕ is convex and differentiable with $\phi'(0) < 0$ (Zhang, 2004b; Bartlett et al., 2006a). This ensures that the loss is classification calibrated and that $\mathcal{R}(g) = \mathbb{E}\left[\ell_g(X, Y)\right]$ is a convex upper bound of $\mathcal{R}^{0-1}(g)$. Various consistency results ensuring that $g^\star = \arg\min_{g \in \mathbb{R}^{\mathcal{X}}} \mathcal{R}(g) = \arg\min_{g \in \mathbb{R}^{\mathcal{X}}} \mathcal{R}^{0-1}(g)$ can be found in Bartlett et al. (2006a). Examples include the logistic ($\phi(u) = \log(1 + e^{-u})$), exponential ($\phi(u) = e^{-u}$), squared ($\phi(u) = (1-u)^2$), and hinge loss ($\phi(u) = \max(0, 1-u)$).

The balanced $0-1$ risk is defined as $\mathcal{R}^{0-1}_{\mathrm{bal}}(g) = (P_+(Y \neq g(X)) + P_-(Y \neq g(X)))/2$ and is referred to as the *AM measure* in existing literature (Menon et al., 2013a). The minimizer of the latter risk, $g^\star_{bal}$, is known as the balanced Bayes classifier. It returns 1 when $\eta(X) = \mathbb{P}(Y = 1 \,|\, X) \geq p$ and -1 otherwise (refer to Theorem 2 or or Proposition 2 in Koyejo et al. (2014)). In the present work we consider a general balanced risk allowing for a real-valued loss function ℓ_g, defined for $g \in \mathcal{G}$ as

$$\mathcal{R}_{\mathrm{bal}}(g) = \frac{1}{2}\left(P_+(\ell_g) + P_-(\ell_g)\right).$$

Given an independent and identically distributed sample $(X_i, Y_i)_{1 \leq i \leq n}$ according to P, we denote by P_n the empirical measure, $P_n(f) = (1/n)\sum_{i=1}^{n} f(X_i, Y_i)$, for any measurable and real-valued function f on $\mathcal{X} \times \mathcal{Y}$. While the standard risk estimate is simply expressed as $P_n(\ell_g)$ for any $g \in \mathcal{G}$, the balanced empirical risk is necessarily defined in terms of empirical conditional measures,

$$P_{n,+}(f) = \frac{P_n(f\mathbb{1}\{Y = 1\})}{p_n},$$

where by convention $P_{n,+}(f) = 0$ when $p_n = P_n(Y = 1) = 0$. The empirical measure of the negative class, $P_{n,-}$, is defined in a similar manner. Finally the balanced empirical risk considered in this thesis is

$$\mathcal{R}_{n,\text{bal}}(g) = \frac{1}{2}\left(P_{n,+}(\ell_g) + P_{n,-}(\ell_g)\right).$$

Motivating Examples We now present two examples where the probability $p \to 0$ as $n \to \infty$:

1. The first example is the problem of contaminated data which is central in the robustness literature. A common theoretical assumption is that the number of anomalies n_0 grows sub-linearly with the sample size, as discussed in (Xu et al., 2012; Staerman et al., 2021a). In this context, $n_0 = n^a$ for some $a < 1$ and consequently, $p = n^{a-1} \to 0$.

2. The second example pertains to Extreme Value Theory (EVT) (Resnick, 2013; Goix et al., 2015; Jalalzai et al., 2018; Aghbalou et al., 2023). Consider a continuous positive random variable T, predicting exceedances over arbitrarily high threshold t may be viewed as a binary classification problem. Indeed for fixed t, consider the binary target $Y = \mathbb{1}\{T > t\} - \mathbb{1}\{T \leq t\}$ with marginal class probability $p = P(T > t)$. The goal is thus to predict Y , by means of the covariate vector X.

   One major goal of EVT is to learn a classifcation model for extremely high tresholds t. In practice, EVT based approaches set the threshold t as the $1 - \alpha$ quantile of T with $\alpha = k/n \to 0$ and $k = o(n)$. This approach essentially assumes that the positive class consists of the $k = o(n)$ largest observations of T so that $P(T > t) = P(Y = 1) = k/n \to 0$.

4.3 Standard learning rates under relative rarity

4.3.1 Concentration bound

The primary goal of this chapter is to assess the error associated with estimating the balanced risk $\mathcal{R}_{\text{bal}}(g)$ using the empirical balanced risk $\mathcal{R}_{n,\text{bal}}(g)$. Given the definition of the balanced risk, the quantity of interest takes the form $(P_{n,+} - P_+)(f)$, and a similar analysis applies to $(P_{n,-} - P_-)(f)$.In this chapter we control the complexity of the function class *via* the following notion of VC-complexity.

Definition 4.1. *The family of functions $\mathcal{F}$ is said to be of VC-type with constant envelop $U > 0$ and parameter (v, A) if $\mathcal{F}$ is bounded by U and for any $0 < \epsilon < 1$ and any probability measure Q on $(S, \mathcal{S})$, we have*

$$\mathcal{N}\left(\mathcal{F}, L_2(Q), \epsilon U\right) \leq (A/\epsilon)^v.$$

The connection between the usual VC definition (Vapnik and Chervonenkis, 1971) and Definition 4.1 can be directly established through Haussler's inequality (Haussler, 1995), which indicates that the covering number of a class of binary classifiers with VC dimension v (in the sense of Vapnik and Chervonenkis (1971)) is given by

$$\mathcal{N}\left(\epsilon, \mathcal{F}, L_2(Q), \epsilon\right) \leq Cv(4e)^v \epsilon^{-2v} = \left(\frac{2\sqrt{e}(Cv)^{1/v}}{\epsilon}\right)^{2v},$$

for some universal constant $C > 0$. Thus a VC-class of functions in the sense of Vapnik and Chervonenkis (1971) is necessarily a VC-type class in the sense of Definition 4.1.

Notice that within a class $\mathcal{F}$ with envelop $U > 0$, the following variance bounds are automatically satisfied:

$$\sigma_+^2, \sigma_-^2 = \sup_{f \in \mathcal{F}} \mathrm{Var}_+(f), \sup_{f \in \mathcal{F}} \mathrm{Var}_-(f) \leq U^2.$$

The following theorem states a uniform generalization bound that incorporates the probability of each class in such a way that the deviations of the empirical measures are controlled by the expected number of examples in each class, np and $n(1-p)$. Interestingly the deviations may be small even for small p, as soon as the product np is large. The bound also incorporates the conditional variance of a class (σ_+^2, σ_-^2), which will play a key role in our application to nearest neighbors.

Theorem 4.2. *Let $\mathcal{F}$ be of VC-type with constant envelop U and parameter (v, A). For any n and δ such that*

$$np \geq \max\left[\frac{U^2}{\sigma_+^2} v \log\left(K'A/\left(2\delta\sqrt{p}\right)\right), 8\log(1/\delta)\right]$$

we have with probability $1 - \delta$,

$$\sup_{f \in \mathcal{F}} \left|P_{n,+}(f) - P_+(f)\right| \leq 4K'\sigma_+ \sqrt{\frac{v}{np} \log\left(K'A/(2\delta\sqrt{p})\right)}$$

For some universal constant $K' > 0$. We also have with probability $1 - \delta$,

$$\sup_{f \in \mathcal{F}} \left|P_{n,-}(f) - P_-(f)\right| \leq 4K'\sigma_- \sqrt{\frac{v}{n(1-p)} \log\left(K'A/(2\delta\sqrt{(1-p)})\right)}.$$

Remark 4.3. *This upper bound extends Theorem 1.11 in Lugosi (2002), which is limited to a binary class of functions characterized by finite shatter coefficients. The extension is possible by utilizing results from Plassier et al. (2023). It is crucial to recognize that all existing non-asymptotic statistical rates in the imbalanced classification literature (Menon et al., 2013a; Koyejo et al., 2014; Xu et al., 2020c) follow the rate $1/(p_n\sqrt{n})$, leading to a trivial upper bound when $p_n \leq 1/\sqrt{n}$. In our analysis, the upper bound remains consistent provided that $np_n \to \infty$, thereby emphasizing the merits of using concentration inequalities incorporating the variance of the positive class $\mathrm{Var}(f\mathbb{1}\{y = 1\}) \leq Up \ll 1$.*

The next corollary, which derives from Theorem 4.2 together with standard arguments, provides generalization guarantees for ERM algorithms based upon the balanced risk. Namely it gives an upper bound on the excess risk of a minimizer of the balanced risk. The proof is provided in the supplementary material for completeness.

Corollary 4.4. *Suppose that* $\{\ell_g : g \in \mathcal{G}\}$ *is VC-type with envelop* U *and parameter* v, A*. Under the conditions of Theorem 4.2, we have, with probability* $1-\delta$*,*

$$\mathcal{R}_{\text{bal}}\left(\hat{g}_{\text{bal}}\right) \leq \mathcal{R}_{\text{bal}}\left(g^{\star}_{\text{bal}}\right) + 4K'\sigma_{\max}\sqrt{\frac{v\log\left(K'A/\left(2\delta\sqrt{p}\right)\right)}{np}},$$

where $\sigma_{\max} = \max\left(\sigma_+, \sigma_-\right) \leq U$ *and* $K' > 0$ *is a universal constant.*

The previous result shows that whenever $np \to \infty$, learning from ERM based on a VC class of functions is consistent. Another application of our result pertains to k-nearest neighbor classification algorithms. In this case the sharpness of our bound is fully exploited by leveraging the variance term σ_+. This is the subject of the next section.

4.3.2 Balanced k-nearest neighbor

In the context of imbalanced classification, we consider here a balanced version of the standard k-nearest neighbor (k-NN for short) rule, which is designed in relation with the balanced risk $R^*_{bal}(g)$. We establish the consistency of the balanced k-NN classifier with respect to the balanced risk.

Let $x \in \mathbb{R}^d$ and $\|\cdot\|$ be the Euclidean norm on $\mathbb{R}^d$. Denote by $B(x,\tau)$ the set of points $z \in \mathbb{R}^d$ such that $\|x - z\| \leq \tau$. For $n \geq 1$ and $k \in \{1, \ldots, n\}$, the k-NN radius at x is defined as

$$\hat{\tau}_{n,k,x} := \inf\left\{\tau \geq 0 : \sum_{i=1}^{n} 1_{B(x,\tau)}(X_i) \geq k\right\}.$$

Let $I_n(x)$ be the set of index i such that $X_i \in B(x, \hat{\tau}_{n,k,x})$ and define the estimate of the regression function $\eta(x)$ as

$$\hat{\eta}_n(x) = \frac{1}{k}\sum_{i \in I_n(x)} \mathbb{1}_{Y_i=1}.$$

While standard NN classification rule is a majority vote following $\hat{\eta}_n(x)$, i.e., predict 1 whenever $\hat{\eta}_n(x) \geq 1/2$, it is natural, in view of well known results recalled in Section 4.2, to consider a balanced classifier $\hat{g}_n$ for imbalanced data predicts 1 whenever $\hat{\eta}_n(x) \geq p_n$, that is $\hat{g}_n = \text{sign}(\hat{\eta}_n(x)/p_n - 1)$.

The analysis of the k-NN classification rule is conducted for covariates X that admit a density with respect to the Lebesgue measure. We will need in addition that the support S_X is well shaped and that the density is lower bounded. These standard regularity conditions in the k-NN literature are recalled below.

(X1) The random variable X admits a density f_X with compact support $S_X \subset \mathbb{R}^d$.

(X2) There is $c > 0$ and $T > 0$ such that

$$\forall \tau \in (0, T],\ \forall x \in S_X,\ \lambda(S_X \cap B(x, \tau)) \geq c\lambda(B(x, \tau)),$$

where λ is the Lebesgue measure.

(X3) There is $0 < b_X \leq U_X < +\infty$ such that

$$b_X \leq f_X(x) \leq U_X, \qquad \forall x \in S_X.$$

In light of Proposition 4.15 (stated in the supplement), we consider the estimation of $\nu^*(x) := \eta(x)/p$ using the k-NN estimate $\hat{\eta}_n/p_n$. The proof, which is postponed to the supplementary file, crucially relies on arguments from the proof of our Theorem 4.2 combined with known results concerning the VC dimension of Euclidean balls (Wenocur and Dudley, 1981).

Theorem 4.5. *Suppose that (X1) (X2) and (X3) are fulfilled and that $x \mapsto \eta(x)/p$ is L-Lipschitz on S_X. Then whenever $pn/\log(n) \to \infty$, $k/\log(n) \to \infty$ and $k/n \to 0$, we have, with probability* 1*,*

$$\sup_{x \in \mathcal{X}} |\hat{\eta}_n(x)/p_n - \nu^*(x)| = O\left(\sqrt{\frac{\log(n)}{kp}} + \left(\frac{k}{n}\right)^{1/d}\right).$$

The consistency of the balanced k-NN with respect to the AM risk, encapsulated in the next corollary, follows from Theorem 4.5 combined with an additional result (Proposition 4.15) relating the deviations of the empirical regression function with the excess balanced risk.

Corollary 4.6. *Suppose that (X1) (X2) and (X3) are fulfilled and that $x \mapsto \eta(x)/p$ is L-Lipschitz on S_X. Then whenever $p \leq 1/2$, $kp/\log(n) \to \infty$ and $k/n \to 0$, we have, with probability* 1*,*

$$\mathcal{R}^*_{bal}(\hat{g}_n) \to \mathcal{R}^*_{bal}(g^*_{\text{bal}}).$$

The principal interest of Corollary 4.6 is that the condition for consistency involves the product of the number of neighbors k with the rare class probability p. The take-home message is that learning nonparametric decision rules is possible with imbalanced data, as soon as kp is large enough. In other words local averaging process should be done carefully to ensure a sufficiently large *expected* number of neighbors from the rare class.

4.4 Fast rates under relative rarity

4.4.1 A concentration bound for balanced measures

We now state and prove a concentration inequality that is key to obtain fast convergence rates for excess risk in the context of balanced ERM. Prior to stating this main result, we define a weighted class $\widetilde{\mathcal{F}}$ as

$$\widetilde{\mathcal{F}} = \left\{ \frac{1}{2} f I_1 + \frac{1}{2} \frac{p}{1-p} f I_{-1} \mid f \in \mathcal{F} \right\},$$

where $I_s(x,y) = \mathbb{1}\{y = s\}$ for $(x,y) \in \mathcal{X} \times \mathcal{Y}$ ans $s \in \{-1,1\}$. Moreover, for a given measure P, we denote a balanced counterpart of P as $P_{\text{bal}}(f) = \frac{1}{2}\Big(P_+(f) + P_-(f)\Big)$.

Theorem 4.7. *Suppose that $\mathcal{F}$ is of VC-type with envelop $U \geq 1$ and parameter $v, A \geq 1$. Assume that there is some constant B such that for every $\tilde{f} \in \widetilde{\mathcal{F}}, P\left(\tilde{f}^2\right) \leq BP\tilde{f}$. Then, with $c_1 = 5$, $c_2 = 22U$, for any $K > 1$ and every $\delta > 0$, with probability at least $1 - 3\delta$,*

$$\forall f \in \mathcal{F} \quad P_{\text{bal}}(f) \leq \frac{K}{K-1} P_{n,\text{bal}}(f)\left(1 + \sqrt{\frac{3\log(1/\delta)}{np}}\right) + \frac{DK}{B}\frac{\log(An)^2}{np} + U_{B,K}\frac{\log(1/\delta)}{np}.$$

Also, with probability at least $1 - 2\delta$,

$$\forall f \in \mathcal{F} \quad P_{n,\text{bal}}(f)\left(1 - \sqrt{\frac{2\log(1/\delta)}{np}}\right) \leq \frac{K+1}{K} P_{\text{bal}}(f) + \frac{DK}{B}\frac{\log(An)^2}{np} + U_{B,K}\frac{\log(1/\delta)}{np},$$

where $D = 8^{\frac{1}{v}}(v+1)CAUC_1C_2$, $C > 0$ is universal constant, $C_1 = 1/\sqrt{\log(8A)}$, $C_2 = \sqrt{2}\left(\max(\log(4AU)/\log(8A), 1) + \sqrt{2}\right)$ and $U_{B,K} = c_2 + c_1BK$.

Proof [Sketch of proof]

The main tool for the proof is Theorem 3.3 in Bartlett et al. (2005) recalled for completeness in the supplementary material (Theorem 4.21). More precisely, the argument from the cited reference relies heavily on a fixed point technique relative to a subroot function upper bounding the a local variance term. We establish that the fixed point $r^\star$ involved in the argument satisfies an inequality of the form

$$r^\star \leq O\left(\frac{\log(A/r^\star)}{\sqrt{n}}\right).$$

Using this inequality along with the latter theorem and Lemma 7 from Cucker et al. (2002) yields, with high probability, for any $\tilde{f} \in \widetilde{\mathcal{F}}$,

$$P(\tilde{f}) \leq \frac{K}{K-1} P_n(\tilde{f}) + O\left(\frac{\log(An)^2}{n}\right).$$

It remains to notice that in our context of imbalanced classification, for

$$\tilde{f} = \frac{1}{2} f I_1 + \frac{1}{2}\frac{p}{1-p} f I_{-1},$$

one has $P(\tilde{f}) = pP_{\text{bal}}(f)$. The result follows by an application of a Chernoff bound (Theorem 4.12) The full proof can be found in the supplement, in Section 4.C. ■

Discussion. *Similar proof techniques can be found in the standard classification literature, for example Corollary 3.7 in Bartlett et al. (2005). Nevertheless, this particular work primarily concentrates on loss functions with binary values, namely $\{0,1\}$. The*

proof is based upon the fact that these functions are positive, and it employs the conventional definition of the VC *dimension. In contrast, other existing works (e.g. Theorem 2.12 in Bartlett and Mendelson (2006) or Example 7.2 in Giné and Koltchinskii (2006)) demonstrate accelerated convergence rates for the typical empirical risk minimizers, which do not extend to their balanced counterparts. The present result is more general, as it is uniformly applicable to a broader range of bounded functions and encompasses a more extensive definition of the* VC *class. This notable extension facilitates the establishment of fast convergence rates for the excess risk of (*ML*) algorithms employed in imbalanced classification scenarios, such as cost-sensitive logistic regression and balanced boosting (Menon et al., 2013a; Koyejo et al., 2014; Tanha et al., 2020; Xu et al., 2020c). In the next section we provide examples of algorithms verifying the assumptions of Theorem 4.7.*

As an application of Theorem 4.7, we derive fast rates for the excess risk of empirical risk minimizers. The following assumption, known as the Bernstein condition, is a prevalent concept within the fast rates literature (Bartlett and Mendelson, 2006; Klochkov and Zhivotovskiy, 2021a).

Definition 4.8. *We say that the triplet $(\mathcal{G}, P, \ell)$ satisfy the Bernstein condition if for some $B > 0$ it holds that,*

$$\forall g \in \mathcal{G}\,,\ \mathbb{E}\left[\Big(\ell_g(X,Y) - \ell_{g^\star}(X,Y)\Big)^2\right] \le B\left(\mathcal{R}(g) - \mathcal{R}(g^\star)\right),$$

where $g^\star = \arg\min_{g\in\mathcal{G}} \mathcal{R}[g] = \arg\min_{g\in\mathcal{G}} \mathbb{E}\left[\ell_g(X,Y)\right]$.

Set

$$\tilde{\ell}_g = \ell_g I_1 + \frac{p}{1-p}\ell_g I_{-1},$$

and notice that $\mathcal{R}_{\mathrm{bal}}(g) = \mathbb{E}\left[\tilde{\ell}_g(X,Y)\right]/p$, so that

$$g^\star_{\mathrm{bal}} = \arg\min_{g\in\mathcal{G}} P(\tilde{\ell}_g) = \arg\min_{g\in\mathcal{G}} \mathcal{R}_{\mathrm{bal}}(g)\,.$$

In the sequel we shall suppose that the latter condition holds for $(\mathcal{G}, P, \tilde{\ell})$ in order to apply Theorem 4.7 and obtain fast convergence rates for the excess risk. The proof is deferred to the supplementary material.

Corollary 4.9. *Suppose that $\mathcal{F} = \{\ell_g : g \in \mathcal{G}\}$ is VC, L-bounded and assume that $(\mathcal{F}, P, \tilde{\ell})$ satisfy the Bernstein condition for some $B > 0$. Then, for any $\delta > 0$, we have with probability $1 - 4\delta$,*

$$\mathcal{R}_{\mathrm{bal}}(\hat{g}_{\mathrm{bal}}) \le \mathcal{R}_{\mathrm{bal}}\left(g^\star_{\mathrm{bal}}\right) + \frac{D}{B}\frac{\log(An)^2}{np} + \frac{\log(1/\delta)\left(c_2 + c_1 B\right)}{np},$$

where the constants appearing in the latter inequality are the same as in Theorem 4.7.

In the following lemma, we provide a sufficient condition for the Bernstein assumption (Definition 4.8). The proof and the definition of strongly convex function is given in the supplement.

Lemma 4.10. *Suppose that the family $\mathcal{G}$ is a normed space. if $g \mapsto \mathbb{E}\left[\ell_g(X,Y)\right]$ is L-Lipschitz and λ-strongly convex, then $(\mathcal{F}, P, \tilde{\ell})$ verifies the Bernstein assumption (Definition 4.8) with $B = 2L^2/\lambda$.*

We conclude this section by an illustration of the significance of our results, through the concrete example of a constrained empirical risk minimization problem over a linear class of classifiers [1]. We show that fast rates of convergence are achieved provided that the covariate space $\mathcal{X} \in \mathcal{R}^d$ is bounded and the loss is twice differentiable with a second derivative lower bounded away from 0. More precisely, we make the following assumption.

Assumption 6. *The space $\mathcal{X}$ is bounded in $\mathbb{R}^d$* i.e. *, there exists some $\Delta_X > 0$ such as $\forall x \in \mathcal{X}, \|x\| \leq \Delta_X$ for a given norm $\|\cdot\|$. Furthermore, the family of classifier and the loss function are chosen as $\mathcal{G}_u = \left\{g(x) = \beta^T x \mid \left\|\beta\right\| \leq u\right\}$ and $\ell_g(X,Y) = \phi(\beta^T XY)$, where $\phi : \mathbb{R} \mapsto \mathbb{R}$ is a twice differentiable function verifying $\inf_{|x|\leq u\Delta_X} \phi''(x) > \lambda$ for some $\lambda > 0$.*

An immediate implication of the aforementioned assumption is that, identifying g with β, we have $\sup_{x,y} \left\| \frac{\partial}{\partial g} \ell_g(x,y) \right\| < \infty$ which ensures that the risk is Liptchitz. In addition this assumption guarantees that the risk is λ-strongly convex with respect to g. The following corollary is a direct consequence of Corollary 4.9 and guarantees fast rates of convergence for constrained ERM, specifically for algorithms of the form $\hat{g}_{u,bal}(x) = \hat{\beta}_u^T x$ with

$$\hat{\beta}_u = \underset{\|\beta\|\leq u}{\arg\min} \frac{1}{n} \sum_{i=1}^{n} \phi(\beta^T X_i Y_i) \left(\frac{\mathbb{1}\{Y=1\}}{p_n} + \frac{\mathbb{1}\{Y=-1\}}{1-p_n} \right). \tag{4.1}$$

Corollary 4.11. *Suppose that Assumption 6 holds for some $\lambda > 0$. Then the excess risk of $\hat{g}_u$ verifies, for any $\delta > 0$, with probability $1 - 4\delta$,*

$$\mathcal{R}_{\text{bal}}\left(\hat{g}_{\text{bal}}\right) - \mathcal{R}_{\text{bal}}\left(g^\star_{\text{bal}}\right) \leq \frac{D\lambda}{L'^2} \frac{\log(An)^2}{np} + \frac{\log(1/\delta)\left(c_2 + 2c_1(L'^2/\lambda)\right)}{np},$$

where $L' = \sup_{|x|\leq u\Delta_X} \phi'(x)$.

Discussion. *In the context of constrained logistic regression, where $\phi(x) = \log(1+e^{-x})$, the latter corollary yields fast convergence rates with constants and $L' = 1$, along with $\lambda = e^{-u}$. Corollary 4.11 further establishes accelerated convergence rates for constrained empirical* balanced *risk minimization with respect to losses such as mean squared error, squared hinge, and exponential loss, among others. This outcome aligns with expectations, as constrained empirical risk minimization is equivalent to penalization (Lee et al., 2006; Homrighausen and McDonald, 2017). Numerous studies have demonstrated the effectiveness of penalization in achieving rapid convergence rates (Koren and Levy, 2015; van Erven et al., 2015). This aspect is particularly significant in the present context, as the standard convergence rate for imbalanced classification is $1/\sqrt{np}$, and accelerating the convergence rate leads to a more pronounced impact.*

[1] Non-linear classifiers can be easily produced with the use of kernels.

4.5 Numerical illustration

In this section, we provide, using synthetic data, a numerical illustration of the theoretical results on k-NN classification (Corollary 4.6) and on logistic regression (Corollary 4.11). In both cases, a particular attention is given to highly imbalanced setting where $p = n^{-a}$ for some $0 < a < 1$. Due to space constraint, the real data based numerical experiments are postponed to the supplementary file.

Synthetic dataset. In the two cases considered below, we use the following simple data generation process. Consider the binary classification dataset $(X_i, Y_i)_{i=1,\dots,n}$ such that $X_i \in \mathbb{R}^2$ and $Y \in \{-1, 1\}$. For each i, the random variable Y_i is such that $P(Y_i = 1) = (1/n^a)$, for some $a < 1$. Then, having generated $Y_i = y$, X_i is drawn according to a t-multivariate-student distribution, with parameters (μ_y, σ_y, ν_y). We set $(\mu_{-1}, \mu_1) = ((0, 0), (1, 1))$, $\sigma_1 = 3\sigma_{-1} = 3I$ and $(\nu_{-1}, \nu_1) = (2.5,\ 1.1)$.

4.5.1 Balanced k-nearest neighbors

Corollary 4.6 gives conditions on k and p to ensure the consistency of the k-NN classification rule. The key condition on which we focus here is that kp should go to ∞. This condition suggests the existence of a learning frontier on the set (k, p) above which consistent learning is ensured. Here we validate empirically this result and we also provide numerical results to support the stronger conclusion that whenever kp is not large enough (we are below the learning frontier), then k-NN is no longer consistent making clear that the choice of the number of neighbors k should be made considering the value of p.

The experiments setup is as follows. The training size is $n = 1e4$. We set $p = 1/n^a$ and $k = n^b$, while varying a, b over the interval $[1/4, 3/4]$ to cover different cases ranging from $pn \to 0$ to $pn \to \infty$. The AM-risk for the classification error associated to the balanced k-NN classifier (estimated with 20 simulations) is displayed as a function of (k, p) in Figure 4.1.

Upon examining the figure, it is observed that the performance of the k-nearest neighbors classifier mirrors that of a random guess, maintaining an AM risk near 0.5, when kp is kept small. This observation illustrates (and extends) the conclusion of Corollary 4.6, supporting that consistency is obtained if (and only if) $kp \to \infty$.

4.5.2 Balanced ERM

Now, keeping in mind the fast convergence rate $1/(np)$ obtained in Corollary 4.11, our goal is to show that such a rate is quite sharp as it can be recovered in practice.

We consider the simple setting of a linear classifier defined as $\hat{\beta}_u = \hat{g}$ introduced in Section 4.4 with logistic loss $\ell_g(X, Y) = \log(1 - e^{-g(X)Y})$, $g(X) = \beta^T X$ and $u = 10$. Here the sample size n is ranging over the grid $[100, 1e4]$ and rare class probability $p = n^{-a}$ while $a \in \{1/3, 1/2, 2/3\}$.

Some Monte Carlo simulation are needed to estimate $g^\star_{\text{bal}}$. We use an $1e5$ simulations according to a well balanced data set ($p = 1/2$) so that the error computing $g^\star_{\text{bal}}$ is sufficiently small. In addition, we use some more Monte Carlo simulation from a balanced

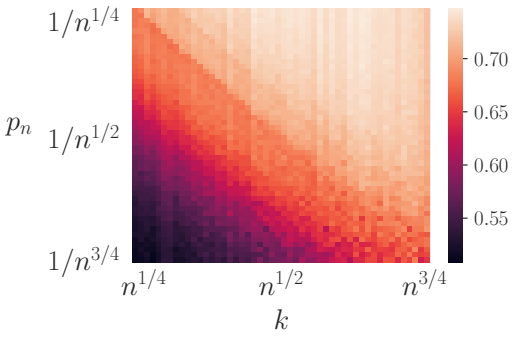
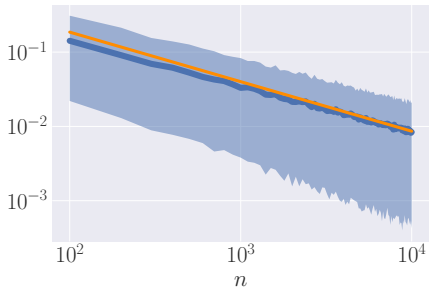
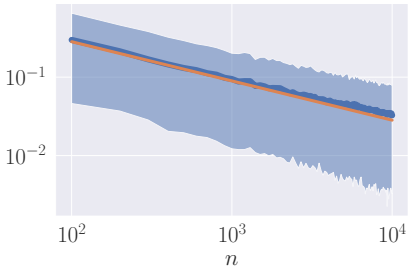


Figure 4.1 – Heatmap showing the AM risk of the balanced k-NN.

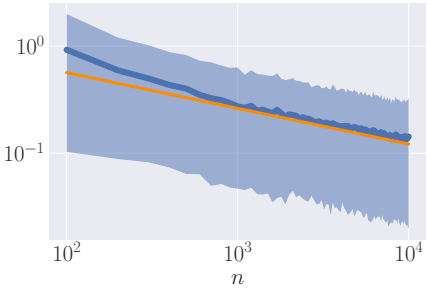
test dataset of size $1e4$, to evaluate without bias the risk function $\mathcal{R}_{\text{bal}}$. Based on this, we can obtain both $\mathcal{R}_{\text{bal}}(g^\star_{\text{bal}})$ and $\mathcal{R}_{\text{bal}}(\hat{g})$ so that an excess risk value can be obtained. We perform $n_{simu} = 1e4$ experiments and we report the average and the upper 0.10 and 0.90 quantile of the absolute error obtained over the n_{simu} experiments.



(a) $p_n = 1/n^{1/3}$.



(b) $p_n = 1/n^{1/2}$.



(c) $p_n = 1/n^{2/3}$.

Figure 4.2 – Excess risk (blue) of logistic regression for different sample size n and the curve $1/np$ (orange). The blue area corresponds to the 0.9-confidence interval.

.

Figure 4.2 displays the excess risk as a function of the sample size n in a logarithmic scale, for $a \in \left\{1/3, 1/2, 2/3\right\}$. We notice that the excess risk vanishes in the same way as the function $n \mapsto 1/np$ confirming the accuracy of the upper bound from Corollary 4.11.

4.6 Conclusion

In this chapter, we have derived upper bounds for the balanced risk in highly imbalanced classification scenarios. Notably, our bounds remain consistent even under severe class imbalance ($p \to 0$), setting our work apart from existing studies in imbalanced classification (Menon et al., 2013a; Koyejo et al., 2014; Xu et al., 2020c). Furthermore, it is worth to highlight that this is the first study to achieve fast rates in imbalanced classification, marking a significant advancement in the field.

Our findings corroborate that both risk-balancing approaches and cost-sensitive learning are consistent across nearly all imbalanced classification scenarios. This aligns with experimental works previously documented in the literature (Elkan, 2001b; Wang et al., 2016; Khan et al., 2018; Wang et al., 2019a; Pathak et al., 2022). We also

Furthermore, the methodologies and proof techniques presented in this chapter are adaptable to other imbalanced classification metrics beyond balanced classification. Potential extensions include demonstrating consistency for metrics such as the F_1 measure, recall, and their respective variants.

4.A Auxiliary results

The following standard Chernoff inequality is stated and proven in Hagerup and Rüb (1990).

Theorem 4.12. *Let $(Z_i)_{i\geq 1}$ be a sequence of i.i.d. random variables valued in $\{0,1\}$. Set $\mu = nP(Z_1)$ and $S = \sum_{i=1}^{n} Z_i$. For any $\delta \in (0,1)$ and all $n \geq 1$, we have with probability $1-\delta$:*

$$S \geq \left(1 - \sqrt{\frac{2\log(1/\delta)}{\mu}}\right)\mu.$$

In addition, for any $\delta \in (0,1)$ and $n \geq 1$, we have with probability $1-\delta$:

$$S \leq \left(1 + \sqrt{\frac{3\log(1/\delta)}{\mu}}\right)\mu.$$

The following is taken from Plassier et al. (2023) (see also Giné and Guillou (2001) for similar uniform bound).

Theorem 4.13. *Let $(Z, Z_1, \ldots, Z_n)$ be an independent and identically distributed collection of random variables in $(S, \mathcal{S})$. Let $\mathcal{G}$ be a VC class of functions with parameters $v \geq 1$, $A \geq 1$ and uniform envelope $U \geq \sup_{g\in\mathcal{G}, x\in S} |g(x)|$. Let σ be such that $\sigma^2 \geq \sup_{g\in\mathcal{G}} var(g(Z))$ and $\sigma \leq 2U$. For any $n \geq 1$ and $\delta \in (0,1)$, it holds, with probability $1-\delta$,*

$$\sup_{g\in\mathcal{G}} \left| \sum_{i=1}^{n} \{g(Z_i) - \mathbb{E}[g(Z)]\} \right| \leq K' \left(\sigma \sqrt{vn \log\left(K'\theta/\delta\right)} + Uv \log\left(K'\theta/\delta\right) \right), \tag{4.2}$$

with $\theta = AU/\sigma$ and $K' > 0$ a universal constant.

Lemma 4.14. *Suppose that $\mathcal{F}$ is of VC-type with envelop U and parameter (v, A), then*

1. *$\{fI_C : f \in \mathcal{F}\}$ is of VC-type with envelop U and parameter (v, A).*
2. *$\{f - P_C(f) : f \in \mathcal{F}\}$ is of VC-type with envelop $2U$ and parameter $(2v, A)$.*

Proof

Let Q be a probability measure on $(S, \mathcal{S})$. Let $(f_k)_{k=1,\dots,K}$ be the center of an ϵU-covering of $(\mathcal{F}, Q)$. The first result follows from the fact that $\|f1_C - f_k1_C\|_{L_2(Q)} \leq \|f - f_k\|_{L_2(Q)}$. Now let $(\tilde{f}_k)_{k=1,\dots,K}$ be the center of an ϵU-covering of $(\mathcal{F}, P_C)$. Consider the covering induces by the centers $(f_k - P_C(\tilde{f}_j))_{1\leq k,j\leq K}$ made of K^2 elements. Suppose that $f \in \mathcal{F}$. Then there is k and j such that

$$\begin{aligned}\|(f - P_C(f)) - (f_k - P_C(\tilde{f}_j)\|_{L_2(Q)} &\leq \|f - f_k\|_{L_2(Q)} + P_C(f - \tilde{f}_j) \\ &\leq \|f - f_k\|_{L_2(Q)} + \|f - \tilde{f}_j\|_{L_2(P)} \\ &\leq 2U\epsilon.\end{aligned}$$

Hence we have found a $2U\epsilon$-covering of size K^2 which by assumption is smaller than $(A/\epsilon)^{2v}$. This implies the second statement of the lemma.

■

The next lemma generalizes Theorem 17.1 from Biau and Devroye (2015) to the balanced type classifiers.

Lemma 4.15. *For any classifier g that writes $g(x) = \text{sign}(\nu(x) - 1)$, $x \in \mathcal{X}$, we have*

$$\mathcal{R}^*_{bal}(g) - \mathcal{R}^*_{bal}(g^*_{bal}) = \mathbb{E}\left[\mathbb{1}_{g(X)\neq g^*(X)}\frac{|\eta(X) - p|}{p(1-p)}\right],$$

where g^*_{bal} is the balanced Bayes classifier (introduced in Section 4.2). Furthermore, whenever $p \leq 1/2$,

$$\mathcal{R}^*_{bal}(g) - \mathcal{R}^*_{bal}(g^*_{bal}) \leq 2\mathbb{E}\left[\left|\nu(X) - \nu^*(X)\right|\right]$$

where $\nu^*(x) = \eta(x)/p$.

Proof The balanced risk writes as

$$\begin{aligned}\mathcal{R}^*_{bal}(g) &= P_+\left(\nu(X) < 1\right) + P_-\left(\nu(X) \geq 1\right) \\ &= \mathbb{E}\left[\frac{\mathbb{I}_{(\nu(X)<1)}\mathbb{I}_{Y=1}}{p} + \frac{\mathbb{I}_{(\nu(X)\geq 1)}\mathbb{I}_{Y=-1}}{1-p}\right].\end{aligned}$$

In addition, using a conditioning argument yields,

$$\mathcal{R}^*_{bal}(g) = \mathbb{E}\left[\frac{\mathbb{I}_{(\nu(X)<1)}\eta(X)}{p} + \frac{\mathbb{I}_{(\nu(X)\geq 1)}(1 - \eta(X))}{1-p}\right].$$

Similarly we have

$$\mathcal{R}^*_{bal}(g^*) = \mathbb{E}\left[\frac{\mathbb{I}_{(\nu^*(X)<1)}\eta(X)}{p} + \frac{\mathbb{1}_{(\nu^*(X)\geq 1)}(1-\eta(X))}{1-p}\right].$$

It follows that

$$\begin{aligned}\mathcal{R}^*_{bal}(g) - \mathcal{R}^*_{bal}(g^*) &= \mathbb{E}\left[\mathbb{I}_{\text{sign}(\nu^*(X)-1)\neq\text{sign}(\nu(X)-1)}\frac{|\eta(X)-p|}{p(1-p)}\right] \\ &= \mathbb{E}\left[\mathbb{I}_{g^*(X)\neq g(X)}\frac{|\eta(X)-p|}{p(1-p)}\right],\end{aligned}$$

This concludes the first part. For the second part, it remains to note that for any real numbers (x, y)

$$\text{sign}(x-1) \neq \text{sign}(y-1) \implies |y-1| \leq |x-y|,$$

so that, using that $\nu^* = \eta^*/p$, we obtain

$$\begin{aligned}\mathcal{R}^*_{bal}(g) - \mathcal{R}^*_{bal}(g^*) &= \mathbb{E}\left[\mathbb{I}_{\text{sign}(\nu^*(X)-1)\neq\text{sign}(\nu(X)-1)}\frac{|\eta(X)-p|}{p(1-p)}\right] \\ &= \mathbb{E}\left[\mathbb{I}_{\text{sign}(\nu^*(X)-1)\neq\text{sign}(\nu(X)-1)}\frac{|\nu^*(X)-1|}{(1-p)}\right] \\ &\leq \frac{\mathbb{E}\left[|\nu^*(X)-\nu(X)|\right]}{1-p},\end{aligned}$$

but since $p \leq 1/2$ we obtain the desired result.

■

4.B Standard rates proof

4.B.1 Proof of Theorem 4.2

Starting with

$$P_{n,+}\left(f\right) - P_+\left(f\right) = \frac{P_n\left(\left(f - P_+\left(f\right)\right)\mathbb{1}_{\{Y=1\}}\right)}{p_n} \tag{4.3}$$

we focus on each term, denominator and numerator, separately. For the numerator, the term $\left(f - P_+(f)\right)\mathbb{1}_{\{Y=1\}}$ has mean 0. In virtue of Lemma 4.14, the class $(f - P_+(f))\mathbb{1}_{\{Y=1\}}$ is still bounded by $2U$ and is still VC with VC parameter $(2v, A)$. As a consequence, we can use Proposition 2 in Plassier et al. (2023), stated as Theorem 4.13 in the present supplementary file. The variance is bounded as follows

$$\text{Var}\left(\left(f - P_+(f)\right)\mathbb{1}_{\{Y=1\}}\right) \leq P\left((f - P_+(f))^2\mathbb{1}_{\{Y=1\}}\right) = \text{Var}_+(f)p = \sigma^2_+ p,$$

by definition of σ_+^2. As a consequence, Theorem 4.13 gives that

$$\begin{aligned}
&P_n\left(\left(f - P_+(f)\right)\mathbb{1}_{\{Y=1\}}\right) \\
&\leq K'\left(\sqrt{\frac{v\sigma_+^2 p}{n}\log\left(K'A/\left(2\delta\sqrt{p}\right)\right)} + \frac{Uv}{n}\log\left(K'A/\left(2\delta\sqrt{p}\right)\right)\right) \\
&\leq 2K'\sqrt{\frac{v\sigma_+^2 p}{n}\log\left(K'A/\left(2\delta\sqrt{p}\right)\right)},
\end{aligned}$$

where the last inequality has been obtained using the stated condition on n and δ. For the denominator, using Theorem 4.12 we have that, with probability $1-\delta$,

$$\frac{p_n}{p} \geq \left(1 - \sqrt{\frac{2\log(1/\delta)}{np}}\right) \geq 1/2$$

where the last inequality has been obtained using the condition on n and δ. Using the union bound, we get, with probability $1-2\delta$,

$$\frac{P_n\left(\left(f - P_+(f)\right)\mathbb{1}_{\{Y=1\}}\right)}{p_n} \leq 4K'\sqrt{\frac{v\sigma_{\mathcal{F},C}^2}{np}\log\left(K'A/\left(2\delta\sqrt{p}\right)\right)}$$

and the proof is complete.

4.B.2 Proof of Corollary 4.4

First, using the definition of $\hat{g}_{\text{bal}}$ yields

$$\mathcal{R}_{n,\text{bal}}\left(\hat{g}_{\text{bal}}\right) - \mathcal{R}_{n,bal}\left(g_{\text{bal}}^*\right) \leq 0,$$

So that,

$$\begin{aligned}
\mathcal{R}_{\text{bal}}(\hat{g}_{\text{bal}}) - \mathcal{R}(g_{\text{bal}}^*) &\leq \mathcal{R}_{\text{bal}}(\hat{g}_{\text{bal}}) - \mathcal{R}_{n,bal}(\hat{g}_{\text{bal}}) - \left(\mathcal{R}_{\text{bal}}(g_{\text{bal}}^*) - \mathcal{R}_{n,\text{bal}}(g_{\text{bal}}^*)\right) \\
&\leq \sup_{g\in\mathcal{G}}\left|\mathcal{R}_{\text{bal}}(g) - \mathcal{R}_{n,\text{bal}}(g)\right| \\
&\leq \sup_{g\in\mathcal{G}}\left|P_{n,-}(g) - P_-(g)\right| + \sup_{g\in\mathcal{G}}\left|P_{n,+}(g) - P_+(g)\right|.
\end{aligned}$$

It remains to use Theorem 4.2 and the proof is complete.

4.B.3 Proof of Theorem 4.5

First we recall three results that will be useful in the proof. The following Lemma (Portier, 2021, Lemma 4) controls the size of the k-NN balls uniformly over all $x \in S_X$.

Lemma 4.16 (Portier (2021, Lemma 4))**.** *Suppose that (X1) (X2) and (X3) hold true. Then for all $n \geq 1$, $\delta \in (0,1)$ and $1 \leq k \leq n$ such that $24d\log(12n/\delta) \leq k \leq T^d n b_X c V_d/2$, it holds, with probability at least $1-\delta$:*

$$\sup_{x \in S_X} \hat{\tau}_{n,k,x} \leq \overline{\tau}_{n,k} := \left(\frac{2k}{n b_X c V_d} \right)^{1/d}, \tag{4.4}$$

where $V_d = \lambda(B(0,1))$.

The following lemma is a simple consequence of Theorem 4.12.

Lemma 4.17. *Let $z_n = \sqrt{2\log(1/\delta)/(np)}$. We have with probability at least $1-\delta$,*

$$\frac{p}{p_n} - 1 \leq \frac{z_n}{1-z_n}. \tag{4.5}$$

The next lemma is a consequence of Theorem 4.13. Let

$$\mathcal{G} = \{ g(Y,X) = (\mathbb{1}_{Y=1} - \eta(X)) \mathbb{I}_{\|X-x\| \leq \tau} : \tau \leq \overline{\tau}_{n,k},\, x \in \mathbb{R}^d \}$$

which is of VC type as shown in Lemma 9 in Portier (2021) (see also Wenocur and Dudley (1981)). Because S_X is compact and η/p continuous, there exists C such that $\eta(x) \leq pC$ for all $x \in S_X$. The variance of each element in the class is bounded as

$$\mathrm{Var}(g(Y,X)) \leq E(\mathbb{1}_{Y=1} \mathbb{I}_{\|X-x\| \leq \tau}) \leq \int \eta(z) \mathbb{I}_{\|z-x\| \leq \tau} f_X(z) dz \leq C p U_X \overline{\tau}_{n,k}^d V_d.$$

Injecting the previous variance bound (which is scales as pk/n) in the upper-bound given in Theorem 4.13 we obtain the following statement.

Lemma 4.18. *We have with probability at least $1-\delta$,*

$$\sup_{g \in \mathcal{G}} \left| \sum_{i=1}^n g(Y_i, X_i) \right| \leq C_1 (\sqrt{kp \log(C_2/\delta)} + \log(C_2/\delta)) \tag{4.6}$$

where C_1 and C_2 are constants that does not depend on n, k and p (but on the dimension d, the VC parameter of $\mathcal{G}$, and the probability measure P_X).

Define the event E_n as the union of (4.4), (4.5) and (4.6). By the previous three lemmas and the union bound $P(E_n) \geq 1 - 3\delta$. In light of Borel-Cantelli Lemma we choose $\delta = 1/n^2$ so that $\sum_n (1 - P(E_n))$ is finite and the event $\liminf_n E_n$ has probability 1. It then suffices to show that E_n implies that $\hat{\eta}_n(x)/p_n - \nu^*(x) = O(\sqrt{\log(n)/kp} + (k/n)^{1/d})$. Note that under E_n, when n is large enough, by (4.5), $p/p_n \leq 2$. Let $M_i = \mathbb{1}_{Y_i=1} - \eta(X_i)$ and $B_i(x) = \eta(X_i) - \eta(x)$. We have

$$\frac{\hat{\eta}_n(x)}{p_n} - \nu^*(x) = \frac{\sum_{i \in I_n(x)} M_i}{k p_n} + \frac{\sum_{i \in I_n(x)} B_i(x)}{k p_n} + \eta(x) \left(\frac{1}{p_n} - \frac{1}{p} \right). \tag{4.7}$$

On the event E_n, the function $(Y,X) \mapsto (\mathbb{1}_{Y=1} - \eta(X)) \mathbb{1}_{\|X-x\| \leq \hat{\tau}_{n,k,x}}$ belongs to the space $\mathcal{G}$. Consequently, the first term in (4.7) is smaller than

$$(k p_n)^{-1} \sup_{g \in \mathcal{G}} \left| \sum_{i=1}^n g(Y_i, X_i) \right|$$

which by (4.5) and (4.6) is $O(\sqrt{\log(n)/kp})$. Using the assumption that $x \mapsto \eta(x)/p$ is L-Lipschitz we get that, on E_n, the second term in (4.7) is such that

$$\frac{\sum_{i \in I_n(x)} B_i(x)}{kp_n} \leq \frac{p}{p_n} L\overline{\tau}_{n,k},$$

which, using (4.5), is $O((k/n)^{1/d})$. The third term in (4.7) is smaller than $\Big(\eta(x)/p\Big)\Big(p/p_n - 1\Big)$ which is, using again the Lipschitz assumption and (4.5), $O(\sqrt{\log(n)/(np)})$. The latter bound is smaller than $\sqrt{\log(n)/(kp)}$ so it does not appear in the stated bound.

4.C Fast rates proofs

Before moving to the main proof we remind some necessary notions. First, let's recall the definitions of *sub-root* functions:

Definition 4.19. *A function $\psi : [0, \infty) \to [0, \infty)$ is sub-root if it is nonnegative, non-decreasing and if $r \mapsto \psi(r)/\sqrt{r}$ is nonincreasing for $r > 0$.*

In the sequel, we will focus on a specific type of functions, given by:

$$\phi_{\mathcal{F}}(r) = P\left(R_n\left\{f \in \mathcal{F} \mid P_n f^2 \leq r\right\}\right)$$

where $R_n\left(\mathcal{F}\right) = \frac{1}{n}\sup_{f\in\mathcal{F}}\sum_{i=1}^n \sigma_i f(X_i)$ denotes the empirical Rademacher complexity for a given realisation of σ_i's, X_i's. The expectation in this formulation is with respect to the budget of samples $(X_i, Y_i)_{1\leq i\leq n}$ and the Rademacher chaos variables σ_i's. Remember that $\sigma_i \in \left\{-1, 1\right\}$ and $P(\sigma_i = \pm 1) = \frac{1}{2}$.
At this point, it is important to mention that if we define $\mathcal{G} = star(\mathcal{F}, 0)$ as the *star-hull* of $\mathcal{F}$ centered around 0, that is,

$$\mathcal{G} = star(\mathcal{F}, 0) = \left\{\alpha f \mid \alpha \in [0, 1], f \in \mathcal{F}\right\},$$

then the function $\phi_{\mathcal{G}}$ is sub-root (see *e.g.* Lemma 3.4 in Bartlett et al. (2005)). In the next lemma we derive an upper bound for $\phi_{\mathcal{G}}(2r)$, which is a crucial quantity in the proof of Theorem 4.7.

Lemma 4.20. *Let $\mathcal{F}$ be a class of functions that is VC-type with envelop $U \geq 1$ and parameter $v, A \geq 1$. Set $\mathcal{G}$ to be the star-hull of $\mathcal{F}$ around 0. Then, the Rademacher complexity $\phi_{\mathcal{G}}(2r)$ verifies*

$$\phi_{\mathcal{G}}(2r) = P\left(R_n\left\{g \in \mathcal{G} \mid P_n g^2 \leq 2r\right\}\right) \leq CD\log\left(\frac{8AU}{r}\right)\sqrt{\frac{v+1}{n}}\sqrt{r}.$$

Where $D_1 = 8^{1/v}CAUC_1C_2$, $C > 0$ is a universal constant, $C_1 = \frac{1}{\sqrt{\log(8A)}}$ and $C_2 = \sqrt{2}\left(\max\left(\frac{\log(4AU)}{\log(8A)}, 1\right) + \sqrt{2}\right)$.

Proof By definition of $\mathcal{G}$, any element $g \in \mathcal{G}$ verifies $P_n g^2 \leq U^2$, thus for $r \geq \frac{U^2}{2}$ the function $\phi_{\mathcal{G}}(r)$ is constant. Therefore if the latter inequality holds for $r \leq \frac{U^2}{2}$ it automatically extends to the case $r \geq \frac{U^2}{2}$. Thus, in the sequel we assume that $r \leq \frac{U^2}{2}$. Using classical results (see *e.g.* the discussion following Lemma 7.3 in Van Handel (2014)) one has, for some universal constant $C > 0$,

$$P_\sigma \left[R_n(\mathcal{G}_{2r}) \right] \leq \frac{C}{\sqrt{n}} \int_0^\infty \sqrt{\log \mathcal{N}\left(\mathcal{G}_{2r}, L_2(P_n), \epsilon\right)} d\epsilon,$$

which yields,

$$P\left(R_n(\mathcal{G}_{2r})\right) \leq P\left(\frac{C}{\sqrt{n}} \int_0^\infty \sqrt{\log \mathcal{N}\left(\mathcal{G}_{2r}, L_2(P_n), \epsilon\right)} d\epsilon\right),$$

where $\mathcal{G}_r = \left\{ g \in \mathcal{G} \mid P_n g^2 \leq r \right\}$ and $C > 0$ is a universal constant. By definition of $\mathcal{G}_r$, the covering number of the latter class verifies $\mathcal{N}\left(\mathcal{G}_r, L_2(P_n), \epsilon\right) = 1$ as soon as $\epsilon \geq \sqrt{r}$. Therefore,

$$\begin{aligned}
P\left(R_n(\mathcal{G}_{2r})\right) &\leq \frac{C}{\sqrt{n}} P\left(\int_0^{\sqrt{2r}} \sqrt{\log \mathcal{N}\left(\mathcal{G}_{2r}, L_2(P_n), \epsilon\right)} d\epsilon\right) \\
(\mathcal{G}_{2r} \subset \mathcal{G}) &\leq \frac{C}{\sqrt{n}} P\left(\int_0^{\sqrt{2r}} \sqrt{\log \mathcal{N}\left(\mathcal{G}, L_2(P_n), \epsilon\right)} d\epsilon\right) \\
&\leq \frac{C}{\sqrt{n}} P\left(\int_0^{\sqrt{2r}} \sqrt{\log \left(\mathcal{N}\left(\mathcal{F}, L_2(P_n), \frac{\epsilon}{2}\right)\left(\frac{4}{\epsilon}\right)\right)} d\epsilon\right). \qquad (4.8)
\end{aligned}$$

Where the last line follows from Lemma 4.5 in Mendelson (2002). On the other hand, using the VC assumption we obtain

$$\forall \epsilon \leq \sqrt{2r}\,, \mathcal{N}\left(\mathcal{F}, L_2(P_n), \epsilon\right) \leq 2\left(\frac{AU}{\epsilon}\right)^v.$$

So that

$$\begin{aligned}
\log\left(\mathcal{N}\left(\mathcal{F}, L_2(P_n), \epsilon/2\right)\left(\frac{4}{\epsilon}\right)\right) &\leq (v+1)\log\left(\frac{(AU)^{\frac{v}{v+1}} 8^{1/v}}{\epsilon}\right) \\
(A \geq 1\,,\, U \geq 1) &\leq (v+1)\log\left(\frac{8AU}{\epsilon}\right).
\end{aligned}$$

Therefore, Inequality (4.8) becomes

$$\begin{aligned} P\left(R_n(\mathcal{G}_{2r})\right) &\leq C\sqrt{\frac{v+1}{n}}\int_0^{\sqrt{2r}}\sqrt{\log\left(\frac{8AU}{\epsilon}\right)}d\epsilon \\ &\leq CC_1\sqrt{\frac{v+1}{n}}\int_0^{\sqrt{2r}}\log\left(\frac{8AU}{\epsilon}\right)d\epsilon \\ &\leq 8^{1/v}AUCC_1\sqrt{\frac{v+1}{n}}\int_0^{\sqrt{2r/(8^{1/v}AU)^2}}\log\left(\frac{1}{\epsilon}\right)d\epsilon. \\ \left(8^{1/v}\leq 8\right) &\leq 8^{1/v}AUCC_1\sqrt{\frac{v+1}{n}}\int_0^{\sqrt{2r/(8AU)^2}}\log\left(\frac{1}{\epsilon}\right)d\epsilon. \end{aligned}$$

Where $C_1 = \frac{1}{\sqrt{\log(8A)}} > 0$. Indeed, the second inequality follows since

$$\forall r \leq \frac{U^2}{2}\,,\ \forall \epsilon \leq \sqrt{2r}\,,\ \sqrt{\log\left(\frac{8AU}{\sqrt{\epsilon}}\right)} \leq C_1\log\left(\frac{8AU}{\sqrt{\epsilon}}\right).$$

Besides, $\int\log\left(\frac{1}{x}\right) = -\int\log(x) = x\log\left(\frac{1}{x}\right) + x$ which yields,

$$P\left(R_n(\mathcal{G}_{2r})\right) \leq 8^{1/v}AUCC_1C_2\log\left(\frac{8AU}{r}\right)\sqrt{\frac{v+1}{n}}\sqrt{r}. \tag{4.9}$$

For some constant $C_2 = \sqrt{2}\left(\max\left(\frac{\log(4AU)}{\log(8A)},1\right)+\sqrt{2}\right) > 0$. Indeed, by considering the cases $r \geq \sqrt{2r}$, $r \leq \sqrt{2r}$ and by simple algebra one has

$$\forall r \leq \frac{U^2}{2}\,,\ \sqrt{2r}\log\left(\frac{8AU}{\sqrt{2r}}\right) + \sqrt{2r} \leq C_2\sqrt{r}\log\left(\frac{8AU}{r}\right).$$

In fact, $r \leq \sqrt{2r}$ implies $\frac{1}{\sqrt{2r}} \leq \frac{1}{r}$ and $r \leq 2$, so that

$$\begin{cases} \sqrt{r}\log\left(\frac{8AU}{r}\right) & \geq \sqrt{r}\log\left(\frac{8AU}{\sqrt{2r}}\right) \\ \sqrt{r}\log\left(\frac{8AU}{r}\right) & \geq \sqrt{r}\log\left(4AU\right) \end{cases}.$$

On the other hand, $r \geq \sqrt{2r}$ implies $r \geq 2$. Now, remind that $r \leq \frac{U^2}{2}$ and write

$$\begin{cases} \log\left(\frac{8AU}{r}\right) & \geq \log\left(8A\right) \\ \log\left(\frac{8AU}{\sqrt{2r}}\right) & \leq \log\left(4AU\right) \leq \log\left(4AU\right)\log\left(\frac{8AU}{r}\right)/\log\left(8A\right) \end{cases},$$

which yields the desired result. ■

The proof of Theorem 4.7 relies on applying Theorem 3.3 in Bartlett et al. (2005) recalled for completeness below.

Theorem 4.21. *Let $\mathcal{F}$ be a class of functions with ranges in $[a,b]$ and assume that there are some functional $T : \mathcal{F} \to \mathbb{R}^+$and some constant B such that for every $f \in \mathcal{F}, \mathrm{Var}[f] \leq T(f) \leq BPf$. Let ψ be a sub-root function and let $r^\star$ be the fixed point of ψ,* i.e. *$\psi(r^\star) = r^\star$. Assume that ψ satisfies, for any $r \geq r^\star$,*

$$\psi(r) \geq BP\left(R_n\{f \in \mathcal{F} : T(f) \leq r\}\right)$$

Then, with $c_1 = 704$ and $c_2 = 26$, for any $K > 1$ and every $x > 0$, with probability at least $1 - e^{-x}$,

$$\forall f \in \mathcal{F} \quad Pf \leq \frac{K}{K-1}P_n f + \frac{c_1 K}{B}r^\star + \frac{x\left(11(b-a) + c_2 BK\right)}{n}.$$

Also, with probability at least $1 - e^{-x}$,

$$\forall f \in \mathcal{F} \quad P_n f \leq \frac{K+1}{K}Pf + \frac{c_1 K}{B}r^\star + \frac{x\left(11(b-a) + c_2 BK\right)}{n}.$$

Furthermore if the functional T verifies $T(\alpha f) \leq \alpha^2 T(f)$ then the same inequalities holds with $c_2 = 6$ and $c_1 = 5$.

Let us now demonstrate a direct corollary of Lemma 7 from Cucker et al. (2002), which will enable us to establish an upper bound for the fixed point $r^\star$ featured in the aforementioned expression.

Corollary 4.22. *Let $x, s, q, a, b > 0$ such as $s \geq q$ and $x^s \leq ax^q + b$. Then,*

$$x \leq \max\left((2a)^{\frac{1}{s-q}}, (2b)^{\frac{1}{s}}\right).$$

Proof By Lemma 7 in Cucker et al. (2002) the equation $x^s - ax^q + b = 0$ has a unique solution $x^\star$ verifying,

$$x^\star \leq \max\left((2a)^{\frac{1}{s-q}}, (2b)^{\frac{1}{s}}\right).$$

Furthermore , since $s \geq q$ the function $x \mapsto x^s - ax^q + b$ is continuous negative at 0 and positive at $+\infty$. Therefore,

$$x^s \leq ax^q + b \implies x \leq x^\star,$$

and the result follows.

■

To conclude this section, Let's provide a useful lemma that establishes a connection between the VC dimension of the class $\widetilde{\mathcal{F}} = \left\{\frac{1}{2}fI_1 + \frac{1}{2}\frac{p}{1-p}fI_{-1} \mid f \in \mathcal{F}\right\}$ and the VC dimension of $\mathcal{F}$.

Lemma 4.23. *If a family $\mathcal{F}$ of functions is of* VC*-type with envelop U and parameter (v, A), then if $p \leq \frac{1}{2}$, the family $\widetilde{\mathcal{F}} = \left\{\frac{1}{2}fI_1 + \frac{1}{2}\frac{p}{1-p}fI_{-1} \mid f \in \mathcal{F}\right\}$ is also* VC*-type with envelop U and parameter (v, A).*

Proof Let $0 < \epsilon \leq 1$, $p = P(C)$ and $f_k, k = 1, \ldots, N_{\mathcal{F}}$ be an ϵ covering of $\mathcal{F}$ for a norm $||$. Take an element g of $\widetilde{\mathcal{F}}$ which writes $g = \frac{1}{2} f I_1 + \frac{1}{2}\frac{p}{1-p} f I_{-1}$ for some $f \in \mathcal{F}$ and notice that there exists some k such as $\left\|f - f_k\right\| \leq \epsilon$. Thus, by setting $g_k = \frac{1}{2} f_k I_1 + \frac{1}{2}\frac{p}{1-p} f_k I_{-1} \in \widetilde{\mathcal{F}}$, one has

$$\left\|g - g_k\right\| \leq \frac{1}{2}\left\|f_k - f\right\| + \frac{1}{2}\frac{p}{1-p}\left\|(f_k - f)\right\| \leq \epsilon\left(\frac{1}{2} + \frac{1}{2}\frac{p}{1-p}\right) \leq \epsilon.$$

Indeed $p \leq \frac{1}{2}$ implies $\frac{p}{1-p} \leq 1$. The latter fact implies that the covering number of $\mathcal{F}$ and $\widetilde{\mathcal{F}}$ is the same and the proof is complete. ■

4.C.1 Proof of Theorem 4.7

This proof follows a line of reasoning similar to the proofs of Corollary 3.7 in Bartlett et al. (2005) and Theorem 2.12 in Bartlett and Mendelson (2006) which holds only for the class of binary functions. The proof consists on using Theorem 4.21 and upper bounding the term r^* appearing in the display of the latter theorem. To do so set

$$\widetilde{\mathcal{F}} = \left\{\frac{1}{2} f I_1 + \frac{1}{2}\frac{p}{1-p} f I_{-1} \mid f \in \mathcal{F}\right\},$$

and let $\mathcal{G} = star\left(\widetilde{\mathcal{F}}, 0\right)$. Now, consider the following sub-root function

$$\psi_{\mathcal{G}}(r) = 10 B_U P\left(R_n\left\{g \in \mathcal{G} \mid P_n g^2 \leq 2r\right\}\right) + \frac{11 B_U^2 \log(n) + B_U^2}{n}, \tag{4.10}$$

with $B_U = \max(B, U)$. Since $\mathcal{F}$ is VC, Lemma 4.23 allows to use Lemma 4.22 on $\mathcal{G}$ to obtain

$$\psi_{\mathcal{G}}(r) \leq 10 B_U D_1 \sqrt{\frac{v+1}{n}}\sqrt{r}\log\left(\frac{8AU}{r}\right) + \frac{11 B_U^2 \log(n) + B_U^2}{n}.$$

Now, remind that by definition $\psi_{\mathcal{G}}(r) \geq \frac{B_U^2}{n}$ thus $r^\star = \psi_{\mathcal{G}}(r^\star) \geq \frac{B_U^2}{n} \geq \frac{U^2}{n}$ and the latter inequality becomes,

$$r^\star = \psi_{\mathcal{G}}(r^\star) \leq 10 B_U D_1 \sqrt{\frac{v+1}{n}}\sqrt{r^\star}\log\left(\frac{8An}{U}\right) + \frac{11 U^2 \log(n) + U^2}{n}.$$

So that by Corollary 4.22

$$r^\star \leq 20 B_U D_1 \frac{v+1}{n}\log\left(\frac{8An}{U}\right)^2. \tag{4.11}$$

But, Corollary 2.2 in Bartlett et al. (2005) states that, for any r such as

$$r \geq u_n := 10 U P\left(R_n\left\{g \in \mathcal{G} \mid P_n g^2 \leq 2r\right\}\right) + \frac{11 U^2 \log(n)}{n},$$

one has with probability $1 - \frac{1}{n}$,

$$\left\{ g \in \mathcal{G} \mid Pg^2 \leq r \right\} \subset \left\{ g \in \mathcal{G} \mid P_n g^2 \leq 2r \right\}.$$

On the other hand, by Assumption 4.8 the family $\mathcal{G}$ is U bounded. Therefore, one has $R_n \left\{ g \in \mathcal{G} \mid Pg^2 \leq r \right\} \leq U$ and for any $r \geq u_n$,

$$P\left(R_n \left\{ g \in \mathcal{G} \mid Pg^2 \leq r \right\} \right) \leq \left(1 - \frac{1}{n} \right) P\left(R_n \left\{ g \in \mathcal{G} \mid P_n g^2 \leq 2r \right\} \right) + \frac{U}{n}.$$

Remind that, by definition of $\psi_{\mathcal{G}}$ (cf. Equation (4.10)) one has $\psi_{\mathcal{G}}(r) \geq u_n$, so that for any $r \geq \psi_{\mathcal{G}}(r) \geq u_n$ it holds

$$\begin{aligned} BP\left(R_n \left\{ g \in \mathcal{G} \mid Pg^2 r \right\} \right) &\leq BP\left(R_n \left\{ g \in \mathcal{G} \mid P_n g^2 \leq 2r \right\} \right) + \frac{BU}{n} \\ \left(B = \max(B, U) \right) &\leq B_U P\left(R_n \left\{ g \in \mathcal{G} \mid P_n g^2 \leq 2r \right\} \right) + \frac{B_U^2}{n} \\ &\leq \psi_{\mathcal{G}}(r). \end{aligned}$$

Since $\psi_{\mathcal{G}}$ is subroot (as discussed in the introduction of the present section) Lemma 3.2 in the latter reference implies that :

$$r \geq r^\star \iff r \geq \psi_{\mathcal{G}}(r).$$

Thus one has,

$$\forall r \geq r^\star \,,\, \psi_{\mathcal{G}}(r) \geq BP\left(R_n \left\{ g \in \mathcal{G} \mid Pg^2 \leq r \right\} \right) \tag{4.12}$$

$$\left(\widetilde{\mathcal{F}} \subset \mathcal{G} \right) \geq BP\left(R_n \left\{ f \in \widetilde{\mathcal{F}} \mid Pg^2 \leq r \right\} \right). \tag{4.13}$$

In addition, it holds that,

$$\forall \tilde{f} \in \widetilde{\mathcal{F}} \,,\, var\left[\tilde{f} \right] \leq P\left(\tilde{f}^2 \right) \leq BP\left(\tilde{f} \right).$$

It remains to use Theorem 4.21 combined with Inequality (4.11) to obtain, with $c_1 = 6$, $c_2 = 5$, with probability $1 - \delta$

$$\begin{aligned} \forall f \in \mathcal{F} \quad P\left(\frac{1}{2} f I_1 + \frac{1}{2} \frac{p}{1-p} f I_{-1} \right) &\leq \frac{K}{K-1} P_n \left(\frac{1}{2} f I_1 + \frac{1}{2} \frac{p}{1-p} f I_{-1} \right) \\ &+ \frac{c_1 K}{B} \frac{D' \log(An)^2}{n} + \frac{\log(1/\delta) \left(22U + c_2 BK \right)}{n}. \end{aligned}$$

With $D' = 20B_U D_1(v+1)$. On the other hand, notice that for any $\tilde{f} = \frac{1}{2}fI_1 + \frac{1}{2}\frac{p}{1-p}fI_{-1}$ one has $\frac{1}{p}P\left(\tilde{f}\right) = P_{\text{bal}}(f)$ thus dividing by p the latter inequality yields,

$$\begin{aligned}\forall f \in \mathcal{F} \quad P_{\text{bal}}(f) &\leq \frac{K}{K-1} \times \frac{1}{2}\left(\frac{1}{p}P_n(fI_1) + \frac{1}{1-p}P_n(fI_{-1})\right) \\ &\qquad + \frac{c_1 K}{B}\frac{D'\log(An)^2}{np} + \frac{\log(1/\delta)\left(22U + c_2 BK\right)}{np}. \\ &= \frac{K}{K-1} \times \frac{1}{2}\left(P_n(f \mid Y=1)\frac{p_n}{p} + P_n(f \mid Y=-1)\frac{1-p_n}{1-p}\right) \\ &\qquad + \frac{c_1 K}{B}\frac{D'\log(An)^2}{np} + \frac{\log(1/\delta)\left(22U + c_2 BK\right)}{np}. \qquad (4.14)\end{aligned}$$

Now notice that one has by Theorem 4.12, with probability $1-2\delta$,

$$\begin{cases} \frac{p_n}{p} \leq 1 + \sqrt{\frac{3\log(1/\delta)}{np}} \\ \frac{1-p_n}{1-p} \leq 1 + \sqrt{\frac{3\log(1/\delta)}{n(1-p)}} \leq 1 + \sqrt{\frac{3\log(1\delta)}{np}} \end{cases}.$$

The last inequaliy follow since $p \leq \frac{1}{2}$. Finally, by union bound and Inequality (4.14) one has, with probability $1-3\delta$,

$$\begin{aligned}\forall f \in \mathcal{F} \quad P_{\text{bal}}(f) &\leq \frac{K}{K-1}P_{n,\text{bal}}(f)\left(1 + \sqrt{\frac{3\log(1/\delta)}{np}}\right) \\ &\qquad + \frac{DK}{B}\frac{\log(An)^2}{n} + \frac{\log(1/\delta)\left(22U + c_2 BK\right)}{n}.\end{aligned}$$

with $D = c_1(v+1)D' = C'8^{\frac{1}{v}}(v+1)AUC_1C_2$, $C' = 20c_1 = 120$. To show the second part and to conclude the proof follow the same reasoning as before and use instead the second statement of Theorem 4.21.

4.C.2 Proof of Corollary 4.9

The proof follows in a straightforward way from Theorem 4.7. Since Assumption 4.8 holds, we can apply Theorem 4.7 to the class of functions $\mathcal{F}_1 = \{\ell_g - \ell_{g^\star_{\text{bal}}} : g \in \mathcal{G}\}$ and obtain, with probability $1-3\delta$,

$$\begin{aligned}\forall g \in \mathcal{G} \quad P_{\text{bal}}(\ell_g - \ell_{g^\star_{\text{bal}}}) &\leq \frac{K}{K-1}P_{n,\text{bal}}(\ell_g - \ell_{g^\star_{\text{bal}}})\left(1 + \sqrt{\frac{3\log(1/\delta)}{np}}\right) \\ &\qquad + \frac{DK}{B}\frac{\log(An)^2}{np} + \frac{\log(1/\delta)\left(c_2 + c_1 BK\right)}{np}.\end{aligned}$$

In particular for $\hat{g}_{\text{bal}} = \arg\min_{g\in\mathcal{G}} \mathcal{R}_{n,\text{bal}}(g) = \arg\min_{g\in\mathcal{G}} P_{n,\text{bal}}(g)$, we have

$$P_{n,\text{bal}}\left(\hat{g}_{\text{bal}} - g^\star_{\text{bal}}\right) = P_{n,\text{bal}}(\hat{g}_{\text{bal}}) - P_{n,\text{bal}}(g^\star_{\text{bal}}) \leq 0$$

and the result follows by reminding that $P_{\text{bal}}(\ell_{\hat{g}_{\text{bal}}} - \ell_{g^\star_{\text{bal}}}) = \mathcal{R}_{\text{bal}}(\hat{g}_{\text{bal}}) - \mathcal{R}_{\text{bal}}(g^\star_{\text{bal}})$ and by taking $K = 1$.

4.C.3 Proof of Lemma 4.10

By definition one has,

$$\tilde{\ell}_g(X,Y) = \ell_g(X,Y)I_A + \frac{p}{1-p}\ell_g(X,Y)I_B,$$

confirming that, if $g \mapsto \mathbb{E}\left[\ell_g(X,Y)\right]$ is λ-strongly convex then $g \mapsto \mathbb{E}\left[\tilde{\ell}_g(X,Y)\right]$ is $p\lambda$-strongly convex.

Note that $g^\star_{\text{bal}}$ minimizes $\mathbb{E}\left[\tilde{\ell}_g(X,Y)\right]$, which implies:

$$\mathbb{E}\left[\tilde{\ell}_g(X,Y)\right] - \mathbb{E}\left[\tilde{\ell}_{g^\star_{\text{bal}}}(X,Y)\right] \geq p\lambda \left\|g - g^\star_{\text{bal}}\right\|^2.$$

Using the Lipschitz assumption, we obtain:

$$\mathbb{E}\left[\left(\tilde{\ell}_g(X,Y) - \tilde{\ell}_{g^\star_{\text{bal}}}(X,Y)\right)^2\right] \leq L^2\left(P(I_A) + \frac{p^2}{(1-p)^2}P(I_B)\right)\left\|g - g^\star_{\text{bal}}\right\|^2,$$
$$(p \leq 1/2) \leq 2pL^2\left\|g - g^\star_{\text{bal}}\right\|^2.$$

The proof is concluded by combining the above inequalities.

4.D Numerical experiments: Real world dataset

Our aim, just as in the main chapter, is to illustrate the decision boundary of the k-nn classifiers on real-world datasets. To do so, we follow the same procedure as Section 4.5, but instead of using synthetic data, we employ six real-world datasets (Pima, Breast, Cardio, Sattelite, Annthyroid, Ionosphere) from the ODDS repository[2]. Figures 4.3 to 4.8 display the balanced accuracy ($1 - \mathcal{R}^{0-1}_{\text{bal}}$) of the balanced k-nn as function of (k,p), we make the proportion of positive class p vary by randomly removing positive examples. Similar to the findings on synthetic data, these experiments suggest that a large number of neighbors k should be chosen relative to p_n to ensure the consistency of the nearest neighbors method. It's important to note, however, that the learning boundary appears somewhat more noisy than in the synthetic data case. This is indeed not surprising since the number of examples available is significantly smaller in comparison to the previous simulation.

[2]http://odds.cs.stonybrook.edu

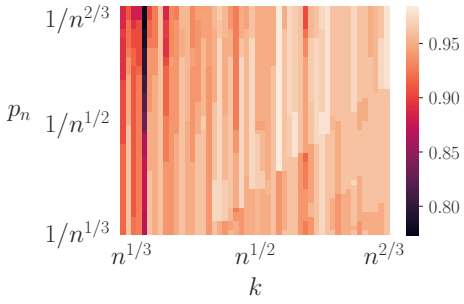


Figure 4.3 – Balanced accuracy heat map for the Breast dataset.

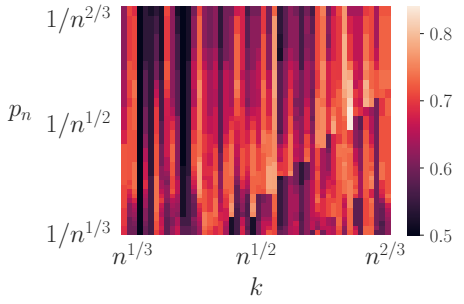


Figure 4.4 – Balanced accuracy heat map for the Ionosphere dataset.

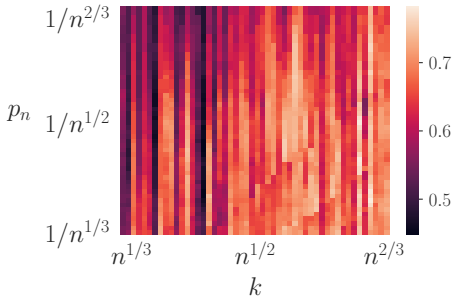


Figure 4.5 – Balanced accuracy heat map for the Pima dataset.

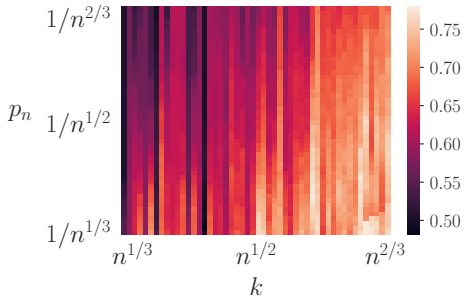


Figure 4.6 – Balanced accuracy heat map for the Annthyroid dataset.

Chapter 5

Hypothesis transfer learning with surrogate classification losses

Contents

5.1 Introduction

Traditional supervised machine learning methods share the common assumption that training data and test data are drawn from the same underlying distribution. However, this assumption is often too restrictive to hold in practice. In many real-world applications, a hypothesis is learnt and deployed in different environments that exhibit a distributional shift. A more realistic assumption is that the marginal distributions of training (*source*) and testing (*target*) domains are different but related. This is the framework of *domain adaptation* (DA), where the learner is provided little or no labeled data from the target domain but a large amount of data from the source domain. This problem arises in various real-world applications like natural language processing Dredze et al. (2007); Ruder et al. (2019), sentiment analysis Blitzer et al. (2007b); Liu et al. (2019), robotics Zhang et al. (2012); Bousmalis et al. (2018) and many other areas.

Several works shed light on the theory of DA Blitzer et al. (2007a); Mansour et al. (2009); Ben-David et al. (2010); Zhang et al. (2012); Cortes et al. (2015); Zhang et al. (2019) and suggest schemes that generally rely on minimizing some similarity distances between the source and the target domains. However, the theoretical analysis shows that a DA procedure needs many unlabeled data from both domains to be efficient. Besides, even when unlabeled data are abundant, minimizing a similarity distance can be time-consuming in many scenarios.

To tackle this practical limitation, a new framework that relies only on the source hypothesis was introduced, the so-called *hypothesis transfer learning* (HTL) Li and Bilmes (2007); Orabona et al. (2009); Kuzborskij and Orabona (2013); Perrot and Habrard (2015); Kuzborskij and Orabona (2017); Du et al. (2017). HTL is tailored to the scenarios where the user has no direct access to the source domain nor to the relatedness between the source and target environments. As a direct consequence, HTL does not

introduce any assumptions about the similarity between the source and target distributions. It has the advantage of not storing abundant source data in practice.

In this work, we analyze HTL through Regularized Empirical Risk Minimization (RERM) in the binary classification framework. Our working assumptions encompass many widely used *surrogate* losses, such as the exponential loss used by several boosting algorithms like AdaBoost Freund and Schapire (1997b), the logistic loss, the softplus loss, which serves as a smooth approximation of the hinge loss Dugas et al. (2000), the mean squared error (MSE) and the squared hinge that represents the default losses for least squares/modified least squares algorithms Rifkin et al. (2003). The attractive quality of these surrogate losses is that they are *classification calibrated* Zhang (2004a); Bartlett et al. (2006b). In other words, they represent a convex upper bound for the classification error and minimizing the expected risk regarding a surrogate loss yields a predictor with sound accuracy.

This chapter's theoretical analysis uses the notion of *algorithmic stability*. Formally, assuming that one has access to a small labeled set, we derive many complexity-free generalisation bounds that depend only on the source hypothesis's quality. In particular, such an analysis allows us to compare the behavior of different losses in different scenarios and to answer some practical questions such as: *which surrogate loss is recommended when the source and target domains are related? Which surrogate loss is robust to heavy distribution shift?*

The notion of algorithmic stability and its consequences in learning theory has received much attention since its introduction in Devroye and Wagner (1979). It allows obtaining complexity-free generalization bounds for a large class of learning algorithms such as k-nearest-neighbours Devroye and Wagner (1979), empirical risk minimizers Kearns and Ron (1999), Support Vector Machine Bousquet and Elisseeff (2002), Bagging Elisseeff et al. (2005), RERM Zhang (2004a); Wibisono et al. (2009b), stochastic gradient descent Hardt et al. (2016b), neural networks with a simple architecture Charles and Papailiopoulos (2018), to name but a few. For an exhaustive review of the different notions of *stability* and their consequences on the generalization risk of a learning algorithm, the reader is referred to Kutin and Niyogi (2002).

Only a few works derive theoretical guarantees for RERM in the HTL framework and are all formalized in a regression setting. A stability analysis has been provided for the HTL algorithm in the case of RLS for regression in Kuzborskij and Orabona (2013) limited to the least-squares loss. Later, Kuzborskij and Orabona (2017) considered the class of smooth losses and obtained statistical rates on the empirical risk, being a particular case of the stability guarantees. However, this smoothness assumption may be considered strong since it is not satisfied for hypotheses learnt from the exponential loss or vacuously satisfied for hypotheses learnt from the softplus loss. Besides, Du et al. (2017) proposed a novel algorithm to adapt the source hypothesis to the target domain. Nonetheless, the theoretical guarantees they derived are obtained with several strong assumptions, unverifiable in practice. The obtained bounds depend on many unknown parameters (for further details, see Section 5.3, where all these assumptions are explicitly listed and discussed). Other theoretical results studying HTL outside the framework of RERM can be found Li and Bilmes (2007); Morvant et al. (2012); Perrot and Habrard (2015); Dhouib and Redko (2018). However, most of these theoretical results depend on a complexity/distance measure or/and are valid on a different framework than classification. For example, Perrot and Habrard (2015) explores the notion of algorithmic stability in *metric learning* with Lipschitz loss functions to study the

excess risk of some algorithms. The obtained bounds are not intuitive as they depend on the Lipschitz constant and cannot be easily extended to many usual classification losses. Furthermore, the proof techniques in the latter work are far from ours.

On the other hand, when the source is known, many theoretical guarantees can be found in the domain adaptation literature, see e.g. Mansour et al. (2009); Ben-David et al. (2010); Zhang et al. (2012); Cortes et al. (2015) and Zhang et al. (2019), among others. Their rates involve the complexity of the hypothesis class and the distance between the source and the target distribution that may be unknown in practice and drastically deteriorate the rates.

Another related subject is *meta learning*, broadly described as leveraging data from pre-existing tasks to derive algorithms or representations that yield superior results on unencountered tasks. Many theoretical works such as Khodak et al. (2019); Balcan et al. (2019); Denevi et al. (2019) or Denevi et al. (2020) have studied this problem. Yet, the obtained theoretical guarantees in the latter works depend on the smoothness parameters of the loss function and the regularizers. The proof techniques from the present chapter can be incorporated into the proof of the latter references to obtain more sharp and intuitive learning bounds, that is, bounds exclusively depending on the quality of the source hypothesis.

Contributions In this chapter, we investigate the statistical risk of the hypothesis transfer learning procedure dedicated to the binary classification task. To that end, we adopt the angle of algorithmic stability that offers an appealing theoretical framework to analyze such a method. This is the first work exploring algorithmic stability for HTL with the usual classification loss functions. In this chapter, we provide a (pointwise) hypothesis stability analysis of the HTL in the classification framework for any losses satisfying mild conditions. Furthermore, we show that our main assumptions are valid for the most popular classification losses and derive their associated constants. Based on these stability results, we investigate the statistical behavior of the generalization gap and the excess risk of the HTL procedure. We provide an intuitive finite-sample analysis of these quantities and highlight the statistical behavior of common losses.

5.2 Background and Preliminaries

In this section, we start by recalling the framework of Hypothesis transfer learning and describe the concept of stability.

5.2.1 Hypothesis Transfer Learning

Considering the source and target domains, hypothesis transfer learning leverages the learnt hypothesis with the source dataset, without having access to the raw source data or any information between source and target domains, to solve a machine learning task on the target domain. Formally, we denote by $\mathcal{Z}_S$ and $\mathcal{Z}_T$ the source and target domains and assume that we have access to $n \in \mathbb{N}, n \geq 1$ i.i.d. observations $\mathcal{D}_T = Z_1, \ldots, Z_n \in \mathcal{Z}_T$ with a distribution P_T lying in the target domain and a source hypothesis h_S learnt from $m \in \mathbb{N}, m \geq 1$ i.i.d. observations $\mathcal{D}_S = Z_1^S, \ldots, Z_m^S \in \mathcal{Z}_S$ drawn from the source distribution P_S. In the HTL framework, we do not have access to the source observations

but only to the resulting source hypothesis h_S. It is worth noting that $n \ll m$ in many practical scenarios. In this chapter, we focus on the binary classification task. Therefore, our domains consist of a Cartesian product of a source/target covariate space $\mathcal{X}_S/\mathcal{X}_T$ and the set $\{-1, 1\}$, i.e. $\mathcal{Z}_S = \mathcal{X}_S \times \{-1, 1\}$ and $\mathcal{Z}_T = \mathcal{X}_T \times \{-1, 1\}$. In addition, we assume that $\mathcal{X}_T \subset \mathcal{X}_S \subset \mathbb{R}^d$. Consider two classes of hypotheses $\mathcal{H}_S$ and $\mathcal{H}_T$, an HTL algorithm aims to use a source hypothesis $h_S \in \mathcal{H}_S$ learnt on $\mathcal{D}_S$ to improve the performance of a classification algorithm over $\mathcal{D}_T$. Precisely, it is defined as a map

$$\begin{aligned} \mathcal{A} : \left(\mathcal{Z}_T\right)^n \times \mathcal{H}_S &\to \mathcal{H}_T \\ \left(\mathcal{D}_T, h_S\right) &\mapsto h_T. \end{aligned}$$

Throughout the chapter, we assume that h_S is given and fixed, and we use the shorthand notation $\mathcal{A}(\mathcal{D}_T)$ instead of $\mathcal{A}(\mathcal{D}_T, h_S)$ for the sake of clarity.

Let $\ell : \mathcal{H}_T \times \mathcal{Z}_T \mapsto \mathbb{R}_+$ denote a loss function so that $\ell(h_T, Z)$ is the error of $h_T \in \mathcal{H}_T$ on the observation $Z = (X, Y) \in \mathcal{Z}_T$. In this work, we assume that $\ell(h_T, Z) = \phi\left(h_T(X)Y\right)$ for some non negative convex function ϕ. The generalization risk of the predictor $\mathcal{A}(D_T)$ is denoted by

$$\begin{aligned} \mathcal{R}\Big[\mathcal{A}\left(\mathcal{D}_T\right)\Big] &= \mathbb{E}_{Z\sim P_T}\left[\ell\left(\mathcal{A}\left(\mathcal{D}_T\right), Z\right)\right] \\ &= \mathbb{E}\left[\ell\left(\mathcal{A}\left(\mathcal{D}_T\right), Z\right) \mid \mathcal{D}_T\right]. \end{aligned}$$

Notice that the randomness in the latter expectation stems from the novel observation Z only while the trained algorithm $\mathcal{A}(\mathcal{D}_T)$ is fixed. Its empirical counterpart, the *training* error of $\mathcal{A}\left(\mathcal{D}_T\right)$ writes as

$$\widehat{\mathcal{R}}\Big[\mathcal{A}(\mathcal{D}_T)\Big] = \frac{1}{n}\sum_{i=1}^{n} \ell(\mathcal{A}(\mathcal{D}_T), Z_i).$$

The latter estimate is known to be optimistic since most learning algorithms are conceived to minimize the training loss. Thus, a more reliable estimate would be the *deleted* estimate or the so-called leave-one-out (*l.o.o.*) estimate:

$$\widehat{\mathcal{R}}_{\text{loo}}\Big[\mathcal{A}(\mathcal{D}_T)\Big] = \frac{1}{n}\sum_{i=1}^{n} \ell\left(\mathcal{A}(\mathcal{D}_T^{\setminus i}), Z_i\right), \tag{5.1}$$

where $\mathcal{D}_T^{\setminus i} = \mathcal{D}_T \setminus \{Z_i\}$ denotes the dataset $\mathcal{D}_T$ with the i'th element removed.

Remark 5.1 (ACCELERATED *l.o.o.*)**.** *At first sight, one can notice that computing the* l.o.o. *risk measure is a heavy task in practice since one needs to train the algorithm n times. However, in our case, one can use the closed form formula of the* l.o.o. *estimate for* RERM *algorithms derived in Wang et al. (2018).*

5.2.2 Algorithmic Stability

In this part, we briefly recall important notions of stability that will be used in the chapter. The notion of *stability* was first introduced in Devroye and Wagner (1979) to derive non-asymptotic guarantees for the leave-one-out estimate. Let denote by $[n]$ the

set of indices $\{1, \ldots, n\}$. The algorithm $\mathcal{A}$ is called stable if removing a training point Z_i, $i \in [n]$, from the $\mathcal{D}_T$ or replacing Z_i with an independent observation Z' drawn from the same distribution does not alter the risk of the output. Later, Bousquet and Elisseeff (2002) introduced the strongest notion of stability, namely *uniform stability*, an assumption used to derive probability upper bounds for the training error and the *l.o.o.* estimate Bousquet and Elisseeff (2002); Elisseeff et al. (2005); Hardt et al. (2016b); Bousquet et al. (2020); Klochkov and Zhivotovskiy (2021b). Equipped with the above notations, *uniform* stability, also called *leave-one-out* stability, can be defined as follows.

Definition 5.2. *The algorithm $\mathcal{A}$ is said to be* $\beta(n)$-uniformly *stable with respect to a loss function ℓ if, for any $i \in [n]$ and $Z \in \mathcal{Z}_T$, it holds:*

$$\left| \ell\left(\mathcal{A}(\mathcal{D}_T), Z\right) - \ell\left(\mathcal{A}(\mathcal{D}_T^{\setminus i}), Z\right) \right| \leq \beta(n).$$

In practice, uniform stability may be too restrictive since the bound above must hold for all Z, irrespective of its marginal distribution. While weaker, the following notion of stability is still enough to control the leave-one-out deviations Devroye and Wagner (1979); Bousquet and Elisseeff (2002); Elisseeff et al. (2005); Kuzborskij and Orabona (2013).

Definition 5.3. *The algorithm $\mathcal{A}$ has a* hypothesis *stability $\beta(n)$ with respect to a loss function ℓ if, for any $i \in [n]$, it holds:*

$$\left\| \ell\left(\mathcal{A}(\mathcal{D}_T), Z\right) - \ell\left(\mathcal{A}(\mathcal{D}_T^{\setminus i}), Z\right) \right\|_1 \leq \beta(n),$$

where $\left\|X\right\|_q = \left(\mathbb{E}\left[\left|X\right|^q\right]\right)^{1/q}$ *is the L_q norm of X.*

We now recall a direct analogue of hypothesis stability: the *pointwise hypothesis stability*. The latter property is used to derive PAC learning bounds for the training error Bousquet and Elisseeff (2002); Elisseeff et al. (2005); Charles and Papailiopoulos (2018).

Definition 5.4. *The algorithm $\mathcal{A}$ has a* pointwise hypothesis *stability $\gamma(n)$ with respect to a loss function ℓ if, for any $i \in [n]$, it holds:*

$$\left\| \ell\left(\mathcal{A}(\mathcal{D}_T), Z_i\right) - \ell\left(\mathcal{A}(\mathcal{D}_T^{\setminus i}), Z_i\right) \right\|_1 \leq \gamma(n).$$

Note that the approach based on stability does not refer to a complexity measure like the VC dimension or the Rademacher complexity. There is no need to prove uniform convergence, and the generalization error (cf. Equation 5.4.1 below) depends directly on the stability parameter. Our work aims to use the notion of algorithmic stability to derive sharper bounds for the HTL problem. More precisely, the magnitude of the obtained bounds is directly related to the quality of h_S on the target domain $\Big($represented by $\mathcal{R}[h_S]\Big)$ instead of the complexity of the hypothesis class Ben-David et al. (2010); Zhang et al. (2012); Cortes et al. (2015); Zhang et al. (2019).

5.2.3 Working Framework

This chapter analyses hypothesis transfer learning through regularised empirical risk minimization (RERM). In particular, it includes the popular Regularized Least Squares (RLS) with biased regularization (Orabona et al., 2009) that has been analyzed in Kuzborskij and Orabona (2013) and Kuzborskij and Orabona (2017). Formally, we consider the following algorithm $\mathcal{A}$ such that:

$$\mathcal{A}(\mathcal{D}_T, h_S) = \hat{h}(\cdot\,; \mathcal{D}_T) + h_S(\cdot), \tag{5.2}$$

where the function $\hat{h} : \mathbb{R}^d \to \mathbb{R}$ is obtained from the target set of data via the minimization problem:

$$\begin{aligned}\hat{h} &= \arg\min_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} \phi\left(\left(h\left(X_i\right) + h_S\left(X_i\right)\right) Y_i\right) + \lambda\|h\|_k^2 \\ &= \arg\min_{h\in\mathcal{H}} \widehat{\mathcal{R}}(h + h_S) + \lambda\|h\|_k^2,\end{aligned} \tag{5.3}$$

with the family of hypotheses $\mathcal{H}$ being a reproducing kernel Hilbert space (RKHS) endowed with a kernel k, an inner product $\langle\cdot,\cdot\rangle$ and a norm $\|\cdot\|_k$. The resulting map arising from the HTL is the sum of the source hypothesis h_S and the target hypothesis $\hat{h}$ where $\hat{h}$ is learnt involving the source map.

It is worth noting that our analysis encompasses the least square with biased regularization (Schölkopf et al., 2001; Orabona et al., 2009) commonly studied in transfer learning (Kuzborskij and Orabona, 2013, 2017), briefly recalled below.

Remark 5.5 (LINK WITH RLS)**.** *The* RLS *with biased regularization is a particular case of the proposed algorithm 5.2. Indeed, by choosing k as the linear kernel $k(x_1, x_2) = x_1^\top x_2$ and the loss $\phi(x) = \left(1 - x\right)^2$, it is equivalent to*

$$\mathcal{A} = \hat{h} + h_S,$$

with $\hat{h}(x) = \hat{u}^\top x$ and

$$\hat{u} = \arg\min_{u\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^{n} \left(u^\top X_i + h_S(X_i) - Y_i\right)^2 + \lambda\|u\|_2^2. \tag{5.4}$$

Furthermore, if $h_S(x) = v^\top x$ is a linear classifier with $v \in \mathbb{R}^d$, then

$$\hat{u} = \arg\min_{u\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^{n} \left(u^\top X_i - Y_i\right)^2 + \lambda\|u - v\|_2^2,$$

which is the original form of biased regularisation algorithms Schölkopf et al. (2001); Orabona et al. (2009). See Appendix 5.A.1 for technical details.

5.3 Stability Analysis

The subsequent analysis requires technical assumptions, listed below. We assume that the source hypothesis and the kernel k are bounded, as stated in the following assumptions.

Assumption 7. *The source hypothesis is bounded on the target space:*

$$\|h_S\|_\infty = \sup_{x \in \mathcal{X}_T} |h_S(x)| < \infty.$$

Assumption 8. *The kernel k is bounded:*

$$\sup_{x_1, x_2 \in \mathcal{X}_T} k(x_1, x_2) \leq \kappa.$$

The boundness of the kernel is a common and mild assumption (see e.g. Bousquet and Elisseeff, 2002; Zhang, 2004a; Wibisono et al., 2009b). It is satisfied by many usual kernels like the Gaussian kernel and the sigmoid kernel. Furthermore, when $\mathcal{X}_T$ is bounded, then polynomial kernels are also bounded.

We now investigate the accuracy of the HTL proposed framework and provide general stability results under slight assumptions. Furthermore, we show that these assumptions are satisfied by most of the popular ML surrogate losses used in practice and derive precisely the associated constants involved in our theoretical results.

5.3.1 Hypothesis Stability

This section analyzes the hypothesis stability of general surrogate ML losses for the proposed HTL framework. To study the stability of Algorithm 5.2, we start by showing that the solution of the optimization problem 5.3 lies in the sphere with a data-driven radius, as stated in the following lemma.

Lemma 5.6. *Suppose that Assumptions 7 and 8 are satisfied. Then the solution of Equation (5.3) lies in the set* $\left\{h \in \mathcal{H},\ \|h\|_\infty \leq \hat{r}_\lambda\right\}$ *with*

$$\hat{r}_\lambda = \kappa \sqrt{\alpha \widehat{\mathcal{R}}\left[h_S\right]},$$

where $\alpha = \kappa/\lambda$.

Proof The proof is postponed in the Appendix 5.B.1. ■

This lemma ensures that the norm of the solution of the optimisation problem 5.3 decreases when the quality of h_S increases. In the rest of the chapter, for a given index $i \in [n]$, we denote by $\hat{r}_\lambda^i = \kappa \sqrt{\alpha \widehat{\mathcal{R}}^{\setminus i}\left[h_S\right]}$, $\widehat{\mathcal{R}}^{\setminus i}$ the training error with the i'th sample removed and $\hat{\rho}_\lambda^i = \max\left(\hat{r}_\lambda, \hat{r}_\lambda^i\right)$.

Before stating our main theorem, we first require an additional assumption involving the empirical radius obtained in Lemma 5.6.

Assumption 9. *The function ϕ is differentiable and convex. Furthermore, $\forall i \in [n]$, it holds:*

$$\mathbb{E}\left[\sup_{|y'|,|y| \leq \hat{\rho}_\lambda^i} \left|\phi'(h_S(X')Y' + y')\phi'(h_S(X)Y + y)\right|\right] \leq \Psi_1\left(\mathcal{R}\left[h_S\right]\right),$$

where $Z = (X, Y)$, $Z' = (X', Y')$ *are two samples drawn from* P_T *independent of* $\mathcal{D}^{\setminus i}$ *and* Ψ_1 *is a decreasing function verifying* $\Psi_1(0) = 0$.

The bound stated in the theorem below reveals the generalisation properties of the presented HTL procedure through the stability framework.

Proposition 5.7. *Suppose that Assumptions 7, 8 and 9 are satisfied. Then the algorithm* $\mathcal{A}$ *(cf. Equation (5.2)) is hypothesis stable with parameter*

$$\beta(n) = \frac{\alpha\left(\Psi_1\left(\mathcal{R}\left[h_S\right]\right) \wedge \|\phi'\|_\infty^2\right)}{n}.$$

Proof The proof is postponed to the Appendix 5.B.2. ■

We obtain a stability rate of order $\mathcal{O}\left(\frac{\Psi_1\left(\mathcal{R}\left[h_S\right]\right)\alpha}{n}\right)$ for any losses satisfying Assumption 9. It naturally depends on the risk of the source classifier, where the expectation is taken on the target data distribution. Therefore, the source task directly influences the rate of the HTL classifier. The standard stability rate of RERM without transfer learning (without source) is of order $\mathcal{O}(\alpha/n)$, see Theorem 4.3 in Zhang (2004a) or Theorem 3.5 in Wibisono et al. (2009b). A relevant source hypothesis allows us to obtain faster rates than in standard RERM. Thus, one can directly notice the benefits of using a *good* source hypothesis on the stability of RERM. The negative transfer, i.e. the source hypothesis has a negative effect and deteriorates the target learner, is analyzed and discussed in Section 5.4.1.

Remark 5.8 (Related Work)**.** *The only existing result studying hypothesis stability in HTL is in Kuzborskij and Orabona (2013). However, the analysis is only in a regression framework with the mean squared error loss. The proof techniques in Kuzborskij and Orabona (2013) rely heavily on the closed-form formulas of the ordinary least square estimate, which does not hold in a general setting like ours. Furthermore, we obtain equivalent (up to constants) stability rates as in Kuzborskij and Orabona (2013). More details are given in Section 5.3.3 where we explicit constants* Ψ_1 *for most of popular losses.*

Existing assumptions in DA and HTL literature Statistical guarantees obtained in these fields generally assume that the loss function verifies a smoothness condition. For example, in Mansour et al. (2009) and Cortes et al. (2015), their analysis supposes that ℓ verifies the triangle inequality, which holds only for the MSE and squared hinge. Moreover, the obtained upper bounds in these works depend on the complexity of $\mathcal{H}$ and some discrepancy distances between the source and target distributions P_S and P_T, which deteriorates the statistical rates. In Kuzborskij and Orabona (2017), they suppose that the derivative of the loss is Lipschitz which is not the case for the exponential. Furthermore, even if the loss satisfies this smoothness assumption, their constants depend heavily on the smoothness parameter, and it would yield vacuous bounds in many practical situations. For example, the softplus function $\psi_s(x) = s\log(1 + e^{\frac{1-x}{s}})$ with small values of s serves as an approximation of the hinge loss $\max(0, 1-x)$ and is

$1/s$ Lipschitz. This function converges to the Hinge loss when $s \to 0$ and usual choices of s are usually close to 0. Therefore, the Lipschitz constant of the derivative $1/s$ verifies $1/s \gg 1$, and the bounds from Kuzborskij and Orabona (2017) become vacuous. Besides, Du et al. (2017) made several assumptions about the true regression function of both the source and target domains. To clarify, by the true regression function, f, we refer to the actual model denoted by $Y = f(X)$. However, these assumptions are challenging to empirically confirm due to their reliance on the real source and target distributions, which generally remain unknown. Moreover, the theoretical guarantees achieved depend on several constants, also derived from the true distribution, that makes quantifying the bounds magnitude a complex task.

To our best knowledge, the vast majority of existing theoretical results from the HTL literature have similar assumptions to those discussed above. However, in this work, our assumptions are flexible: we only require the differentiability of the loss and a *local* majorant of the derivative, which will make the analysis more flexible and more suited for the usual classification losses.

To understand the intuition behind Assumption 9 notice that, when $\mathcal{R}[h_S] \to 0$, $\phi(h_S(X)Y)$ approaches the minimum then $\phi'(h_S(X)Y)$ approaches 0 (in expectation). Thus, the function Ψ_1 can be seen as a function that dictates the rate of convergence of the derivative to 0 as h_S approaches the optimal hypothesis. One must note that the latter assumption is verified for many loss functions, namely any loss satisfying the following inequality $|\phi'(x)| \leq \Psi(\phi(x))$ for some concave loss function Ψ. The function Ψ effectively mediates between ϕ and ϕ'. As an example, in the context of Mean Squared Error (MSE) loss, it is straightforwardly observable that $|\phi'(x)| \leq \sqrt{\phi(x)}$. Thus ϕ' is directly linked to $\phi(x)$ via the square root function.

Remark 5.9 (SCORE SCALING)**.** RERM *for regression (cf. Equation 5.4) is equivalent to fitting a predictor on the residuals* $Y_i - h_S(X_i)$*. However, in the classification case, if we follow the standard approach that* $h_S : \mathcal{X} \mapsto \mathcal{Y} = \{-1, 1\}$ *is a binary classifier Mansour et al. (2009); Cortes et al. (2015), then latter residuals are either* 1 *or* 0*. Thus, this won't provide enough information for many losses to improve the training. To see this, see the example of the logistic loss and notice that* $\phi(1) = \log(1 + e^{-1})$ *and* $\phi(-1) = \log(1+e^{1})$*. Therefore, in the best case scenario,* $\mathcal{R}[h_S] = \log(1+e^{-1})$*, which is far from the minimum (that is zero). To tackle this problem, we suggest taking the score learned on the source, which is more informative, especially when the loss function used to train the algorithm on the source has the same minimum as the loss used to train on the target. Note that one can also think of transforming the score, for example, if* ϕ *is the logistic loss* $\phi(x) = \log(1 + e^{-x})$ *and* $h_S \in]-1, 1[$ *we can use an increasing transformation function to an interval* $]-C, C[$ *with* $C >> 1$ *in order to adapt to the target loss which is nearly 0 for large values* x*.*

5.3.2 Pointwise Hypothesis Stability

To go further than the widely used hypothesis stability, we analyze our HTL problem through the angle of pointwise hypothesis stability. Results presented in this part will be the cornerstone of those shown in Section 5.4. To analyze the pointwise hypothesis stability of Algorithm 5.2, we require a direct analogue of Assumption 9, involving the data-driven radius provided in Lemma 5.6.

Assumption 10. *The function ϕ is differentiable and convex. Furthermore, $\forall i \in [n]$, it holds:*

$$\mathbb{E}\left[\sup_{|y'|,|y|\leq\hat{\rho}_\lambda^i}\left|\phi'(h_S(X)Y+y')\phi'(h_S(X)Y+y)\right|\right] \leq \Psi_2\left(\mathcal{R}\left[h_S\right]\right).$$

where $Z=(X,Y)$ is a sample drawn from P_T independent of $\mathcal{D}^{\setminus i}$ and Ψ_2 is a decreasing function verifying $\Psi_2(0)=0$.

Under the latter assumption, the following proposition is obtained in a similar manner to Proposition 5.7.

Proposition 5.10. *Suppose that Assumptions 7, 8 and 10 are satisfied. Then the algorithm $\mathcal{A}$ (cf. Equation (5.2)) is pointwise hypothesis stable with parameter*

$$\gamma(n) = \frac{\alpha\left(\Psi_2\left(\mathcal{R}\left[h_S\right]\right) \wedge \|\phi'\|_\infty^2\right)}{n}.$$

Proof The proof is postponed to the Appendix 5.B.3. ■

Again, this result shows the benefits of using a good hypothesis on the pointwise hypothesis stability of RERM. This stability result, combined with that of Proposition 5.7, can be leveraged to propose new convergence results on the generalisation gap and the excess risk of this HTL problem for a wide class of losses, as shown in Section 5.4. In the sequel, we explicitly compute the functions Ψ_1 and Ψ_2 for many widely used classification losses.

5.3.3 Deriving Constants for Popular Losses

As the results of Propositions 5.7 and 5.10 are general and stated for any losses satisfying Assumptions 9 and 10, it is the purpose of this part to investigate our results with widespread machine learning losses. To that end, we first show that these Assumptions are satisfied for the most popular losses. Second, we derive constants involved in these two statistical rates. In particular, we focus on the five following losses:

- Exponential: $\phi(x) = e^{-x}$.
- Logistic: $\phi(x) = \log\left(1+e^{-x}\right)$.
- Mean Squared Error: $\phi(x) = (1-x)^2$.
- Squared Hinge: $\phi(x) = \max(0,1-x)^2$.
- Softplus: $\phi_s(x) = s\log\left(1+e^{\frac{1-x}{s}}\right)$, for some $s>0$.

In the next proposition, we show that most of classical losses verifies Assumptions 9, 10 and we detail their associated functions Ψ_1 and Ψ_2.

Loss	$\Psi_1(x)$	$\Psi_2(x)$
Sq. hinge	$8x(4\alpha+1)$	$8x(4\alpha+1)$
MSE	$8x(4\alpha+1)$	$8x(4\alpha+1)$
Exponential	$C_S x^2 e^{2\alpha x}$	$M_S C_S x e^{2\alpha x}$
Logistic	$C_S e^{2\alpha x}(e^{\sqrt{x}}-1)^2$	$C_S e^{2\alpha x}(e^{\sqrt{x}}-1)$
Softplus	$C_S e^{2\alpha x}(e^{\sqrt{\frac{x}{s}}}-1)^2$	$C_S e^{2\alpha x}(e^{\sqrt{\frac{x}{s}}}-1)$

Table 5.1 – Examples of losses verifying Assumptions 9, 10 and their corresponding functions. The constants M_S and C_S are given by $M_S = \sup_{z\in\mathcal{Z}_T} \ell(h_S, z)$, $C_S = \exp\left\{2 + \frac{2\alpha M_S}{n} + \frac{4\alpha^2 M_S^2}{n-1}\right\}$.

Proposition 5.11. *The exponential, logistic, squared hinge,* MSE *and softplus losses satisfy Assumptions 9 and 10 with corresponding functions Ψ_1 and Ψ_2 listed in Table 5.1.*

Proof The proof is postponed to the Appendix 5.B.4. ■

This result shows that bounds derived in Propositions 5.7 and 5.10 are therefore valid under mild assumptions. Indeed, our results only require the kernel and the source hypothesis to be bounded, classical in the HTL framework. Thus, we obtain the first stability result in HTL without limiting assumptions, which remains valid in a practical setting.

As shown in Table 5.1, functions Ψ_1 and Ψ_2 are linear for the square hinge and the MSE losses. Besides, for the softplus and logistic losses, we have $\|\phi'\|_\infty = 1$ and their stability parameters capped by α/n. Thus, the impact of an irrelevant source hypothesis h_S with large $\mathcal{R}[h_S]$ remains negligible on the stability of RERM when using these losses. In contrast, for the exponential loss, the functions Ψ_1 and Ψ_2 are roughly exponential, and the corresponding convergence rate deteriorates quickly as $\mathcal{R}[h_S]$ increases. This is indeed not surprising since a prediction in the wrong direction ($\text{sign}(h_S(X)) \neq Y$) would increase the loss $e^{-h_S(X)Y}$ exponentially fast. In the particular case of the MSE, we obtain the same stability rate $\mathcal{O}\left(\frac{\alpha\mathcal{R}[h_S]}{n}\right)$ as in the regression framework Kuzborskij and Orabona (2013). In the next section, we shall discuss the implications of these stability rates on the *generalization gap* Hardt et al. (2016b); Charles and Papailiopoulos (2018), cross-validation schemes and the excess risk of Algorithm 5.2.

5.4 Generalisation guarantees for HTL with surrogate losses

In this part, we leverage the stability results provided in Section 5.3 in several statistical errors commonly used.

5.4.1 Generalization Gap

Here we investigate the accuracy of the algorithm $\mathcal{A}$ through the generalization gap. Precisely, this gap is defined as the expected error between the empirical risk and the theoretical risk of the algorithm $\mathcal{A}$:

$$\mathcal{E}_{\text{gen}} = |\mathbb{E}\left[\widehat{\mathcal{R}}\left[\mathcal{A}(\mathcal{D}_T)\right] - \mathcal{R}\left[\mathcal{A}(\mathcal{D}_T)\right]\right]|.$$

To discuss the impact of h_S on the generalization gap, it suffices to analyse the stability parameters $\beta(n)$ and $\gamma(n)$. Indeed, $\mathcal{E}_{\text{gen}}$ is directly linked to these quantities, as stated in the following theorem.

Theorem 5.12. *Suppose that $\mathcal{A}$ has a hypothesis stability $\beta(n)$ and a pointwise hypothesis stability $\gamma(n)$. Then, it holds:*

$$\mathcal{E}_{gen} \leq \beta(n) + \gamma(n).$$

Furthermore, suppose that Assumptions 7, 8, 9 and 10 are satisfied. Thus, $\beta(n)$ and $\gamma(n)$ are given by Propositions 5.7 and 5.10 and the generalization gap of $\mathcal{A}$ (cf. Equation (5.2)) is upper-bounded as:

$$\mathcal{E}_{gen} \leq \alpha \frac{\left(\Psi_1\left(\mathcal{R}\left[h_S\right]\right) + \Psi_2\left(\mathcal{R}\left[h_S\right]\right)\right) \wedge \left(2\|\phi'\|_\infty^2\right)}{n}.$$

Proof The proof is postponed to the Appendix 5.B.5. ■

When the source hypothesis is relevant, the risk $\mathcal{R}[h_S]$ is close to zero so that $e^{\mathcal{R}[h_S]} - 1 \approx \mathcal{R}[h_S]$ and $e^{\alpha\mathcal{R}[h_S]} \approx 1$. Equipped with Table 5.1, this theorem yields the following upper bounds for $\mathcal{E}_{\text{gen}}$:

- MSE, Sq. hinge: $\mathcal{E}_{\text{gen}} = \mathcal{O}\left(\frac{\alpha\mathcal{R}[h_S]}{n}\right)$.
- Logistic: $\mathcal{E}_{\text{gen}} = \mathcal{O}\left(\alpha\frac{\sqrt{\mathcal{R}[h_S]}\wedge 2}{n}\right)$.
- Softplus: $\mathcal{E}_{\text{gen}} = \mathcal{O}\left(\alpha\frac{\left(\sqrt{\mathcal{R}[h_S]/s}\right)\wedge 2}{n}\right)$.
- Exponential: $\mathcal{E}_{\text{gen}} = \mathcal{O}\left(\frac{\alpha M_S\mathcal{R}[h_S]}{n}\right)$.

Thus, if $\mathcal{R}[h_S]$ is small, the exponential, the squared hinge and the MSE losses have the fastest generalization gap rate. Therefore, our analysis suggests that the user should privilege using the latter losses if one disposes of a good hypothesis h_S.

Negative learning The phenomenon of negative transfer occurs when the hypothesis h_S learned from the source domain has a detrimental effect on the target learner. In such a case, training without using h_S on the target domain would yield a better learner.

We refer the reader to Weiss et al. (2016) and Wang et al. (2019b) for further details about this topic. For the softplus and the logistic losses, the generalization gap remains bounded by $\mathcal{O}(\alpha/n)$ even if $\mathcal{R}[h_S] \to \infty$. As a consequence, Algorithm 5.2 with the sofplus and logistic losses is robust to negative learning since the generalization gap still achieves the same rate of convergence $\mathcal{O}(\alpha/n)$ as a standard RERM algorithm with no source information *i.e.* $h_S = 0$ (see *e.g.* Zhang, 2004a; Wibisono et al., 2009b). Finally, we must highlight that one should avoid using the exponential loss when the source and target domains are unrelated due to the presence of the term $e^{\alpha\mathcal{R}[h_S]}$ in the corresponding upper bound.

Remark 5.13 (CROSS VALIDATION PROCEDURES)**.** *The notion of stability has many attractive qualities. In particular, it yields complexity-free bounds for cross-validation methods. (see* e.g. *Bousquet and Elisseeff, 2002; Kumar et al., 2013; Celisse and Mary-Huard, 2018). For example, one can easily show that*

$$\mathbb{E}\left[\left|\widehat{\mathcal{R}}_{\text{loo}}\left[\mathcal{A}\left(\mathcal{D}_T\right)\right] - \mathcal{R}\left[\mathcal{A}\left(\mathcal{D}_T\right)\right]\right|\right] \leq \beta(n).$$

Proposition 5.7 shows that the quality of risk estimation with l.o.o. *depends directly on the quality of the source predictor h_S. Note that the same conclusion holds for model selection with* l.o.o. *cross-validation: Given a family of source hypotheses, the quality of the model selection procedure depends directly on the quality of the provided learners independently of the complexity of $\mathcal{H}_T$. Besides, using the same proof techniques, we can show that Algorithm 5.2 is L_2 stable with stability parameter depending on $\Psi\left(\mathcal{R}\left[h_S\right]\right)$. L_2 stability is similar to hypothesis stability, where the L_1 moment is replaced by the L_2 moment in Definition 5.3. The latter notion allows obtaining theoretical guarantees regarding the K-fold and the* l.o.o. *schemes. It also derives asymptotic confidence intervals for cross-validation procedures in risk estimation and model selection Bayle et al. (2020); Austern and Zhou (2020b). In our particular case, Proposition 5.7 implies that the tightness of the confidence intervals of cross-validation methods depends only on the quality of h_S.*

5.4.2 Excess Risk

In this section we analyse the excess risk of Algorithm 5.2 defined as:

$$\mathcal{E}_{\text{ex}} = \mathbb{E}\left[\mathcal{R}\left[\mathcal{A}\right] - \mathcal{R}\left[h^* + h_S\right]\right],$$

where $h^* = \arg\min_{h\in\mathcal{H}} \mathcal{R}\left[h_S + h\right]$. To this end, we start by showing that $\mathcal{E}_{\text{ex}}$ depends on the upper bounds of the *(pointwise) hypothesis stability* and the regularization parameter λ. Further, we derive precise finite-sample rates for the surrogate losses introduced in Section 5.3.3.

Theorem 5.14. *Suppose that $\|h^*\|_k < \infty$. Then, the excess risk of algorithm 5.2 verifies,*

$$\mathcal{E}_{ex} \leq \gamma(n) + \beta(n) + \lambda\|h^*\|_k^2.$$

Making λ varying with the sample size n, we obtain various consistent bounds for different losses. In the sequel, we assume that $\kappa \leq 1$ and $M_S \leq 1$ to avoid notional burden. When ϕ is either the MSE *or the squared hinge and $\lambda = \sqrt{\frac{\mathcal{R}[h_S]}{\sqrt{n}}}$, it holds:*

$$\mathcal{E}_{ex} \leq \mathcal{O}\left(\sqrt{\frac{\mathcal{R}[h_S]}{\sqrt{n}}}\right).$$

Furthermore, if ϕ is the exponential loss and $n \geq \frac{M_S^2 \ln(n)^2}{\mathcal{R}[h_S]}$, picking $\lambda = 4\frac{\sqrt{\mathcal{R}[h_S]} \wedge 1}{\ln(n)}$ yields:

$$\mathcal{E}_{ex} \leq \mathcal{O}\left(\frac{\sqrt{\mathcal{R}[h_S]} \wedge 1}{\ln(n)}\right),$$

otherwise picking $\lambda = \frac{\ln(n)^2}{\sqrt{n}}$ gives:

$$\mathcal{E}_{ex} \leq \mathcal{O}\left(\frac{\ln(n)^2}{\sqrt{n}}\right).$$

Suppose that the function ϕ is the logistic loss or the softplus. Then the choice $\lambda = \frac{1}{\sqrt{n}}$ yields:

$$\mathcal{E}_{ex} \leq \mathcal{O}\left(\frac{1}{\sqrt{n}}\right).$$

In particular, Theorem 5.14 yields the consistency of RERM. Furthermore, the Remark 5.13 regarding the generalization gap still holds for the excess risk. First, when $\mathcal{R}[h_S]$ is small, Algorithm 5.3 with MSE or squared hinge would have the fastest convergence rate. Second, when $\mathcal{R}[h_S]$ is large compared to the sample size n, then the safest option is to use the logistic or the softplus losses with $\lambda = \frac{1}{\sqrt{n}}$. Note that, if $\mathcal{R}[h_S]$ is small an improved convergence rate $\left(1/\sqrt{-n \ln\left(\mathcal{R}[h_s]\right)}\right)$ can be achieved for the latter losses (see Appendix 5.B.6 for further details). Finally, Algorithm 5.2 with the exponential loss is likely to suffer from negative learning. Indeed, if $\mathcal{R}[h_S]$ is large, one needs a large amount of data to ensure the non-triviality of the rate $\mathcal{R}[h_S]/\ln(n)$. It is worth noting that the rate of convergence with the exponential loss is naturally logarithmic even without a source hypothesis; see, for instance, Corollary 4.1 and Theorem 4.4 in Zhang (2004a). To conclude, using a good source hypothesis improves convergence rates of RERM compared to those derived without transfer Zhang (2004a).

Remark 5.15 (ON THE UNIVERSAL CONSISTENCY)**.** *If we assume that the kernel k is non-polynomial, h_S is continuous and the distribution of $X \in \mathcal{X}_T$ is* regular *(see* e.g. *Definition 4.2 in Zhang, 2004a). Then, one can use any* universal approximation theorem *(see for instance Theorem 4.1 in Zhang, 2004a) to obtain*

$$h^* = \operatorname*{arg\,min}_{h \in \mathcal{H}} \mathcal{R}\left[h_S + h\right] = \operatorname*{arg\,min}_{h \in \mathcal{L}(\mathcal{X}_T, \mathbb{R})} \mathcal{R}\left[h_S + h\right],$$

where $\mathcal{L}(\mathcal{X}_T, \mathbb{R})$ *is the space of real-valued functions defined on* $\mathcal{X}_T$*. The universal consistency of* $\mathcal{A}$ *follows immediately from Theorem 5.14. Further, all the losses presented in this chapter are* classification calibrated *(Bartlett et al., 2006b) meaning that:*

$$\underset{h \in \mathcal{L}(\mathcal{X}_T, \mathbb{R})}{\arg\min}\ \mathcal{R}\left[h\right] = \underset{h \in \mathcal{L}(\mathcal{X}_T, \mathbb{R})}{\arg\min}\ \mathcal{R}^{0\text{-}1}\left[h\right],$$

where $\mathcal{R}^{0\text{-}1}\left[h\right] = P_T\left(\operatorname{sign}\left(h(X)\right) \neq Y\right)$ *is the usual classification accuracy. Thus, minimizing the excess risk would likely yield a classifier with good accuracy.*

5.5 Numerical experiments

We illustrate our analysis by providing some results using simulated data that aim to underscore the robustness of each loss to negative learning scenarios. The experiment is conducted as follows. A source domain is considered with random variables $(X_S, Y_S) \in \mathbb{R}^2 \times \{-1, 1\}$, where the positive and negative classes are respectively drawn from two multivariate t-distributions $\mathcal{T}((r, 0), 3I_2, 2.5)$ and $\mathcal{T}((-r, 0), 3I_2, 2.5)$. We train a linear classifier h_S on a source dataset of size 10000 using the SVM algorithm.

To emphasize the impact of negative learning on each loss, we generate a smaller target dataset of size 100. The distributions for positive and negative classes are given by $\mathcal{T}(((r+d)cos(\theta), (r+d)sin(\theta)), I_2, 2.5)$ and $\mathcal{T}((-(r+d)cos(\theta), -(r+d)sin(\theta)), I_2, 2.5)$, respectively. For different values of θ, the target risk $\mathcal{R}\left[\hat{h} + h_s\right]$ of the analyzed RERM algorithm (with $\lambda = 1$) trained on the small size dataset is estimated using a test set of size 10000.

It is important to note that when $\theta = 0$, it corresponds to the scenario of positive learning since the decision boundaries of both domains are similar. On the other hand, the case where $\theta = \pi$ corresponds to negative learning since the true decision functions of the source and the target domain are pointing to opposite directions.

Figure 5.1 presents the median true risk of the HTL algorithm (cf. Equation 5.3) as a function of θ for $(r, d) = (5, 5)$ computed over 1000 simulations. The parameter s of the softplus loss is set to 0.1. Consistent with our theoretical analysis, the softplus and logistic functions exhibit significant robustness to negative transfer.

5.6 Conclusion

In this chapter, we study hypothesis transfer learning through the angle of Algorithmic Stability. Following the work of Kuzborskij and Orabona (2013), where hypothesis stability is shown for the MSE in the regression setting, we derive similar hypothesis stability rates in classification with general losses under slight assumptions. Furthermore, we show that our assumptions are satisfied for the most popular machine learning losses, making our work valuable for practitioners. Moreover, we leverage our stability results to provide finite-sample analysis on the generalization gap and the excess risk. We show that HTL framework is efficient and explicit (fast) rates for these popular losses. Our theoretical analysis will help practitioners better understand the benefits of HTL and give insight into the loss choices.

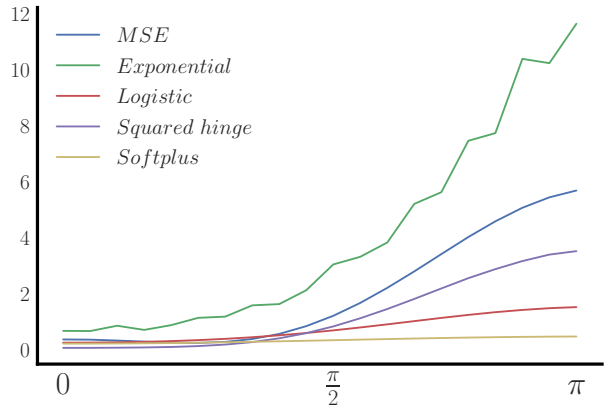


Figure 5.1 – Target risk of Algorithm 5.2 as a function of θ.

The proposed work is general and may fit with many other domains. Future work may involve our analysis for different Machine Learning tasks where transfer learning procedures can be beneficial such as robust learning (Shafahi et al., 2020; Laforgue et al., 2021; Staerman et al., 2021b), anomaly detection (Andrews et al., 2016; Chandola et al., 2009; Staerman et al., 2020, 2022a), speech Campi et al. (2021, 2023), automatic language generation (Staerman et al., 2021c; Golovanov et al., 2019), knowledge distillation (Cho and Hariharan, 2019), events-based modelling Staerman et al. (2022b), fairness Colombo et al. (2022b) or general neural-networks based tasks (Colombo et al., 2022a; Picot et al., 2023; Darrin et al., 2023).

5.A preliminary results

In this section, we show some useful technical lemmas used in the subsequent proofs.

Lemma 5.16. *Suppose that X, Y, Z are three mutually independent random variables such that $E(X) = E(Y)$. Then it holds:*

$$\mathbb{E}\left[(X+Z)(Y+Z)\right] \leq 2\left(\mathbb{E}\left[X\right]^2 + \mathbb{E}\left[Z^2\right]\right).$$

Proof Since X, Y, Z are mutually independent one has the following identities,

$$\begin{aligned}\mathbb{E}\left[(X+Z)(Y+Z)\right] &= \mathbb{E}\left[X\right]\mathbb{E}\left[Y\right] + \mathbb{E}\left[X\right]\mathbb{E}\left[Z\right] + \mathbb{E}\left[Z\right]\mathbb{E}\left[Y\right] + \mathbb{E}\left[Z^2\right] \\ &= \mathbb{E}\left[X\right]^2 + 2\mathbb{E}\left[X\right]\mathbb{E}\left[Z\right] + \mathbb{E}\left[Z^2\right].\end{aligned}$$

Now, noticing that $\left(E[Z]^2 \leq E[Z^2]\right)$ we get:

$$\begin{aligned}\mathbb{E}\left[X\right]^2 + 2\mathbb{E}\left[X\right]\mathbb{E}\left[Z\right] + \mathbb{E}\left[Z^2\right] &\leq 2\mathbb{E}\left[X\right]^2 + \mathbb{E}\left[Z\right]^2 + \mathbb{E}\left[Z^2\right] \\ &\leq 2\left(\mathbb{E}\left[X\right]^2 + \mathbb{E}\left[Z^2\right]\right),\end{aligned}$$

which is the desired result. ■

In the sequel, we shall provide an upper bound for the exponential of $\hat{\tau}_\lambda^i$ defined as:

$$\hat{\tau}_\lambda^i = \sqrt{\alpha\left(\widehat{\mathcal{R}}^{\setminus i}[h_S] + \frac{M_S}{n}\right)}, \tag{5.5}$$

with $M_S = \sup_{z\in\mathcal{Z}_T} \ell(h_S, z)$ and

$$\widehat{\mathcal{R}}^{\setminus i}[h] = \frac{1}{n-1}\sum_{j\neq i} \ell(h, Z_j), \tag{5.6}$$

the training error of a hypothesis h with the i'th datum removed. The quantity $\hat{\tau}_\lambda^i$ will serve as an upper bound of $\hat{\rho}_\lambda^i = \max(\hat{r}_\lambda, \hat{r}_\lambda^i)$ independent of the observation $Z_i \in \mathcal{D}_T$. Indeed, by definition:

$$\hat{\tau}_\lambda^i \geq \hat{r}_\lambda^i = \sqrt{\alpha\left(\widehat{\mathcal{R}}^{\setminus i}[h_S]\right)}.$$

Moreover, it holds:

$$\widehat{\mathcal{R}}\left[h\right] \leq \widehat{\mathcal{R}}^{\setminus i}\left[h\right] + \frac{\ell(h, Z_i)}{n} \leq \widehat{\mathcal{R}}^{\setminus i}\left[h\right] + \frac{M_S}{n},$$

so that $\hat{\tau}_\lambda^i \geq \hat{r}_\lambda^i$. Thus, we have $\hat{\tau}_\lambda^i \geq \hat{\rho}_\lambda^i$.

Lemma 5.17. *Let $W_1, W_2, \ldots, W_n$ be a sequence of i.i.d. random variables bounded by $C > 0$. Then one has*

$$\mathbb{E}\left[e^{\hat{\mu}}\right] \leq e^{\mu + \frac{C^2}{n}},$$

where $\mu = \mathbb{E}\left[W_1\right]$ and $\hat{\mu} = \frac{1}{n}\sum_{i=1}^n W_i$.

Proof The proof follows in two steps. First, we apply Hoeffding's inequality to obtain:

$$\mathbb{P}\left(|\hat{\mu} - \mu| \geq t\right) \leq e^{\frac{-nt^2}{C^2}}.$$

Second, applying Theorem 2.5.2 in Vershynin (2018) yields:

$$\mathbb{E}\left[e^{\hat{\mu}-\mu}\right] \leq e^{\frac{C^2}{n}},$$

which leads to the desired result. ■

Lemma 5.18. *For all $i \in [n]$ and $p \in \mathbb{N}$, the quantity $e^{\hat{\tau}_\lambda^i}$ verifies:*

$$\mathbb{E}\left[e^{p\hat{\tau}_\lambda^i}\right] \leq e^{p + \frac{\alpha p M_S}{n} + \frac{\alpha^2 p^2 M_S^2}{n-1}} e^{p\alpha\mathcal{R}[g]}.$$

Proof First, using the fact that $\sqrt{x} \leq x + 1$, one has:

$$e^{\hat{\tau}_\lambda^i} \leq e^{p\alpha\widehat{\mathcal{R}}^{\setminus i}[h_S] + \alpha\frac{pM_S}{n} + p}.$$

Since $p\alpha\widehat{\mathcal{R}}^{\setminus i}\left[h_S\right] = \frac{1}{n-1}\sum_{i\neq j} p\alpha\ell(h, Z_i)$, applying Lemma 5.17 with $W_i = p\alpha\ell(h, Z_i)$ and $C = p\alpha M_S$ yields the desired result.

■

To prove Propositions 5.7 and 5.10, we extend Theorem 4.3 in Zhang (2004a), that gives an upper bound for standard RERM to the HTL framework. This extension leads to the next lemma.

Lemma 5.19. *The leave one out deviations of the algorithm $\mathcal{A}$ (cf. Equation (5.2)) verifies:*

$$\|\mathcal{A}\left(\mathcal{D}_T\right) - \mathcal{A}\left(\mathcal{D}_T^{\setminus i}\right)\|_k \leq \frac{k\left(X_i, X_i\right)^{1/2}\left|\phi'\left(\mathcal{A}\left(\mathcal{D}_T, X_i\right) Y_i\right)\right|}{\lambda n}.$$

Proof Since ϕ is convex, the Bregman divergence of ϕ is non negative. More precisely,

$$d_\phi(x, y) = \phi(x) - \phi(y) - (x - y)\phi'(y) \geq 0,$$

so that, for any $Z_i = (X_i, Y_i) \in \mathcal{D}_T$ one has:

$$\ell\left(\mathcal{A}\left(\mathcal{D}_T^{\setminus i}\right), Z_i\right) - d_\phi\left(\mathcal{A}\left(\mathcal{D}_T^{\setminus i}, X_i\right) Y_i, \mathcal{A}\left(\mathcal{D}_T, X_i\right) Y_i\right) \leq \ell\left(\mathcal{A}\left(\mathcal{D}_T^{\setminus i}\right), Z_i\right),$$

where $\mathcal{A}\left(\mathcal{D}_T, X_i\right)$ is the prediction of the input X_i by the algorithm $\mathcal{A}$. Also, the term on the left side in the above inequality can be written as follows:

$$\begin{aligned}\ell\left(\mathcal{A}\left(\mathcal{D}_T^{\setminus i}\right), Z_i\right) - d_\phi\left(\mathcal{A}\left(\mathcal{D}_T^{\setminus i}, X_i\right) Y_i, \mathcal{A}\left(\mathcal{D}_T, X_i\right) Y_i\right) &= \ell\left(\mathcal{A}(\mathcal{D}_T), Z_i\right) \\ &+ \phi'\left(\mathcal{A}(\mathcal{D}_T, X_i) Y_i\right)\left(\mathcal{A}\left(\mathcal{D}_T^{\setminus i}, X_i\right) - \mathcal{A}\left(\mathcal{D}_T, X_i\right)\right) Y_i,\end{aligned}$$

so that:

$$\ell\left(\mathcal{A}(\mathcal{D}_T), Z_i\right) + \phi'\left(\mathcal{A}(\mathcal{D}_T, X_i) Y_i\right)\left(\mathcal{A}\left(\mathcal{D}_T^{\setminus i}, X_i\right) - \mathcal{A}\left(\mathcal{D}_T, X_i\right)\right) Y_i \leq \ell\left(\mathcal{A}(\mathcal{D}_T^{\setminus i}), Z_i\right).$$

Thus, we get:

$$\widehat{\mathcal{R}}^{\setminus i}\left[\mathcal{A}\left(\mathcal{D}_T\right)\right] + S_i \leq \widehat{\mathcal{R}}^{\setminus i}\left[\mathcal{A}\left(\mathcal{D}_T^{\setminus i}\right)\right], \tag{5.7}$$

where $\widehat{\mathcal{R}}^{\setminus i}$ is defined previously in Equation (5.6) and

$$S_i = \frac{1}{n}\sum_{j\neq i} \phi'\left(\mathcal{A}(\mathcal{D}_T, X_j) Y_j\right)\left(\mathcal{A}\left(\mathcal{D}_T^{\setminus i}, X_j\right) - \mathcal{A}(\mathcal{D}_T, X_j)\right).$$

Let $\hat{h}^{\setminus i}$ denote the solution of the optimization problem 5.3 with the i'th datum removed. One gets by definition of $\mathcal{A}$ (cf. Equation (5.2)),

$$\widehat{\mathcal{R}}^{\setminus i}\left[\mathcal{A}\left(\mathcal{D}_T^{\setminus i}\right)\right] + \lambda\|\hat{h}^{\setminus i}\|_k^2 \leq \widehat{\mathcal{R}}^{\setminus i}\left[\mathcal{A}\left(\mathcal{D}_T\right)\right] + \lambda\|\hat{h}\|_k^2.$$

Using (5.7), it yields:

$$\begin{aligned} S_i &\leq \lambda \left(\|\hat{h}\|_k^2 - \|\hat{h}^{\setminus i}\|_k^2 \right) \\ &\leq -\lambda \|\hat{h} - \hat{h}^{\setminus i}\|_k^2 - 2\lambda \langle \hat{h}, \hat{h}^{\setminus i} - \hat{h} \rangle, \end{aligned}$$

where the second line follows from $\|x\| - \|y\| = \|x - y\|^2 + 2\langle x - y, y \rangle$. Reverting the inequality leads to:

$$\begin{aligned} \lambda \|\hat{h} - \hat{h}^{\setminus i}\|_k^2 &\leq -\frac{1}{n_T} \sum_{j \in T^{\setminus i}} \phi' \left(\mathcal{A}(\mathcal{D}_T, X_j) Y_j \right) \langle \hat{h}^{\setminus i} - \hat{h}, k\left(X_i, \cdot\right) \rangle - 2\lambda \langle \hat{h}, \hat{h}^{\setminus i} - \hat{h} \rangle \\ &\leq \left\| \frac{1}{n_T} \sum_{j \in T^{\setminus i}} \phi' \left(\mathcal{A}(\mathcal{D}_T, X_j) Y_j \right) k\left(X_i, \cdot\right) + 2\lambda g \right\|_k \|\hat{h}^{\setminus i} - \hat{h}\|_k. \end{aligned} \tag{5.8}$$

The last inequalities hold because of the definition of S_i:

$$\begin{aligned} S_i &= \frac{1}{n} \sum_{j \neq i} \phi' \left(\mathcal{A}(\mathcal{D}_T, X_j) Y_j \right) \left(\mathcal{A}\left(\mathcal{D}_T^{\setminus i}, X_j \right) - \mathcal{A}(\mathcal{D}_T, X_j) \right) \\ &= \frac{1}{n} \sum_{j \neq i} \phi' \left(\mathcal{A}(\mathcal{D}_T, X_j) Y_j \right) \left(\hat{h}^{\setminus i}\left(X_j\right) - \hat{h}\left(X_j\right) \right) \\ &= \frac{1}{n} \sum_{j \neq i} \phi' \left(\mathcal{A}(\mathcal{D}_T, X_j) Y_j \right) \langle \hat{h}^{\setminus i} - \hat{h}, k(X_j, \cdot) \rangle. \end{aligned}$$

On the other hand, since $A(\mathcal{D}_T, X_j) = h_S(X_j) + \langle \hat{h}, k(X_j, \cdot) \rangle$ and by Theorem 3.1.20 in Nesterov et al. (2018), we know that the following optimality condition holds:

$$\frac{1}{n} \sum_{j=1}^{n} \phi' \left(\mathcal{A}(\mathcal{D}_T, X_j) Y_j \right) k\left(X_j, \cdot\right) + 2\lambda \hat{h} = 0.$$

Therefore Inequality (5.8) becomes:

$$\lambda \|\hat{h} - \hat{h}^{\setminus i}\|^2 \leq \left\| \frac{1}{n} \phi' \left(\mathcal{A}(T, X_i) Y_i \right) \right\|_k \|k(X_i, \cdot)\|_k \|\hat{h}^{\setminus i} - \hat{h}\|_k.$$

it remains to remind that $\|k\left(X_i, \cdot\right)\|^2 = k\left(X_i, X_i\right)$ and $\|\mathcal{A}\left(\mathcal{D}_T\right) - \mathcal{A}\left(\mathcal{D}_T^{\setminus i}\right)\|_k = \|\hat{h}^{\setminus i} - \hat{h}\|_k$ to complete the proof. ■

Before highlighting the link between Algorithm 5.2 with RLS, let's remind a useful lemma (representer theorem) that allows simplifying the optimization problem 5.3 in practice.

Lemma 5.20. *The learning rule $\hat{h}$ (cf. Equation 5.3) lies in the linear span in $\mathcal{H}$ of the vectors $\left(k\left(X_i, \cdot\right)\right)_{1 \leq i \leq n}$,* i.e.

$$\hat{h} \in \mathcal{H}_{\mathcal{D}},$$

with $\mathcal{H}_{\mathcal{D}} = \left\{ \sum_1^n \alpha_i k\left(X_i, \cdot\right) \mid \alpha_1, \ldots, \alpha_n \in \mathbb{R} \right\}$.

Proof Since $\mathcal{H}_{\mathcal{D}}$ is a finite dimensionnal subspace of $\mathcal{H}$, any $h \in \mathcal{H}$ can be decomposed as:

$$h = h_{\mathcal{D}} + h^{\perp},$$

with $h_{\mathcal{D}} \in \mathcal{H}_{\mathcal{D}}$ and $h^{\perp} \perp \mathcal{H}_{\mathcal{D}}$. Furthermore using the fact that $h(x) = \langle h, k(x, \cdot)\rangle_k$, for all $i \in [n]$, one obtains:

$$h(X_i) = \langle h, k(X_i, \cdot)Y_i\rangle = \langle h_{\mathcal{D}}, k(X_i, \cdot)Y_i\rangle = h_{\mathcal{D}}(X_i)Y_i.$$

Thus, for any $Z_i \in \mathcal{D}_T$, it holds:

$$\ell(h{+}h_S, Z_i) = \phi\left(\Big(h(X_i) + h_S(X_i)\Big)\, Y_i\right) = \phi\left(\Big(h_{\mathcal{D}}\left(X_i\right) + h_S\left(X_i\right)\Big)\, Y_i\right) = \ell(h_{\mathcal{D}}{+}h_S, Z_i),$$

which gives

$$\widehat{\mathcal{R}}(h + h_S) = \widehat{\mathcal{R}}(h_{\mathcal{D}} + h_S).$$

On the other hand, by the Pythagorean theorem,

$$\|h_{\mathcal{D}}\|_k^2 \leq \|h\|_k^2,$$

and

$$\widehat{\mathcal{R}}(h + h_S) + \lambda\|h_{\mathcal{D}}\|_k^2 \leq \widehat{\mathcal{R}}(h_{\mathcal{D}} + h_S) + \lambda\|h_{\mathcal{D}}\|_k^2.$$

Thus, the solution of the minimization problem 5.3 must lie in $\mathcal{H}_{\mathcal{D}}$. ■

5.A.1 Link with Least Squares with Biased Regularisation

To begin, it is a well know fact that, when the kernel k is linear then the RKHS space consists of the set of linear classifiers:

$$\mathcal{H} = \left\{h(x) = u^{\top}x \mid u \in \mathbb{R}^d\right\}.$$

In this case, the solution of the optimization problem with the mean square loss $\ell(h, Z) = (1 - h(X)Y)^2$, writes as $\hat{h} = \hat{u}^{\top}x$ with

$$\begin{aligned}
\hat{u} &= \operatorname*{arg\,min}_{u\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^{n}\left(u^{\top}X_iY_i + h_S(X_i)Y_i - 1\right)^2 + \lambda\|u\|_2^2 \\
&= \operatorname*{arg\,min}_{u\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^{n} Y_i^2\left(u^{\top}X_i + h_S(X_i) - \frac{1}{Y_i}\right)^2 + \lambda\|u\|_2^2 \\
&= \operatorname*{arg\,min}_{u\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^{n}\left(u^{\top}X_i + h_S(X_i) - Y_i\right)^2 + \lambda\|u\|_2^2,
\end{aligned}$$

where the last inequality follows from the facts that $Y_i^2 = 1$ and $\frac{1}{Y_i} = Y_i$. Furthermore, if $h_S(x) = v^T x$ for some $v \in \mathbb{R}^d$ one has:

$$\begin{aligned}
\hat{u} &= \operatorname*{arg\,min}_{u\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^{n}\left((u+v)^{\top}X_i - Y_i\right)^2 + \lambda\|u\|_2^2 \\
&= \operatorname*{arg\,min}_{u\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^{n}\left(u^{\top}X_i - Y_i\right)^2 + \lambda\|u - v\|_2^2.
\end{aligned}$$

This is the original form of biased regularisation algorithms.

5.B Technical proofs of the main results

Before starting the proof of our main results, we remind two properties of RKHS spaces that are:

$$\forall x, y \in \mathcal{X}_T \ , \quad \langle k(y,\cdot), k(x,\cdot)\rangle = k(x,y),$$

and

$$\forall h \in \mathcal{H} \ , \ \forall x \in \mathcal{X}_T \ , \quad h(x) = \langle h, k(x,\cdot)\rangle.$$

Under Assumption 8, using Cauchy Schwartz-inequality yields:

$$\forall h \in \mathcal{H} \ , \quad \|h\|_\infty \leq \sqrt{\kappa}\|h\|_k.$$

5.B.1 Proof of Lemma 5.6

This lemma follows from our assumptions and a simple fact. Indeed, notice that by definition of $\hat{h}$

$$\widehat{\mathcal{R}}(\hat{h} + h_S) + \lambda\|\hat{h}\|^2 \leq \widehat{\mathcal{R}}(\mathbf{0} + h_S).$$

Furthermore, $\widehat{\mathcal{R}}(h_S + \hat{h})$ is non-negative since ϕ is non-negative which concludes the proof.

5.B.2 Proof of Proposition 5.7

Let $Z = (X, Y) \in \mathcal{Z}_T$ and remind that, by definition of $\mathcal{A}$, one has:

$$\left|\ell\left(\mathcal{A}\left(\mathcal{D}_T\right), Z\right) - \ell\left(\mathcal{A}\left(\mathcal{D}_T^{\setminus i}\right), Z\right)\right| = \left|\phi\left(\left(\hat{h}(X) + h_S(X)\right)Y\right) - \phi\left(\left(\hat{h}^{\setminus i}(X) + h_S(X)\right)Y\right)\right|,$$

where $\hat{h}$ is the solution of the optimization problem 5.3. Moreover, since ϕ is differentiable, one can apply the mean value theory to obtain:

$$\begin{aligned}\left|\ell\left(\mathcal{A}\left(\mathcal{D}_T\right), Z\right) - \ell\left(\mathcal{A}\left(\mathcal{D}_T^{\setminus i}\right), Z\right)\right| &= |\phi'\left(\left(y_{\mathcal{D}} + h_S(X)\right)Y\right)||\hat{h}(X) - \hat{h}^{\setminus i}(X)| \\ &\leq \sqrt{\kappa}|\phi'\left(\left(y_{\mathcal{D}} + h_S(X)\right)Y\right)|\|\hat{h} - \hat{h}^{\setminus i}\|_k \\ &= \sqrt{\kappa}|\phi'\left(\left(y_{\mathcal{D}} + h_S(X)\right)Y\right)|\|\mathcal{A}(\mathcal{D}_T) - \mathcal{A}\left(\mathcal{D}_T^{\setminus i}\right)\|_k,\end{aligned}$$

for some $|y_{\mathcal{D}}| \leq \max\left(\hat{h}(X), \hat{h}^{\setminus i}(X)\right)$. By Lemma 5.6, we have $|y_{\mathcal{D}}| \leq \hat{\rho}_\lambda^i = \max\left(\hat{r}_\lambda, \hat{r}_\lambda^{\setminus i}\right)$. Now, Using Theorem 5.19 with Assumption 8 yields:

$$\left|\ell\left(\mathcal{A}\left(\mathcal{D}_T\right), Z\right) - \ell\left(\mathcal{A}\left(\mathcal{D}_T^{\setminus i}\right), Z\right)\right| \leq \kappa\frac{|\phi'\left(\left(y_{\mathcal{D}} + h_S(X)\right)Y\right)\phi'\left(\left(\hat{h}(X_i) + h_S(X_i)\right)Y_i\right)|}{\lambda n}, \tag{5.9}$$

which gives using the fact that $\|\hat{h}\|_\infty \leq \hat{r}_\lambda \leq \hat{\rho}_\lambda$:

$$\left|\ell\left(\mathcal{A}\left(\mathcal{D}_T\right), Z\right) - \ell\left(\mathcal{A}\left(\mathcal{D}_T^{\setminus i}\right), Z\right)\right| \leq \sup_{|y|,|y'|\leq\hat{\rho}_\lambda^i} \frac{\alpha\left|\phi'(h_S(X_i)Y_i + y)\phi'(h_S(X)Y + y')\right|}{n}, \tag{5.10}$$

with $\alpha = \frac{\kappa}{\lambda}$. Now, by taking the expectation and using the fact that ϕ verifies assumption 9, Inequality (5.10) becomes:

$$\mathbb{E}\left[\left|\ell\left(\mathcal{A}\left(\mathcal{D}_T\right), Z\right) - \ell\left(\mathcal{A}\left(\mathcal{D}_T^{\setminus i}\right), Z\right)\right|\right] \leq \alpha\frac{\Psi_1(\mathcal{R}\left[h_S\right])}{n}.$$

Besides, notice that by Equation (5.9),

$$\forall \mathcal{D}_T \in \mathcal{Z}_T^n,\ \forall Z \in \mathcal{Z}_T\ ,\ \left|\ell\left(\mathcal{A}\left(\mathcal{D}_T\right), Z\right) - \ell\left(\mathcal{A}\left(\mathcal{D}_T^{\setminus i}\right), Z\right)\right| \leq \alpha\frac{\|\phi'\|_\infty^2}{n}.$$

It remains to take the expectation to complete the proof.

5.B.3 Proof of Proposition 5.10

The proof is similar to the previous one thus we will only give the key step: replace $Z = (X, Y)$ by $Z_i = (X_i, Y_i)$ in Equation (5.10) to obtain:

$$\left|\ell\left(\mathcal{A}\left(\mathcal{D}\right), Z_i\right) - \ell\left(\mathcal{A}\left(\mathcal{D}^{\setminus i}\right), Z_i\right)\right| \leq \sup_{|y|,|y'|\leq\hat{\rho}_\lambda^i} \frac{\alpha\left|\phi'(h_S(X_i)Y_i + y)\phi'(h_S(X_i)Y_i + y')\right|}{n}.$$

To conclude the proof, take the expectation of both sides of the last inequality and use the Assumption 10.

5.B.4 proof of Proposition 5.11

First, let $i \in [n]$ and $|y|, |y'| \leq \hat{\rho}_\lambda^i$. Furthermore let $Z = (X, Y)$ and $Z = (X', Y')$ be two observations independent of $\mathcal{D}^{\setminus i}$. We start by showing that the MSE and squared hinge verify Assumptions 9, 10 and explicit their corresponding function Ψ_1, Ψ_2 . To do so, remind that:

$$\begin{aligned}
\left(\hat{\rho}_\lambda^i\right)^2 &= \max(\hat{r}_\lambda^i, \hat{r}_\lambda)^2 \\
&\leq (\hat{r}_\lambda^i + \hat{r}_\lambda)^2 \\
&\leq 2\left(\hat{r}_\lambda^i\right)^2 + 2\left(\hat{r}_\lambda\right)^2 \\
&= 2\alpha\left(\widehat{\mathcal{R}} + \widehat{\mathcal{R}}^{\setminus i}[h_S]\right).
\end{aligned} \tag{5.11}$$

MSE

Recall the MSE loss $\phi(x) = (1-x)^2$. For all $x \in \mathbb{R}$, one has:

$$\begin{aligned} |\phi'(x+y)| &= 2|1-x-y| \\ &\leq 2|1-x| + 2|y| \\ &\leq 2\sqrt{\phi(x)} + 2\hat{\rho}_\lambda^i. \end{aligned} \tag{5.12}$$

Thus,

$$\sup_{|y'|,|y|\leq\hat{\rho}_\lambda^i} \left|\phi'(h_S(X')Y' + y')\phi'(h_S(X)Y + y)\right| \leq 4\left(\sqrt{\phi\left(h_S(X')Y'\right)} + \hat{\rho}_\lambda^i\right)\left(\sqrt{\phi\left(h_S(X)Y\right)} + \hat{\rho}_\lambda^i\right).$$

Taking the expectation of the latter inequality and using Lemma 5.16 with $X = \sqrt{\phi\left(h_S(X')Y'\right)}$, $Y = \sqrt{\phi\left(h_S(X)Y\right)}$ and $Z = \hat{\rho}_\lambda^i$ yields:

$$\begin{aligned} \mathbb{E}\left[\sup_{|y'|,|y|\leq\hat{\rho}_\lambda^i} \left|\phi'(h_S(X')Y' + y')\phi'(h_S(X)Y + y)\right|\right] &\leq 8\left(\mathbb{E}\left[\sqrt{\phi\left(h_S(X)Y\right)}\right]^2 + \mathbb{E}\left[\left(\hat{\rho}_\lambda^i\right)^2\right]\right) \\ \text{(by Jensen's Inequality)} &\leq 8\left(\mathbb{E}\left[\phi\left(h_S(X)Y\right)\right] + \mathbb{E}\left[\left(\hat{\rho}_\lambda^i\right)^2\right]\right) \\ \left(\phi\left(h_S(X)Y\right) = \ell(h_S, Z)\right) &\leq 8\left(\mathcal{R}[h_S] + \mathbb{E}\left[\left(\hat{\rho}_\lambda^i\right)^2\right]\right) \\ \text{(Inequality (5.11))} &\leq 8\left(\mathcal{R}[h_S] + 4\alpha\mathcal{R}[h_S]\right). \end{aligned}$$

This means that the MSE verifies Assumption 9 with $\Psi_1(x) = 8x(1+4\alpha)$. Now using Inequality (5.12) again yields:

$$\sup_{|y'|,|y|\leq\hat{\rho}_\lambda^i} \left|\phi'(h_S(X)Y + y')\phi'(h_S(X)Y + y)\right| \leq 4\left(\sqrt{\phi\left(h_S(X)Y\right)} + \hat{\rho}_\lambda^i\right)^2.$$

By taking the expectation and mimicking the previous step one can show that the MSE verifies Assumption 10 with $\Psi_2(x) = 8x(1+4\alpha)$.

Squared hinge

First recall the loss function $\phi(x) = \max\left(0, 1-x\right)^2$. By simple calculation we obtain:

$$|\phi'(x+y)| = 2\max\left(0, 1-x-y\right).$$

On the other hand, one has:

$$\begin{cases} 0 \leq \max(0, 1-x) + |y|, \\ 1-x-y \leq \max(0, 1-x) + |y|. \end{cases}$$

Thus, it holds:

$$\begin{aligned}|\phi'(x+y)| &\leq 2\max(0,1-x)+2|y| \\ &\leq 2\sqrt{\phi(x)}+2\hat{\rho}_\lambda^i.\end{aligned}$$

The result follows using the same steps as in the MSE case.

Exponential

Recalling the loss function $\phi(x)=e^{-x}$, first notice that the exponential loss verifies:

$$|\phi'(x+y)| = e^{-x}e^{-y} = \phi(x)e^{-y} \leq \phi(x)e^{\hat{\rho}_\lambda^i} \leq \phi(x)e^{\hat{\tau}_\lambda^i}, \tag{5.13}$$

where $\hat{\tau}_\lambda^i$ is given by Equation (5.5). Thus, we get:

$$\begin{aligned}\mathbb{E}\left[\sup_{|y'|,|y|\leq\hat{\rho}_\lambda^i}\left|\phi'(h_S(X')Y'+y')\phi'(h_S(X)Y+y)\right|\right] &\leq \mathbb{E}\left[\phi(h_S(X')Y')\phi(h_S(X)Y)e^{2\hat{\tau}_\lambda^i}\right] \\ (Z \perp\!\!\!\perp Z' \perp\!\!\!\perp \hat{\tau}_\lambda^i) &\leq \mathcal{R}[h_S]^2\mathbb{E}\left[e^{2\hat{\tau}_\lambda^i}\right]. \\ \text{(By Lemma 5.18 with } p=2) &\leq \mathcal{R}[h_S]^2 e^{2+\frac{2\alpha M_S}{n}+\frac{4\alpha^2 M_S^2}{n-1}} e^{2\alpha\mathcal{R}[g]}\end{aligned}$$

Thus the exponential loss verifies Assumption 9 with $\Psi_1(x)=C_S x^2 e^{2\alpha x}$ and $C_S = e^{2+\frac{2\alpha M_S}{n}+\frac{4\alpha^2 M_S^2}{n-1}}$. Besides, using (5.13) again yields:

$$\begin{aligned}\mathbb{E}\left[\sup_{|y'|,|y|\leq\hat{\rho}_\lambda^i}\left|\phi'(h_S(X)Y+y')\phi'(h_S(X)Y+y)\right|\right] &\leq \mathbb{E}\left[\phi(h_S(X)Y)^2 e^{2\hat{\tau}_\lambda^i}\right] \\ (M_S = \sup_{Z\in\mathcal{Z}_T}\ell(h_S,Z)) &\leq M_S\mathbb{E}\left[\phi(h_S(X)Y)e^{2\hat{\tau}_\lambda^i}\right] \\ (Z \perp\!\!\!\perp \hat{\tau}_\lambda^i) &\leq M_S\mathcal{R}[h_S]\mathbb{E}\left[e^{2\hat{\tau}_\lambda^i}\right] \\ \text{(By Lemma 5.18 with } p=2) &\leq M_S\mathcal{R}[h_S]e^{2+\frac{2\alpha M_S}{n}+\frac{4\alpha^2 M_S^2}{n-1}} e^{2\alpha\mathcal{R}[g]}.\end{aligned}$$

Therefore the exponential loss verifies Assumption 10 with $\Psi_2(x)=C_S M_S x e^{2\alpha x}$.

Logistic

Recall the loss function $\phi(x)=\log(1+e^{-x})$ and its derivative:

$$|\phi'(x)| = \frac{e^{-x}}{e^{-x}+1}.$$

Thus, we have:

$$\begin{aligned}|\phi'(x+y)| &= \frac{e^{-x-y}}{e^{-x-y}+1}\\ &\leq e^{-y}e^{-x}\\ &= e^{-y}\left(e^{\phi(x)}-1\right)\\ &\leq e^{\hat{\rho}_\lambda^i}\left(e^{\phi(x)}-1\right)\\ &\leq e^{\hat{\tau}_\lambda^i}\left(e^{\phi(x)}-1\right),\end{aligned}$$

where the two last inequalities result from the facts that $y \leq \hat{\rho}_\lambda^i$ and $\hat{\rho}_\lambda^i \leq \hat{\tau}_\lambda^i$ respectively. Using the facts that $\|\phi'\|_\infty \leq 1$ and $e^{\hat{\tau}_\lambda^i} \leq 1$, one obtains:

$$\begin{aligned}\left|\phi'(h_S(X)Y+y)\right| &\leq \min\left(e^{\hat{\tau}_\lambda^i}\left(e^{\phi\left(h_S(X)Y\right)}-1\right),1\right)\\ &= \min\left(e^{\hat{\tau}_\lambda^i}\left(e^{\ell(h_S,Z)}-1\right),1\right)\\ &\leq \min\left(e^{\hat{\tau}_\lambda^i}\left(e^{\ell(h_S,Z)}-1\right),e^{\hat{\tau}_\lambda^i}\right)\\ &\leq e^{\hat{\tau}_\lambda^i}\min\left(\left(e^{\ell(h_S,Z)}-1\right),1\right). \qquad (5.14)\end{aligned}$$

The latter inequality yields:

$$\sup_{|y'|,|y|\leq\hat{\rho}_\lambda^i}\left|\phi'(h_S(X')Y'+y')\phi'(h_S(X)Y+y)\right| \leq e^{2\hat{\tau}_\lambda^i}\min\left(e^{\ell\left(h_S,Z\right)}-1,1\right)\min\left(e^{\ell\left(h_S,Z'\right)}-1,1\right).$$

Thus, since Z, Z' are independent of $\mathcal{D}_T^{\setminus i}$, they are also independent of $\hat{\tau}_\lambda^i$. It follows:

$$\begin{aligned}\mathbb{E}\left[\sup_{|y'|,|y|\leq\hat{\rho}_\lambda^i}\left|\phi'(h_S(X')Y'+y')\phi'(h_S(X)Y+y)\right|\right] &\leq \mathbb{E}\left[e^{2\hat{\tau}_\lambda^i}\right]\mathbb{E}\left[\min\left(e^{\ell(h_S,Z)}-1,1\right)\right]^2\\ \text{(By Lemma 5.18)} &\leq C_S e^{2\alpha\mathcal{R}[h_S]}\mathbb{E}\left[\min\left(e^{\ell(h_S,Z)}-1,1\right)\right]^2. \qquad (5.15)\end{aligned}$$

Now using the fact that:

$$e^{\ell(h_S,Z)}-1 \leq 1 \implies \ell(h_S,Z) \leq 1 \implies \ell(h_S,Z) \leq \sqrt{\ell(h_S,Z)},$$

we have:

$$\mathbb{E}\left[\min\left(e^{\ell(h_S,Z)}-1,1\right)\right] \leq \mathbb{E}\left[\min\left(e^{\sqrt{\ell(h_S,Z)}}-1,1\right)\right].$$

In addition, notice that:

$$(e^{\sqrt{x}}-1)\wedge 1 = \begin{cases} e^{\sqrt{x}}-1 & \text{if } x \leq \ln(2)^2,\\ 1 & \text{otherwise,}\end{cases}$$

which is concave. Therefore, it holds:

$$\mathbb{E}\left[\min\left(e^{\ell(h_S,Z)}-1,1\right)\right] \leq \min\left(e^{\sqrt{\mathcal{R}[h_S]}}-1,1\right).$$

To show that the logistic loss verifies Assumption 9 with $\Psi_1(x) = C_S e^{2\alpha\mathcal{R}[h_S]}(e^{\sqrt{x}}-1)^2$, it suffices to plug the latter inequality in (5.15). Now, using (5.14) again yields:

$$\begin{aligned}\mathbb{E}\left[\sup_{|y'|,|y|\leq\hat{\rho}_\lambda^i}\left|\phi'(h_S(X)Y+y')\phi'(h_S(X)Y+y)\right|\right] &\leq \mathbb{E}\left[e^{2\hat{\tau}_\lambda^i}\min\left(e^{\ell(h_S,Z)}-1,1\right)^2\right]\\ &\leq \mathbb{E}\left[e^{2\hat{\tau}_\lambda^i}\right]\mathbb{E}\left[\min\left(e^{\ell(h_S,Z)}-1,1\right)\right].\end{aligned}$$

Finally, using the same steps as before, we show that the logistic loss verifies Assumption 10 with $\Psi_1(x) = C_S e^{2\alpha\mathcal{R}[h_S]}(e^{\sqrt{x}}-1)$.

Softplus

The proof is similar to that of the logistic loss and is left for the reader.

5.B.5 Proof of Theorem 5.12

First, notice that:

$$\begin{aligned}\mathcal{E}_{\text{gen}} = |\mathbb{E}\left[\widehat{\mathcal{R}}\left[\mathcal{A}(\mathcal{D}_T)\right]-\mathcal{R}\left[\mathcal{A}(\mathcal{D}_T)\right]\right]| &= |\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n \ell\left(\mathcal{A}\left(\mathcal{D}_T\right),Z_i\right)-\ell\left(\mathcal{A}\left(\mathcal{D}_T^{\setminus i}\right),Z\right)\right]|\\ &= |\mathbb{E}\left[\ell\left(\mathcal{A}\left(\mathcal{D}_T\right),Z_1\right)-\ell\left(\mathcal{A}\left(\mathcal{D}_T^{\setminus i}\right),Z\right)\right]|.\end{aligned}$$

Using triangle inequality and the fact that Z and Z_1 have the same distributions, we obtains:

$$\begin{aligned}\mathcal{E}_{\text{gen}} &\leq |\mathbb{E}\left[\ell\left(\mathcal{A}\left(\mathcal{D}_T\right),Z_1\right)-\ell\left(\mathcal{A}\left(\mathcal{D}_T^{\setminus i}\right),Z_1\right)\right]| + |\mathbb{E}\left[\ell\left(\mathcal{A}\left(\mathcal{D}_T^{\setminus i}\right),Z_1\right)-\ell\left(\mathcal{A}\left(\mathcal{D}_T^{\setminus i}\right),Z\right)\right]|\\ &= |\mathbb{E}\left[\ell\left(\mathcal{A}\left(\mathcal{D}_T\right),Z_1\right)-\ell\left(\mathcal{A}\left(\mathcal{D}_T^{\setminus i}\right),Z_1\right)\right]| + |\mathbb{E}\left[\ell\left(\mathcal{A}\left(\mathcal{D}_T^{\setminus i}\right),Z\right)-\ell\left(\mathcal{A}\left(\mathcal{D}_T^{\setminus i}\right),Z\right)\right]|.\end{aligned}$$

The desired result follows from Propositions 5.7 and 5.10.

5.B.6 Proof of Theorem 5.14

First introduce

$$h_\lambda = \operatorname*{arg\,min}_{h\in\mathcal{H}} \mathcal{R}\left[h_S+h\right]+\lambda\|h\|_k^2,$$

and write

$$\mathcal{R}\left[\mathcal{A}\right] - \mathcal{R}[h^* + h_S] = \mathcal{R}\left[\mathcal{A}\right] - \widehat{\mathcal{R}}\left[\mathcal{A}\right] + \widehat{\mathcal{R}}\left[\mathcal{A}\right] + \lambda\|f\|_k^2 - \mathcal{R}[h_\lambda + h_S] + \mathcal{R}[h_\lambda + g'] - \mathcal{R}[h^* + h_S].$$

Now by rearranging and reminding that:

$$\widehat{\mathcal{R}}\left[\mathcal{A}\right] + \lambda\|f\|_k^2 \leq \widehat{\mathcal{R}}\left[h_\lambda + h_S\right] + \lambda\|h_\lambda\|_k^2,$$

we obtain:

$$\begin{aligned}\mathcal{R}\left[\mathcal{A}\right] - \mathcal{R}[h^* + h_S] \leq \mathcal{R}\left[\mathcal{A}\right] - \widehat{\mathcal{R}}\left[\mathcal{A}\right] + \widehat{\mathcal{R}}\left[h_\lambda + h_S\right] - \mathcal{R}[h_\lambda + h_S] \\ + \mathcal{R}[h_\lambda + h_S] + \lambda\|h_\lambda\|_k^2 - \mathcal{R}[h^* + h_S].\end{aligned}$$

For the first term notice that:

$$\mathbb{E}\left[\mathcal{R}\left[\mathcal{A}\right] - \widehat{\mathcal{R}}\left[\mathcal{A}\right]\right] \leq \mathcal{E}_{\text{gen}} \leq \beta(n) + \gamma(n).$$

Regarding the second term, since h_λ is independent of $\mathcal{D}_T$ we have:

$$\mathbb{E}\left[\widehat{\mathcal{R}}\left[h_\lambda + h_S\right] - \mathcal{R}[h_\lambda + h_S]\right] = 0.$$

Finally, notice that by definition of g_λ that:

$$\mathcal{R}[h_\lambda + h_S] + \lambda\|h_\lambda\|_k^2 - \mathcal{R}[h^* + h_S] \leq \lambda\|h^*\|_k^2.$$

Combining the latter four inequalities yields:

$$\mathcal{E}_{\text{ex}} = \mathcal{R}\left[\mathcal{A}\right] - \mathcal{R}[h^* + h_S] \leq \beta(n) + \gamma(n) + \lambda\|h^*\|_k, \tag{5.16}$$

which concludes the first part. For the second part we shall use Table 5.1 and the fact that

$$\gamma(n) + \beta(n) \leq \alpha\frac{\left(\Psi_1\left(\mathcal{R}\left[h_S\right]\right) + \Psi_2\left(\mathcal{R}\left[h_S\right]\right)\right) \wedge \left(2\|\phi'\|_\infty^2\right)}{n}. \tag{5.17}$$

MSE and Squared hinge

For these two losses, $\Psi_1(x) = \Psi_2(x) = 8x(4\alpha + 1)$, so that by inequality (5.17) we get:

$$\begin{aligned}\gamma(n) + \beta(n) &\leq \alpha\frac{16\mathcal{R}\left[h_S\right](4\alpha + 1)}{n} \\ &= \frac{16\kappa\mathcal{R}\left[h_S\right](4\frac{\kappa}{\lambda} + 1)}{\lambda n}.\end{aligned}$$

Thus for small λ one has:

$$\gamma(n) + \beta(n) = \mathcal{O}\left(\frac{\mathcal{R}[h_S]}{\lambda^2 n}\right).$$

To conclude, set $\lambda = \sqrt{\frac{\mathcal{R}[h_S]}{\sqrt{n}}}$ and use Inequality (5.16) to obtain:

$$\mathcal{E}_{\text{ex}} = \mathcal{O}\left(\sqrt{\frac{\mathcal{R}[h_S]}{\sqrt{n}}}\right).$$

Exponential

Using Table 5.1, remind that the functions $\Psi_1(x)$ and Ψ_2 are given by: $\Psi_1(x) = C_S x^2 e^{2\alpha x}$, $\Psi_2(x) = C_S M_S x e^{2\alpha x}$ with $M_S = \sup_{z\in\mathcal{Z}_T} \ell(h_S, z)$ and

$$C_S = \exp\left\{2 + \frac{2\alpha M_S}{n} + \frac{4\alpha^2 M_S^2}{n-1}\right\} = \exp\left\{2 + \frac{2\kappa M_S}{\lambda n} + \frac{4\kappa^2 M_S^2}{\lambda^2(n-1)}\right\}.$$

Assume that $n \geq \max\left(\frac{M_S^2 \ln(n)^2}{\mathcal{R}[h_S]}, 2\right)$ and $\lambda = 4\frac{\sqrt{\mathcal{R}[h_S]}\wedge 1}{\ln(n)} = 4\frac{\sqrt{\mathcal{R}[h_S]}}{\ln(n)}$. The case where $\mathcal{R}[h_S] \geq 1$ is similar and thus omitted. Now, write

$$n \geq \frac{M_S^2 \ln(n)^2}{\mathcal{R}[h_S]} = \frac{M_S^2}{\lambda^2} \implies \frac{M_S^2}{\lambda^2(n-1)} \leq \frac{n}{n-1} \leq 2.$$

The latter condition also implies that $\frac{M_S}{\lambda n} \leq \frac{\lambda}{M_S} \leq \frac{\sqrt{\mathcal{R}[h_S]}}{M_S \ln(n)} \leq \frac{1}{2M_S}$. By these two facts, we deduce that C_S can be bounded independently of n. Thus, using (5.17) yields:

$$\begin{aligned}\gamma(n) + \beta(n) &\leq \alpha \frac{C_S\left(\mathcal{R}[h_S]^2 + M_S\mathcal{R}[h_S]\right)e^{2\alpha\mathcal{R}[h_S]}}{n} \\ &= \ln(n)\frac{C_S\left(\mathcal{R}[h_S]^{3/2} + M_S\mathcal{R}[h_S]^{1/2}\right)(\sqrt{n})^{\kappa\sqrt{\mathcal{R}[h_S]}}}{n} \\ &= \mathcal{O}\left(\frac{\sqrt{\mathcal{R}[h_S]}}{\sqrt{n}}\right),\end{aligned} \tag{5.18}$$

where the two last inequalities follow from the facts that $\alpha = \frac{\kappa}{\lambda} = \frac{\kappa\ln(n)}{\sqrt{\mathcal{R}[h_S]}}$ and $\kappa \leq 1$. It remains to use the (5.16) to conclude the first part. For the second part, set $\lambda = \frac{\ln(n)^2}{\sqrt{n}}$ and notice that, if $n \leq \frac{M_S^2 \ln(n)^2}{\mathcal{R}[h_S]}$ then $\mathcal{R}[h_S] \leq \frac{M_S^2 \ln(n)^2}{n} \leq M_S^2\lambda$ and $\alpha\mathcal{R}[h_s] \leq M_S^2\kappa \leq 1$. Furthermore, the constant C_S can be bounded independently of n with such a choice of λ. Inequality (5.18) becomes:

$$\gamma(n) + \beta(n) = \mathcal{O}\left(\frac{\mathcal{R}[h_S]}{\sqrt{n}\ln(n)^2}\right).$$

It remains to use Inequality (5.16) to complete the proof.

Logistic

For this loss, we have $\|\phi'\|_\infty = 1$ and Inequality (5.17) becomes:

$$\beta(n) + \gamma(n) \leq \frac{2\alpha}{n} = \frac{2\kappa}{\lambda n} \leq \frac{2}{\lambda n}.$$

Thus, setting $\lambda = \frac{1}{\sqrt{n}}$ and using Inequality (5.16) yields:

$$\mathcal{E}_{\text{ex}} = \mathcal{O}\left(\sqrt{\frac{1}{n}}\right).$$

Furthermore, if $n \geq 9$ and $\mathcal{R}[h_S] \leq \frac{1}{\sqrt{n}} \leq \frac{1}{e}$, then with the choice $\lambda = \frac{8}{\sqrt{-n\ln(\mathcal{R}[h_S])}}$ one has:

$$\mathcal{E}_{\text{ex}} = \mathcal{O}\left(\frac{1}{\sqrt{-n\ln\left(\mathcal{R}[h_S]\right)}}\right).$$

Indeed, in the setting above, it leads to that:

$$\begin{aligned} e^{\alpha\mathcal{R}[h_S]} = e^{\frac{\kappa}{\lambda}\mathcal{R}[h_S]} &= e^{\frac{\kappa\sqrt{-\ln(\mathcal{R}[h_S])}}{8}} \\ (\kappa \leq 1) &\leq e^{\frac{\sqrt{-\ln(\mathcal{R}[h_S])}}{8}} \\ \left(-\ln(\mathcal{R}[h_S]) \geq 1\right) &\leq e^{\frac{-\ln(\mathcal{R}[h_S])}{8}} = \mathcal{R}[h_S]^{-1/8}, \end{aligned}$$

and

$$e^{\frac{2\kappa M_S}{\lambda n}} \leq e^{\frac{2}{\lambda n}} = e^{-\frac{\sqrt{\ln(\mathcal{R}[h_S])}}{4\sqrt{n}}} \leq e^{\sqrt{\frac{\ln(n)}{4n}}} \leq e^{1/4}.$$

Besides, since $\frac{n}{n-1} \leq 2$,

$$e^{\frac{4\kappa M_S}{\lambda^2 n}} \leq e^{\frac{4}{\lambda^2(n-1)}} = e^{-\frac{\ln(\mathcal{R}[h_S])(n)}{16(n-1)}} \leq \mathcal{R}[h_S]^{-1/8}.$$

Moreover, using Inequality (5.17) and Table 5.1 gives:

$$\begin{aligned} \gamma(n) + \beta(n) &\leq \alpha\frac{C_S e^{\sqrt{\mathcal{R}[h_S]}} e^{2\alpha\mathcal{R}[h_S]}\left(e^{\sqrt{\mathcal{R}[h_S]}} - 1\right)}{n} \\ &= \kappa\frac{\exp\left\{2 + \frac{2\alpha M_S}{n} + \frac{4\alpha^2 M_S^2}{n-1} + \sqrt{\mathcal{R}[h_S]} + 2\alpha\mathcal{R}[h_S]\right\}\left(e^{\sqrt{\mathcal{R}[h_S]}} - 1\right)}{\lambda n} \\ &= \mathcal{O}\left(\frac{\sqrt{-\ln\left(\mathcal{R}[h_S]\right)}\left(e^{\sqrt{\mathcal{R}[h_S]}} - 1\right)}{\mathcal{R}[h_S]^{1/4}\sqrt{n}}\right). \end{aligned}$$

Now, since the function $e^{\sqrt{x}} - 1 \leq 2\sqrt{x}$ for all $x \leq \ln(2)^2$ and $\mathcal{R}[h_S] \leq \frac{1}{n} \leq \frac{1}{3} \leq \ln(2)^2$ the latter inequality becomes:

$$\gamma(n) + \beta(n) = \mathcal{O}\left(\frac{\sqrt{-\ln\left(\mathcal{R}[h_S]\right)}\mathcal{R}[h_S]^{1/4}}{\sqrt{n}}\right).$$

To conclude the proof notice that, for all $x \leq 1$, we have $\ln(x^{-1/4}) \leq x^{-1/4}$ and thus $x^{1/4}(-\ln(x)) \leq 4$. This leads to :

$$x^{1/4}\sqrt{-\ln(x)} \leq \frac{4}{\sqrt{-\ln(x)}}.$$

Therefore,

$$\gamma(n) + \beta(n) = \mathcal{O}\left(\frac{1}{\sqrt{-n\ln\left(\mathcal{R}[h_S]\right)}}\right).$$

It remains to use Inequality (5.16) to complete the proof.

Softplus

For the softplus, the choice $\lambda = 1/\sqrt{n}$ yields:

$$\mathcal{E}_{\text{ex}} = \mathcal{O}\left(\sqrt{\frac{1}{n}}\right).$$

Furthermore, if $n \geq 9$ and $\mathcal{R}[h_S] \leq \frac{1}{\sqrt{n}}$ and $\frac{1}{s} \leq -\ln(\mathcal{R}[h_S])$, then with the choice $\lambda = \frac{8}{\sqrt{-n\ln(\mathcal{R}[h_S])}}$, one has:

$$\mathcal{E}_{\text{ex}} = \mathcal{O}\left(\frac{1}{\sqrt{-sn\ln\left(\mathcal{R}[h_S]\right)}}\right).$$

The proof is identical to the previous one and thus omitted.

Chapter 6

On the bias of K-fold cross validation with stable learners

Contents

6.1 Introduction

Introduced in Stone (1974), cross-validation (CV) is a popular tool in statistics for estimating the generalization risk of a learning algorithm. It is also the mainstream approach for model and parameter selection. Despite its widespread use, it has been shown in several contexts that CV schemes fail to select the correct model unless the test fraction is negligible in front of the sample size. Unfortunately, this excludes the widely used K-fold CV. This suboptimality has been pinned in the linear regression framework by Burman (1989); Shao (1997); Yang (2007), then in other specific frameworks such as density estimation (Arlot, 2008a) and classification (Yang, 2006). The theoretical properties of CV procedures for model selection in wider settings are notoriously difficult to establish and remain the subject of active research (Bayle et al., 2020; Wager, 2020).

To tackle the suboptimality of K-fold, Burman (1989); Fushiki (2011) have proposed to add some debiasing correction terms to the K-fold CV estimate in order to improve the convergence rate. However, the analysis conducted in these works is purely asymptotic and focuses only on ordinary linear regression. More recently Arlot and Lerasle (2016) conduct a non asymptotic study for the bias corrected K-fold in the density estimation framework and show the benefits of such a correction. Nonetheless, the latter study relies on closed-form formulas for risk estimates which are valid only for histogram rules. To summarize, the statistical consistency of the debiased version has been established only in specific frameworks pertaining to classical statistics. In addition, to our best knowledge, the consistency of K-fold without correction has not been proved nor disproved in the existing literature.

The main purpose of this chapter is to establish non-asymptotic results (upper and lower bounds) regarding the error of the K-fold risk estimate, under realistic assumptions which are valid for a wide class of modern algorithms (regularized empirical risk minimization, neural networks, bagging, SGD, *etc.*...), namely an *algorithmic stability* assumption discussed below. In other words, the question we seek to answer is as follows:

- Is K-fold cross-validation consistent under algorithmic stability assumptions? If not, how about the bias corrected K-fold?

The notion of *algorithmic stability* and its consequences in learning theory has received much attention since its introduction in Devroye and Wagner (1979). This property allows to obtain generalization bounds for a large class of a learning algorithms such as k-nearest-neighbors (Devroye and Wagner, 1979), empirical risk minimizers (Kearns and Ron, 1999), regularization networks (Bousquet and Elisseeff, 2001), bagging (Elisseeff et al., 2005) to name but a few. For an exhaustive review of the different notions of *stability* and their consequences on the generalization risk of a learning algorithms, the reader is referred to Kutin and Niyogi (2002). Our working assumption in this chapter is *uniform stability*, which encompasses many algorithms such as Support Vector Machine (Bousquet and Elisseeff, 2002), regularized empirical risk minimization (Zhang, 2004a; Wibisono et al., 2009a), stochastic gradient descent (Hardt et al., 2016a) and neural networks with a simple architecture (Charles and Papailiopoulos, 2018).

Related Work on K-fold CV with Stable Learners. In Kale et al. (2011); Kumar et al. (2013), K-fold CV for risk estimation is envisioned under stability assumptions regarding the algorithm. It is shown that the K-fold risk estimate has a much smaller variance than the simple hold-out estimate and the amount of variance reduction is quantified. Another related work is Abou-Moustafa and Szepesvári (2017) who builds upon a variant of algorithmic stability, namely L^q*-stability* to derive PAC upper bounds for K-fold CV error estimates. Other results regarding the asymptotic behavior of K-fold estimates can be found in Austern and Zhou (2020a); Bayle et al. (2020).

However, none of the results mentioned above imply a universal upper bound regarding the K-fold neither for risk estimation nor for model selection. Indeed their focus is on the the *variance term* of the K-fold error, while they do not take into account the high *bias* generally induced by this CV scheme (see Shao (1997); Arlot and Lerasle (2016) for instance). To our best knowledge, the literature on algorithmic stability is silent about the consistency of K-fold CV – the most widely used CV scheme – in a generic stability setting. Filling this gap is the main purpose of the present chapter.

Contributions and Outline. We introduce the necessary background and notations about CV risk estimation and algorithmic stability in Section 6.2. Section 6.3 is intended to give some context about provable guarantees regarding K-fold CV scheme, namely we state and prove a generic upper bound on the error of the generalization risk estimate for uniformly stable algorithms. However, with realistic stability constants, the obtained upper bound is not satisfactory, in so far as it does not vanish as the sample size $n \to \infty$.

Our main contributions are gathered in sections 6.4 to 6.6 and may be summarized as follows:

1. One may wonder whether the looseness of the bound for the K-fold CV error is just an artifact from our proof. We answer in the negative by deriving a lower bound on the K-fold error (Section 6.4) in two different contexts, specifically, regularized empirical risk minimization and stochastic gradient optimization. The latter bound shows that under the uniform stability assumption alone, K-fold CV is inefficient in so far as it can fail in estimating the generalization risk of a uniformly stable algorithm.

2. We analyze a corrected K-fold procedure and prove a PAC generalization upper bound covering the general case of uniformly stable learners. As a consequence, the corrected version of the K-fold is shown to be efficient in contrast to the standard version. The corrected K-fold scheme has been investigated in Burman (1989, 1990); Fushiki (2011); Arlot and Lerasle (2016) in the particular frameworks of ordinary linear regression and density estimation. Furthermore, the analysis in the latter references relies on strong regularity assumptions (further details are given in Section 6.5) which aren't satisfied by many modern learning rules like Support Vector Machine (SVM), stochastic gradient descent methods, bagging, etc. Instead our upper bound covers the general case of uniformly stable learners. As an example of application, we show that the debiased K-fold permits to select a model within a finite collection in a risk consistent manner (Section 6.6). In other words, the excess risk of the selected model tends to 0 as $n \to \infty$. Finally we demonstrate empirically the added value of the debiased K-fold compared with the standard one in terms of the test error of the selected model.

6.2 Background, notations and working assumptions

6.2.1 Notations

We place ourselves in the following general learning setting. One receives a collection of independent and identically distributed random vectors $\mathcal{D} = (O_1, \ldots, O_n)$ lying in a sample space $\mathcal{Z}$, with common distribution P. For any $n \in \mathbb{N}$, let $[n]$ denote the set of integers $\{1, 2, \ldots, n\}$. Consider a class of predictors $\mathcal{G}$ and a loss function $\ell : \mathcal{G} \times \mathcal{Z} \to \mathbb{R}$, so that $\ell(g, O)$ be the error of g on the observation $O \in \mathcal{Z}$. As an example, in the supervised learning setting $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, g is a mapping $\mathcal{X} \to \mathcal{Y}$ and for $o = (x, y)$ the loss function writes as $\ell(g, o) = \ell(g(x), y)$. However our results are not limited to the supervised setting. Given a subsample $\mathcal{D}_T = \{O_i \mid i \in T\}$ indexed by $T \subset [n]$ and an algorithm (or learning rule) $\mathcal{A}$, we denote by $\mathcal{A}(T) \in \mathcal{G}$ the predictor obtained by training $\mathcal{A}$ on $\mathcal{D}_T$. We consider in this chapter deterministic algorithms, that is, given a subsample $\mathcal{D}_T$, the output of the algorithm $\mathcal{A}(T)$ is non random. We thereby neglect the randomness brought *e.g.* by optimization routines. The case of random algorithms can be covered at the price of additional notational burden. For the sake of readability we restrict ourselves to deterministic algorithms in the main chapter and show how to relax it in the supplementary (Section 6.D), in order to cover the case of random algorithms such as stochastic gradient descent (SGD) or neural networks. This extension is in particular necessary to one of our counter-examples (Section 6.8) where we build a lower bound for the SGD algorithm.

The generalization risk of the predictor $\mathcal{A}(T)$ is then

$$\mathcal{R}\Big[\mathcal{A}(T)\Big] = \mathbb{E}\left[\ell\left(\mathcal{A}\left(T\right), O\right) \mid \mathcal{D}_T\right],$$

where O is independent from $\mathcal{D}_T$. Notice that the randomness in the latter expectation stems from the novel observation O only while the trained algorithm $\mathcal{A}(T)$ is fixed. The quantity of interest here is the generalization risk of the learning rule trained on the full dataset, $\mathcal{R}\Big[\mathcal{A}([n])\Big]$. The hold-out estimate of the latter involves a validation index set V disjoint from T and writes as

$$\widehat{\mathcal{R}}\Big[\mathcal{A}(T), V\Big] = \frac{1}{n_V}\sum_{i\in V}\ell(\mathcal{A}(T), O_i),$$

where $n_V = card(V)$.
Given a family of validation sets in $[n]$, $V_{1:K} = (V_j)_{j=1,\dots,K}$, the K-fold CV estimator of the generalization risk of $\mathcal{A}([n])$ is

$$\widehat{\mathcal{R}}_{\mathrm{CV}}\left[\mathcal{A}, V_{1:K}\right] = \frac{1}{K}\sum_{j=1}^{K}\widehat{\mathcal{R}}\Big[\mathcal{A}(T_j), V_j\Big], \tag{6.1}$$

where $T_j = [n]\backslash V_j$. For clarity reasons, we suppose further that n is divisible by K so that n/K is an integer. This condition guarantees, that all validation sets have the same cardinal $n_V = n/K$.

6.2.2 Algorithmic Stability

An algorithm $\mathcal{A}$ is called stable if removing a training point O_i from $\mathcal{D}_T$ ($i \in T$) or replacing O_i with an independent observation O' drawn from the same distribution does not change much the risk of the output. Formally, for $i \in T \subset [n]$ as above, let $T^{\backslash i} = T \setminus \{i\}$, so that $\mathcal{A}(T^{\backslash i})$ is the output of $\mathcal{A}$ trained on $\mathcal{D}_T \setminus \{O_i\}$. Denote similarly $\mathcal{A}(T^i)$ the output of $\mathcal{A}$ trained on $\mathcal{D}_T \setminus \{O_i\} \cup \{O'\}$. The notion of *hypothesis stability* was first introduced in Devroye and Wagner (1979) to derive non asymptotic guarantees for the leave-one-out CV (*l.o.o.*). In this chapter, we consider instead *uniform stability*, an assumption used in Bousquet and Elisseeff (2002); Wibisono et al. (2009a); Hardt et al. (2016a); Feldman and Vondrak (2019) to derive probability upper bounds for the training error and *l.o.o.* estimates. With the above notations, uniform stability is defined as follows.

Definition 6.1. *An algorithm $\mathcal{A}$ is said to be $(\beta_t)_{1\le t\le n}$ uniformly stable with respect to a loss function ℓ if, for any $T \subset [n]$, $i \in T$ it holds that*

$$\left|\ell\left(\mathcal{A}(T), O\right) - \ell\left(\mathcal{A}(T^{\backslash i}), O\right)\right| \le \beta_{n_T}, \tag{6.2}$$

with P-probability one.

Many widely used Machine Learning algorithms are uniformly stable in the sense of Definition 6.1. In particular $\beta_n \le \frac{C}{n}$ for SVM and least square regression with the usual mean squared error, and for SVM classification with the soft margin loss (Bousquet and

Elisseeff, 2002). Up to a minor definition of uniform stability accounting for randomness (see Section 6.D in the supplement), many extensively used stochastic gradient methods are also uniformly stable, such as *e.g.* SGD with convex and non convex losses (Hardt et al., 2016a) or RGD (randomized coordinate descent) and SVRG (stochastic variance reduced gradient method) with loss functions verifying the Polyak-Łojasiewicz condition (Charles and Papailiopoulos, 2018).

The following simple fact concerns the effect of removing n' training points on uniformly stable algorithms.

Fact 6.2.1. *Let $\mathcal{A}$ be a decision rule which is $(\beta_t)_{1\leq t\leq n}$ uniformly stable, additionally suppose that the sequence $(\beta_t)_{1\leq t\leq n}$ is decreasing, then for any $T \subset [n]$, one has,*

$$\left|\ell\left(\mathcal{A}\left([n]\right), O\right) - \ell\Big(\mathcal{A}(T), O\Big)\right| \leq \sum_{i=n_T+1}^{n} \beta_i.$$

Remark 6.2. *Definition 6.1 and Fact 6.2.1 play a key role in our proofs. Namely we use Fact 6.2.1 to control the bias of the CV risk estimate and Definition 6.1 to derive a probability upper bound on its deviations via McDiarmid's inequality.*

We also rely on the fact that the training and validation sets of K-fold CV verify a certain balance condition.

Fact 6.2.2. *For the K-fold* CV *the validation sets $V_1, V_2, \dots V_K$ satisfies*

$$card(V_j) = n_V \quad \forall j \in [\![1, K]\!], \tag{6.3}$$

for some $n_V \in [\![1, n]\!]$. Moreover it holds that

$$\frac{1}{K}\sum_{j=1}^{K} \mathbb{1}\Big\{l \in V_j\Big\} = \frac{n_V}{n} \quad \forall l \in [n]. \tag{6.4}$$

Because $T_j = [n]\backslash V_j$, if (6.4) holds, then the training sets T_j verify a similar equation, that is,

$$\frac{1}{K}\sum_{j=1}^{K} \mathbb{1}\Big\{l \in T_j\Big\} = \frac{n_T}{n} \quad \forall l \in [n].$$

We prove Fact 6.2.2 in the supplement (Lemma 6.14).

Throughout this chapter we work under the following uniform stability assumption combined with a boundedness assumption regarding the cost function.

Assumption 11 (Stable algorithm)**.** *The algorithm $\mathcal{A}$ is $(\beta_t)_{1\leq t\leq n}$ uniformly stable with respect to a cost function ℓ that satisfies*

$$\forall O \in \mathcal{Z}\,, \forall T \subset [n], \left|\ell\left(\mathcal{A}\left(T\right), O\right)\right| \leq L.$$

6.3 Upper bounds for K-fold risk estimation

Our first result Theorem 6.3 is a generic upper bound on the error of the generalization risk estimate for stable algorithms satisfying Assumption 11. Our upper bound is of the same order of magnitude as existing results in the literature which apply to other

contexts, *e.g.* in Bousquet and Elisseeff (2002) for *l.o.o.* CV or in Cornec (2017) for the specific case empirical risk minimizers, although our techniques of proof are different. The fact that the upper bound does not vanish with large sample sizes n with realistic stability constants (Corollary 6.4) gives some context and motivates the rest of this work.

Theorem 6.3. *Consider a stable learning algorithm $\mathcal{A}$ satisfying Assumption 11. Then, we have with probability $1-2\delta$,*

$$\left|\widehat{\mathcal{R}}_{\mathrm{CV}}\left[\mathcal{A}, V_{1:K}\right] - \mathcal{R}\Big[\mathcal{A}([n])\Big]\right| \leq \sum_{i=n_T+1}^{n} \beta_i + (4\beta_{n_T} n_T + 2L)\sqrt{\frac{\log(1/\delta)}{2n}}.$$

Where L is the upper bound on the cost function ℓ from Assumption 11.

Proof [Sketch of proof] Define the average risk of the family $\left(\mathcal{A}(\mathcal{D}_{T_j})\right)_{1\leq j\leq K}$

$$\mathcal{R}_{CV}\left[\mathcal{A}, V_{1:K}\right] = \frac{1}{K}\sum_{j=1}^{K} \mathcal{R}\Big[\mathcal{A}(T_j)\Big], \tag{6.5}$$

then write the following decomposition

$$\widehat{\mathcal{R}}_{\mathrm{CV}}\left[\mathcal{A}, V_{1:K}\right] - \mathcal{R}\Big[\mathcal{A}([n])\Big] = \mathrm{D}_{\mathrm{cv}} + \mathrm{Bias}, \tag{6.6}$$

with

$$\mathrm{D}_{\mathrm{cv}} = \widehat{\mathcal{R}}_{\mathrm{CV}}\left[\mathcal{A}, V_{1:K}\right] - \mathcal{R}_{CV}\left[\mathcal{A}, V_{1:K}\right], \tag{6.7}$$

$$\mathrm{Bias} = \mathcal{R}_{CV}\left[\mathcal{A}, V_{1:K}\right] - \mathcal{R}\Big[\mathcal{A}([n])\Big]. \tag{6.8}$$

The proof consists in bounding each term of the above decomposition independently. The term D_{cv} measure the deviations of $\widehat{\mathcal{R}}_{\mathrm{CV}}$ from its mean and it can be controlled using McDiarmid's inequality (Proposition 6.A.1). The second term Bias is controlled using Fact 6.2.1. The detailed proof is deferred to the appendix. ■

As discussed in the background section 6.2.2, typical uniform stability constants β_n for standard algorithms satisfy $\beta_n \leq \frac{C}{n}$. In this case, Theorem 6.3 yields the following corollary.

Corollary 6.4. *Consider a stable learning algorithm $\mathcal{A}$ satisfying Assumption 11 with stability parameter $\beta_n \leq \frac{C}{n}$. Then, we have with probability $1-2\delta$,*

$$\left|\widehat{\mathcal{R}}_{\mathrm{CV}}\left[\mathcal{A}, V_{1:K}\right] - \mathcal{R}\Big[\mathcal{A}([n])\Big]\right| \leq C\log\left(\frac{K}{K-1}\right) + (4C + 2L)\sqrt{\frac{\log(1/\delta)}{2n}},$$

wich does not converge to 0 as $n \to \infty$.

Proof Recall that $\frac{n}{n_T} = \frac{K}{K-1}$ and write

$$\sum_{i=n_T+1}^{n} \beta_i \leq C \sum_{i=n_T+1}^{n} \frac{1}{i} \leq C \log\left(\frac{K}{K-1}\right).$$

The result follows from Theorem 6.3. ■

Remark 6.5 (Bias and Variance of K-fold CV)**.** *The two terms of the sum in the upper bound of Theorem 6.3 correspond respectively to a variance and a bias term, namely* D_{cv} *and* Bias *in the error decomposition* (6.6)*.*

On the one hand, the term $(4\beta_{n_T} n_T + 2L)\sqrt{\frac{\log(1/\delta)}{2n}}$ *reflects the variance of the* CV *procedure. When* $\beta_n \leq \frac{C}{n}$ *it yields the usual rate* $1/\sqrt{n}$*. On the other hand, the term* $\sum_{i=n_T+1}^{n} \beta_i$ *reflects the bias of K-fold* CV*. Contrarily to the variance term, it does not vanish as* $n \to \infty$*, even when* $\beta_n \leq C/n$*. Finally, Notice that, as the number of folds* K *increases, the training size* n_T *gets closer to the sample size* n *and the bias of K-fold vanishes. However, for computational efficiency, increasing the number of folds is not always desirable.*

One may wonder whether the looseness of the bias term $\sum_{i=n_T+1}^{n} \beta_i$ is just an artifact from our proof. In the next theorem we answer in the negative by deriving a lower bound for the K-fold. The latter bound shows that under the uniform stability assumption alone, K-fold CV is inefficient in so far as it can fail in estimating the generalization risk of a uniformly stable algorithm.

6.4 Lower bound for the K-fold error under algorithmic stability

To construct lower bounds on K-fold CV error, we consider two families of algorithms that satisfy the uniform stability hypothesis with parameter β_n scaling as $1/n$. Namely, regularized empirical risk minimizers (RERM) and stochastic gradient descent (SGD).

6.4.1 Regularized Empirical Risk Minimization

The first counter-example that we build to prove a lower bound is formulated in a regression framework. In particular, we consider L^2-regularized empirical risk minimization (L^2-RERM)

$$\mathcal{A} = \operatorname*{arg\,min}_{g \in \mathcal{G}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \ell\left(g, O_i\right) + \lambda_n \left\|g\right\|_{\mathcal{G}}^2 \right\}, \tag{6.9}$$

where ℓ is a convex loss function and $\mathcal{G}$ is a hypothesis space.
Under some mild assumptions on ℓ and the input space $\mathcal{Z}$, an L^2-RERM algorithm is uniformly stable with $\beta_n \leq \frac{C}{n}$ (see *e.g.* Zhang (2004a); Wibisono et al. (2009a); Liu et al. (2017) for further details).
The next result confirms that uniform stability alone is not sufficient to ensure the consistency of K-fold CV for RERM.

Theorem 6.6 (Non vanishing lower bound on the K-fold error)**.** *Consider the regression problem for a random pair $O = (X, Y) \in \mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, loss function $\ell(g, o) = (g(x) - y)^2$ and hypothesis class $\mathcal{G}$ consisting of linear predictors . Set $M \geq 1$ and $n \in [\![2, e^M]\!]$. Then, there exists an input space $\mathcal{Z}$ with distribution P and a regularization parameter λ_n, such that the RERM algorithm $\mathcal{A}$ (Equation 6.9) satisfies Assumption 11 with $L = M$ and $\beta_n \leq 2/n$.*
Furthermore the K-fold CV *error satisfies,*

$$\mathbb{E}\left[\left|\widehat{\mathcal{R}}_{\mathrm{CV}}\left[\mathcal{A}, V_{1:K}\right] - \mathcal{R}\left[\mathcal{A}([n])\right]\right|\right] \geq 2\log\left(\frac{K}{K-1}\right)\left(1 - \frac{1}{M}\right),$$

and

$$\mathbb{E}\left[\left|\widehat{\mathcal{R}}_{\mathrm{CV}}\left[\mathcal{A}, V_{1:K}\right] - \mathcal{R}\left[\mathcal{A}([n])\right]\right|\right] \leq 2\log\left(\frac{K}{K-1}\right).$$

Proof [Sketch of proof] First set $\mathcal{Z} = \mathcal{X} \times \mathcal{Y} \subset \mathbb{R} \times \mathbb{R}$ and consider the hypothesis class of linear regressors,

$$\mathcal{G} = \left\{g_b \mid b \in \mathbb{R}\right\},$$

where $g_b(x) = bx$ for $x \in \mathbb{R}$. With the loss $\ell(g_b, o) = (y - bx)^2$, it can be shown that the solution to problem 6.9 writes as $\mathcal{A}\left([n]\right) = g_{b_n}$ with

$$b_n = f(\lambda_n), \tag{6.10}$$

for some decreasing function f. Now, by carefully picking $\mathcal{Z}$ and λ_n we construct a problem such that, for any $O \in \mathcal{Z}$ and $T \subset T' \subset [n]$

1. λ_n is decreasing.
2. $\ell\Big(\mathcal{A}\left(T\right), O\Big) \geq \ell\Big(\mathcal{A}\left(T'\right), O\Big)$.
3. $\ell\Big(\mathcal{A}(T^{\setminus i}), O\Big) - \ell\Big(\mathcal{A}\left(T\right), O\Big) \leq f\left(\lambda_n\right) - f\left(\lambda_{n-1}\right)$.
4. $f(\lambda_n) - f(\lambda_{n-1}) \leq \frac{2}{n}$.

The result easily follows from the three latter facts. For the detailed proof, the reader is deferred to Appendix 6.C.2. ■

6.4.2 Stochastic Gradient Descent

In this section, we consider the SGD update rule

$$\forall t \geq 0\,,\ \mathcal{A}_{t+1} = \mathcal{A}_t - \alpha_{t,n}\nabla_{\mathcal{A}}\ell\left(\mathcal{A}_t, X_{i_t}\right), \tag{6.11}$$

where $\nabla_{\mathcal{A}}\ell$ denotes the derivative of ℓ with respect to the first argument, i_t is the index picked by SGD at step t and $\alpha_{t,n} \geq 0$ is the step size.
It is well known that SGD verifies the uniform stability assumption with respect to many losses (Hardt et al., 2016a; Liu et al., 2017; Charles and Papailiopoulos, 2018). For instance, when the loss function is convex, β smooth, and σ-Lipschitz, it has been shown that SGD algorithm (Hardt et al., 2016a) is uniformly stable with parameter $\beta_n \leq \frac{2\sigma^2}{n}\sum_{k=1}^{t}\alpha_{k,n}$.

Remark 6.7. *For a fixed data sequence $\mathcal{D}$, the output of SGD is random. Hence, the definition of* random uniform stability *is introduced in Appendix 6.D and is slightly different than Definition 6.1. However, most if not all the properties of deterministic uniformly stable rules (discussed before) are preserved by random uniformly stable learners (see* e.g. *Elisseeff et al. (2005); Hardt et al. (2016a); Liu et al. (2017)).*

Theorem 6.8. *Let $M > 1$ be a real number , $2 \leq n \leq e^M$ an integer, and $t \geq 1$ a maximum number of iterations . Set the initialization $\mathcal{A}_0 = 0$. Then, There exists an input space $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, a convex loss function ℓ and a sequence of step sizes $(\alpha_{k,n})_{k\leq t}$ such as,*

$$\mathbb{E}\left[\left|\widehat{\mathcal{R}}_{\mathrm{CV}}\left[\mathcal{A}_t, V_{1:K}\right] - \mathcal{R}\left[\mathcal{A}_t([n])\right]\right|\right] \geq \frac{1}{3}\log\left(\frac{K}{K-1}\right).$$

Furthermore, $\mathcal{A}_t$ fulfills Assumption 11 with $L = M$ and

$$\beta_n \leq \frac{3}{n-1}\sum_{k=1}^{t}\alpha_{k,n} \leq \frac{3M}{n-1}.$$

Proof The proof is deferred to the appendix (section 6.D.2). ■

Remark 6.9. *In both examples, we suppose that the input is a binary random variable. Such a restrictive setting serves as a corner case to derive lower bounds (Bousquet et al., 2020; Zhang et al., 2022), suggesting the necessity of additional assumptions to ensure the consistency of K-fold scheme.*

Remark 6.10 (BOUNDEDNESS OF n)**.** *With the current assumptions, the lower bounds don't ensure the inconsistency as the sample size n grows to infinity. However, Theorems 6.6 and 6.8 in their current state prove that, one cannot obtain a standard universal vanishing upper bound on the expected estimation error of K-fold ($\mathbb{E}[\mathrm{error}_{CV}]$) which would be valid for all sample spaces ($\mathcal{Z}$), all distributions (P) and all stable algorithms. In other words, one cannot construct a function $h(n)$ such that*

$$\exists n_0 \in \mathbb{N}\,,\ \forall n \geq n_0\,,\ \forall(\mathcal{Z}, P),$$
$$\forall \mathcal{A} \text{ stable with parameter } \beta_n \leq \frac{1}{n}\,, \mathbb{E}\left[\mathrm{error}_{CV}\right] \leq h(n),$$

and $h(n) \xrightarrow[n\to\infty]{} 0$.
An additional remark that ought to be made: if we relax Assumption 11 into : "the loss function $\ell(\mathcal{A}\left([n]\right),\cdot)$ is $L\log(n)$-bounded and the algorithm $\mathcal{A}$ is $\frac{\log(n)}{n}$ uniformly stable", then all the upper bounds from the present paper remain valid (up to a $\log(n)$

factor), in particular the upper bound on the bias-corrected K-fold estimation error (Corollary 6.12 below). Furthermore, under this new assumption, the lower bound in Theorem 6.8 is valid $\forall n \in \mathbb{N}$*, which yields the inconsistency of K-fold CV. More precisely, Theorem 6.8 becomes*

$$\exists (\mathcal{Z}, P), \forall n_0 \in \mathbb{N}\,,\ \forall n \geq n_0\,,$$
$$\exists \mathcal{A} \text{ stable with parameter } \beta_n \leq \frac{\log(n)}{n},\ \mathbb{E}\left[\text{error}_{CV}\right] \geq \kappa,$$

with $\kappa = \frac{1}{3}\log\left(\frac{K}{K-1}\right)$*. One must note that the additional* $\log(n)$ *factor in the upper bound weakens the tightness of the lower bound, however this is compatible with existing lower bounds from the stability literature (Bousquet et al., 2020).*

Discussion. *Theorems 6.6 and 6.8 reveals the suboptimality of K-fold CV in the general stability framework for risk estimation. For the purpose of model selection, which is arguably a harder problem, existing works have shown the suboptimality of K-fold* CV*. For example, in a regression framework, under the linear model assumption, Yang (2007) (See also Yang (2006) for classification problems) has shown that the usual K-fold procedure may not select the best model. For efficient model selection, where the performance is measured in terms of generalization risk of the selected model, Arlot (2008a) shows (See Theorem 1 in the latter reference) that K-fold* CV *can be suboptimal,* i.e. *an example is provided where the risk ratio between the selected model and the optimal one is uniformly greater (for all* n*) than* $1 + \kappa$ *for some* $\kappa > 0$*. In Theorem 6.6 (*resp. *6.8) of the present paper we show a stronger result, in so far as we consider the easier task of risk estimation, and we show that K-fold* CV *with fixed* K *does not enjoy sanity-check guarantees because for any* n*, there exists a regression (*resp. *optimization) problem where the error is at least* $2\log(\frac{K}{K-1})$ $\left(\text{resp. } \frac{1}{3}\log\frac{K}{K-1}\right)$*. It is worth mentioning at this stage that even if the uniform stability ensures the low variance of K-fold (Kumar et al. (2013); Bayle et al. (2020), etc.), it is not sufficient to control the bias.*

In the next section, we show that adding correction terms (Burman, 1989) to K-fold CV addresses the inconsistency issues underlined by Theorems 6.6 and 6.8. The resulting CV scheme enjoys both the computational efficiency of the K-fold (compared with the *l.o.o.*) and finite sample guarantees comparable to those of the *l.o.o.* in view of the upper bound stated in Theorem 6.11.

6.5 Bias corrected K-fold with stable learners

A key ingredient of the lack of guarantee regarding the K-fold risk estimate is its bias for finite sample sizes, an issue pointed out by Burman (1989) who proposes a corrected version of the standard K-fold with a reduced bias, see also Burman (1990) for applications to model selection. In the present work we follow in the footsteps of Burman (1989) and consider the same corrected CV estimate of the generalization risk

$$\begin{aligned}\widehat{\mathcal{R}}^{corr}_{CV}\left[\mathcal{A}, V_{1:K}\right] &= \widehat{\mathcal{R}}_{\text{CV}}\left[\mathcal{A}, V_{1:K}\right] \\ &+ \widehat{\mathcal{R}}\left[\mathcal{A}([n]), [n]\right] - \frac{1}{K}\sum_{j=1}^{K}\widehat{\mathcal{R}}\left[\mathcal{A}(T_j), [n]\right].\end{aligned} \tag{6.12}$$

The correcting term (second line in the above display) is the average difference between $\mathcal{R}\left[\mathcal{A}([n])\right]$ and the empirical risks $\mathcal{R}\left[\mathcal{A}(T_j)\right]$'s. In Burman (1989) the analysis is carried out in an asymptotic framework and focuses on the asymptotic bias and variance of the estimator for different CV schemes. The results are obtained under differentiability assumptions regarding the loss function (See also the appendix section in Fushiki (2011) where these assumptions are explicitly listed) which are typically not satisfied by SVM, stochastic gradient methods or L^1-regularized risk minimization algorithms. In Burman (1990) the working assumption is that of a linear model and the specific task is ordinary linear regression. In contrast, we conduct here a non asymptotic analysis (valid for any sample size) which applies to any uniformly stable algorithm.

Theorem 6.11. *Suppose that Assumption 11 holds. Then, we have, with probability* $1-6\delta$,

$$\left|\widehat{\mathcal{R}}_{CV}^{corr}\left[\mathcal{A}, V_{1:K}\right] - \mathcal{R}\left[\mathcal{A}([n])\right]\right| \leq 2\left(\beta_n + \beta_{n_T}\right) + 3(4\beta_{n_T} n_T + 2L)\sqrt{\frac{\log(1/\delta)}{2n}}.$$

Proof [Sketch of proof] By simple algebra write the corrected CV estimator as

$$\widehat{\mathcal{R}}_{CV}^{corr}\left[\mathcal{A}, V_{1:K}\right] = \widehat{\mathcal{R}}\left[\mathcal{A}([n]), [n]\right] + \frac{n_T}{nK}\sum_{j=1}^{K}\left[\widehat{\mathcal{R}}\left[\mathcal{A}(T_j), V_j\right] - \widehat{\mathcal{R}}\left[\mathcal{A}(T_j), T_j\right]\right].$$

From this, deduce the following error decomposition

$$\widehat{\mathcal{R}}_{CV}^{corr}\left[\mathcal{A}, V_{1:K}\right] - \mathcal{R}\left[\mathcal{A}[n]\right] = D_{\mathcal{A}[n]} + \frac{n_T}{n}(\mathrm{D_{cv}} - D_{\mathcal{A}[T_{1:K}]}), \tag{6.13}$$

where $\mathrm{D_{cv}}$ is defined in (6.7) and

$$D_{\mathcal{A}[n]} = \widehat{\mathcal{R}}\left[\mathcal{A}[n], [n]\right] - \mathcal{R}\left[\mathcal{A}[n]\right], \tag{6.14}$$

$$D_{\mathcal{A}[T_{1:K}]} = \frac{1}{K}\sum_{j=1}^{K}\left(\widehat{\mathcal{R}}\left[\mathcal{A}(T_j), T_j\right] - \mathcal{R}\left[\mathcal{A}(T_j)\right]\right). \tag{6.15}$$

Notice the absence of a bias term in the above display, contrarily to the error decomposition (6.6) for the standard CV estimate in the proof of Theorem 6.3. The remaining technical arguments for bounding each term of the decomposition above are gathered in the supplement. Namely, using McDiarmid's inequality (Proposition 6.A.1), we derive concentration bound both for $\mathrm{D_{cv}}$ as stated in Lemma 6.15 and for $D_{\mathcal{A}[T_{1:K}]}$ as given in Lemma 6.16. Finally the deviations of the empirical risk $D_{\mathcal{A}[n]}$ are controlled using Remark 6.17, a consequence of Lemma 6.16, in the supplement. ■

Theorem 6.11 yields immediately a consistent upper bound for the K-fold CV error when $\beta_n \leq C/n$ for some $C > 0$. For the proof it suffices to notice that $n_T = \frac{K-1}{K}n$ for the K-fold scheme and to use that $(2K-1)/(K-1) \leq 3$ whenever $K \geq 2$.

Corollary 6.12. *Suppose that Assumption 11 holds with $\beta_n \leq \frac{C}{n}$, for some $C > 0$. Then, for $K \geq 2$, the error of the corrected K-fold* CV *estimate of the generalization risk satisfies with probability at least $1 - 6\delta$,*

$$\left| \widehat{\mathcal{R}}_{CV}^{corr} \left[\mathcal{A}, V_{1:K} \right] - \mathcal{R}\Big[\mathcal{A}([n]) \Big] \right| \leq \frac{6C}{n} + 3(4C + 2L)\sqrt{\frac{\log(1/\delta)}{2n}}.$$

Discussion. *Corollary 6.12 confirms the relevance of the bias corrected K-fold for risk estimation in the broad context of uniformly stable learners satisfing $\beta_n \leq \frac{C}{n}$, such as SVM, stochastic gradient methods or regularized empirical risk minimizers.*

Our main results Theorem 6.11 and Corollary 6.12 concern the problem of risk estimation by means of K-fold CV. However in practice, CV is widely used in the context of model- or parameter *selection*. It is precisely the purpose of the next section to illustrate how our results shed light on such practice. Namely we consider the practical problem of selecting a penalty parameter among a finite collection of candidates for Support Vector Regression and Classification.

6.6 Application to hyper-parameter selection and numerical experiments

Selecting a penalty parameter in regularized empirical risk minimization algorithms such as SVM's may be viewed as a particular instance of a model selection problem, where one identifies a model with an algorithm equipped with a particular choice of regularization parameter. We start-off this section with a brief introduction to this topic. We provide minimal theoretical guarantees (Proposition 6.13) regarding model selection within a finite collection of models by means of the debiased K-fold procedure. We describe our experimental setting in Subsections 6.6.2, 6.6.3 and we report our results in Section 6.6.4.

6.6.1 Cross-Validation for Parameter Selection

Following the terminology of Arlot and Lerasle (2016), we consider here the problem of *efficient* model selection, aiming at selecting a model for which the generalization risk of the learnt predictor is close to the smallest possible risk. The goal of efficient model selection is in general easier to attain than *identification* of the best possible model. It is a known fact that cross-validation is in general sub-optimal for model *identification* purpose. A major reason for this is that different models (or algorithms) may have comparable performance. However if one aims only at selecting a model for which the generalization risk is close to that of the optimal one, lack of identifiability is not necessarily an issue anymore, which is precisely the approach we take here.

Model selection is a prominent topic in statistical learning theory which is by far too broad to be extensively covered here. We refer the reader to the monograph of Massart (2007). Cross-validation is one of several possible candidate methods for this problem,

which has been the subject of a wealth of literature as discussed in the introduction, see also Arlot and Celisse (2010) for a review.

Given a family of models (or algorithms) $\mathcal{A}^{(m)}$ indexed by $m \in \mathcal{M}$ and a dataset $\mathcal{D}$ of size n, an optimal model $\mathcal{A}^{(m^*)}$ called an oracle is any model such that

$$\mathcal{A}^{(m^*)} \in \arg\min_{m \in \mathcal{M}} \mathcal{R}\left[\mathcal{A}^{(m)}([n])\right]. \tag{6.16}$$

Since the true risk is unknown, an empirical criterion must be used instead to select an efficient model. One of the most popular tools for model selection is K-fold CV. However, as discussed earlier, the latter procedure may not be consistent. To tackle this issue, we propose to use the corrected K-fold and select a model $\mathcal{A}^{\widehat{m}}$ such that

$$\mathcal{A}^{(\widehat{m})} \in \arg\min_{m \in \mathcal{M}} \widehat{\mathcal{R}}^{corr}_{\text{Kfold}}\left[\mathcal{A}^{(m)}, V_{1:K}\right]. \tag{6.17}$$

Proposition 6.13 provides a sanity check guarantee regarding the consistency of the corrected K-fold CV procedure for this purpose, in the form of an upper bound in probability for the excess risk $\mathcal{R}\left[\mathcal{A}^{(\hat{m})}[n]\right] - \mathcal{R}\left[\mathcal{A}^{(m^*)}([n])\right]$. The result applies to the case where the family of models $\mathcal{M}$ is finite.

Proposition 6.13. *Let $(\mathcal{A}^{(m)})_{m \in \mathcal{M}}$ be a family of algorithms where each learner $\mathcal{A}^{(m)}$ is $(\beta_{m,t})_{1 \leq t \leq n}$ uniformly stable with respect to a loss function $0 \leq \ell(g, O) \leq L$. Additionally, assume that,*

$$\forall m \in \mathcal{M}\,;\, \beta_{m,t} \leq \frac{M}{t},$$

for some $M > 0$. Then one has, with probability at least $1 - 6\delta$,

$$\mathcal{R}\Big[\mathcal{A}^{(\hat{m})}([n])\Big] - \mathcal{R}\Big[\mathcal{A}^{(m^*)}([n])\Big] \leq \frac{12M}{n} + 6(4M + 2L)\sqrt{\frac{\log(|\mathcal{M}|/\delta)}{n}}.$$

Proof The proof consists in applying a union bound combined with the exponential tail bound of Theorem 6.11, which yields a multiplicative constant which only depends logarithmically on the number of models. The details are gathered in the supplementary material, Section 6.18. ■

Discussion (bounded stability parameters $\beta_{m,t}$'s)**.** *The assumption $\beta_{m,t} \leq \frac{M}{t}$ for all m is indeed verified in many applications. For example, in regularized SVM, where each learner $\mathcal{A}^{(m)}$ is trained using a regularization parameter λ_m, Bousquet and Elisseeff (2002) show that $\beta_{m,t} = \frac{1}{t}\sqrt{\frac{C}{\lambda_m}}$ where C is a positive constant. Thus, if one performs a grid search for λ_m on a grid $[a, b]$ with $a > 0$, then $\beta_{m,t} \leq \frac{1}{t}\sqrt{\frac{C}{a}}$ for any $\lambda_m \in [a, b]$. In other words, since the search space is generally bounded, the boundedness assumption regarding the $\beta_{m,t}$'s is not too restrictive in practice.*

6.6.2 Support Vector Machines and Experimental Setting

The aim of our experiments is to illustrate empirically the added value of the corrected K-fold compared with the standard one in terms of efficiency in model selection. In other words, we perform model selection with the K-fold and the corrected version that we promote and we compare the generalization risks of the selected trained models.

We consider a finite family of SVM regressors and classifiers trained with a regularization parameter λ_m ranging in a finite grid in an interval $[a, b]$, where $m \in \mathcal{M}$ a finite index set. Namely we set $[a, b] = [0.1, 100]$, and the grid is constructed with a constant step size equal to $\Delta = 0.1$, so that

$$(\lambda_m)_{m\in\mathcal{M}} = \{a + j\Delta \mid 0 \leq j < 1000\}.$$

In this SVM framework,

$$\mathcal{A}^{(m)}(T) = \arg\min_{f\in\mathcal{F}} \frac{1}{n_T} \sum_{i\in T} \ell(f, O_i) + \lambda_m \|f\|_k^2$$

where $\mathcal{F}$ is a reproducing kernel Hilbert space with kernel k. The kernel k is chosen here as the sigmoid kernel $\tanh(\tau\langle x, \rangle)$. Following standard practice we set $\tau = \frac{1}{d}$, where d is the dimension of the dataset. We use the quadratic loss for regression problems and the hinge loss $\ell(g, (x, y)) = (1 - yg(x))_+$ for classification, where $(f(x))_+ = max(0, f(x))$. Since the training datasets are bounded, we may consider that both these losses as bounded as well. For both these losses the algorithm $\mathcal{A}^{(m)}$ is $\frac{C}{n}$ uniformly stable (see Bousquet and Elisseeff (2002) for further details). The assumptions of Theorem 6.13 are thus satisfied, as pointed out in the discussion following the theorem's statement.

6.6.3 Datasets

Eight reference datasets from UCI [1] are considered: four classification datasets and four regression datasets listed below.

Regression datasets *Real estate valuation* (REV, 414 house price of unit area with 5 covariates); *QSAR fish toxicity* (906 toxic chemical concentration with 6 attributes); *Energy efficiency* (EE, 768 heating loads with 8 features.); *Concrete Compressive Strength* (CS, 1030 measure of the compressive strength with 8 attributes).

Classification datasets *Ionosphere dataset* (IO, 351 radar returns with 34 attributes), *Raisin dataset* (RS, 900 Keciman/Besni raisin with 7 attributes), *Audit risk dataset* (AR, 777 firm evaluation (fraudulent/non fraudulent) with 18 risk factors) and *QSAR bio degradation Data Set* (BIODEG, 1055 chemicals categorization with 12 descriptive features).

For each dataset one third of the data are removed ($\mathcal{S}$) and reserved for testing, *i.e.* for evaluating the generalization risk of the model selected using the remaining two thirds ($\mathcal{D}$). Let $\hat{m}_{\text{Kf}}$ (*resp.* $\hat{m}_{\text{Kf}-\text{corr}}$) denote the model selected using K-fold CV (*resp.* corrected K-fold CV) on the train set $\mathcal{D}$. In other words

$$\mathcal{A}^{(\widehat{m}_{\text{Kf}-\text{corr}})} = \arg\min_{m\in\mathcal{M}} \widehat{\mathcal{R}}^{corr}_{\text{Kfold}}\left[\mathcal{A}^{(m)}, V_{1:K}\right],$$
$$\mathcal{A}^{(\widehat{m}_{\text{Kf}})} = \arg\min_{m\in\mathcal{M}} \widehat{\mathcal{R}}_{\text{Kfold}}\left[\mathcal{A}^{(m)}, V_{1:K}\right].$$

In the end, in line with Section 6.6, the performance of both models are compared in terms of the mean squared error (or hinge loss for classification) and its estimated standard deviation on the test set $\mathcal{S}$.

[1] https://archive.ics.uci.edu

6.6.4 Numerical Results

We use the implementation provided by the python library `scikit-learn`. Tables 6.1 and 6.2 gather the results obtained respectively with the regression and classification datasets, for different numbers of folds K varying between 3 and 5. In all cases, the model selected by the bias corrected K-fold has a lower generalization risk than the one selected by the standard K-fold. As expected, the standard K-fold procedure behaves generally better with $K = 5$ than with $K = 3$. Indeed larger values of K decrease the bias of the K-fold CV. The benefit of the bias correction is thus all the more important for small values of K.

Table 6.1 – Regression mean squared errors for the K-fold and the bias corrected K-fold on various data sets. Estimated standard deviations are reported between parentheses.

Dataset	K-fold	Bias corrected K-fold
REV; K=3	74.198 (12.57)	**68.958 (11.63)**
K=4	74.189 (12.48)	**68.958 (11.63)**
K=5	73.359 (12.38)	**68.958 (11.63)**
EE; K=3	15.501 (1.81)	**14.405 (1.86)**
K=4	15.825 (1.84)	**14.350 (1.77)**
K=5	14.730 (1.73)	**14.298 (1.79)**
QSAR; K=3	1.183 (0.16)	**1.035 (0.14)**
K=4	1.112 (0.15)	**1.035 (0.14)**
K=5	1.112 (0.15)	**1.035 (0.14)**
CS; K=3	146.881 (13.81)	**126.492(10.23)**
K=4	144.195 (13.16)	**124.205 (9.46)**
K=5	137.060 (11.48)	**123.641 (9.30)**

Table 6.2 – Hinge losses for the K-fold and the bias corrected K-fold on various data sets. Estimated standard deviations are reported between parentheses.

Dataset	K-fold	Bias corrected K-fold
RS; K=3	0.470 (0.048)	**0.419 (0.038)**
K=4	0.420 (0.039)	**0.418 (0.039)**
K=5	0.420 (0.039)	**0.419 (0.038)**
IO; K=3	0.454 (0.081)	**0.414 (0.072)**
K=4	0.447 (0.092)	**0.425 (0.072)**
K=5	0.477 (0.091)	**0.464 (0.095)**
BIODEG; K=3	0.361 (0.039)	**0.357 (0.036)**
K=4	0.363 (0.037)	**0.357 (0.036)**
K=5	0.381 (0.041)	**0.357 (0.036)**
AR; K=3	0.112 (0.023)	**0.109 (0.021)**
K=4	0.108 (0.027)	**0.105 (0.025)**
K=5	0.107 (0.025)	**0.106 (0.024)**

6.7 Conclusion

This chapter demonstrates the limitations of the standard K-fold procedure for risk estimation *via* a lower bound on its error with a uniformly stable learner. We show that the corrected version of the K-fold for uniformly stable algorithms does not suffer the same drawbacks through a sanity-check upper bound and leverage this result to obtain guarantees regarding efficient model selection. This paves the way towards two possible research directions. A relevant follow-up would be to relax our uniform stability assumption in order to cover still a wider class of algorithms such as k-nearest-neighbors (Devroye and Wagner, 1979), Adaboost (Freund and Schapire, 1997a) and Lasso regression (Celisse and Guedj, 2016). A second promising avenue would be to consider an extension of the K-fold *penalization* proposed by Arlot and Lerasle (2016) to the class of stable learners.

6.A Main taools

First we recall McDiarmid's inequality (Theorem 3.1 in McDiarmid (1998)).

Proposition 6.A.1 (McDiarmid's inequality)**.** *let $Z = f(\mathcal{D})$ for some measurable function f and define*

$$\Delta_l(\mathcal{D}, O') = f(\mathcal{D}) - f(\mathcal{D}^l),$$

where $\mathcal{D}^l$ is obtained by replacing the l'th element of $\mathcal{D}$ by a sample $O' \in \mathcal{Z}$. In addition, suppose that

$$\forall l \in [n]\,, \sup_{\mathcal{D}\in\mathcal{Z}^n} \sup_{O'\in\mathcal{Z}} |\Delta_l(\mathcal{D}, O')| \leq c_l.$$

Then for any $t \geq 0$,

$$\mathbb{P}(Z - \mathbb{E}(Z) \geq t) \leq \exp\left(\frac{-2t^2}{\sum_{l=1}^n c_l^2}\right).$$

The following lemma guarantees that Fact 6.2.2 is verified for K-fold CV.

Lemma 6.14. *For the K-fold procedure, the training samples $T_{1:K}$ and validation samples $V_{1:K}$ satisfy Fact 6.2.2 i.e*

$$\frac{1}{K}\sum_{j=1}^{K} \frac{\mathbb{1}\Big\{l \in T_j\Big\}}{n_T} = \frac{1}{K}\sum_{j=1}^{K} \frac{\mathbb{1}\Big\{l \in V_j\Big\}}{n_V} = \frac{1}{n} \quad \forall l \in [\![1, n]\!].$$

Proof

The validation sets of K-fold verify the following property

$$\bigcup_{j=1}^{K} V_j = [\![1, n]\!] \text{ and } V_j \bigcap V_k = \varnothing\,,\ \forall j \neq k \in [\![1, K]\!]. \tag{6.18}$$

Under the hypothesis that $card(V_j) = n_V$ for all the validation sets, (6.18) implies that

$$n = \sum_{j=1}^{K} card(V_j) = K n_V. \tag{6.19}$$

Furthermore, under (6.18), an index $l \in [\![1, n]\!]$ belongs to a unique validation test V_j' and to all the train sets $T_j = V_j^c$ with $j \neq j'$. Hence, we have

$$\begin{cases} \sum_{j=1}^{K} \mathbb{1}\Big\{ l \in T_j \Big\} = K - 1, \\ \sum_{j=1}^{K} \mathbb{1}\Big\{ l \in V_j \Big\} = 1. \end{cases}$$

Using (6.18) and the fact that $n_T = n - n_V = (K-1)n_V$ yields the desired result.

■

6.B Intermediate results

First we provide a concentration bound for $\mathrm{D_{cv}}$ which has been defined in (6.7) as

$$\mathrm{D_{cv}} = \widehat{\mathcal{R}}_{\mathrm{CV}}\left[\mathcal{A}, V_{1:K}\right] - \mathcal{R}_{CV}\left[\mathcal{A}, V_{1:K}\right].$$

Lemma 6.15. *Suppose that Assumption 11 holds. Then, we have*

$$\mathbb{P}(|\mathrm{D_{cv}}| \geq t) \leq \exp\left(\frac{-2nt^2}{(4\beta_{n_T} n_T + 2L)^2}\right).$$

Proof Let $O' \in \mathcal{X}$ be an independent copy of $O_1, O_2, \ldots, O_n$, for any $l \in [\![1, n]\!]$ define

$$\Delta_l(\mathcal{D}, O') = \Big|\mathrm{D_{cv}}(\mathcal{D}) - \mathrm{D_{cv}}(\mathcal{D}^l)\Big|,$$

where $\mathcal{D}^l$ is obtained by replacing the l'th element of $\mathcal{D}$ by O'. Now we derive an upper bound on $\mathbb{P}(|\mathrm{D_{cv}}| \geq t)$ using Proposition 6.A.1. Namely, we will bound the maximum deviation of Δ_l by $\Delta_l \leq \frac{4\beta_{n_T} n_T}{n} + \frac{L}{n}$. To do so, write

$$\begin{aligned} \mathrm{D_{cv}} &= \widehat{\mathcal{R}}_{\mathrm{CV}}\left[\mathcal{A}, V_{1:K}\right] - \mathcal{R}_{CV}\left[\mathcal{A}, V_{1:K}\right] \\ &= \frac{1}{K n_{val}} \sum_{j=1}^{K} \sum_{i \in V_j} \left(\ell(\mathcal{A}(T_j), O_i) - \mathbb{E}_O\Big[\ell(\mathcal{A}(T_j), O) \mid \mathcal{D}_{T_j}\Big] \right) \\ &= \frac{1}{K n_{val}} \sum_{j=1}^{K} \sum_{i \in V_j} h(\mathcal{A}(T_j), O_i), \end{aligned}$$

where the last is used to define h. For a training set $T_j \subset [n]$, let $\mathcal{A}(T_{j,l})$ denote the algorithm $\mathcal{A}$ trained on the sequence $\mathcal{D}_{T_j,l} = \{O_i \in \mathcal{D}^l \mid i \in T_j\}$. Note that, for all $\mathcal{D}_{T_j}, o \in \mathcal{Z}^{n_T} \times \mathcal{Z}$ one has

$$\begin{cases} \ell\Big(\mathcal{A}(T_j), o\Big) = \ell\Big(\mathcal{A}(T_{j,l}), o\Big) \text{ if } l \notin T_j\,, \\ \left|\ell\Big(\mathcal{A}(T_j), o\Big) - \ell\Big(\mathcal{A}(T_{j,l}), o\Big)\right| \leq 2\beta_{n_T} \text{ otherwise.} \end{cases} \tag{6.20}$$

The first equation follows from the fact that $\mathcal{D}_{T_j,l} = \mathcal{D}_{T_j}$ if $l \notin T_j$, indeed, if the training set $\mathcal{D}_{T_j}$ doesn't contain the index l then changing the l'th element of $\mathcal{D}$ won't affect $\mathcal{D}_{T_j}$. The second inequality is obtained using the *uniform* stability of $\mathcal{A}$. Furthermore, using Equation 6.20 write

$$\begin{cases} \mathbb{E}\Big[\ell\Big(\mathcal{A}(T_j), O\Big) \mid \mathcal{D}_{T_j}\Big] = \mathbb{E}\Big[\ell\Big(\mathcal{A}(T_{j,l}), O\Big) \mid \mathcal{D}_{T_{j,l}}\Big] \text{ if } l \notin T_j \\ \Big|\mathbb{E}\Big[\ell\Big(\mathcal{A}(T_j), O\Big) \mid \mathcal{D}_{T_j}\Big] - \mathbb{E}\Big[\ell\Big(\mathcal{A}(T_{j,l}), O\Big) \mid \mathcal{D}_{T_{j,l}}\Big]\Big| \leq 2\beta_{n_T}, \text{ otherwise.} \end{cases} \tag{6.21}$$

Combining (6.21) and (6.20) gives

$$\begin{cases} |\mathbb{1}\Big\{l \in T_j\Big\}(h(\mathcal{A}(T_j), O_i)) - h(\mathcal{A}(T_{j,l}), O_i))| \leq 4\beta_{n_T}\,, \\ |\mathbb{1}\Big\{l \notin T_j\Big\}(h(\mathcal{A}(T_j), O_i)) - h(\mathcal{A}(T_{j,l}), O_i))| \leq 2L\mathbb{1}\{i = l\}, \end{cases} \tag{6.22}$$

so that

$$\begin{aligned} |\Delta_l(\mathcal{D}, O')| &\leq \frac{1}{Kn_{val}} \sum_{j=1}^{K} \sum_{i \in V_j} |h(\mathcal{A}(T_j), O_i)) - h(\mathcal{A}(T_{j,l}), O_i))| \\ \text{(From the fact that } [n] \backslash T_j = V_j) \ &= \frac{1}{Kn_{val}} \sum_{j=1}^{K} \sum_{i \in V_j} |h(\mathcal{A}(T_j), O_i)) - h(\mathcal{A}(T_{j,l}), O_i))| \left(\mathbb{1}\Big\{l \in T_j\Big\} + \mathbb{1}\Big\{l \in V_j\Big\}\right) \\ \text{(By Equation 6.22)} \ &\leq \frac{4\beta_{n_T}}{K} \sum_{j=1}^{K} \mathbb{1}\Big\{l \in T_j\Big\} + \frac{2L}{n_{val}K} \sum_{j=1}^{K} \mathbb{1}\Big\{l \in V_j\Big\} \\ \text{(By Fact 6.2.2)} \ &\leq \frac{4\beta_{n_T} n_T}{n} + \frac{2L}{n}. \end{aligned}$$

Using Mcdiarmid's inequality (Proposition 6.A.1) gives

$$\mathbb{P}(\mathrm{D_{cv}} \geq t) \leq \exp\left(\frac{-2nt^2}{(4\beta_{n_T} n_T + 2L)^2}\right).$$

Symmetrically, one has,

$$\mathbb{P}(\mathrm{D_{cv}} \leq -t) \leq \exp\left(\frac{-2nt^2}{(4\beta_{n_T} n_T + 2L)^2}\right).$$

Thus,

$$\mathbb{P}(|\mathrm{D_{cv}}| \geq t) \leq 2\exp\left(\frac{-2nt^2}{(4\beta_{n_T} n_T + 2L)^2}\right),$$

which is the desired result. ■

In the next lemma we obtain a similar concentration bound for the term $D_{\mathcal{A}[T_{1:K}]}$ defined in Eq 6.15 as

$$D_{\mathcal{A}[T_{1:K}]} = \frac{1}{K} \sum_{j=1}^{K} \left(\widehat{\mathcal{R}}\left[\mathcal{A}(T_j), T_j\right] - \mathcal{R}\Big[\mathcal{A}(T_j)\Big]\right).$$

Lemma 6.16. *Suppose that Assumption 11 holds. Then, one has,*

$$\mathbb{P}(|D_{\mathcal{A}[T_{1:K}]}| \geq t + 2\beta_{n_T}) \leq \exp\left(\frac{-2nt^2}{(4\beta_{n_T} n_T + 2L)^2}\right),$$

where L is the upper bound of the cost function defined in Assumption 11.

Proof Though the proof bears resemblance with the one of Lemma 6.15, we provide the full details details for compeleteness. We use McDiarmid's inequality with $f(\mathcal{D}) = D_{\mathcal{A}[T_{1:K}]}$, that is,

$$f(\mathcal{D}) = \frac{1}{K}\sum_{j=1}^{K}\left(\widehat{\mathcal{R}}\left[\mathcal{A}(T_j), T_j\right] - \mathcal{R}\Big[\mathcal{A}(T_j)\Big]\right)$$
$$f(\mathcal{D}^l) = \frac{1}{K}\sum_{j=1}^{K}\left(\widehat{\mathcal{R}}\left[\mathcal{A}(T_{j,l}), T_{j,l}\right] - \mathcal{R}\Big[\mathcal{A}(T_{j,l})\Big]\right).$$

Since for $l \notin T_j$, $T_j = T_{j,l}$, we find that

$$\begin{aligned}|\Delta_l(\mathcal{D}, O')| &\leq \frac{1}{Kn_T}\sum_{j=1}^{K}\mathbb{1}\Big\{l \in T_j\Big\}\left(\widehat{\mathcal{R}}\left[\mathcal{A}(T_j), T_j\right] - \widehat{\mathcal{R}}\left[\mathcal{A}(T_{j,l}), T_{j,l}\right] + \mathcal{R}\Big[\mathcal{A}(T_{j,l})\Big] - \mathcal{R}\Big[\mathcal{A}(T_j)\Big]\right)\\ &\leq \frac{1}{Kn_T}\sum_{j=1}^{K}\mathbb{1}\Big\{l \in T_j\Big\}\sum_{i\in T_j}|h(\mathcal{A}(T_j), O_i) - h(\mathcal{A}(T_{j,l}), O_i^l)|,\end{aligned}$$

with $h(\mathcal{A}(T), o) = \ell(\mathcal{A}(T), o) - \mathbb{E}_O\Big[\ell(\mathcal{A}(T), O) \mid \mathcal{D}_T\Big]$ and $(O_i^l)_{i=1,\dots,n}$ is the same as $(O_i)_{i=1,\dots,n}$ except the l-th element, O_l, which is replaced by O'. Whenever $l \in T_j$, it holds that

$$\begin{aligned}|\ell(\mathcal{A}(T_j), O_i) - \ell(\mathcal{A}(T_{j,l}), O_i^l)| &= |\ell(\mathcal{A}(T_j), O_i) - \ell(\mathcal{A}(T_{j,l}), O_i)|\mathbb{1}\big\{i \neq l\big\}\\ &\quad + |\ell(\mathcal{A}(T_j), O_l) - \ell(\mathcal{A}(T_{j,l}), O')|\mathbb{1}\big\{i = l\big\}\\ &\leq 2\beta_{n_T} + L\mathbb{1}\big\{i = l\big\},\end{aligned}$$

and that

$$\mathbb{E}\Big[|\ell(\mathcal{A}(T_j), O) - \ell(\mathcal{A}(T_{j,l}), O)| \mid \mathcal{D}_T, O'\Big] \leq 2\beta_{n_T}.$$

It follows from the definition of h that

$$|h(\mathcal{A}(T_j), O_i) - h(\mathcal{A}(T_{j,l}), O_i^l)| \leq 4\beta_{n_T} + 2L\mathbb{1}\big\{i = l\big\}.$$

By using the identity $\frac{1}{K}\sum_{j=1}^{K}\mathbb{1}\Big\{l \in T_j\Big\} = \frac{n_T}{n}$ we get

$$\Delta_l(\mathcal{D}, O') \leq \frac{4\beta_{n_T} n_T}{n} + \frac{2L}{n}.$$

Thus by Mcdiarmid's (Proposition 6.A.1), we obtain, for $Z = f(\mathcal{D}) - \mathbb{E}[f(\mathcal{D})]$, that

$$\mathbb{P}(Z \geq t) \leq \exp\left(\frac{-2nt^2}{(4\beta_{n_T} n_T + 2L)^2}\right).$$

Symmetrically, the event $-Z \geq t$ is subject to the same probability bound. It follows that

$$\mathbb{P}(|Z| \geq t) \leq 2 \exp\left(\frac{-2nt^2}{(4\beta_{n_T} n_T + 2L)^2}\right). \tag{6.23}$$

To derive an upper bound for $\mathbb{E}\left[D_{\mathcal{A}[T_{1:K}]}\right]$, we use the fact that all the training sets have the same length so that

$$\mathbb{E}(D_{\mathcal{A}[T_{1:K}]}) = \mathbb{E}\left[\mathcal{R}\left[\mathcal{A}(T_1)\right] - \widehat{\mathcal{R}}\left[\mathcal{A}(T_1), T_1\right]\right].$$

Then, we use Lemma 7 from Bousquet and Elisseeff (2002) ensuring that

$$\forall T \subset [n] \quad , \quad \mathbb{E}\left[\mathcal{R}\left[\mathcal{A}(T)\right] - \widehat{\mathcal{R}}\left[\mathcal{A}(T), T\right]\right] = \mathbb{E}\left[\ell\left(\mathcal{A}(T), O'\right) - \ell\left(\mathcal{A}(T^l), O'\right)\right].$$

Where $\mathcal{A}(T^l)$ is the learning rule $\mathcal{A}$ trained on the sample $\mathcal{D}_T \setminus \{O_l\} \cup \{O'\}$. Replacing the left side term by $\mathbb{E}\left[D_{\mathcal{A}[T_{1:K}]}\right]$ and using Definition 6.1 gives

$$|\mathbb{E}\left[f(\mathcal{D})\right]| = |\mathbb{E}\left[D_{\mathcal{A}[T_{1:K}]}\right]| \leq 2\beta_{n_T}.$$

Since $|f(\mathcal{D})| \leq |Z| + 2\beta_{n_T}$, we simply use (6.23) to reach the conclusion. ■

Remark 6.17. *Applying Lemma 6.16 with $T = [n]$ and $K = 1$ gives the following probability upper bound*

$$\mathbb{P}\left(\left|\widehat{\mathcal{R}}\Big[\mathcal{A}([n]), [n]\Big] - \mathcal{R}\Big[\mathcal{A}([n])\Big]\right| \geq t + 2\beta_n\right) \leq 2 \exp\left(\frac{-2nt^2}{(2L + 4\beta_n n)^2}\right).$$

Thus we retrieve the bound of Theorem 12 in Bousquet and Elisseeff (2002).

6.C Detailed proofs

6.C.1 Proof of Theorem 6.3

We proceed as described in the sketch of proof. Using Equation 6.6, write

$$\left|\widehat{\mathcal{R}}_{\mathrm{CV}}\left[\mathcal{A}, V_{1:K}\right] - \mathcal{R}\Big[\mathcal{A}([n])\Big]\right| \leq |\mathrm{D_{cv}}| + |\mathrm{Bias}|.$$

It remains to combine Lemma 6.15 with Fact 6.2.1 to obtain the desired result.

6.C.2 Proof of Theorem 6.6

Consider the regression problem for a random pair $O = (X, Y)$ consisting of a covariate $X = \frac{\epsilon}{M}$ where ϵ is a Rademacher variable and a response $Y = M \operatorname{sign} X$ for some $M > 1$. Namely $\epsilon \in \{+1, -1\}$ and $P(\epsilon = \pm 1) = 1/2$.

Now, let $n \leq e^M$ and define the algorithm $\mathcal{A}$ as a regularized empirical risk minimizer, more precisely $\mathcal{A}(\mathcal{D}, X) = \hat{\beta}\left(\mathcal{D}\right) X$ with

$$\hat{\beta}\left(\mathcal{D}\right) = \arg\min_{\beta \in \mathbb{R}} \frac{1}{n} \sum_{i=1}^{n} (Y_i - \beta X_i)^2 + \lambda_n |\beta|^2,$$

and $\lambda_n = \frac{1}{\log(n)} - \frac{1}{M^2}$.

Algorithmic Stability It's easy to check that $\hat{\beta} = \frac{\overline{X_n Y_n}}{\overline{(X_n)^2} + \lambda_n}$ where $\overline{X_n Y_n} = \frac{1}{n}\sum_{i=1}^{n} X_i Y_i$. Moreover, using the fact that $X_i Y_i = 1$, one obtains

$$\hat{\beta}\left(\mathcal{D}\right) = \frac{1}{1/M^2 + \lambda_n} = \log(n).$$

On the other hand, write

$$\begin{aligned}
\ell\Big(\mathcal{A}\left(\mathcal{D}\right), O\Big) - \ell\Big(\mathcal{A}(\mathcal{D}^{\setminus i}), O\Big) &= (\beta_{n-1} - \beta_n) X \Big(2Y - (\beta_n + \beta_{n-1}) X\Big) \\
&= \Big(\beta_{n-1} - \beta_n\Big) \left(2 - \frac{(\beta_n + \beta_{n-1})}{M^2}\right), \\
&= \Big(\log(n-1) - \log(n)\Big) \left(2 - \frac{\log(n) + \log(n-1)}{M^2}\right),
\end{aligned}$$

where the second line follows from the fact that $XY = 1$ and the last follows by replacing β and λ by their expression.
To conclude this part, we use the fact that $\log(1+x) \leq x$ for all $x \leq 1$, to obtain

$$\left|\ell\Big(\mathcal{A}\left(\mathcal{D}\right), O\Big) - \ell\Big(\mathcal{A}(\mathcal{D}^{\setminus i}), O\Big)\right| \leq \frac{2}{n}.$$

Bias Lower Bound Using the same equation as before we have

$$\begin{aligned}
\ell\Big(\mathcal{A}\left(\mathcal{D}\right), O\Big) - \ell\Big(\mathcal{A}(\mathcal{D}_T), O\Big) &= \Big(\log(n_T) - \log(n)\Big) \left(2 - \frac{\log(n) + \log(n_T)}{M^2}\right) \\
&= \log(\frac{K-1}{K}) \left(2 - \frac{\log(n) + \log(n_T)}{M^2}\right).
\end{aligned}$$

Thus, since $n_T \leq n \leq e^M$ we obtain

$$\mathcal{R}\left[\mathcal{A}\left(\mathcal{D}_T\right)\right] - \mathcal{R}\left[\mathcal{A}\left(\mathcal{D}\right)\right] \geq 2\log(\frac{K}{K-1}) \left(1 - \frac{1}{M}\right).$$

It remains to notice that

$$\mathbb{E}\left[\left|\widehat{\mathcal{R}}_{\mathrm{CV}}\left[\mathcal{A}, V_{1:K}\right]-\mathcal{R}\left[\mathcal{A}[n]\right]\right|\right] \geq\left|\mathbb{E}\left[\widehat{\mathcal{R}}_{\mathrm{CV}}\left[\mathcal{A}, V_{1:K}\right]-\mathcal{R}\left[\mathcal{A}[n]\right]\right]\right| \\ =\mathcal{R}\left[\mathcal{A}\left(\mathcal{D}_{T}\right)\right]-\mathcal{R}\left[\mathcal{A}\left(\mathcal{D}\right)\right],$$

and the proof is complete.

6.C.3 Proof of Theorem 6.11

We proceed as described in the sketch of proof. First remind the error decomposition 6.13

$$\widehat{\mathcal{R}}_{CV}^{corr}\left[\mathcal{A}, V_{1:K}\right]-\mathcal{R}\left[\mathcal{A}([n])\right]=D_{\mathcal{A}[n]}+\frac{n_{T}}{n}(\mathrm{D}_{\mathrm{cv}}-D_{\mathcal{A}[T_{1:K}]}),$$

where D_{cv}, $D_{\mathcal{A}[T_{1:K}]}$ and $D_{\mathcal{A}[n]}$ are defined in 6.7, 6.15 and 6.14 respectively. Since $n_T \leq n$, using the triangular inequality yields

$$\left|\widehat{\mathcal{R}}_{CV}^{corr}\left[\mathcal{A}, V_{1:K}\right]-\mathcal{R}\left[\mathcal{A}[n]\right]\right| \leq|D_{\mathcal{A}[n]}|+|\mathrm{D}_{\mathrm{cv}}|+|D_{\mathcal{A}[T_{1:K}]}|.$$

Combining lemma 6.15 and 6.16 regarding D_{cv} and $D_{\mathcal{A}[T_{1:K}]}$ with Remark 6.17 regarding $D_{\mathcal{A}[n]}$, one obtains

$$\begin{aligned}
&\mathbb{P}\left(\left|\widehat{\mathcal{R}}_{CV}^{corr}\left[\mathcal{A}, V_{1:K}\right]-\mathcal{R}\left[\mathcal{A}[n]\right]\right| \geq t+2(\beta_{n_T}+\beta_n)\right) \\
&\leq \mathbb{P}\left(|\mathrm{D}_{\mathrm{cv}}| \geq t/3\right)+\mathbb{P}\left(|D_{\mathcal{A}[n]}| \geq t/3+2\beta_n\right)+\mathbb{P}\left(|D_{\mathcal{A}[T_{1:K}]}| \geq t/3+2\beta_{n_T}\right) \\
&\leq 6\exp\left(\frac{-2nt^2}{9(4\beta_{n_T}n_T+L)^2}\right). \qquad (6.24)
\end{aligned}$$

By inverting, and using the assumption $\beta_t \leq \frac{\lambda}{t}$ one gets, with probability $1-6\delta$,

$$|\widehat{\mathcal{R}}_{CV}^{corr}\left[\mathcal{A}, V_{1:K}\right]-\mathcal{R}\left[\mathcal{A}[n]\right]| \leq 2\lambda(\frac{1}{n}+\frac{1}{n_T})+3(4\lambda+L)\sqrt{\frac{\log(1/\delta)}{2n}},$$

which is the desired result.

6.C.4 Proof of Theorem 6.13

The proof of Theorem 6.13 relies on the following proposition,

Proposition 6.18. *Let $(\mathcal{A}^{(m)})_{m\in\mathcal{M}}$ be a family of algorithms where each learner $\mathcal{A}^{(m)}$ is $(\beta_{m,t})_{1\leq i\leq n}$ uniform stable with respect to loss function $0 \leq \ell(g, O) \leq L$. Additionally, assume that, $|\mathcal{M}| < \infty$ and that*

$$\forall m \in \mathcal{M}\,;\ \beta_{m,t} \leq \frac{M}{t},$$

for some $M > 0$. Then one has, with probability at least $1-6\delta$,

$$\sup_{m\in\mathcal{M}}\left|\widehat{\mathcal{R}}_{K\text{-}fold}^{corr}\left[\mathcal{A}^{(m)}, V_{1:K}\right]-\mathcal{R}\left[\mathcal{A}^{(m)}([n])\right]\right| \leq \frac{6M}{n}+4(M+L)\sqrt{\frac{\log(|\mathcal{M}|/\delta)}{n}}.$$

Proof Using Fact 6.2.2 for the K-fold scheme ($n_T = n(K-1)/K$ with $K \geq 2$), and the fact that $\beta_{m,n} + \beta_{m,n_T} \leq (M/n)((2K-1)/(K-1))$ as well as $(2K-1)/(K-1) \leq 3$, one gets

$$\forall m \in \mathcal{M}, \mathbb{P}\left(\left|\widehat{\mathcal{R}}^{corr}_{\text{K-fold}}\left[\mathcal{A}^{(m)}, V_{1:K}\right] - \mathcal{R}\left[\mathcal{A}^{(m)}([n])\right]\right| \geq t + (6M/n)\right) \leq 6\exp\left(\frac{-2nt^2}{9(4M+L)^2}\right),$$

which gives by a union bound

$$\mathbb{P}\left(\sup_{m\in\mathcal{M}}\left|\widehat{\mathcal{R}}^{corr}_{\text{K-fold}}\left[\mathcal{A}^{(m)}, V_{1:K}\right] - \mathcal{R}\left[\mathcal{A}^{(m)}([n])\right]\right| \geq t + (6M/n)\right) \leq 6|\mathcal{M}|\exp\left(\frac{-2nt^2}{9(4M+L)^2}\right).$$

Thus, by inverting, we obtain the desired result. ■

Proof of Theorem 6.13 First, using the definition of $\hat{m}$ (eq. 6.17), write

$$\widehat{\mathcal{R}}^{corr}_{\text{Kfold}}\left[\mathcal{A}^{(\hat{m})}, V_{1:K}\right] - \widehat{\mathcal{R}}^{corr}_{\text{Kfold}}\left[\mathcal{A}^{(m^*)}, V_{1:K}\right] \leq 0.$$

It follows that

$$\begin{aligned}\mathcal{R}\left[\mathcal{A}^{(\hat{m})}([n])\right] - \mathcal{R}\left[\mathcal{A}^{(m^*)}([n])\right] &\leq \mathcal{R}\left[\mathcal{A}^{(\hat{m})}([n])\right] - \widehat{\mathcal{R}}^{corr}_{\text{Kfold}}\left[\mathcal{A}^{(\hat{m})}, V_{1:K}\right] \\ &\quad + \widehat{\mathcal{R}}^{corr}_{\text{Kfold}}\left[\mathcal{A}^{(m^*)}, V_{1:K}\right] - \mathcal{R}\left[\mathcal{A}^{(m^*)}([n])\right] \qquad (6.25)\\ &\leq 2\sup_{m\in\mathcal{M}}\left|\widehat{\mathcal{R}}^{corr}_{\text{K-fold}}\left[\mathcal{A}^{(m)}, V_{1:K}\right] - \mathcal{R}\left[\mathcal{A}^{(m)}([n])\right]\right|.\end{aligned}$$

It remains to use proposition 6.18 and the proof is complete.

6.D Uniform stability for randomized algorithms

In this section we generalize the results from the main chapter to the case of randomized algorithms. Let us start with reminding the concept of uniform stability for randomized learning algorithms introduced in Elisseeff et al. (2005).

Definition 6.19. *An algorithm $\mathcal{A}$ is said to be* $(\beta_t)_{1\leq t\leq n}$-uniform *stable with respect to a loss function ℓ if, for any $\mathcal{D} \in \mathcal{Z}^n$, $T \subset [n]$, $i \in T$ and $o \in \mathcal{Z}$, the following holds*

$$\left|\mathbb{E}_{\mathcal{A}}\left[\ell\left(\mathcal{A}(T), o\right)\right] - \mathbb{E}_{\mathcal{A}}\left[\ell\left(\mathcal{A}(T^{\setminus i}), o\right)\right]\right| \leq \beta_{n_T}. \qquad (6.26)$$

Where the randomness in the latter expectation stems from the algorithm $\mathcal{A}$ while the observation o and the data sequence $\mathcal{D}$ are fixed. Equivalently,

$$\left|\mathbb{E}\left[\ell\left(\mathcal{A}(T), O\right) - \ell\left(\mathcal{A}(T^{\setminus i}), O\right)\middle|\mathcal{D}_T, O\right]\right| \leq \beta_{n_T}. \qquad (6.27)$$

Note that a similar version of Fact 6.2.1 still holds, more precisely, one has

Fact 6.D.1. *Let $\mathcal{A}$ be a decision rule which is $(\beta_t)_{1\leq t\leq n}$ uniformly stable, additionally suppose that the sequence $(\beta_t)_{1\leq t\leq n}$ is decreasing, then for any $T\subset[n]$, and $o\in\mathcal{Z}$, one has*

$$\left|\mathbb{E}\left[\ell\Big(\mathcal{A}([n]),o\Big)-\ell\Big(\mathcal{A}(T),o\Big)\mid\mathcal{D}\right]\right|\leq\sum_{i=n_T}^{n}\beta_i.$$

6.D.1 Upper Bounds for K-fold CV under random uniform stability

We now derive upper bounds -*in expectation*- on the error induced by $\widehat{\mathcal{R}}_{\mathrm{CV}}$ and $\widehat{\mathcal{R}}_{CV}^{corr}$. As highlighted in Remark 6.5, the main problem with K-fold CV is it's bias. Therefore, for the sake of brevity, we will focus only on the expectation of the estimate .

Theorem 6.20. *Suppose that $\mathcal{A}$ is $(\beta_t)_{1\leq t\leq n}$ uniformly stable. Then we have*

$$\left|\mathbb{E}\left[\widehat{\mathcal{R}}_{\mathrm{CV}}\left[\mathcal{A},V_{1:K}\right]-\mathcal{R}\Big[\mathcal{A}([n])\Big]\right]\right|\leq\sum_{i=n_T}^{n}\beta_i.$$

Proof First, by applying the tower rule one has

$$\begin{aligned}
\mathbb{E}\left[\widehat{\mathcal{R}}_{\mathrm{CV}}\left[\mathcal{A},V_{1:K}\right]-\mathcal{R}\Big[\mathcal{A}([n])\Big]\right] &= \mathbb{E}\left[\frac{1}{K}\sum_{j=1}^{K}\widehat{\mathcal{R}}\Big[\mathcal{A}(T_j),V_j\Big]-\mathcal{R}\left[\mathcal{A}([n])\right]\right]\\
&= \frac{1}{K}\sum_{j=1}^{K}\mathbb{E}\left[\mathbb{E}\left[\widehat{\mathcal{R}}\Big[\mathcal{A}(T_j),V_j\Big]\mid\mathcal{D}_{T_j}\right]-\mathcal{R}\Big[\mathcal{A}([n])\Big]\right]\\
&= \frac{1}{K}\sum_{j=1}^{K}\mathbb{E}\left[\mathbb{E}\left[\ell(\mathcal{A}(T_j,O)\mid\mathcal{D}_{T_j}\right]-\mathcal{R}\Big[\mathcal{A}([n])\Big]\right]\\
&= \frac{1}{K}\sum_{j=1}^{K}\mathbb{E}\left[\mathbb{E}\left[\ell(\mathcal{A}(T_j,O)-\ell(\mathcal{A}([n],O))\mid\mathcal{D}\right]\right].
\end{aligned}\tag{6.28}$$

The third line follows from the fact that $\ell(\mathcal{A}(T),O_j)$ and $\ell(\mathcal{A}(T),O)$ has the same law for all $j\in V$. This indeed verified since all the training sets T_j's has the same length and the O_j's are independent from $\mathcal{D}_T$. To obtain the desired result, it remains to combine Equation 6.28 with fact 6.D.1. ■

Now let us prove that the bias corrected K-fold (cf. Eq (6.12)) has a vanishing bias for randomized algorithms.

Theorem 6.21 (Corrected K-fold bias)**.** *Suppose that $\mathcal{A}$ is $(\beta_t)_{1\leq t\leq n}$ uniformly stable and $\beta_t\leq\frac{\lambda}{t}$, for some $\lambda>0$. Then, we have*

$$\left|\mathbb{E}\left[\widehat{\mathcal{R}}_{\mathrm{Kfold}}^{corr}\left[\mathcal{A},V_{1:K}\right]-\mathcal{R}\Big[\mathcal{A}([n])\Big]\right]\right|\leq\frac{2\lambda(2K-1)}{(K-1)n}.$$

Proof Combining the error decomposition (6.13) with the fact that $n_T = \frac{K-1}{K}n$ we obtain

$$\widehat{\mathcal{R}}^{corr}_{\text{Kfold}}\left[\mathcal{A}, V_{1:K}\right] - \mathcal{R}\left[\mathcal{A}[n]\right] = D_{\mathcal{A}[n]} + \frac{K-1}{K}(\text{D}_{\text{cv}} - D_{\mathcal{A}[T_{1:K}]}). \tag{6.29}$$

Using Theorem 2.2 in Hardt et al. (2016a), one obtains the twin inequality

$$\left|\mathbb{E}\left[D_{\mathcal{A}[n]}\right]\right| \leq 2\beta_n$$

$$\left|\mathbb{E}\left[D_{\mathcal{A}[T_{1:K}]}\right]\right| \leq 2\beta_{n_T}.$$

Since $\mathbb{E}\left[\text{D}_{\text{cv}}\right] = 0$, Equation 6.29 combined with the triangular inequality gives

$$\begin{aligned}\left|\mathbb{E}\left[\widehat{\mathcal{R}}^{corr}_{\text{Kfold}}\left[\mathcal{A}, V_{1:K}\right] - \mathcal{R}\Big[\mathcal{A}([n])\Big]\right]\right| &\leq 2\left(\beta_n + \frac{(K-1)\beta_{n_T}}{K}\right) \\ &\leq 2\left(\beta_n + \beta_{n_T}\right). \end{aligned} \tag{6.30}$$

It remains to use the assumption $\beta_t \leq \frac{\lambda}{t}$ and the proof is complete. ■

We conclude this section by deriving an upper bound for the model selection problem,

Theorem 6.22. *Let $(\mathcal{A}^{(m)})_{m\in\mathcal{M}}$ be a family of algorithms where each learner $\mathcal{A}^{(m)}$ is $(\beta_{m,t})_{1\leq i\leq n}$ uniform stable with respect to loss function ℓ. Additionally, assume that*

$$\beta_{m,t} \leq \frac{M}{t}$$

for some $M > 0$. Then one has

$$\mathbb{E}\left[\mathcal{R}\Big[\mathcal{A}^{(\hat{m})}([n])\Big] - \mathcal{R}\Big[\mathcal{A}^{(m^*)}([n])\Big]\right] \leq \frac{4M(2K-1)}{(K-1)n},$$

where m^* and $\hat{m}$ are defined by Equations 6.16, 6.17 respectively.

Proof First, using Equation 6.30, one obtains, for all $m \in \mathcal{M}$,

$$\begin{aligned}\left|\mathbb{E}\left[\widehat{\mathcal{R}}^{corr}_{\text{Kfold}}\Big[\mathcal{A}^{(m)}, V_{1:K}\Big] - \mathcal{R}\Big[\mathcal{A}^{(m)}([n])\Big]\right]\right| &\leq 2\left(\beta_{m,n} + \beta_{m,n_T}\right) \\ &\leq \frac{2M(2K-1)}{(K-1)n}.\end{aligned}$$

So that

$$\sup_{m\in\mathcal{M}}\left|\mathbb{E}\left[\widehat{\mathcal{R}}^{corr}_{\text{Kfold}}\left[\mathcal{A}^{(m)}, V_{1:K}\right] - \mathcal{R}\Big[\mathcal{A}^{(m)}([n])\Big]\right]\right| \leq \frac{2M(2K-1)}{(K-1)n}. \tag{6.31}$$

On the other hand, Inequality 6.25 yields Thus, by Inequality 6.31, one has

$$\mathbb{E}\left[\mathcal{R}\Big[\mathcal{A}^{(\hat{m})}([n])\Big] - \mathcal{R}\Big[\mathcal{A}^{(m^*)}([n])\Big]\right] \leq \frac{4M(2K-1)}{(K-1)n},$$

which concludes the proof. ■

6.D.2 Proof of Theorem 6.8

Following the line of Zhang et al. (2022), consider the following convex function ,

$$f(w, o) = \frac{1}{2} w^\top A w - y x^\top w,$$

where A is a positive semi definite penalization matrix (PSD) in $\mathbb{R}^{d\times d}$ with rank $p < d$, $w \in \mathbb{R}^d$ and $x, y \in \mathbb{R}^d \times \mathbb{R}$. Such a rank deficient penalization matrix can be found in multiple contexts like fused lasso (Tibshirani and Taylor, 2011), fused ridge (Bilgrau et al., 2020), etc.
In the next lemma, by picking carefully the input space $\mathcal{X} \times \mathcal{Y}$ and the distribution P, we construct an example where we control exactly the amount of *instability* of SGD . Theorem 6.8 follows directly from the following proposition.

Proposition 6.23. *Let $M > 1$, $n \in [\![1, e^M]\!]$. Suppose that $O = (X, Y) \in \mathcal{X} \times \mathcal{Y} = \{v, -v\} \times \{1\}$ and $P(X = v) = \frac{2}{3}$ where v is a unit vector in $\mathbb{R}^d$ such as $Av = 0$. For $t \geq 1$, there exist a sequence of step sizes $(\alpha_{k,n})_{1\leq k\leq t}$, such as the SGD algorithm (Definition 6.11) with $\ell = f$ and $\mathcal{A}_0 = 0$ verifies,*

$$\mathbb{E}\left[\left|\widehat{\mathcal{R}}_{\mathrm{CV}}\left[\mathcal{A}, V_{1:K}\right] - \mathcal{R}\left[\mathcal{A}[n]\right]\right|\right] \geq \frac{\log\left(K/K-1\right)}{3}.$$

Furthermore SGD satisfies Assumption 11 with respect to ℓ with $L = M$ and

$$\beta_n \leq \frac{3}{n-1}\sum_{k=1}^{t} \alpha_{k,n} \leq \frac{3M}{n-1}.$$

Proof Computing the gradient of $f(\cdot, o)$ for $o = (x, y)$ yields,

$$\nabla_w f(w, o) = Aw - yx.$$

Now set $w_t = \mathcal{A}_t\left([n]\right)$ and write using the definition of SGD (Equation 6.11) :

$$w_{t+1} = \left(I - \alpha_t A\right) w_t + \alpha_{t,n} y X_{it}. \tag{6.32}$$

Thus by induction we obtain $w_t = \theta_t v$ for some $\theta_t \in \mathbb{R}$. Consequently, Equation 6.32 yields,

$$w_{t+1} = w_t + \alpha_{t,n} y X_{it},$$

so that,

$$w_t = \sum_{k=1}^{t} \alpha_{k,n} y X_{it}. \tag{6.33}$$

Furthermore SGD picks $X_{it} = v$ with probability n_+/n where n_+ (*resp.* n_-) is the number of samples such as $X = v$ (*resp.* $-v$), hence

$$\mathbb{E}_{\mathcal{A}}\left[w_t\right] = \sum_{k=1}^{t} \alpha_{k,n}\left(\frac{n_+}{n} - \frac{n_-}{n}\right) v. \tag{6.34}$$

On the other hand, since $w_t = \theta_t v$ we obtain,

$$f(w_t, o) = -yx^\top w_t. \tag{6.35}$$

Set $w_t' = \mathcal{A}\left([n]^{\setminus j}\right)$ and consider the case where j is such as $X_j = v$. Note that the other case is similar and thus omitted. Since $\|v\| = 1$, one has by Equations 6.34 and 6.35

$$\forall o \in \mathcal{Z}, \left|\mathbb{E}_{\mathcal{A}}\left[f\left(w_t', o\right)\right] - \mathbb{E}_{\mathcal{A}} f\left[\left(w_t, o\right)\right]\right| = \left|\frac{n_+ - n_-}{n}\sum_{k=1}^{t}\alpha_{k,n} - \frac{n_+ - n_- - 1}{n-1}\sum_{k=1}^{t}\alpha_{k,n-1}\right|.$$

Now for $t \geq 0$, take $\alpha_{k,n} = \dfrac{\log(n)}{t}$ for all $k \leq t$ so that,

$$\forall o \in \mathcal{Z}, \left|\mathbb{E}_{\mathcal{A}}\left[f\left(w_t', o\right)\right] - \mathbb{E}_{\mathcal{A}} f\left[\left(w_t, o\right)\right]\right| = \left|\frac{n_+ - n_-}{n}\log(n) - \frac{n_+ - n_- - 1}{n-1}\log(n-1)\right|,$$

which yields by simple algebra

$$\begin{aligned}
\forall o \in \mathcal{Z}\ , \left|\mathbb{E}_{\mathcal{A}}\left[f\left(w_t', o\right)\right] - \mathbb{E}_{\mathcal{A}} f\left[\left(w_t, o\right)\right]\right| &\leq \frac{2\log(n) + 1}{n-1} \\
&\leq \frac{3\log(n)}{n-1} \\
&= \frac{3}{n-1}\sum_{k=1}^{t}\alpha_{k,n} \\
&\leq \frac{3M}{n-1}.
\end{aligned} \tag{6.36}$$

Now, using Equations 6.33 and 6.35 and the expression of $\alpha_{k,n}$ we get,

$$\forall o \in \mathcal{Z}\ , |f(w_t, o)| = \sum_{k=1}^{t}\alpha_{k,n} = \log(n) \leq M,$$

The latter equation combined with Equation 6.36 confirms that SGD verifies Assumption 11with $L = M$ and

$$\beta_n \leq \frac{3}{n-1}\sum_{k=1}^{t}\alpha_{k,n} \leq \frac{3M}{n-1}.$$

For the lower bound, let $T \subset [n]$ and set $w_t^T = \mathcal{A}_t\left(T\right)$. Using 6.34 yields

$$\mathbb{E}\left[\mathbb{E}_{\mathcal{A}}\left[w_t^T\right]\right] = \sum_{k=1}^{t}\alpha_{k,n_T}\frac{v}{3},$$

so that by 6.35 ,

$$\begin{aligned}
\left|\mathbb{E}\left[\mathbb{E}_{\mathcal{A}}\left[f\left(w_t, o\right)\right]\right] - \left[\mathbb{E}_{\mathcal{A}}\left[f\left(w_t^T, o\right)\right]\right]\right| &= \frac{\sum_{k=1}^{t}\alpha_{k,n} - \sum_{k=1}^{t}\alpha_{k,n_T}}{3} \\
\left(\alpha_{k,n} = \log(n)/t\right) &= \frac{\log\left(K/K-1\right)}{3}.
\end{aligned}$$

It remains to notice that

$$\begin{aligned}\mathbb{E}\left[\left|\widehat{\mathcal{R}}_{\mathrm{CV}}\left[\mathcal{A}, V_{1:K}\right]-\mathcal{R}\left[\mathcal{A}[n]\right]\right|\right] &\geq\left|\mathbb{E}\left[\widehat{\mathcal{R}}_{\mathrm{CV}}\left[\mathcal{A}, V_{1:K}\right]-\mathcal{R}\left[\mathcal{A}[n]\right]\right]\right| \\ &=\left|\mathcal{R}\left[\mathcal{A}\left(\mathcal{D}_T\right)\right]-\mathcal{R}\left[\mathcal{A}\left(\mathcal{D}\right)\right]\right|,\end{aligned}$$

and the proof is complete. ■

Conclusion

As we look towards the future of our research, two key areas stand out for further exploration:

Algorithmic Stability as a Tool for Deriving Probability Upper Bounds in EVT We aim to delve into the concept of algorithmic stability, investigating its potential in deriving more refined and reliable probability upper bounds for EVT. This approach may offer a novel perspective, differing from traditional complexity-based measures, and could lead to sharper more interpretable bounds.

Optimizing Extreme Threshold α Through Cross-Validation Another promising direction for our research involves the use of cross-validation techniques to tune the optimal extreme threshold α. This threshold is critical in predictions pertaining to extreme regions. By leveraging cross-validation, we can potentially develop an adaptable approach, allowing for the adjustment of α in response to varying data sets and prediction requirements.

Bibliography

K. Abou-Moustafa and C. Szepesvári. An a priori exponential tail bound for k-folds cross-validation. *arXiv preprint*, 2017. pages 29, 188

K. Abou-Moustafa and C. Szepesvári. An exponential efron-stein inequality for l_q stable learning rules. In A. Garivier and S. Kale, editors, *Proceedings of the 30th International Conference on Algorithmic Learning Theory*, volume 98 of *Proceedings of Machine Learning Research*, pages 31–63, 22–24 Mar 2019. page 55

A. Agarwal, A. Beygelzimer, M. Dudík, J. Langford, and H. Wallach. A reductions approach to fair classification. In *International Conference on Machine Learning*, pages 60–69. PMLR, 2018. page 128

A. Aghbalou, F. Portier, A. Sabourin, and C. Zhou. Supplement to "Tail inverse regression for dimension reduction with extreme response", 2021. pages 78, 80, 102

A. Aghbalou, F. Portier, A. Sabourin, and C. Zhou. Tail inverse regression for dimension reduction with extreme response, 2023. page 131

M. M. Ahsan and Z. Siddique. Machine learning-based heart disease diagnosis: A systematic literature review. *Artificial Intelligence in Medicine*, 128:102289, 2022. ISSN 0933-3657. doi: https://doi.org/10.1016/j.artmed.2022.102289. URL https://www.sciencedirect.com/science/article/pii/S0933365722000549. page 17

S. Al-Stouhi and C. K. Reddy. Transfer learning for class imbalance problems with inadequate data. *Knowledge and Information Systems*, 48(1):201–228, Aug. 2015. doi: 10.1007/s10115-015-0870-3. URL https://doi.org/10.1007/s10115-015-0870-3. page 33

J. Andrews, T. Tanay, E. J. Morton, and L. D. Griffin. Transfer representation-learning for anomaly detection. JMLR, 2016. page 170

M. Anthony and S. B. Holden. Cross-validation for binary classification by real-valued functions: theoretical analysis. In *Proceedings of the eleventh annual conference on Computational learning theory*, pages 218–229, 1998. pages 25, 41

S. Arlot. V-fold cross-validation improved: V-fold penalization. 40 pages, plus a separate technical appendix., Feb. 2008a. pages 187, 196

S. Arlot. V-fold cross-validation improved: V-fold penalization. Feb. 2008b. pages 24, 42

S. Arlot and A. Celisse. A survey of cross-validation procedures for model selection. *Statistics surveys*, 4:40–79, 2010. pages 41, 199

S. Arlot and M. Lerasle. Choice of v for v-fold cross-validation in least-squares density estimation. *Journal of Machine Learning Research*, 17(208):1–50, 2016. pages 24, 25, 26, 29, 42, 187, 188, 189, 198, 202

S. Asenova, G. Mazo, and J. Segers. Inference on extremal dependence in the domain of attraction of a structured hüsler–reiss distribution motivated by a markov tree with latent variables. *Extremes*, pages 1–40, 2021. pages 20, 78

M. Austern and W. Zhou. Asymptotics of cross-validation. *arXiv preprint*, 2020a. pages 29, 188

M. Austern and W. Zhou. Asymptotics of cross-validation. *arXiv preprint arXiv:2001.11111*, 2020b. pages 25, 167

D. Babichev, F. Bach, et al. Slice inverse regression with score functions. *Electronic journal of statistics*, 12(1):1507–1543, 2018. pages 21, 80

M.-F. Balcan, M. Khodak, and A. Talwalkar. Provable guarantees for gradient-based meta-learning. In K. Chaudhuri and R. Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 424–433. PMLR, 09–15 Jun 2019. page 157

P. L. Bartlett and S. Mendelson. Empirical minimization. *Probability theory and related fields*, 135(3):311–334, 2006. pages 129, 136, 149

P. L. Bartlett, O. Bousquet, and S. Mendelson. Local Rademacher complexities. *The Annals of Statistics*, 33(4):1497 – 1537, 2005. doi: 10.1214/009053605000000282. URL https://doi.org/10.1214/009053605000000282. pages 35, 129, 135, 145, 147, 149

P. L. Bartlett, M. I. Jordan, and J. D. Mcauliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006a. ISSN 01621459. pages 74, 130

P. L. Bartlett, M. I. Jordan, and J. D. McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006b. pages 36, 156, 169

S. Bates, T. Hastie, and R. Tibshirani. Cross-validation: what does it estimate and how well does it do it? *arXiv preprint*, 2021. page 41

P. Bayle, A. Bayle, L. Janson, and L. Mackey. Cross-validation confidence intervals for test error. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 16339–16350, 2020. pages 25, 29, 167, 187, 188, 196

J. Beirlant, Y. Goegebeur, J. Segers, and J. L. Teugels. *Statistics of extremes: theory and applications*. John Wiley & Sons, 2006. pages 18, 77

S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. W. Vaughan. A theory of learning from different domains. *Machine learning*, 79(1):151–175, 2010. pages 37, 155, 157, 159

Y. Bengio, J.-f. Paiement, P. Vincent, O. Delalleau, N. Roux, and M. Ouimet. Out-of-sample extensions for lle, isomap, mds, eigenmaps, and spectral clustering. In *NeurIPS proceedings*, volume 16. MIT Press, 2003. page 106

P. Bertail, S. Clémençon, Y. Guyonvarch, and N. Noiry. Learning from biased data: A semi-parametric approach. In M. Meila and T. Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 803–812. PMLR, 18–24 Jul 2021. pages 33, 128

S. Bhatia, A. Jain, and B. Hooi. Exgan: Adversarial generation of extreme samples. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(8):6750–6758, May 2021. pages 19, 40

G. Biau and L. Devroye. *Lectures on the nearest neighbor method*, volume 246. Springer, 2015. page 141

A. E. Bilgrau, C. F. Peeters, P. S. Eriksen, M. Bøgsted, and W. N. van Wieringen. Targeted fused ridge estimation of inverse covariance matrices from multiple high-dimensional data classes. *J. Mach. Learn. Res.*, 21(26):1–52, 2020. page 212

J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. Wortman. Learning bounds for domain adaptation. *Advances in neural information processing systems*, 20, 2007a. page 155

J. Blitzer, M. Dredze, and F. Pereira. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Proceedings of the 45th annual meeting of the association of computational linguistics*, pages 440–447, 2007b. page 155

A. Blum, A. Kalai, and J. Langford. Beating the hold-out: bounds for k-fold and progressive cross-validation. In *COLT '99*, 1999. pages 24, 25, 42

B. Bobbia, C. Dombry, and D. Varron. A donsker and glivenko-cantelli theorem for random measures linked to extreme value theory. *HAL preprint hal-03402380*, 2021. page 43

S. Boucheron and M. Thomas. Concentration inequalities for order statistics. *Electronic Communications in Probability*, 17:1–12, 2012. pages 19, 40

S. Boucheron and M. Thomas. Tail index estimation, concentration and adaptivity. *Electronic Journal of Statistics*, 9(2):2751–2792, 2015. pages 19, 40

S. Boucheron, O. Bousquet, and G. Lugosi. Theory of classification: A survey of some recent advances. *ESAIM: probability and statistics*, 9:323–375, 2005. pages 33, 45, 129

M. Bousebata, G. Enjolras, and S. Girard. Extreme Partial Least-Squares regression. working paper or preprint, 2021. URL https://hal.inria.fr/hal-03165399. pages 78, 79

K. Bousmalis, A. Irpan, P. Wohlhart, Y. Bai, M. Kelcey, M. Kalakrishnan, L. Downs, J. Ibarz, P. Pastor, K. Konolige, et al. Using simulation and domain adaptation to improve efficiency of deep robotic grasping. In *2018 IEEE international conference on robotics and automation (ICRA)*, pages 4243–4250. IEEE, 2018. page 155

O. Bousquet and A. Elisseeff. Algorithmic stability and generalization performance. In T. Leen, T. Dietterich, and V. Tresp, editors, *Advances in Neural Information Processing Systems*, volume 13. MIT Press, 2001. page 188

O. Bousquet and A. Elisseeff. Stability and generalization. *The Journal of Machine Learning Research*, 2:499–526, 2002. pages 25, 29, 30, 41, 55, 156, 159, 161, 167, 188, 190, 192, 199, 200, 206

O. Bousquet, Y. Klochkov, and N. Zhivotovskiy. Sharper bounds for uniformly stable algorithms. In *Conference on Learning Theory*, pages 610–626. PMLR, 2020. pages 30, 159, 195, 196

W. Bryc. *The normal distribution: characterizations with applications*, volume 100. Springer Science & Business Media, 2012. page 82

P. Burman. A comparative study of ordinary cross-validation, v-fold cross-validation and the repeated learning-testing methods. *Biometrika*, 76(3):503–514, 1989. pages 24, 29, 187, 189, 196, 197

P. Burman. Estimation of optimal transformations using v-fold cross validation and repeated learning-testing methods. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 314–345, 1990. pages 29, 189, 196, 197

J.-J. Cai, J. H. J. Einmahl, and L. de Haan. Estimation of extreme risk regions under multivariate regular variation. *The Annals of Statistics*, 39(3):1803 – 1826, 2011. page 73

M. Campi, G. W. Peters, N. Azzaoui, and T. Matsui. Machine learning mitigants for speech based cyber risk. *IEEE Access*, 9:136831–136860, 2021. page 170

M. Campi, G. W. Peters, and D. Toczydlowska. Ataxic speech disorders and parkinson's disease diagnostics via stochastic embedding of empirical mode decomposition. *Plos one*, 18(4):e0284667, 2023. page 170

A. Carpentier and A. K. Kim. Adaptive and minimax optimal estimation of the tail coefficient. *Statistica Sinica*, pages 1133–1144, 2015. page 40

A. Celisse and B. Guedj. Stability revisited: new generalisation bounds for the leave-one-out, 2016. pages 55, 202

A. Celisse and T. Mary-Huard. Theoretical analysis of cross-validation for estimating the risk of the k-nearest neighbor classifier. *The Journal of Machine Learning Research*, 19(1):2373–2426, 2018. page 167

V. Chandola, A. Banerjee, and V. Kumar. Anomaly detection: A survey. *ACM computing surveys (CSUR)*, 41(3):1–58, 2009. page 170

Z. Charles and D. Papailiopoulos. Stability and generalization of learning algorithms that converge to global optima. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pages 745–754. PMLR, 10–15 Jul 2018. pages 30, 156, 159, 165, 188, 191, 195

E. Chautru. Dimension reduction in multivariate extreme value analysis. *Electronic journal of statistics*, 9(1):383–418, 2015. pages 20, 78

N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002. pages 31, 127

D. Chetverikov, Z. Liao, and V. Chernozhukov. On cross-validated Lasso in high dimensions. *The Annals of Statistics*, 49(3):1300 – 1317, 2021a. page 55

D. Chetverikov, Z. Liao, and V. Chernozhukov. On cross-validated lasso in high dimensions. *The Annals of Statistics*, 49(3):1300–1317, 2021b. page 54

M. Chiapino and A. Sabourin. Feature clustering for extreme events analysis, with application to extreme stream-flow data. In *International Workshop on New Frontiers in Mining Complex Patterns*, pages 132–147. Springer, 2016. pages 19, 20, 40, 78

M. Chiapino, S. Clémençon, V. Feuillard, and A. Sabourin. A multivariate extreme value theory approach to anomaly clustering and visualization. *Computational Statistics*, pages 1–22, 2019a. pages 20, 78

M. Chiapino, A. Sabourin, and J. Segers. Identifying groups of variables with the potential of being large simultaneously. *Extremes*, 22(2):193–222, 2019b. pages 20, 78

M. Chiapino, S. Clémençon, V. Feuillard, and A. Sabourin. A multivariate extreme value theory approach to anomaly clustering and visualization. *Computational Statistics*, 35(2):607–628, 2020. page 40

J. H. Cho and B. Hariharan. On the efficacy of knowledge distillation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4794–4802, 2019. page 170

W. Chojnacki and M. J. Brooks. A note on the locally linear embedding algorithm. *Intern. J. Pattern Recognit. Artif. Intell.*, 23(08):1739–1752, 2009. page 106

E. Christou. Central quantile subspace. *Statistics and Computing*, 30(3):677–695, 2020. page 90

S. Clémençon, H. Jalalzai, S. Lhaut, A. Sabourin, and J. Segers. Concentration bounds for the empirical angular measure with statistical learning applications, 2022. pages 19, 40, 43, 44, 45, 49

P. Colombo, E. Dadalto, G. Staerman, N. Noiry, and P. Piantanida. Beyond mahalanobis distance for textual ood detection. *Advances in Neural Information Processing Systems*, 35:17744–17759, 2022a. page 170

P. Colombo, G. Staerman, N. Noiry, and P. Piantanida. Learning disentangled textual representations via statistical measures of similarity. *arXiv preprint arXiv:2205.03589*, 2022b. page 170

P. Constantinou and A. P. Dawid. Extended conditional independence and applications in causal inference. *The Annals of Statistics*, 45(6):2618–2653, 2017. pages 20, 79, 80, 81

R. D. Cook. *Regression graphics: Ideas for studying regressions through graphics*, volume 482. John Wiley & Sons, 2009. pages 20, 21, 79, 80, 81

R. D. Cook and B. Li. Dimension reduction for conditional mean in regression. *The Annals of Statistics*, 30(2):455–474, 2002. page 90

R. D. Cook and L. Ni. Sufficient dimension reduction via inverse regression: A minimum discrepancy approach. *Journal of the American Statistical Association*, 100(470):410–428, 2005. pages 81, 82

R. D. Cook and S. Weisberg. Comment. *Journal of the American Statistical Association*, 86(414):328–332, 1991. pages 21, 79, 81, 83, 121

D. Cooley and E. Thibaud. Decompositions of dependence for high-dimensional extremes. *Biometrika*, 106(3):587–604, 2019a. pages 20, 78

D. Cooley and E. Thibaud. Decompositions of dependence for high-dimensional extremes. *Biometrika*, 106(3):587–604, 2019b. pages 18, 39

M. Cornec. *Probability bounds for the cross-validation estimate in the context of the statistical learning theory and statistical models applied to economics and finance*. Thesis, Université de Paris-Nanterre, June 2009. pages 25, 26, 40, 41

M. Cornec. Concentration inequalities of the cross-validation estimator for empirical risk minimizer. *Statistics*, 51(1):43–60, 2017. pages 25, 26, 27, 28, 40, 41, 42, 49, 51, 192

C. Cortes, M. Mohri, and A. Muñoz Medina. Adaptation algorithm and theory based on generalized discrepancy. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 169–178, 2015. pages 37, 155, 157, 159, 162, 163

M. Csorgo, S. Csorgo, L. Horváth, and D. M. Mason. Weighted empirical and quantile processes. *The Annals of Probability*, pages 31–85, 1986. page 97

F. Cucker, S. Smale, et al. Best choices for regularization parameters in learning theory: on the bias-variance problem. *Foundations of computational Mathematics*, 2(4):413–428, 2002. pages 135, 148

A. S. Dalalyan, A. Juditsky, and V. Spokoiny. A new algorithm for estimating the effective dimension-reduction subspace. *The Journal of Machine Learning Research*, 9:1647–1678, 2008. pages 21, 79

M. Darrin, G. Staerman, E. D. C. Gomes, J. C. Cheung, P. Piantanida, and P. Colombo. Unsupervised layer-wise score aggregation for textual ood detection. *arXiv preprint arXiv:2302.09852*, 2023. page 170

A. P. Dawid. Conditional independence in statistical theory. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 1–31, 1979. pages 20, 79

L. de Haan. *On Regular Variation and Its Application to the Weak Convergence of Sample Extremes, by L. de Haan*. Mathematical Centre tracts. 1970. URL https://books.google.fr/books?id=cJ3uswEACAAJ. page 18

L. De Haan and A. Ferreira. *Extreme value theory: an introduction*, volume 21. Springer, 2006. page 40

L. De Haan and A. Ferreira. *Extreme value theory: an introduction*. Springer Science & Business Media, 2007. page 77

M. Delecroix, M. Hristache, and V. Patilea. On semiparametric m-estimation in single-index regression. *Journal of Statistical Planning and Inference*, 136(3):730–769, 2006. pages 21, 79

G. Denevi, C. Ciliberto, R. Grazzi, and M. Pontil. Learning-to-learn stochastic gradient descent with biased regularization. In K. Chaudhuri and R. Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 1566–1575. PMLR, 09–15 Jun 2019. page 157

G. Denevi, M. Pontil, and C. Ciliberto. The advantage of conditional meta-learning for biased regularization and fine tuning. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 964–974. Curran Associates, Inc., 2020. page 157

L. Devroye and T. Wagner. Distribution-free performance bounds for potential function rules. *IEEE Transactions on Information Theory*, 25(5):601–604, 1979. pages 25, 29, 30, 41, 156, 158, 159, 188, 190, 202

S. Dhouib and I. Redko. Revisiting similarity learning for domain adaptation. *Advances in Neural Information Processing Systems*, 31, 2018. pages 36, 156

M. Dredze, J. Blitzer, P. P. Talukdar, K. Ganchev, J. V. Graça, and F. Pereira. Frustratingly hard domain adaptation for dependency parsing. 2007. page 155

H. Drees and A. Sabourin. Principal component analysis for multivariate extremes. *Electronic Journal of Statistics*, 15(1):908–943, 2021. pages 18, 19, 20, 39, 43, 78

S. S. Du, J. Koushik, A. Singh, and B. Póczos. Hypothesis transfer learning via transformation functions. *Advances in neural information processing systems*, 30, 2017. pages 36, 155, 156, 163

C. Dugas, Y. Bengio, F. Bélisle, C. Nadeau, and R. Garcia. Incorporating second-order functional knowledge for better option pricing. *Advances in neural information processing systems*, 13, 2000. page 156

M. L. Eaton. A characterization of spherical distributions. *Journal of Multivariate Analysis*, 20(2):272–276, 1986. pages 21, 80, 82

J. H. Einmahl. Limit theorems for tail processes with application to intermediate quantile estimation. *Journal of Statistical Planning and Inference*, 32(1):137–145, 1992. page 43

J. H. Einmahl and D. M. Mason. Strong limit theorems for weighted quantile processes. *The Annals of Probability*, pages 1623–1643, 1988. page 97

J. H. Einmahl, A. Krajina, and J. Segers. An m-estimator for tail dependence in arbitrary dimensions. *The Annals of Statistics*, 40(3):1764–1793, 2012. page 39

J. H. Einmahl, A. Kiriliouk, A. Krajina, and J. Segers. An m-estimator of spatial tail dependence. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(1):275–298, 2016. page 39

J. H. Einmahl, A. Kiriliouk, and J. Segers. A continuous updating weighted least squares estimator of tail dependence in high dimensions. *Extremes*, 21(2):205–233, 2018. page 39

A. Elgammal and C.-S. Lee. Inferring 3d body pose from silhouettes using activity manifold learning. In *CVPR'04 proceedings*, page 681–688, USA, 2004. IEEE Computer Society. page 106

A. Elisseeff, T. Evgeniou, and M. Pontil. Stability of randomized learning algorithms. *Journal of Machine Learning Research*, 6(3):55–79, 2005. pages 30, 156, 159, 188, 195, 209

C. Elkan. The foundations of cost-sensitive learning. *Proceedings of the Seventeenth International Conference on Artificial Intelligence: 4-10 August 2001; Seattle*, 1, 05 2001a. pages 32, 128

C. Elkan. The foundations of cost-sensitive learning. In *International joint conference on artificial intelligence*, volume 17-1, pages 973–978. Lawrence Erlbaum Associates Ltd, 2001b. pages 128, 140

S. Engelke and A. S. Hitz. Graphical models for extremes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(4):871–932, 2020. pages 20, 78, 84, 93

S. Engelke and J. Ivanovs. Sparse structures for multivariate extremes. *arXiv preprint arXiv:2004.12182*, 2020. pages 20, 78

S. Engelke and J. Ivanovs. Sparse structures for multivariate extremes. *Annual Review of Statistics and Its Application*, 8:241–270, 2021. pages 19, 39

S. Engelke and S. Volgushev. Structure learning for extremal tree models. *Journal of the Royal Statistical Society Series B*, 84(5):2055–2087, 2022. pages 19, 43

S. Engelke, M. Lalancette, and S. Volgushev. Learning extremal graphical structures in high dimensions. *arXiv preprint arXiv:2111.00840*, 2021. pages 19, 20, 43

E. F. Fama and K. R. French. Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics*, 33:3–56, 1993. page 77

E. F. Fama and K. R. French. A five-factor asset pricing model. *Journal of Financial Economics*, 116(1):1–22, 2015. page 77

S. Farkas, O. Lopez, and M. Thomas. Cyber claim analysis using generalized pareto regression trees with applications to insurance. *Insurance: Mathematics and Economics*, 98:92–105, 2021. page 40

V. Feldman and J. Vondrak. High probability generalization bounds for uniformly stable algorithms with nearly optimal rate. In *Conference on Learning Theory*, pages 1270–1279. PMLR, 2019. page 190

J.-D. Fermanian, D. Radulovic, and M. Wegkamp. Weak convergence of empirical copula processes. *Bernoulli*, 10(5):847–860, 10 2004. doi: 10.3150/bj/1099579158. page 95

Y. Freund and R. E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1): 119–139, 1997a. page 202

Y. Freund and R. E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55(1): 119–139, 1997b. pages 36, 156

S. Fu, X. Yu, and Y. Tian. Cost sensitive ν-support vector machine with linex loss. *Information Processing & Management*, 59(2):102809, 2022. page 128

K. Fukumizu, F. R. Bach, and M. I. Jordan. Dimensionality reduction for supervised learning with reproducing kernel hilbert spaces. *Journal of Machine Learning Research*, 5(Jan):73–99, 2004. pages 21, 79

K. Fukumizu, F. R. Bach, M. I. Jordan, et al. Kernel dimension reduction in regression. *The Annals of Statistics*, 37(4):1871–1905, 2009. pages 21, 79

T. Fushiki. Estimation of prediction error by using k-fold cross-validation. *Statistics and Computing*, 21(2):137–146, 2011. pages 24, 29, 187, 189, 197

L. Gardes. Tail dimension reduction for extreme quantile estimation. *Extremes*, 21(1): 57–95, 2018. pages 22, 78, 80, 83, 84, 86, 101, 102, 103, 104, 106, 110

S. Geisser. The predictive sample reuse method with applications. *Journal of the American Statistical Association*, 70(350):320–328, 1975. ISSN 01621459. URL http://www.jstor.org/stable/2285815. page 24

E. Giné and A. Guillou. On consistency of kernel density estimators for randomly censored data: rates holding uniformly over adaptive intervals. In *Annales de l'IHP Probabilités et statistiques*, volume 37, pages 503–522, 2001. pages 47, 49, 61, 62, 129, 140

E. Giné and V. Koltchinskii. Concentration inequalities and asymptotic results for ratio type empirical processes. *The Annals of Probability*, 34(3):1143 – 1216, 2006. doi: 10.1214/009117906000000070. URL https://doi.org/10.1214/009117906000000070. page 136

K. Gkillas and P. Katsiampa. An application of extreme value theory to cryptocurrencies. *Economics Letters*, 164:109–111, 2018. page 40

N. Gnecco, E. M. Terefe, and S. Engelke. Extremal random forests. *arXiv preprint arXiv:2201.12865*, 2022. page 40

N. Goix, A. Sabourin, and S. Clémençon. Learning the dependence structure of rare events: a non-asymptotic study. 05 2015. pages 19, 26, 27, 28, 40, 41, 42, 43, 48, 61, 62, 63, 131

N. Goix, A. Sabourin, and S. Clémençon. Sparse representation of multivariate extremes with applications to anomaly ranking. In *Artificial Intelligence and Statistics*, pages 75–83. PMLR, 2016. pages 19, 20, 39, 78

N. Goix, A. Sabourin, and S. Clémençon. Sparse representation of multivariate extremes with applications to anomaly detection. *Journal of Multivariate Analysis*, 161:12–31, 2017. pages 19, 20, 39, 43, 78

S. Golovanov, R. Kurbanov, S. Nikolenko, K. Truskovskyi, A. Tselousov, and T. Wolf. Large-scale transfer learning for natural language generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6053–6058, 2019. page 170

L. Gyorfi, M. Kohler, A. Krzyzak, and H. Walk. *A Distribution-Free Theory of Nonparametric Regression*. Springer Series in Statistics. Springer New York, 2010. ISBN 9781441929983. URL https://books.google.fr/books?id=_RoFkgAACAAJ. page 25

T. Hagerup and C. Rüb. A guided tour of chernoff bounds. *Information processing letters*, 33(6):305–308, 1990. page 140

P. Hall and K.-C. Li. On almost linearity of low dimensional projections from high dimensional data. *The annals of Statistics*, pages 867–889, 1993. pages 21, 80

W. Härdle and T. M. Stoker. Investigating smooth multiple regression by the method of average derivatives. *Journal of the American statistical Association*, 84(408):986–995, 1989. pages 21, 79

M. Hardt, B. Recht, and Y. Singer. Train faster, generalize better: Stability of stochastic gradient descent. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48, pages 1225–1234. PMLR, 20–22 Jun 2016a. pages 188, 190, 191, 195, 211

M. Hardt, B. Recht, and Y. Singer. Train faster, generalize better: Stability of stochastic gradient descent. In *International conference on machine learning*, pages 1225–1234. PMLR, 2016b. pages 30, 156, 159, 165

D. Haussler. Sphere packing numbers for subsets of the boolean n-cube with bounded vapnik-chervonenkis dimension. *Journal of Combinatorial Theory, Series A*, 69(2):217–232, 1995. ISSN 0097-3165. doi: https://doi.org/10.1016/0097-3165(95)90052-7. URL https://www.sciencedirect.com/science/article/pii/0097316595900527. page 132

A. Hitz and R. Evans. One-component regular variation and graphical modeling of extremes. *Journal of Applied Probability*, 53(3):733–746, 2016. pages 20, 78, 84, 92

D. Homrighausen and D. McDonald. The lasso, persistence, and cross-validation. In *International Conference on Machine Learning*, pages 1031–1039. PMLR, 2013. pages 42, 55

D. Homrighausen and D. J. McDonald. Risk consistency of cross-validation with lasso-type procedures. *Statistica Sinica*, pages 1017–1036, 2017. pages 54, 137

H. Hotelling. The relations of the newer multivariate statistical methods to factor analysis. *British Journal of Statistical Psychology*, 10(2):69–79, 1957. pages 20, 79

M. Hristache, A. Juditsky, J. Polzehl, and V. Spokoiny. Structure adaptive approach for dimension reduction. *Annals of Statistics*, 29(6):1537–1566, 2001. pages 21, 79

H. Jalalzai and R. Leluc. Feature clustering for support identification in extreme regions. In *International Conference on Machine Learning*, pages 4733–4743. PMLR, 2021. pages 19, 40

H. Jalalzai, S. Clémençon, and A. Sabourin. On binary classification in extreme regions. In *NeurIPS*, pages 3096–3104, 2018. pages 19, 40, 43, 44, 45, 47, 48, 63, 72, 73, 131

H. Jalalzai, P. Colombo, C. Clavel, E. Gaussier, G. Varni, E. Vignon, and A. Sabourin. Heavy-tailed representations, text polarity classification & data augmentation. *Advances in Neural Information Processing Systems*, 33, 2020. pages 19, 40

A. Janßen and P. Wan. k-means clustering of extremes. *Electronic Journal of Statistics*, 14(1):1211–1233, 2020a. pages 20, 78

A. Janßen and P. Wan. k-means clustering of extremes. *Electronic Journal of Statistics*, 14(1):1211–1233, 2020b. page 40

R. Jenatton, J.-Y. Audibert, and F. Bach. Structured variable selection with sparsity-inducing norms. *The Journal of Machine Learning Research*, 12:2777–2824, 2011. pages 20, 79

H. A. Jessen and T. Mikosch. Regularly varying functions. *Publications de L'Institut Mathématique*, 80(94):171–192, 2006. page 87

Y. Jiang, D. Cooley, and M. F. Wehner. Principal component analysis for extremes and application to us precipitation. *Journal of Climate*, 33(15):6441–6451, 2020. pages 20, 39, 78

S. Kale, R. Kumar, and S. Vassilvitskii. Cross-validation and mean-square stability. In *In Proceedings of the Second Symposium on Innovations in Computer Science (ICS2011*. Citeseer, 2011. pages 25, 29, 42, 188

M. Kearns and D. Ron. Algorithmic stability and sanity-check bounds for leave-one-out cross-validation. *Neural computation*, 11(6):1427–1453, 1999. pages 25, 26, 27, 28, 40, 41, 42, 49, 51, 68, 156, 188

S. H. Khan, M. Hayat, M. Bennamoun, F. A. Sohel, and R. Togneri. Cost-sensitive learning of deep feature representations from imbalanced data. *IEEE Transactions on Neural Networks and Learning Systems*, 29(8):3573–3587, 2018. doi: 10.1109/TNNLS.2017.2732482. pages 128, 140

M. Khodak, M.-F. F. Balcan, and A. S. Talwalkar. Adaptive gradient-based meta-learning methods. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d' Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. page 157

A. Kiriliouk, H. Rootzén, J. Segers, and J. L. Wadsworth. Peaks over thresholds modeling with multivariate generalized pareto distributions. *Technometrics*, 61(1):123–135, 2019. page 39

Y. Klochkov and N. Zhivotovskiy. Stability and deviation optimal risk bounds with convergence rate $o(1/n)$. *Advances in Neural Information Processing Systems*, 34: 5065–5076, 2021a. page 136

Y. Klochkov and N. Zhivotovskiy. Stability and deviation optimal risk bounds with convergence rate $o(1/n)$. *Advances in Neural Information Processing Systems*, 34: 5065–5076, 2021b. pages 30, 159

T. Koren and K. Levy. Fast rates for exp-concave empirical risk minimization. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015. URL https://proceedings.neurips.cc/paper_files/paper/2015/file/acf4b89d3d503d8252c9c4ba75ddbf6d-Paper.pdf. pages 34, 137

O. O. Koyejo, N. Natarajan, P. K. Ravikumar, and I. S. Dhillon. Consistent binary classification with generalized performance metrics. *Advances in neural information processing systems*, 27, 2014. pages 32, 33, 128, 130, 132, 136, 140

R. Kumar, D. Lokshtanov, S. Vassilvitskii, and A. Vattani. Near-optimal bounds for cross-validation via loss stability. In *International Conference on Machine Learning*, pages 27–35. PMLR, 2013. pages 24, 25, 29, 42, 49, 167, 188, 196

S. Kutin and P. Niyogi. Almost-everywhere algorithmic stability and generalization error. In *Proceedings of the Eighteenth Conference on Uncertainty in Artificial Intelligence*, page 275–282, 2002. pages 156, 188

I. Kuzborskij and F. Orabona. Stability and hypothesis transfer learning. In *International Conference on Machine Learning*, pages 942–950. PMLR, 2013. pages 30, 36, 155, 156, 159, 160, 162, 165, 169

I. Kuzborskij and F. Orabona. Fast rates by transferring from auxiliary hypotheses. *Machine Learning*, 106(2):171–195, 2017. pages 36, 155, 156, 160, 162, 163

P. Laforgue, G. Staerman, and S. Clémençon. Generalization bounds in the presence of outliers: a median-of-means study. In *International Conference on Machine Learning*, pages 5937–5947. PMLR, 2021. page 170

M. Laurer, W. van Atteveldt, A. Casas, and K. Welbers. Less annotating, more classifying: Addressing the data scarcity issue of supervised machine learning with deep transfer learning and bert-nli. *Political Analysis*, page 1–17, 2023. doi: 10.1017/pan.2023.20. page 17

S.-I. Lee, H. Lee, P. Abbeel, and A. Y. Ng. Efficient l˜ 1 regularized logistic regression. In *Aaai*, volume 6, pages 401–408, 2006. pages 54, 137

S. Lhaut, A. Sabourin, and J. Segers. Uniform concentration bounds for frequencies of rare events. *arXiv preprint arXiv:2110.05826, to appear in Statistics and probability Letters*, 2021. pages 19, 26, 27, 28, 40, 41, 42, 43, 49, 62

K.-C. Li. Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, 86(414):316–327, 1991. pages 21, 79, 81, 82, 83

L. Li and X. Yin. Sliced inverse regression with regularizations. *Biometrics*, 64(1): 124–131, 2008. pages 21, 80

X. Li and J. Bilmes. A bayesian divergence prior for classiffier adaptation. In *Artificial Intelligence and Statistics*, pages 275–282. PMLR, 2007. pages 36, 155, 156

R. Liu, Y. Shi, C. Ji, and M. Jia. A survey of sentiment analysis based on transfer learning. *IEEE Access*, 7:85401–85412, 2019. page 155

T. Liu, G. Lugosi, G. Neu, and D. Tao. Algorithmic stability and hypothesis complexity. In *International Conference on Machine Learning*, pages 2159–2167. PMLR, 2017. pages 193, 195

X. Liu, D. Lu, A. Zhang, Q. Liu, and G. Jiang. Data-driven machine learning in environmental pollution: Gains and problems. *Environmental Science & Technology*, 56(4):2124–2133, 2022. doi: 10.1021/acs.est.1c06157. URL https://doi.org/10.1021/acs.est.1c06157. PMID: 35084840. page 17

X.-Y. Liu, J. Wu, and Z.-H. Zhou. Exploratory undersampling for class-imbalance learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 39 (2):539–550, 2009. doi: 10.1109/TSMCB.2008.2007853. page 128

F. M. Longin. From value at risk to stress testing: The extreme value approach. *Journal of Banking and Finance*, 24(7):1097–1130, 2000. ISSN 0378-4266. page 40

G. Lugosi. Pattern classification and learning theory. In *Principles of nonparametric learning*, pages 1–56. Springer, 2002. pages 33, 129, 132

Y. Mansour, M. Mohri, and A. Rostamizadeh. Domain adaptation: Learning bounds and algorithms. *arXiv preprint arXiv:0902.3430*, 2009. pages 155, 157, 162, 163

G. Mariani, F. Scheidegger, R. Istrate, C. Bekas, and C. Malossi. Bagan: Data augmentation with balancing gan. *arXiv preprint arXiv:1803.09655*, 2018. pages 31, 127

P. Massart. *Concentration inequalities and model selection: Ecole d'Eté de Probabilités de Saint-Flour XXXIII-2003*. Springer, 2007. page 198

C. McDiarmid. Concentration. In M. Habib, C. McDiarmid, J. Ramirez-Alfonsen, and B. Reed, editors, *Probabilistic Methods for Algorithmic Discrete Mathematics*, pages 195–248. 1998. pages 28, 42, 49, 55, 58, 202

S. Mendelson. Improving the sample complexity using global data. *IEEE transactions on Information Theory*, 48(7):1977–1991, 2002. page 146

A. Menon, H. Narasimhan, S. Agarwal, and S. Chawla. On the statistical consistency of algorithms for binary classification under class imbalance. In S. Dasgupta and D. McAllester, editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 603–611, Atlanta, Georgia, USA, 17–19 Jun 2013a. PMLR. pages 32, 33, 128, 130, 132, 136, 140

A. Menon, H. Narasimhan, S. Agarwal, and S. Chawla. On the statistical consistency of algorithms for binary classification under class imbalance. In *ICML proceedings*, pages 603–611. PMLR, 2013b. page 85

A. K. Menon and R. C. Williamson. The cost of fairness in binary classification. In *Conference on Fairness, accountability and transparency*, pages 107–118. PMLR, 2018. page 128

N. Meyer and O. Wintenberger. Sparse regular variation. *arXiv preprint arXiv:1907.00686*, 2019. pages 20, 78

E. Morvant, A. Habrard, and S. Ayache. Parsimonious unsupervised and semi-supervised domain adaptation with good similarity functions. *Knowledge and Information Systems*, 33(2):309–349, 2012. pages 36, 156

Y. Nesterov et al. *Lectures on convex optimization*, volume 137. Springer, 2018. page 173

F. Orabona, C. Castellini, B. Caputo, A. E. Fiorilla, and G. Sandini. Model adaptation with least-squares svm for adaptive hand prosthetics. In *2009 IEEE International Conference on Robotics and Automation*, pages 2897–2903. IEEE, 2009. pages 36, 155, 160

B. Pandeya, W. Buytaert, Z. Zulkafli, T. Karpouzoglou, F. Mao, and D. Hannah. A comparative analysis of ecosystem services valuation approaches for application at the local scale and in data scarce regions. *Ecosystem Services*, 22:250–259, 2016. ISSN 2212-0416. doi: https://doi.org/10.1016/j.ecoser.2016.10.015. URL https://www.sciencedirect.com/science/article/pii/S2212041616304259. Integrated valuation of ecosystem services: challenges and solutions. page 17

Y. Pathak, P. Shukla, A. Tiwari, S. Stalin, S. Singh, and P. Shukla. Deep transfer learning based classification model for covid-19 disease. *IRBM*, 43(2):87–92, 2022. ISSN 1959-0318. doi: https://doi.org/10.1016/j.irbm.2020.05.003. URL https://www.sciencedirect.com/science/article/pii/S1959031820300993. pages 128, 140

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Mach. Learn. Res.*, 12:2825–2830, 2011. page 106

M. Perrot and A. Habrard. A theoretical analysis of metric hypothesis transfer learning. In *International Conference on Machine Learning*, pages 1708–1717. PMLR, 2015. pages 36, 155, 156

M. Picot, F. Granese, G. Staerman, M. Romanelli, F. Messina, P. Piantanida, and P. Colombo. A halfspace-mass depth-based method for adversarial attack detection. *Transactions on Machine Learning Research*, 2023. page 170

V. Plassier, F. Portier, and J. Segers. Risk bounds when learning infinitely many response functions by ordinary linear regression. In *Annales de l'Institut Henri Poincare (B) Probabilites et statistiques*, volume 59-1, pages 53–78. Institut Henri Poincaré, 2023. pages 34, 132, 140, 142

F. Portier. An empirical process view of inverse regression. *Scandinavian Journal of Statistics*, 43(3):827–844, 2016. pages 21, 80, 95, 99, 101

F. Portier. Nearest neighbor process: weak convergence and non-asymptotic bound, 2021. pages 143, 144

F. Portier and B. Delyon. Optimal transformation: A new approach for covering the central subspace. *Journal of Multivariate Analysis*, 115:84–107, 2013. page 81

F. Portier and B. Delyon. Bootstrap testing of the rank of a matrix via least-squared constrained estimation. *Journal of the American Statistical Association*, 109(505): 160–172, 2014. page 101

S. I. Resnick. *Heavy-tail phenomena: probabilistic and statistical modeling*. Springer Science & Business Media, 2007. pages 18, 77

S. I. Resnick. *Extreme values, regular variation and point processes*. Springer, 2013. pages 77, 131

R. Rifkin, G. Yeo, T. Poggio, et al. Regularized least-squares classification. *Nato Science Series Sub Series III Computer and Systems Sciences*, 190:131–154, 2003. page 156

W. H. Rogers and T. J. Wagner. A finite sample distribution-free performance bound for local discrimination rules. *The Annals of Statistics*, 6(3), May 1978. page 41

S. T. Roweis and L. K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, 2000. ISSN 0036-8075. doi: 10.1126/science.290.5500.2323. URL https://science.sciencemag.org/content/290/5500/2323. page 106

E. M. Rudd, L. P. Jain, W. J. Scheirer, and T. E. Boult. The extreme value machine. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(3):762–768, 2018. page 40

S. Ruder, M. E. Peters, S. Swayamdipta, and T. Wolf. Transfer learning in natural language processing. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: Tutorials*, pages 15–18, 2019. page 155

D. Ryu, J.-I. Jang, and J. Baik. A transfer cost-sensitive boosting approach for cross-project defect prediction. *Software Quality Journal*, 25:235–272, 2017. page 128

B. Schölkopf, R. Herbrich, and A. J. Smola. A generalized representer theorem. In *International conference on computational learning theory*, pages 416–426. Springer, 2001. page 160

C. Scott. Calibrated asymmetric surrogate losses. *Electronic Journal of Statistics*, 6 (none):958 – 992, 2012. doi: 10.1214/12-EJS699. URL https://doi.org/10.1214/12-EJS699. page 128

J. Segers. Hybrid copula estimators. *Journal of Statistical Planning and Inference*, 160: 23–34, 2015. page 117

A. Shafahi, P. Saadatpanah, C. Zhu, A. Ghiasi, C. Studer, D. Jacobs, and T. Goldstein. Adversarially robust transfer learning. In *8th International Conference on Learning Representations (ICLR 2020)(virtual)*. International Conference on Learning Representations, 2020. page 170

J. Shao. Linear model selection by cross-validation. *Journal of the American Statistical Association*, 88(422):486–494, 1993. page 24

J. Shao. An asymptotic theory for linear model selection. *Statistica Sinica*, 7(2):221–242, 1997. pages 24, 25, 29, 187, 188

A. Siffer, P.-A. Fouque, A. Termier, and C. Largouet. Anomaly detection in streams with extreme value theory. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1067–1075, 2017. page 40

E. S. Simpson, J. L. Wadsworth, and J. A. Tawn. Determining the dependence structure of multivariate extremes. *Biometrika*, 107(3):513–532, 2020. pages 20, 78

G. Staerman, P. Mozharovskyi, S. Clémen, et al. The area of the convex hull of sampled curves: a robust functional statistical depth measure. In *International Conference on Artificial Intelligence and Statistics*, pages 570–579. PMLR, 2020. page 170

G. Staerman, P. Laforgue, P. Mozharovskyi, and F. d'Alché Buc. When ot meets mom: Robust estimation of wasserstein distance. In A. Banerjee and K. Fukumizu, editors, *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 136–144. PMLR, 13–15 Apr 2021a. URL https://proceedings.mlr.press/v130/staerman21a.html. page 131

G. Staerman, P. Laforgue, P. Mozharovskyi, and F. d'Alché Buc. When ot meets mom: Robust estimation of wasserstein distance. In *International Conference on Artificial Intelligence and Statistics*, pages 136–144. PMLR, 2021b. page 170

G. Staerman, P. Mozharovskyi, P. Colombo, S. Clémençon, and F. d'Alché Buc. A pseudo-metric between probability distributions based on depth-trimmed regions. *arXiv preprint arXiv:2103.12711*, 2021c. page 170

G. Staerman, E. Adjakossa, P. Mozharovskyi, V. Hofer, J. Sen Gupta, and S. Clémençon. Functional anomaly detection: a benchmark study. *International Journal of Data Science and Analytics*, pages 1–17, 2022a. page 170

G. Staerman, C. Allain, A. Gramfort, and T. Moreau. Fadin: Fast discretized inference for hawkes processes with general parametric kernels. *arXiv preprint arXiv:2210.04635*, 2022b. page 170

M. Stone. Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society. Series B (Methodological)*, 36(2):111–147, 1974. pages 24, 187

S. Suboh and I. A. Aziz. Anomaly detection with machine learning in the presence of extreme value - a review paper. In *2020 IEEE Conference on Big Data and Analytics (ICBDA)*, pages 66–72, 2020. pages 19, 40

Y. Sun, M. S. Kamel, A. K. Wong, and Y. Wang. Cost-sensitive boosting for classification of imbalanced data. *Pattern Recognition*, 40(12):3358–3378, 2007. ISSN 0031-3203. doi: https://doi.org/10.1016/j.patcog.2007.04.009. URL https://www.sciencedirect.com/science/article/pii/S0031320307001835. page 128

J. Tanha, Y. Abdi, N. Samadi, N. Razzaghi, and M. Asadpour. Boosting methods for multi-class imbalanced data classification: an experimental review. *Journal of Big Data*, 7(1), Sept. 2020. doi: 10.1186/s40537-020-00349-y. URL https://doi.org/10.1186/s40537-020-00349-y. page 136

J. B. Tenenbaum, V. d. Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000. ISSN 0036-8075. doi: 10.1126/science.290.5500.2319. URL https://science.sciencemag.org/content/290/5500/2319. page 106

A. Thomas, S. Clémençon, A. Gramfort, and A. Sabourin. Anomaly detection in extreme regions via empirical mv-sets on the sphere. In *Artificial Intelligence and Statistics*, pages 1011–1019. PMLR, 2017. pages 19, 40, 43

B. Thompson. *Canonical correlation analysis: Uses and interpretation*. Number 47. Sage, 1984. pages 20, 79

R. J. Tibshirani and J. Taylor. The solution path of the generalized lasso. *The annals of statistics*, 39(3):1335–1371, 2011. page 212

I. Triguero, M. Galar, S. Vluymans, C. Cornelis, H. Bustince, F. Herrera, and Y. Saeys. Evolutionary undersampling for imbalanced big data classification. In *2015 IEEE Congress on Evolutionary Computation (CEC)*, pages 715–722, 2015. doi: 10.1109/CEC.2015.7256961. page 128

M. J. van der Laan, S. Dudoit, and S. Keles. Asymptotic optimality of likelihood-based cross-validation. *Statistical Applications in Genetics and Molecular Biology*, 3 (1):1–23, Jan. 2004. doi: 10.2202/1544-6115.1036. URL https://doi.org/10.2202/1544-6115.1036. page 25

A. W. Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 1998. page 92

A. W. Van Der Vaart and J. A. Wellner. Weak convergence. In *Weak convergence and empirical processes*, pages 16–28. Springer, 1996. pages 97, 117, 118

A. W. van der Vaart, S. Dudoit, and M. J. van der Laan. Oracle inequalities for multi-fold cross validation:. *Statistics and Decisions*, 24(3):351–371, 2006. page 41

T. van Erven, P. D. Grünwald, N. A. Mehta, M. D. Reid, and R. C. Williamson. Fast rates in statistical and online learning. *Journal of Machine Learning Research*, 16(54):1793–1861, 2015. URL http://jmlr.org/papers/v16/vanerven15a.html. pages 34, 137

R. Van Handel. Probability in high dimension. Technical report, PRINCETON UNIV NJ, 2014. page 146

V. N. Vapnik and A. Y. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability & Its Applications*, 16(2): 264–280, 1971. doi: 10.1137/1116025. URL https://doi.org/10.1137/1116025. page 132

J. Velthoen, C. Dombry, J.-J. Cai, and S. Engelke. Gradient boosting for extreme quantile regression. *arXiv preprint arXiv:2103.00808*, 2021. page 40

R. Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018. pages 50, 56, 171

S. Viaene and G. Dedene. Cost-sensitive learning and decision making revisited. *European Journal of Operational Research*, 166(1):212–220, 2005. ISSN 0377-2217. doi: https://doi.org/10.1016/j.ejor.2004.03.031. URL https://www.sciencedirect.com/science/article/pii/S0377221704002978. Metaheuristics and Worst-Case Guarantee Algorithms: Relations, Provable Properties and Applications. page 128

E. Vignotto and S. Engelke. Extreme value theory for anomaly detection–the gpd classifier. *Extremes*, 23(4):501–520, 2020. page 40

M. Vlachos, C. Domeniconi, D. Gunopulos, G. Kollios, and N. Koudas. Non-linear dimensionality reduction techniques for classification and visualization. In *ACM SIGKDD proceedings*, pages 645–651, 2002. page 106

R. Vogel, M. Achab, S. Clémencon, and C. Tillier. Weighted empirical risk minimization: Transfer learning based on importance sampling. In *28th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, pages 515–520. i6doc. com, 2020. pages 33, 128

S. Wager. Cross-validation, risk estimation, and model selection: Comment on a paper by rosset and tibshirani. *Journal of the American Statistical Association*, 115(529): 157–160, 2020. pages 25, 41, 42, 187

S. Wang, W. Liu, J. Wu, L. Cao, Q. Meng, and P. J. Kennedy. Training deep neural networks on imbalanced data sets. In *2016 International Joint Conference on Neural Networks (IJCNN)*, pages 4368–4374, 2016. doi: 10.1109/IJCNN.2016.7727770. pages 128, 140

S. Wang, W. Zhou, H. Lu, A. Maleki, and V. Mirrokni. Approximate leave-one-out for fast parameter tuning in high dimensions. In J. Dy and A. Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 5228–5237. PMLR, 10–15 Jul 2018. page 158

X. Wang, H. H. Zhang, and Y. Wu. Multiclass probability estimation with support vector machines. *Journal of Computational and Graphical Statistics*, 28(3):586–595, 2019a. doi: 10.1080/10618600.2019.1585260. URL https://doi.org/10.1080/10618600.2019.1585260. pages 128, 140

Z. Wang, Z. Dai, B. Póczos, and J. Carbonell. Characterizing and avoiding negative transfer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11293–11302, 2019b. page 167

K. Weiss, T. M. Khoshgoftaar, and D. Wang. A survey of transfer learning. *Journal of Big data*, 3(1):1–40, 2016. page 167

R. S. Wenocur and R. M. Dudley. Some special vapnik-chervonenkis classes. *Discrete Mathematics*, 33(3):313–318, 1981. pages 134, 144

A. Wibisono, L. Rosasco, and T. Poggio. Sufficient conditions for uniform stability of regularization algorithms. 2009a. pages 188, 190, 193

A. Wibisono, L. Rosasco, and T. Poggio. Sufficient conditions for uniform stability of regularization algorithms. *Computer Science and Artificial Intelligence Laboratory Technical Report, MIT-CSAIL-TR-2009-060*, 2009b. pages 156, 161, 162, 167

H. Wold. Estimation of principal components and related models by iterative least squares. *Multivariate analysis*, pages 391–420, 1966. pages 20, 79

H.-M. Wu. Kernel sliced inverse regression with applications to classification. *Journal of Computational and Graphical Statistics*, 17(3):590–610, 2008. pages 21, 80

Y. Xia et al. A constructive approach to the estimation of dimension reduction directions. *The Annals of Statistics*, 35(6):2654–2690, 2007. pages 21, 79

P. Xiong, Y. Chi, S. Zhu, H. J. Moon, C. Pu, and H. Hacgümüş. Smartsla: Cost-sensitive management of virtualized resources for cpu-bound database services. *IEEE Transactions on Parallel and Distributed Systems*, 26(5):1441–1451, 2015. doi: 10.1109/TPDS.2014.2319095. page 128

H. Xu, C. Caramanis, and S. Mannor. Outlier-robust pca: The high-dimensional case. *IEEE transactions on information theory*, 59(1):546–572, 2012. page 131

N. Xu, T. C. Fisher, and J. Hong. Rademacher upper bounds for cross-validation errors with an application to the lasso. *arXiv preprint*, 2020a. pages 33, 42

Z. Xu, C. Dan, J. Khim, and P. Ravikumar. Class-weighted classification: Trade-offs and robust approaches. In *ICML proceedings*, pages 10544–10554. PMLR, 2020b. page 85

Z. Xu, C. Dan, J. Khim, and P. Ravikumar. Class-weighted classification: Trade-offs and robust approaches. In *International Conference on Machine Learning*, pages 10544–10554. PMLR, 2020c. pages 34, 128, 129, 132, 136, 140

Y. Yang. Comparing learning methods for classification. *Statistica Sinica*, 16(2):635–657, 2006. pages 24, 187, 196

Y. Yang. Consistency of cross validation for comparing regression procedures. *The Annals of Statistics*, 35(6):2450 – 2473, 2007. pages 187, 196

Y.-R. Yeh, S.-Y. Huang, and Y.-J. Lee. Nonlinear dimension reduction with kernel sliced inverse regression. *IEEE transactions on Knowledge and Data Engineering*, 21 (11):1590–1603, 2008. pages 21, 80

C. Zhang, L. Zhang, and J. Ye. Generalization bounds for domain adaptation. *Advances in neural information processing systems*, 25, 2012. pages 37, 155, 157, 159

T. Zhang. Statistical behavior and consistency of classification methods based on convex risk minimization. *The Annals of Statistics*, 32(1):56–85, 2004a. pages 36, 156, 161, 162, 167, 168, 172, 188, 193

T. Zhang. Statistical behavior and consistency of classification methods based on convex risk minimization. *The Annals of Statistics*, 32(1):56–85, 2004b. pages 45, 74, 130

Y. Zhang, T. Liu, M. Long, and M. Jordan. Bridging theory and algorithm for domain adaptation. In *International Conference on Machine Learning*, pages 7404–7413. PMLR, 2019. pages 37, 155, 157, 159

Y. Zhang, W. Zhang, S. Bald, V. Pingali, C. Chen, and M. Goswami. Stability of sgd: Tightness analysis and improved bounds. In *Uncertainty in Artificial Intelligence*, pages 2364–2373. PMLR, 2022. pages 195, 212

L.-P. Zhu, L.-X. Zhu, and Z.-H. Feng. Dimension reduction in regressions through cumulative slicing estimation. *Journal of the American Statistical Association*, 105 (492):1455–1466, 2010. pages 21, 79, 82, 83, 99

L. Zwald and G. Blanchard. On the convergence of eigenspaces in kernel principal component analysis. *NeurIPS proceedings*, 18, 2005. page 101

ECOLE
DOCTORALE
DE MATHEMATIQUES
HADAMARD

Titre : Analyse statistique des algorithmes dédiés aux événements rares

Mots clés : Théorie des valeurs extrêmes, théorie de l'apprentissage statistique, bornes d'erreur non asymptotiques, validation croisée, apprentissage par transfert

Résumé : Cette thèse se concentre sur l'établissement de garanties statistiques pour l'efficacité des algorithmes d'apprentissage automatique dans des environnements pauvres en données, en particulier dans les contextes d'analyse des valeurs extrêmes, d'apprentissage par transfert et de classification déséquilibrée. Nous développons des bornes supérieures de probabilité qui servent de garanties théoriques pour l'efficacité des algorithmes adaptés à ces scénarios spécifiques. Notre approche commence par une critique des méthodes statistiques actuelles dans des contextes limités en données. Nous identifions les limitations dans les cadres existants et introduisons de nouvelles bornes de probabilité spécifiquement conçues pour fournir des garanties de performance d'algorithme sous contrainte de données. Ces bornes ne sont pas seulement rigoureuses sur le plan théorique, mais sont également directement applicables aux défis pratiques de l'apprentissage automatique. Nous validons nos résultats théoriques avec des études empiriques dans chacun des trois domaines ciblés. Les résultats confirment que nos bornes dérivées sont efficaces pour certifier l'efficacité des algorithmes dans la gestion des valeurs extrêmes, le transfert de connaissances dans des domaines de données éparses et la classification de jeux de données déséquilibrés. En conclusion, la thèse fait progresser le domaine de l'apprentissage statistique en four- nissant des garanties théoriques précises pour la performance des algorithmes dans des situations pauvres en données. Ce travail est particulièrement pertinent pour les appli- cations où il est critique de faire des inférences précises avec des données limitées.

Title : A statistical analysis of algorithms dedicated for rare events

Keywords : Cross Validation, statistical learning theory, non-asymptotic error bounds, extreme value theory, transfer learning

Abstract : This thesis focuses on establishing statistical guarantees for the efficiency of machine learning algorithms in data-scarce environments, particularly within the contexts of extreme value analysis, transfer learning, and imbalanced classification. We develop probability upper bounds that serve as theoretical assurances for the effectiveness of algorithms tailored to these specific scenarios. Our approach begins with a critique of current statistical methods in data-limited set- tings. We identify limitations in existing frameworks and introduce new probability bounds that are specifically designed to provide guarantees for algorithm performance under data scarcity. These bounds are not just theoretically rigorous but are also di- rectly applicable to practical machine learning challenges. We validate our theoretical findings with empirical studies in each of the three focused areas. The results confirm that our derived bounds are effective in certifying the effi- ciency of algorithms in handling extreme values, transferring knowledge in sparse data domains, and classifying imbalanced datasets. Conclusively, the thesis advances the field of statistical learning by providing precise theoretical guarantees for the performance of algorithms in data-scarce situations. This work is particularly relevant for applications where making accurate inferences with limited data is critical.

Institut Polytechnique de Paris
91120 Palaiseau, France