Pontifícia Universidade Católica
do Rio de Janeiro

Cláudio Cardoso Flores

Essays in Econometrics: Online Learning in High-Dimensional Contexts and Treatment Effects with Complex and Unknown Assignment Rules

Tese de Doutorado

Thesis presented to the Programa de Pós–graduação em Economia, do Departamento de Economia da PUC-Rio in partial fulfillment of the requirements for the degree of Doutor em Economia.

Advisor: Prof. Marcelo Cunha Medeiros

Rio de Janeiro
March 2021

Pontifícia Universidade Católica
do Rio de Janeiro

Cláudio Cardoso Flores

Essays in Econometrics: Online Learning in High-Dimensional Contexts and Treatment Effects with Complex and Unknown Assignment Rules

Thesis presented to the Programa de Pós–graduação em Economia da PUC-Rio in partial fulfillment of the requirements for the degree of Doutor em Economia. Approved by the Examination Committee:

Prof. Marcelo Cunha Medeiros
Advisor
Departamento de Economia – PUC-Rio

Prof. Bruno Ferman
Escola de Economia de São Paulo - EESP/FGV

Prof. Eduardo Fonseca Mendes
Escola de Matemática Aplicada - EMAp/FGV

Prof. Marcelo Fernandes
Escola de Economia de São Paulo - EESP/FGV

Prof. Ricardo Pereira Masini
Escola de Economia de São Paulo - EESP/FGV

Prof. Pedro Carvalho Loureiro de Souza
Department of Economics - Warwick University

Rio de Janeiro, March the 24th, 2021

Cláudio Cardoso Flores

Graduated in Chemical Engineering at Universidade Federal do Rio de Janeiro (UFRJ) and obtained his Master of Science degree in Administration from Instituto Coppead de Administração da Universidade Federal do Rio de Janeiro (COPPEAD-UFRJ). Also obtained a Master degree in Mathematical Finance from Instituto de Matemática Pura e Aplicada (IMPA). Now holds a PhD degree in Economics from PUC-Rio.

Bibliographic data

Flores, Cláudio Cardoso

Essays in Econometrics: Online Learning in High-Dimensional Contexts and Treatment Effects with Complex and Unknown Assignment Rules / Cláudio Cardoso Flores; advisor: Marcelo Cunha Medeiros. – 2021.

106 f: il. color. ; 30 cm

Tese (doutorado) - Pontifícia Universidade Católica do Rio de Janeiro, Departamento de Economia, 2021.

Inclui bibliografia

1. Economia – Teses. 2. Economia – Teses. 3. Aprendizado Online. 4. Bandit. 5. Lasso. 6. Aprendizado por Máquina. 7. Regressão Discontínua. 8. Efeitos de Tratamento. 9. Regras de Alocação Desconhecidas. 10. Árvores de Classificação. 11. Florestas Aleatórias. I. Medeiros, Marcelo Cunha. II. Pontifícia Universidade Católica do Rio de Janeiro. Departamento de Economia. III. Título.

CDD: 004

To my son.

Acknowledgments

I would like to express my sincere gratitude to my advisor, Prof. Marcelo Medeiros, for his guidence, friendship and continuous support during the last three years. His knowledge and experience in high-dimensional econometrics in unquestionable and, without his help this work would not be possible.

Besides the excellent professors at PUC-Rio, I would also like to thank to my classmates, especially to my good friends João Veloso and Gustavo Pinto. Without them, certainly my trajectory at PUC-Rio would have been much harder.

I gratefully acknowledge Banco Central do Brasil for the financial support during the course.

I would like to thank to my family, specially my wife and my father (*in memorian*), for their love and support in all periods of my life. Also thanks for your patience and understanding when I was absent in the last years. Finally, I would like to thank to my son for being the source of motivation, inspiration, love and everything good in my life.

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001.

Abstract

Flores, Cláudio Cardoso; Medeiros, Marcelo Cunha (Advisor). **Essays in Econometrics: Online Learning in High-Dimensional Contexts and Treatment Effects with Complex and Unknown Assignment Rules**. Rio de Janeiro, 2021. 106p. Tese de Doutorado – Departamento de Economia, Pontifícia Universidade Católica do Rio de Janeiro.

Sequential learning problems are common in several fields of research and practical applications. Examples include dynamic pricing and assortment, design of auctions and incentives and permeate a large number of sequential treatment experiments. In this essay, we extend one of the most popular learning solutions, the ϵ_t-greedy heuristics, to high-dimensional contexts considering a conservative directive. We do this by allocating part of the time the original rule uses to adopt completely new actions to a more focused search in a restrictive set of promising actions. The resulting rule might be useful for practical applications that still values surprises, although at a decreasing rate, while also has restrictions on the adoption of unusual actions. With high probability, we find reasonable bounds for the cumulative regret of a conservative high-dimensional decaying ϵ_t-greedy rule. Also, we provide a lower bound for the cardinality of the set of viable actions that implies in an improved regret bound for the conservative version when compared to its non-conservative counterpart. Additionally, we show that end-users have sufficient flexibility when establishing how much safety they want, since it can be tuned without impacting theoretical properties. We illustrate our proposal both in a simulation exercise and using a real dataset. The second essay studies deterministic treatment effects when the assignment rule is both more complex than traditional ones and unknown to the public perhaps, among many possible causes, due to ethical reasons, to avoid data manipulation or unnecessary competition. More specifically, sticking to the well-known sharp RDD methodology, we circumvent the lack of knowledge of true cutoffs by employing a forest of classification trees which also uses sequential learning, as in the last essay, to guarantee that, asymptotically, the true unknown assignment rule is correctly identified. The tree structure also turns out to be suitable if the program's rule is more sophisticated than traditional univariate ones. Motivated by real world examples, we show in this essay that, with high probability and based on reasonable assumptions, it is possible to consistently estimate treatment effects under this setup. For practical implementation we propose an algorithm that not only sheds light on the previously unknown assignment rule but also is capable to robustly estimate treatment effects regarding different specifications

imputed by end-users. Moreover, we exemplify the benefits of our methodology by employing it on part of the Chilean P900 school assistance program, which proves to be suitable for our framework.

Keywords

Online Learning; Bandit; Lasso; Machine Learning; Regression Discontinuity Design; Assignment Rules; Classification Trees; Random Forest.

Resumo

Flores, Cláudio Cardoso; Medeiros, Marcelo Cunha. **Estudos em Econometria: Aprendizado Online em Ambientes de Alta Dimensão e Efeitos de Tratamento com Regras de Alocação Complexas e Desconhecidas**. Rio de Janeiro, 2021. 106p. Tese de Doutorado – Departamento de Economia, Pontifícia Universidade Católica do Rio de Janeiro.

Essa tese é composta por dois capítulos. O primeiro deles refere-se ao problema de aprendizado sequencial, útil em diversos campos de pesquisa e aplicações práticas. Exemplos incluem problemas de apreçamento dinâmico, desenhos de leilões e de incentivos, além de programas e tratamentos sequenciais. Neste capítulo, propomos a extensão de uma das mais populares regras de aprendizado, ϵ_t-greedy, para contextos de alta-dimensão, levando em consideração uma diretriz conservadora. Em particular, nossa proposta consiste em alocar parte do tempo que a regra original utiliza na adoção de ações completamente novas em uma busca focada em um conjunto restrito de ações promissoras. A regra resultante pode ser útil para aplicações práticas nas quais existem restrições suaves à adoção de ações não-usuais, mas que eventualmente, valorize surpresas positivas, ainda que a uma taxa decrescente. Como parte dos resultados, encontramos limites plausíveis, com alta probabilidade, para o remorso cumulativo para a regra ϵ_t-greedy conservadora em alta-dimensão. Também, mostramos a existência de um limite inferior para a cardinalidade do conjunto de ações viáveis que implica em um limite superior menor para o remorso da regra conservadora, comparativamente a sua versão não-conservadora. Adicionalmente, usuários finais possuem suficiente flexibilidade em estabelecer o nível de segurança que desejam, uma vez que tal nível não impacta as propriedades teóricas da regra de aprendizado proposta. Ilustramos nossa proposta tanto por meio de simulação, quanto por meio de um exercício utilizando base de dados de um problema real de sistemas de classificação. Por sua vez, no segundo capítulo, investigamos efeitos de tratamento determinísticos quando a regra de aloção é complexa e desconhecida, talvez por razões éticas, ou para evitar manipulação ou competição desnecessária. Mais especificamente, com foco na metodologia de regressão discontínua *sharp*, superamos a falta de conhecimento de pontos de corte na alocação de unidades, pela implementação de uma floresta de árvores de classificação, que também utiliza aprendizado sequencial na sua construção, para garantir que, assintoticamente, as regras de alocação desconhecidas sejam identificadas corretamente. A estrutura de árvore também é útil nos casos em que a regra de alocação desconhecida é mais

complexa que as tradicionais univariadas. Motivado por exemplos da vida prática, nós mostramos nesse capítulo que, com alta probabilidade e baseado em premissas razoáveis, é possível estimar consistentemente os efeitos de tratamento sob esse cenário. Propomos ainda um algoritmo útil para usuários finais que se mostrou robusto para diferentes especificações e que revela com relativa confiança a regra de alocação anteriormente desconhecida. Ainda, exemplificamos os benefícios da metodologia proposta pela sua aplicação em parte do P900, um programa governamental Chileno de suporte para escolas, que se mostrou adequado ao cenário aqui estudado.

Palavras-chave

Aprendizado Online; Bandit; Lasso; Aprendizado por Máquina; Regressão Discontínua; Efeitos de Tratamento; Regras de Alocação Desconhecidas; Árvores de Classificação; Florestas Aleatórias.

Table of contents

List of figures

List of tables

1
Online Action Learning in High Dimensions: A Conservative Perspective

1.1
Introduction

In this chapter we propose modifications to the original ϵ_t-greedy heuristics to work with high-dimensional contexts considering a more conservative perspective. Our framework can be especially useful for practical applications where an agent uses the experience and the repeated observation of a large pool of covariates to conservatively learn the best course of action relatively to some reward.

More specifically, consider a simple example in consumer behavior where few accidental discoveries may have a positive impact on users experience but, as time passes, they tend to increasingly remain loyal to a set of similar vendors. An intuitive example relates to the case where a consumer may be impressed after visiting a completely new restaurant but, after some disappointments, she becomes more and more reluctant to accept suggestions outside her set of preferred restaurants. Other examples from different markets could be provided as well. Now consider a recommendation system that wants to employ a learning rule looking to explore profitable opportunities in this example. One feature that stands out from the above-mentioned pattern is the fact that a proper learning rule should not stop to suggest completely new restaurants (maybe at random) since, eventually, a pleasant discovery can be made. However, it should progressively discourage randomness and stimulate exploitation of not the best action but in a set of preferred similar vendors. We understand such rule under the conservative philosophy introduced in Wu et al. (2016) since, for the algorithm to be useful, exploration of new actions should be done with caution.

The original ϵ_t-greedy, described in Auer et al. (2002), seems to be a valid substrate to be used in setups similar to the example above, since the selection of actions fully at random is already performed at a decreasing rate (ϵ_t). However, it exploits only the best empirical action, which has some drawbacks, such as to be time consuming in contexts when the difference in payoff between

the best and second bests strategies is small (Wu et al. (2016)). In our example, it would also be meaningless to suggest only the best vendor. Therefore, while remaining loyal to the ϵ_t-greedy philosophy, in this chapter we augment it with a new exploitation option that brings more safety to its decision-making.

1.1.1 Motivation and Comparison with the Literature

A multiarmed bandit problem can be interpreted as a sequential problem, where a limited set of resources must be allocated between alternative choices to maximize utility. The properties of the choices are not fully known at the time of allocation and may become better understood as time passes, provided a learning rule with theoretical guarantees is available. A particularly useful extension of the bandit problem is called the contextual multiarmed bandit problem, where observed covariates yield important information to the learning process in the sense that the supposed best policy may be predicted; see, for instance, Auer (2003), Li et al. (2010) and Langford and Zhang (2008).

Contextual multiarmed bandit problems have applications in various areas. For instance, large online retailers must decide on real-time prices for products and differentiate among distinct areas without complete demand information; see, for example, den Boer (2015). Arriving customers may take purchase decisions among offered products based on maximizing their utility. If information on consumers' utility is not available, the seller could learn which subset of products to offer (Saure and Zeevi, 2013). Further, the reserve price of auctions could be better designed to maximize revenue (Cesa-Bianchi et al., 2013). Mechanisms design in the case where agents may not know their true value functions but the mechanism is repeated for multiple rounds can take advantage of accumulated experience (Kandasamy et al., 2020). Also, sequential experiments or programs, including public policies (Tran-Thanh, 2010, devises an algorithm that consider costly policies), may be assigned under the scope of learning problems. In this regard, excellent works can be found in Kock and Thyrsgaard (2017), Kock et al. (2018) and Kock et al. (2020).

Designing a sequence of actions to minimize error is a difficult task and, for a considerable period in the past, was also a computationally intractable goal. In this respect, several heuristics with well-behaved properties have emerged in the literature, such as Thompson sampling (Agrawal and Goyal, 2012; Russo and van Roy, 2016), upper confidence bounds (Abbasi-Yadkori et al., 2011) and greedy algorithms (Auer, 2003; Bastani et al., 2020; Goldenshluger and Zeevi, 2013).

It is very well documented the recent growing availability of large datasets

and their usage for different sectors of society. One of the drivers of this recent popularity can be assigned to shrinkage estimators applied to sparse setups, relative to their potential to catalyze the benefits of huge information sets into strong predictive power. This superior performance is certainly useful for learning problems based on the observation of large pools of covariates, but there is not an extensive list of papers providing theoretical properties of such bandits. Among few others, we cite Carpentier and Munos (2012), Abbasi-Yadkori et al. (2012), Deshpande and Montanari (2012), Bouneffouf et al. (2017), Bastani and Bayati (2020), Kim and Paik (2019) and Krishnamurthy and Athey (2020).

Our work is mostly related to Bastani and Bayati (2020) and to Kim and Paik (2019). Both papers provide theoretical properties for algorithms that are similar to the high-dimensional ϵ_t-greedy we provide in this chapter. Minor differences appear as, for example, the fact that exploration at random in the algorithm studied in Bastani and Bayati (2020) is performed at pre-determined specific instants of time and its frequency does not decrease as time passes. The later is also present in Kim and Paik (2019). However, our work completely distinguishes from these papers by adapting the high-dimensional ϵ_t-greedy to work in a more conservative fashion. In this regard, one can understand our work as an extension of Bastani and Bayati (2020) and Kim and Paik (2019) in directions that respect, in a flexible way, restrictions and particularities imposed by practical applications.

1.1.2 Main Takeaways

We add to the high-dimensional bandit literature, by showing that distinct levels of restrictions in the exploration of new actions can be settled by using variations of the original multiarmed ϵ_t-greedy heuristic. We first expand the ϵ_t-greedy rule to high-dimensional contexts leading to an algorithm that is similar to the ones used in Kim and Paik (2019) and Bastani and Bayati (2020). Then, we combine it with a competing exploitation mechanism that comprises, at each time step, a number of actions that lead to the best predicted rewards. We call it as the order statistics searching set.

From the empirical perspective, we provide an algorithm that can be used to implement the main rule proposed in this chapter. In general terms, it is equipped with an initialization phase where information about parameters is gathered by attempting distinct actions and, after that, the rule properly said starts with the main exloration-exploitation phase. In a simulation study, we show its robustness with respect to a reasonable range for the variables

imputed by end-users.

From the theoretical point of view, we show that the cumulative regret of the conservative high-dimensional ϵ_t-greedy algorithm is reasonably bounded. Aside the benefit that an order statistics searching set can be less time consuming when rewards are very close to each other (Wu et al., 2016), we also show in this chapter that even with separable payoffs (in probability), there are conditions related to the amount of viable actions that being conservative leads to a stricter bound on regret.

We prove that the cumulative regret of the proposed algorithm is sub-linearly bounded, respecting an upper bound growing at no more than $\mathcal{O}\{s_0\sqrt{Tlog(2p)}\}$, and seems to be an improvement on the bound found in Kim and Paik (2019) on a similar non-conservative algorithm ($\mathcal{O}\{s_0 log(pT)\sqrt{T}\}$). The work in Bastani and Bayati (2020) found an upper bound of $\mathcal{O}\{s_0^2(log(T)+log(p))^2\}$ which seems to have a worse dependence on s_0 than ours. When the order statistics searching set is considered as an alternative exploration mechanism, in addition to the benefits already mentioned regarding harmful actions, we show that the $\sqrt{T}$-growing rate above mentioned can be reduced by a factor of $\mathcal{O}\{log(T)\}$.

In addition, we show that it is viable to pick any cardinality for the order statistics searching set and still respect the theoretical limits established in this chapter. Recall an important trade-off: under the conservative approach one should be caution to explore new actions but, a learning rule should explore to be accurate in the long run. In this sense, allowing end-users to choose any cardinality for the order statistics searching set is the same as letting them to choose the "size" of safety, tuning the algorithm to the specifics of the environment/market they are inserted.

The algorithm proposed in this chapter outperforms simple and adapted (to the high-dimensional context) counterparts in a simulation exercise, while seems to be effective, presenting good learning properties, when applied to a a real recommendation system dataset.

1.1.3 Organization of this chapter

The rest of this chapter is structured as follows. Section 1.2 establishes the framework and the main assumptions for the regret analysis, while Section 1.3 depicts the main algorithm. In section 1.4 we derive the theoretical results of the methodology proposed and in section 1.5 we provide a sensitivity analysis of the algorithm with relation to parameters set by the user and a comparison among simple and adapted algorithms. Section 1.6 exhibits the performance of

our proposed learning rule when applied on a real recommendation system data set. Section 1.7 concludes this work. All proofs are relegated to the Appendix.

1.1.4 Notation

Regarding the notation used in this chapter, we provide in this subsection general guidelines. Definitions and particularities are presented throughout the chapter. Bold capital letters $\boldsymbol{X}$ represent matrices, small bold letters $\boldsymbol{x}$ represent vectors and small standard letters x represent scalars. Matrices or vectors followed by subscript or superscript parentheses denote specific elements. For example, $\boldsymbol{X}^{(j)}$ is the j-th column of matrix $\boldsymbol{X}$, $\boldsymbol{X}_{(i,j)}$ is the (i,j) element of $\boldsymbol{X}$, while $\boldsymbol{x}_{(j)}$ is the j-th scalar element of vector $\boldsymbol{x}$. Let M be an arbitrary vector space. The symbol $\|\cdot\|$ is the usual vector norm on M, while $\mathcal{B}(\boldsymbol{x_0}, \tau)$ is the ball defined in M around a point $\boldsymbol{x}_0$, the set $\{\boldsymbol{x} \in M | d(\boldsymbol{x}, \boldsymbol{x}_0) \leq \tau\}$. Let Y be an arbitrary set. Then, $\#Y$ is used to represent its cardinality, while $\lfloor\cdot\rfloor$ and $\mathbb{1}\{y \in Y\}$ are the traditional floor and indicator function, respectively, the later taking the value of 1 when $y \in Y$.

1.2 Setup and Assumptions

Consider an institution, a central planner or a firm, in this chapter simply called the planner, that wants to maximize some variable (reward). In order to do that, it has to choose at each instant of time $t \in \mathcal{T} \equiv \{1, 2, \ldots, T\}$ an action (arm) among some alternatives inside a finite set $\mathcal{W} \equiv \{\boldsymbol{\omega}_0, \ldots, \boldsymbol{\omega}_{w-1}\}$. We consider each $\boldsymbol{\omega}_k \in \mathbb{R}^g$, $k \in \{0, w-1\}$, $g > 0$ arbitrary.

Define the action function $I : \mathcal{T} \to \mathcal{W}$, such that for each $t \in \mathcal{T}$, $I(t) = \boldsymbol{\omega}_k$ informs that at time t the action selected by the planner was $\boldsymbol{\omega}_k$. Let $\vartheta \in \mathcal{T}$ and define $\mathcal{A}_{k\vartheta} \equiv \{t \in \mathcal{T} | I(t) = \boldsymbol{\omega}_k, t < \vartheta\}$ to be the set that stores all values of $t < \vartheta$ when the action $\boldsymbol{\omega}_k$ was adopted and let $n_{k\vartheta} \equiv \#\mathcal{A}_{k\vartheta}$ to be its cardinality.

Let $(\boldsymbol{\Omega}, \mathcal{F}, \mathbb{P})$ to be a probability space. When choosing actions, the planner also observes covariates $\boldsymbol{x}_t$ at each time step, e.g., individual characteristics of its target sample, as well as the sequence of its past realizations which we consider to be identically and independently distributed (iid) draws from $\mathbb{P}$. It also knows the past rewards[1] $\{y_{kt}\}_{t \in \mathcal{A}_{kt}}$, only when $\boldsymbol{\omega}_k$ was implemented before t. The range[2] of y_{kt} is a subset of $\mathcal{Y} \subset \mathbb{R}$, while that of $\boldsymbol{x}_t$ is a subset

[1] At time t, the planner observes $\boldsymbol{x}_t$ but does not yet know y_{kt}.

[2] For ease of notation, in our setup, y_t is a scalar random variable, but the reader will recognize throughout this chapter that this choice is not restrictive. Multivariate versions are allowed.

of $\mathcal{X} \subset \mathbb{R}^p$, where p may grow with the sample size. To simplify notation, in the rest of this chapter we do not exhibit this dependence (between p and t) explicitly.

The connection between covariates and rewards is stated as follows:

Assumption 1 (Contextual Linear Bandit) *There is a linear relationship between rewards and covariates of the form:*

$$y_{kt} = \boldsymbol{\beta}_k' \boldsymbol{x}_t + \epsilon_{kt}, \tag{1-1}$$

where ϵ_{kt} is an idiosyncratic shock and $\forall k$, $\boldsymbol{\beta}_k$ belongs to the parametric space $\mathcal{C} \subset \mathbb{R}^p$. Furthermore:

i. $\forall t \in \mathcal{T}$, $|\boldsymbol{x}_{t,(j)}| \leq \theta_x$, $j \in \{1, ..., p\}$.

ii. $\forall k \in \{0, ..., w-1\}, t \in \mathcal{T}$, *the sequence* $\{\epsilon_{kt}\}$ *is composed of independent centered random variables with variance* $\mathbb{E}(\epsilon_{kt}^2) < \sigma^2$.

Remark 1 *Assumption 1 restrains our setup to linear bandit problems. Rewards are action/time-dependent, in the sense that not only the dynamics of $\boldsymbol{x}_t$ impacts the level of rewards but, depending on the chosen policy $\boldsymbol{\omega}_k$, the mechanism ($\boldsymbol{\beta}_k$) that "links" covariates to rewards is different. Moreover, we require covariates to be bounded in absolute terms and a sequence of errors, with finite variance. Both assumptions are easier to be satisfied in most practical applications and are necessary to guarantee that instantaneous regrets (defined below) do not have explosive behavior.*

Two pieces of nomenclature have been used: actions chosen by the planner and "mechanisms" (parameters) through which these actions operate impacting rewards. Assumption 2 connects them.

Assumption 2 (Metric Spaces) *There is an h-Lipschitz continuous function $f : \mathcal{W} \to \mathcal{C}$, such that $\forall k \in \{0, ..., w-1\}$, $f(\boldsymbol{\omega}_k) = \boldsymbol{\beta}_k$.*

Remark 2 *Assumption 2 is a restriction on the joint behavior of the two relevant metric spaces we are working with, the set of actions and the parametric space. Its main purpose is to avoid the possibility that small changes in actions could result in too different mechanisms, which would not be reasonable in several practical situations. In the case considered by Assumption 2 we have that for two actions $\boldsymbol{\omega}_{k_1}$, $\boldsymbol{\omega}_{k_2}$, both belonging to $\mathcal{W}$, $d_{\mathcal{C}}(\boldsymbol{\beta}_{k_1} - \boldsymbol{\beta}_{k_2}) \leq h d_{\mathcal{W}}(\boldsymbol{\omega}_{k_1} - \boldsymbol{\omega}_{k_2})$, for $h \in \mathbb{R}^+$ the Lipschitz constant and $d_{\mathcal{C}}$ and $d_{\mathcal{B}}$, relevant metrics for the two spaces.*

One of the most useful instruments to assess the effectiveness of online learning algorithms is the regret function, which, in general, may be studied in its instantaneous or cumulative version. In general, a regret function represents the difference (in expected value) between the reward obtained by choosing an arbitrary action $\boldsymbol{\omega}_j \in \mathcal{W}$ and the one that would be obtained if the best action were adopted. Clearly, the term best action does not have an absolute meaning, but relative to alternatives that belong to the same set of actions. A sequential rule is not learning at all to choose actions if its cumulative regret grows linearly or at a more convex rate ($R_t \geq \mathcal{O}\{T\}$). Good algorithms achieve, for example, $R_t \leq \mathcal{O}\{\sqrt{T}\}$ (Wu et al., 2016). Definition 1 formalizes these concepts.

Definition 1 *(Regret Functions) The instantaneous (r_t) regret function of implementing any policy $\boldsymbol{\omega}_j \in \mathcal{W}$ at time $t \in \mathcal{T}$, leading to the reward y_{jt}, and the respective cumulative (R_T) regret until time T are defined as:*

$$r_t = \mathbb{E}\left[\max_{k\in\{0,\ldots,w-1\}} (y_{kt} - y_{jt})\right] \quad \text{and} \quad R_T = \sum_{t=1}^{T-1} r_t$$

Motivated by the high-dimensional context, we perform Lasso estimation in the following sections. This estimator operates on the well-known sparsity assumption, i.e., that in the true model, not all covariates are relevant to explain a given dependent variable. Regarding this aspect, we define the sparsity index in Definition 2 and impose the compatibility condition for random matrices in the Assumption 3, which is standard in the high-dimensional literature.

Definition 2 *(Sparsity Index) For any $p > 0$ and $k \in \{1, \ldots, p\}$, define $S_0 \equiv \{k|\beta_k \neq 0\}$ and the sparsity index as $s_0 = \#S_0$.*

Assumption 3 *(Compatibility Condition) Define $\beta_k[S_0] \equiv \beta_k \mathbb{1}_{\{k\in S_0\}}$ and $\beta_k[S_0^c] \equiv \beta_k \mathbb{1}_{\{k\notin S_0\}}$. For an arbitrary $(n \times p)$-matrix $\boldsymbol{X}$ and $\forall \boldsymbol{\beta} \in \mathbb{R}^p$, such that $\|\boldsymbol{\beta}[S_0^c]\|_1 \leq 3\|\boldsymbol{\beta}[S_0]\|_1$, for some S_0, $\exists\, \phi_0 > \sqrt{32bs_0} > 0$, with $b \geq \max_{j,k}|(\widehat{\boldsymbol{\Sigma}})_{j,k} - (\boldsymbol{\Sigma})_{j,k}|$ such that:*

$$\|\boldsymbol{\beta}[S_0]\|_1^2 \leq \frac{s_0 \boldsymbol{\beta}' \boldsymbol{\Sigma} \boldsymbol{\beta}}{\phi_0^2},$$

where $\widehat{\boldsymbol{\Sigma}}$ and $\boldsymbol{\Sigma}$ are the empirical and population covariance matrices of $\boldsymbol{X}$, respectively.

Finally, we impose a bounding condition for the density of covariates near a decision boundary, as in Tsybakov (2004), Goldenshluger and Zeevi (2013) and Bastani and Bayati (2020), among others.

Assumption 4 *(Margin Condition) Consider the Lasso penalty parameter chosen at t,* $\lambda_t \in [\lambda_{min}, \lambda_{max}]$*. For* $\delta \in \mathbb{R}^+$*,* $\exists C_m \in \mathbb{R}^+$*,* $C_m \leq \frac{\phi_0^2}{8\theta_x s_0 \lambda_{min}}$*, such that for any* $k_1, k_2 \in \{0, \ldots, w-1\}$*:*

$$\mathbb{P}\left[\boldsymbol{x}_t'(\boldsymbol{\beta}_{k_1} - \boldsymbol{\beta}_{k_2}) \leq \delta\right] \leq C_m \delta$$

Remark 3 *Assumption 4 is related to the behavior of the distribution of* $\boldsymbol{x}_t$ *"near" a decision boundary. In these cases, there is a possibility for rewards to be so similar that small deviations in estimation procedures could lead to suboptimal policies being selected by the algorithms. With this assumption, we impose some separability (in probability) for the rewards, in the sense that there is a strictly positive probability that the reward* $(\boldsymbol{x}_t'\boldsymbol{\beta}_{k_1})$ *for a given policy* $\boldsymbol{\omega}_{k_1} \in \mathcal{W}$ *is strictly greater than that of any other policy* $\boldsymbol{\omega}_{k_2} \in \mathcal{W}$*. In contrast to other papers, we establish an upper bound for the constant* C_m *as a function of the intrinsic parameters of the problem.*

1.3 Algorithms and Estimation Procedures

One of the most important feature of every learning rule relates to the way it sequentially selects actions. In general, at each time step, an algorithm should decide between: exploit and adopt the action that leads to the most profitable reward, in a predicted sense in the case of contextual learning, or explore and implement a different one, according to some criteria. The expected outcome of exploitation is to reduce regret by adopting actions that empirically have been outperforming the available alternatives. However, besides the fact that the future may eventually not reflect the past, eroding the intrinsic value of past good actions, if an algorithm does not explore it simply does not discover good actions that have not been sufficiently tested in the past. As a consequence the learning rule could remain stuck, for long periods of time, on suboptimal actions (best only in the past). This exploitation-exploration trade-off is well-known in the learning literature and dictates the properties of the regret function. See, for example, Auer (2003) and Langford and Zhang (2008).

The vast majority of learning algorithms take the above mentioned problem into consideration while pursuing the main goal of a "well-behaved"

bound for their regret functions. The ϵ_t-greedy heuristic is no different. In the way it is presented in Auer et al. (2002), first one should define a decaying exploration weight ϵ_t, for example the one suggested by the authors, $\epsilon_t \equiv \min\left\{1, \frac{cw}{d^2 t}\right\}$, $c > 0$ and $0 < d < 1$. Then, at each time step, the rule should explore with probability ϵ_t and choose a random action inside the set of possible actions, $I(t) = \boldsymbol{\omega}_{a_t}$, a_t drawn from $\mathsf{U}(0, w)$. With probability $1 - \epsilon_t$ it exploits choosing the action that leads to the best empirical average reward, $I(t) = \boldsymbol{\omega}_{e_t}$, $e_t = \underset{j \in \{0,\ldots,w-1\}}{\arg\max} \frac{1}{t-1} \sum_{i=1}^{t-1} y_{ji}$.

The above-defined rule is appropriate for multiarmed bandits (without context). To extend it to cases where covariates play an important role, one should simply change the criterion for exploitation and replace the best empirical average reward for the best predicted reward, that is, $I(t) = \boldsymbol{\omega}_{e_t}$, but now, $e_t = \underset{k \in \{0,\ldots,w-1\}}{\arg\max}\ \widehat{y}_{kt}$ where, considering Assumption 1, $\widehat{y}_{kt} = \widehat{\beta}_k \boldsymbol{x}_t$. We compute each $\widehat{\beta}_k$ considering a high-dimensional context and we call the resulting learning rule as the HD ϵ_t-Greedy.

The algorithm updates $\widehat{\boldsymbol{\beta}}_k$ considering available information when $\boldsymbol{\omega}_k$ was implemented in the past. Specifically, at an arbitrary instant of time $\vartheta \in \mathcal{T}$, let $\boldsymbol{X}_{k\vartheta}$ to be an $n_{k\vartheta} \times p$ matrix containing observed covariates at instants of time t, provided that $t \in \mathcal{A}_{k\vartheta}$. Likewise $\boldsymbol{y}_{k\vartheta}$ and $\boldsymbol{\epsilon}_{k\vartheta}$ are the $n_{k\vartheta} \times 1$ observed rewards and error terms, respectively. Then, we update $\widehat{\boldsymbol{\beta}}_k$ by Lasso:

$$\widehat{\boldsymbol{\beta}}_k = \underset{\boldsymbol{\beta} \in \mathcal{C}}{\arg\min} \frac{1}{n_{k\vartheta}} \left\| \boldsymbol{y}_{k\vartheta} - \boldsymbol{X}_{k\vartheta} \boldsymbol{\beta} \right\|_2^2 + \lambda \left\| \boldsymbol{\beta} \right\|_1, \tag{1-2}$$

where λ is the lasso penalty parameter. Following the suggestions in Kim and Paik (2019) and in Bastani and Bayati (2020), as already introduced in Assumption 4, we consider λ to be time-dependent (λ_t) and bounded by $\lambda_{min} \leq \lambda_t \leq \lambda_{max}$ (see Corolary 1 for further details).

Without previous knowledge, in order to have initial estimates of each $\widehat{\boldsymbol{\beta}}_k$ at the beginning of the learning problem, we impose an initialization phase to the HD ϵ_t-Greedy, which consists to try every action, observe the respective outcomes and estimate parameters according to equation (1-2). We require it to last vw, which implies that every action in $\mathcal{W}$ is implemented v times, $v \in \mathbb{N}^+$. The initialization phase is also present in the conservative version of the HD ϵ_t-Greedy.

The properties of similar versions of the HD ϵ_t-Greedy were investigated in Bastani and Bayati (2020) and Kim and Paik (2019). In the former, the authors employ the forced-sampling exploration, which prescribes, in a deterministic way, a set of times when each action must be adopted. Although it may be similar to our initialization phase, the HD ϵ_t-Greedy

does not explore actions at fixed instants of time, but in a ex-ante unknown frequency, remaining literal to the learning rule defined in Auer et al. (2002). Consequently, comparing the post-initialization (most important) phase of the HD ϵ_t-Greedy with the algorithm in Bastani and Bayati (2020) we see that, differently from the forced-sampling, exploration in our algorithm may occur at a low (high) rate, depending on a future unknown sequence of trials. Theoretical properties of HD ϵ_t-Greedy are, in some sense, a generalization of those provided in Bastani and Bayati (2020), regarding different exploration regimes. Also, the HD ϵ_t-Greedy adopts a decreasing weight for exploration, which seems to be not used both in Bastani and Bayati (2020) and in Kim and Paik (2019). We do not deeply investigate the possible impacts that both these differences may have on the performance of algorithms, because the HD ϵ_t-Greedy is treated in this chapter as a benchmark for its conservative variant (our main contribution). However, we infer that they may not exert great influence, since the theoretical properties of the HD ϵ_t-Greedy are comparable and, in some cases, better than the results obtained by the cited authors (See Section 1.4 for further details).

The conservative version of the above described algorithm, called in this chapter as CHD ϵ_t-Greedy, inserts a competing exploitation mechanism in the HD ϵ_t-Greedy learning rule that comprises, at each time step, a number of actions that lead to the best predicted rewards. We call it as the order statistics searching set.

Recall the standard definition of order statistics, which for the case of predicted rewards computed at each time step considering all w possible actions, takes the form:

$$\widehat{y}_{(0:w-1)t} \equiv \min_{k\in\{0,...,w-1\}} \widehat{y}_{kt} \leq \widehat{y}_{(2:w-1)t} \leq \cdots \leq \widehat{y}_{(w-1:w-1)t} \equiv \max_{k\in\{0,...,w-1\}} \widehat{y}_{kt}$$

For $k \in \{0, ..., w-1\}$, let $\mathcal{H}_t^{(\kappa_t)} \equiv \{\widehat{y}_{kt} | \widehat{y}_{kt} \geq \widehat{y}_{(w-1-\kappa_t:w-1)t}\}$ be the set of the κ_t best predictions at time t which we consider as new option for exploitation, such that $\forall t > vw$, $\kappa_t \equiv \#\mathcal{H}_t^{(\kappa_t)}$. $\kappa_t \in (1, \lfloor w/2 \rfloor]$ to avoid extremes. In fact, if for some t, $\kappa_t = 1$, the overall effect would be to increase the weight to exploit the action with the best estimated reward, and we would be under the scope of the (non-conservative) HD ϵ_t-Greedy algorithm. On the other hand, when κ_t is higher, possibly close to w, the learning rule would be encouraging random exploration and, again the non-conservative version would be dictating the learning properties. In this sense, the bounds on κ_t serve the purpose to guarantee that the resulting learning rule is more conservative than its precursor. Definition 3 presents the CHD ϵ_t-Greedy algorithm.

Definition 3 (CHD ϵ_t-Greedy Algorithm) *Let $c > 0$, $0 < d < 1$, $w > 1$ and ϵ_t be defined as in Auer et al. (2002), $\epsilon_t \equiv \min\left\{1, \frac{cw}{d^2 t}\right\}$. Let $v \in \mathbb{N}^+$, $1 < \kappa_t \leq \lfloor w/2 \rfloor$ and $s_t \in (0,1)$ be the weight for the conservative exploitation. Then, the CHD ϵ_t-Greedy algorithm is:*

Algorithm 1: CLG-κHOS algorithm

input parameters: c, d, w, v, κ_t, s_t
Initialization;
for $i \in \{1, 2, ..., v\}$ **do**
  for $j \in \{1, 2, ..., w\}$ **do**
    $I(t) \longleftarrow \boldsymbol{\omega}_j$;
    Update $\widehat{\boldsymbol{\beta}}_j$;
  end
end
Exploration-Exploitation;
for $t > vw$ **do**
  $\epsilon_t \longleftarrow \min\left\{1, \frac{cw}{d^2 t}\right\}$; $q_t \longleftarrow \mathsf{U}(0,1)$; $r_t \longleftarrow \mathsf{U}(0,1)$;
  if $q_t \leq \epsilon_t$ **then**
    if $r_t \leq s_t$ **then**
      Build $\mathcal{H}_t^{(\kappa_t)}$;
      $u_t \longleftarrow \mathsf{U}(0, \kappa_t)$; $I(t) \longleftarrow \boldsymbol{\omega}_{u_t}$ *in* $\mathcal{H}_t^{(\kappa_t)}$;
      Update $\widehat{\boldsymbol{\beta}}_{u_t}$;
    else
      $a_t \longleftarrow \mathsf{U}(0, w)$; $I(t) \longleftarrow \boldsymbol{\omega}_{a_t}$;
      Update $\widehat{\boldsymbol{\beta}}_{a_t}$;
    end
  else
    $b_t \longleftarrow \underset{j \in \{0,...,w-1\}}{\arg\max}\ \widehat{y}_{jt}$; $I(t) \longleftarrow \boldsymbol{\omega}_{b_t}$;
    Update $\widehat{\boldsymbol{\beta}}_{b_t}$;
  end
end

1.4 Finite Sample Properties of Regret Functions

In this section we present the main contributions of this work in the form of two theorems. Both proofs are developed in the Appendix, as well as the proofs of related Auxiliary Lemmas. More specifically, the proof strategy is as follows: Lemmas 3 and 4 establish the finite-sample properties of the parameters estimated by Lasso in equation (1-2), taking into consideration the framework proposed, while Lemmas 5, 6 and 7 provide theoretical properties for the cumulative regret in the initialization phase and for the instantaneous regret of both HD and CHD ϵ_t-Greedy algorithms, respectively. Theorem 1 is a compilation of the above results and is the main contribution of the chapter,

as it provides the bound for the cumulative regret functions of the CHD ϵ_t-Greedy algorithm. Theorem 2 extends the main result and provides conditions that guarantees the conservative version to be the best course of action.

Theorem 1 (Cumulative Regret of CHD ϵ_t-Greedy algorithms) *Provided that the conditions required by Lemmas 5, 6, 7 in the Appendix are satisfied, at least with probability* $1 - P_{\beta_{max}}$, $P_{\beta_{max}} \equiv \max_{vw<\vartheta<T} P_{\beta_\vartheta}$, *for* $s_\vartheta \equiv s$ *imputed by end-user, the cumulative regret until time* T *of the CHD* ϵ_t*-Greedy learning rule can be bounded as:*

$$R_{T-1}^{CHD} \leq R_{T-1}^{HD} + w\theta_x h\tau_{\mathcal{W}} \left[vs \log\left(\frac{T-1}{vw}\right) \left(w \exp\left\{ -\frac{2}{w} \left[\left[w\left(1 - P_{\beta_\vartheta}\right) - \mathcal{X}_\vartheta \right]^2 \right\} - 1 \right) \right]$$
$$= \mathcal{O}\{s_0\sqrt{Tlog(2p)}\}.$$

where P_{β_ϑ}, $\mathcal{X}_\vartheta$ *and* C_m *are provided in Lemmas 4, 7 and Assumption 4, respectively.*

Notice in the proof of Theorem 1 that the cumulative regret of the HD ϵ_t-Greedy algorithm is the most important term in the bound of its conservative version. This fact is a recognition that the second term of R_{T-1}^{CHD} does not grow at a faster rate than the first one. As already mentioned, the suggested upper bound growing at no more than $\mathcal{O}\{s_0\sqrt{Tlog(2p)}\}$ seems to be an improvement on the bound found of a non-conservative similar algorithm in Kim and Paik (2019) ($\mathcal{O}\{s_0 log(pT)\sqrt{T}\}$). The work in Bastani and Bayati (2020) found an upper bound of $\mathcal{O}\{s_0^2(log(T) + log(p))^2\}$ which has a worse dependence on s_0 than ours.

Theorem 2 (Flexibility and Dominance of CHD ϵ_t-Greedy algorithm) *Provided that the conditions required by Lemmas 6 and 7 are satisfied, the upper bound for the CHD* ϵ_t*-Greedy algorithm does not depend on* κ_ϑ *and, at least with probability* $1 - P_{\beta_{max}}$, *for an increasing sequence* $\{\lambda_\vartheta\}_{\vartheta>vw}$, *if* $w \geq (12 + 2\sqrt{2})\mathcal{X}_{max}$*:*

$$\sup_{\vartheta \in \mathcal{T} \cap \{\vartheta | \vartheta > vw\}} r_\vartheta^{CHD} < \sup_{\vartheta \in \mathcal{T} \cap \{\vartheta | \vartheta > vw\}} r_\vartheta^{HD},$$

where P_{β_ϑ} *is defined in Lemma 4,* r_ϑ^{CHD} *is provided in Lemma 6, while* r_ϑ^{HD} *and* $\mathcal{X}_{max}$ *are defined in Lemma 7, where* $\mathcal{X}_{max}$ *is the usual* $\mathcal{X}_\vartheta$ *plugged with* λ_{max}*.*

Theorem 2 represents our additional contribution to the high-dimensional online learning literature by providing supplemental guarantees for practitioners with mild restrictions in exploration of new actions. In these cases, limitations imposed by practical applications naturally bound exploration to be

confined in a restrictive, possibly time-varying, set of actions. In these cases, it would be preferable to have some flexibility in the action screening process without impacting the algorithm's properties. Theorem 2 can be helpful in this regard since it provides the flexibility to explore groups of different sizes according to the users' needs. Additionally, it suggests that this approach (being conservative) would not only be advisable (by operational limitations in real applications), but also the best course of action in terms of stricter bounds, provided that the set of viable actions is sufficiently large. In these cases the bound of the HD ϵ_t-Greedy can be reduced by a factor that grows at no more than $\mathcal{O}\{log(T)\}$ (See Theorem 1 for negative second term on R_{T-1}^{CHD}).

1.5 Simulations and Sensitivity Analysis

Choosing any policy at each instant of time generates the well-known problem of bandit feedback, which in general terms, relates to the fact that a planner following an arbitrary algorithm obtains feedback for only the chosen action. Rewards from the adoption of other actions are not observable and the best possible one, at each time t, remains unknown to the planner, which impairs the online evaluation of regrets. Also, this characteristic can lead to incorrect premature conclusions, for example, in cases when a action had not been frequently tested in the past. In this case, it may be labeled as suboptimal, while in fact, it simply did not have sufficient opportunity to prove itself. In general, bandit feedback poses serious problems for the evaluation of different learning rules and comparison of algorithms using real data sets. If a target action, different than the implemented one, is to be evaluated, difficulties arise, leading, for example, to alternatives such as counterfactual estimation (Agarwal et al., 2017). In this section we circumvent this problem by analyzing the properties of the CHD ϵ_t-Greedy algorithm in a simulated exercise.

First we evaluate the sensitivity of the algorithm with respect to changes in: (1) the number of available actions, w; (2) the weight attributed for the exploitation in the higher-order statistics searching set, s; and (3) the cardinality of the higher-order statistics searching set, κ. Notice that, since the CHD ϵ_t-Greedy algorithm inherits most of its characteristics from the non-conservative version, it would be expected for the HD version to present similar behavior, at least for changes in w.[3] Second, we also compare both CHD and the HD algorithms with a few related alternatives.

General Setup: We set $T = 2000$; covariates $\boldsymbol{x}_t$ are generated from a truncated Gaussian distribution such that Assumption 1.i translates to

[3]Recall that the HD ϵ_t-Greedy algorithm does not count with s or κ.

$|\boldsymbol{x}_{t(j)}| \leq 1$, $\forall t$. The dimension of $\boldsymbol{x}_t$ is $p = 200$, and the sparsity index is $s_0 = 5$. $\epsilon_{kt} \sim \mathsf{N}(0, 0.05)$, $\forall k \in \{0, \ldots, w-1\}$ and $\forall t$. We consider $v = 30$ as the number of times that each action is implemented in the initialization phase.[4] Each action $\boldsymbol{\omega}_k$ has its own parameter $\boldsymbol{\beta}_k$ drawn independently from a $\mathsf{U}(0, 1)$ probability distribution. The simulation is repeated $n_{sim} = 50$ times, and the results are presented as the average regret. That is, the instantaneous regret at a specific time t is the average of 50 simulated instantaneous regrets at this same time.

Sensitivity to w: We set $w \in \{5, 10, 15\}$, $\kappa_t \equiv \kappa = 2$ and $s_t \equiv s = 0.2$.

From the proof of Theorem 1, one can verify that the cumulative regret function is increasing in w. This can be attributed to the specifics of our framework, since the higher the value of w, the longer the initialization phase, implying that the sub-linear growth of the exploration vs. exploitation phase bound would take longer to operate. Consequently, the levels of cumulative regret increase with w. Recall, however, that low values of w may not guarantee that the CHD ϵ_t-Greedy algorithm respects a stricter bound than its non-conservative alternative. These arguments are illustrated in Figure 2.7.

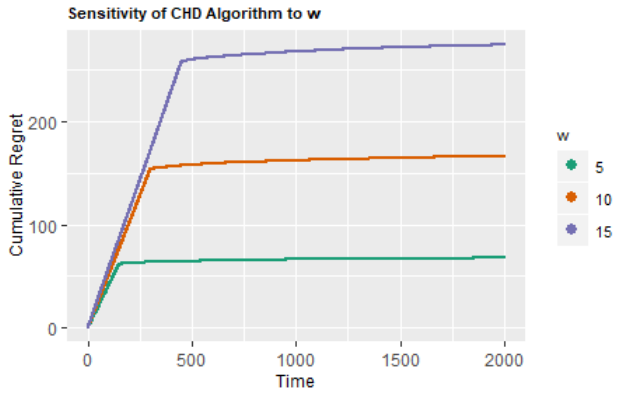

Figure 1.1: Comparison of Cumulative Regrets of the CHD ϵ_t-Greedy algorithm for values of $w \in \{5, 10, 15\}$, $s_t \equiv s = 0.2$ and $\kappa_t \equiv \kappa = 2$.

Sensitivity to s_t: Figure 2.8 illustrates that the performance of CHD ϵ_t-Greedy algorithm is highly robust to small variations in s and Figure 2.9 is just an amplification of Figure 2.8 for $t > vw$. Simulations are conducted for $w = 10$, $\kappa_t \equiv \kappa = 2$ and $s_t \equiv s \in \{0.05, 0.1, 0.015\}$.

Sensitivity to κ_t: Figures 1.4 and 1.5 present the sensitivity of the algorithm to values of $\kappa_t \equiv \kappa \in \{2, 3, 5\}$, $w = 10$ and $s_t \equiv s = 0.2$. The first panel comprises all time steps, and the second is for $t > vw$. The results can

[4]We do not explicitly test the sensitivity of the algorithm to v since, given our specification, this variable affects only the duration of the initialization phase and the precision of the parameters estimates right after this stage. For the first, we use w, since the duration of the initialization phase is set to vw.

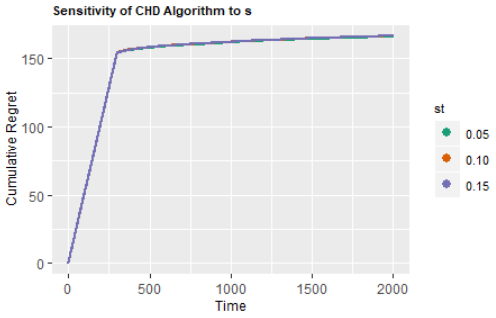

Figure 1.2: Comparison of Cumulative Regrets of the CHD ϵ_t-Greedy algorithm for values of $w = 10$, $s_t \equiv s \in \{0.05, 0.01, 0.015\}$ and $\kappa_t \equiv \kappa = 2$.

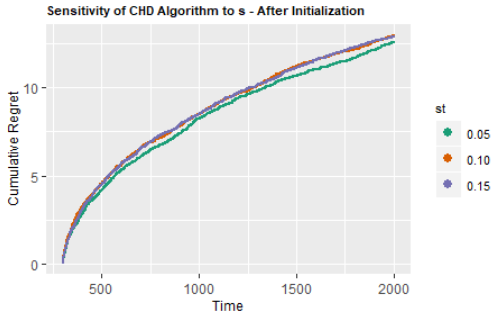

Figure 1.3: Comparison of Cumulative Regrets of the CHD ϵ_t-Greedy algorithm, from $t = vw + 1$ to $t = T$, for values of $w = 10$, $s_t \equiv s \in \{0.05, 0.01, 0.015\}$ and $\kappa_t \equiv \kappa = 2$.

be associated to the first part of Theorem 2 that implies that bounds are not κ_t-dependent. In fact, these figures present the simulated performance of the CHD ϵ_t-Greedy algorithm and not its bounds for different values of κ_t but, in a similar fashion, observe in Figure 1.5 that cumulative regrets intersect each other and none of the curves dominate the others for the entire period tested.

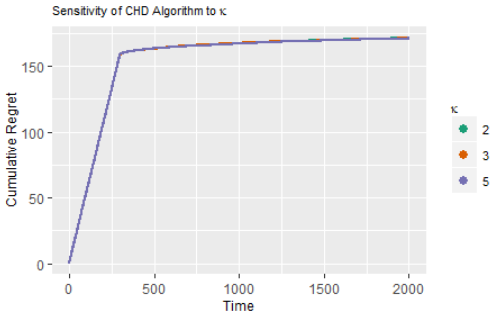

Figure 1.4: Comparison of Cumulative Regrets of the CHD ϵ_t-Greedy algorithm for values of $w = 10$, $s_t \equiv s = 0.2$ and $\kappa_t \equiv \kappa \in \{2, 3, 5\}$.

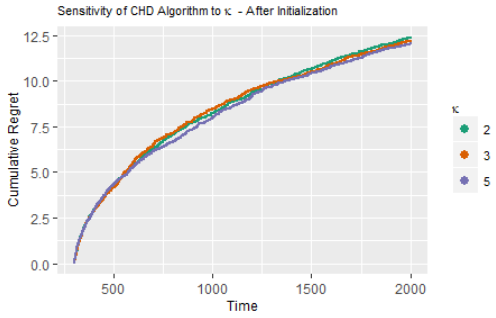

Figure 1.5: Comparison of Cumulative Regrets of the CHD ϵ_t-Greedy algorithm, from $t = vw + 1$ to $t = T$, for values of $w = 10$, $s_t \equiv s = 0.2$ and $\kappa_t \equiv \kappa \in \{2, 3, 5\}$.

Figures 1.6 and 1.7 explore different ways of visualizing the performance of the CHD ϵ_t-Greedy algorithm. The first presents the difference in the action selected by our learning rule at each time step and the respective best one, exemplified by a simulation with $w = 10$, $\kappa_t \equiv \kappa = 2$ and $s_t \equiv s = 0.2$. In this figure, a point with a difference at zero means that the algorithm selected the best action, while any other value for difference implies a sub-optimal action adopted. Compared to the initialization period, the exploration-exploitation phase makes fewer mistakes, qualitatively attesting the learning process.

Figure 1.7 exhibits the average (across simulations and across the time horizon) frequency of hits, restricted to the post-initialization period, of the CHD algorithm for varying parameters: $w \in \{5, 10, 15\}$, $s_t \equiv s \in \{0.05, 0.1, 0.15\}$ and $\kappa_t \equiv \kappa \in \{2, 3, 5\}$. Performance seems to be fairly robust

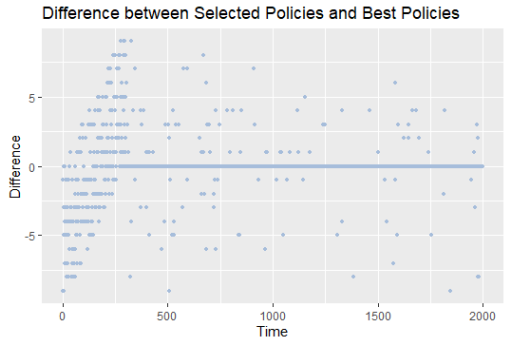

Figure 1.6: Differences between the selected policy and the best policy for the CHD ϵ_t-Greedy algorithm for values of $w = 10$, $\kappa_t \equiv \kappa = 2$ and $s_t \equiv s = 0.2$.

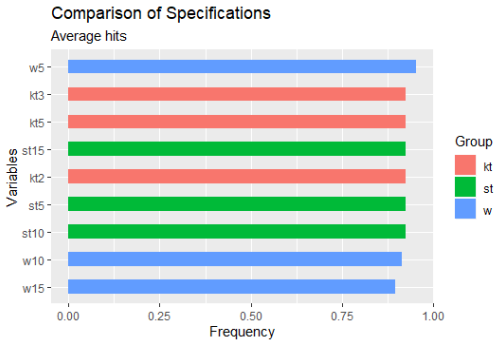

Figure 1.7: Comparison of frequency of hits for the CHD ϵ_t-Greedy algorithm, computed from $t = vw + 1$ to $t = T$, for different specifications of s_t, κ_t and w.

and, considering 50 simulations and 2000 time steps for each one of them, the worst specification adopts the best action approximately 90% of the time on average, while the best one reaches 95%.

1.5.1 Comparison to Related and Adapted Algorithms

As far as we know, the CHD ϵ_t-Greedy algorithm is the first conservative high-dimensional learning rule, which impairs a proper comparison exercise. Therefore, in a simulation exercise, we contrast it with its non-conservative version (HD ϵ_t-Greedy) and three additional adapted algorithms, named in this work as: HDO ϵ_t-Greedy, CHDO ϵ_t-Greedy and Expfirst. The general setup assumed in the beginning of this section is expanded to consider $w = 10$, $s_t \equiv s = 0.1$ and $\kappa_t \equiv \kappa = 5$ and the same initialization phase is implemented for all algorithms. We briefly discuss each algorithm in the sequence.

HDO and CHDO ϵ_t-Greedy algorithms: These are the counterparts of HD and CHD ϵ_t-Greedy algorithms, but using OLS as the estimation methodology to update estimated parameters $\hat{\boldsymbol{\beta}}_k$ when $\boldsymbol{\omega}_k$ is selected. In a high-dimensional sparse context, we would expect lasso to outperform a poorly defined OLS estimator. Inclusion of these algorithms in the comparison set serves to corroborate one of the motivations of this work, by contrasting the differences in performance in a online learning problem, resulting from distinct estimation procedures in a high-dimensional context.

ExpFirst: This is a kind of exploitation-only algorithm. The initialization phase is the same as in the other algorithms and estimation of $\boldsymbol{\beta}_k$ for selected actions in this stage is carried out as in the high-dimensional case, employing lasso. However, after initialization, the algorithm does not explore anymore. That is, it always selects the policy that presented the minimum regret in the initialization. In a different setting, provided that some new as-

sumptions are in place, Bastani et al. (2020) have shown that exploitation-only algorithms can achieve logarithmic growth in the OLS-estimation context.

Figure 2.20 compares the average cumulative regrets (across 50 simulations) of the CHD ϵ_t-Greedy with those of its peers, above discussed. Notice that the CHD algorithm largely outperforms, except when compared to its non-conservative version, in which case the improvement in the average cumulative regret is modest. Figure 2.21 amplifies Figure 2.20 and focus only on these two algorithms, considering the post-initialization phase ($t > vw$).

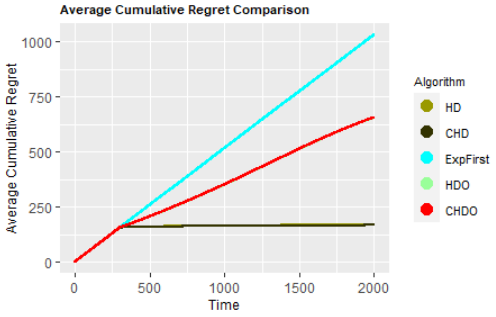

Figure 1.8: Comparison of average cumulative regrets of the CHD ϵ_t-Greedy with HD, HDO, CHDO ϵ_t-Greedy and ExpFirst algorithms for values of $\kappa_t \equiv \kappa = 2$, $w = 10$ and $s_t \equiv s = 0.1$.

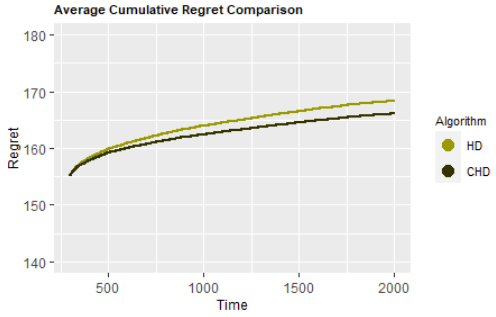

Figure 1.9: Comparison of average cumulative regrets between the CHD and HD ϵ_t-Greedy algorithm, in the post-initialization period: $t > vw$, for values of $\kappa_t \equiv \kappa = 2$, $w = 10$ and $s_t \equiv s = 0.1$.

1.6 Application: Recommendation System

1.6.1 Data and Exploratory Analysis

The dataset is obtained from Kaggle Database Repository[5] and we use it to build a recommendation engine to predict which restaurants customers are most likely to order from, given their characteristics. Information on this data set was initially gathered by Akeed[6], an app-based food delivery service in Oman.

[5]Restaurant Recommendation Challenge, Version 2, from https://www.kaggle.com/mrmorj/restaurant-recommendation-challenge. The dataset is provided public under the license CC BY-NC-SA 4.0, which provides users the right to share, copy, redistribute the material in any medium or format and adapt, remix, transform, and build upon the material for noncommercial purposes.

[6]Akeed is a mobile application that customers can download to their smart phones. It will allow customers in Oman to order food from their favorite vendors and have it delivered to their addresses.

We work with a share of the original dataset considering 15 features[7] and 16,043 observations comprising customers and their respective transactions with 8 vendors. The training sample has 1,605 observations for 927 different customers, used for the initialization stage of our algorithm. The testing sample has 14,438 observations containing 3,095 customers which can be new ones or repeated when compared to the training sample.

Table 2.1 provides descriptive statistics for each of the 15 variables we consider. Notice that in the column labeled as "type" only two categories appear: "original" if the variable came from the original data set or "new" if it were created under the scope of this work. Regarding the former type, we find labels created quite self-explanatory but any question about the meaning of any variable can be settled by visiting https://www.kaggle.com/mrmorj/restaurant-recommendation-challenge. We define the new variables as: Age of Customer Register – number of days since the customer has first registered; Age of Vendor Register – number of days since the vendor has first registered; Frequent Vendor – total number of transactions made by a specific vendor with any customer; Frequent Customer – total number of transactions made by a specific customer with any vendor; and Distance from Customer to Vendor – euclidean distance between a customer and a vendor based on latitude and longitude values.

Table 1.1: Descriptive Statistics for features used in the training sample with 1100 observations and 8 vendors.

Variables	Mean	Std	Maximum	Minimum	type
Amount of items purchased	2.23	1.99	38	1	Original
Total Cost	12.09	9.45	131	0	Original
Payment Mode	1.35	0.76	5	1	Original
Driver Rating	0.56	1.51	5	0	Original
Delivery Distance	3.70	4.02	14.97	0	Original
Gender	0.90	0.30	1	0	Original
Delivery Charge	0.40	0.35	0.7	0	Original
Serving Distance	14.13	2.82	15	5	Original
Preparation Time	14.04	2.25	20	10	Original
Vendor Rating	4.36	0.21	4.8	4	Original
Frequent Vendor	4.36	0.21	4.8	4	New
Frequent Customer	4.36	0.21	4.8	4	New
Age of Vendor Register	642.46	125.91	805	429	New
Age of Customer Register	616.71	160.69	952	255	New
Distance from Customer to Vendor	0.54	0.34	1.75	0	New

[7]Several features of the original data set were discarded since they have the vast majority of their entries as missing values or they are categorical variables with only one category.

1.6.2 Framework and Results

The framework proposed in previous sections is suitable for sequential problems of decision making. In order to comply with this, we interpret the dataset as being sequential, in the sense that at each time step only a customer arises, chooses a vendor and purchases items. We allow the same customer to arise multiple times both in the training and in the testing samples. Also, as defined in Section 1.2, the cardinality of the set of possible actions (w) is, in this context, the number of selected vendors we would like to recommend, indexed by $k = \{1, ..., w\}$. At every instant of time a customer appears and our algorithm should recommend a vendor, based on observed covariates.

To gather information about customers preferences we run a initialization stage by offering every vendor to every customer in the training sample. Since we know every customers' choices, we compute a unit reward if we are right in our recommendation or a zero reward otherwise. Therefore, after this phase we have sufficient information to estimate the sequence $\{\hat{\boldsymbol{\beta}}_k\}_{k=1}^{w}$ in equation (1-2). In the long run, each customer in the test sample also arises at each time step when we are supposed to recommend a vendor following the online learning rule proposed in this chapter. Recall that each $\hat{\boldsymbol{\beta}}_k$ is updated once the action $\boldsymbol{\omega}_k$ is adopted.

Figures 1.10 and 1.11 exhibit two related measures for the variable selection of the Lasso estimation, equation (1-2). Fix an arbitrary variable $\boldsymbol{x}_{(j)}$. Define the **relevance** (r_j) and **strength** (s_j) of the j-th variable as: $r_j = \sum_{k=1}^{w} \mathbb{1}_{\widehat{\boldsymbol{\beta}}_{k(j)} \neq 0}$ and $s_j = \sum_{k=1}^{w} |\widehat{\boldsymbol{\beta}}_{k(j)}|$, where $\hat{\boldsymbol{\beta}}_{k(j)}$ is the j-th entry of estimated beta for vendor k. Intuitively, r_j is about how frequent across vendors a variable is selected as relevant to explain the customers's preferences, and s_j relates to the potential to impact rewards.

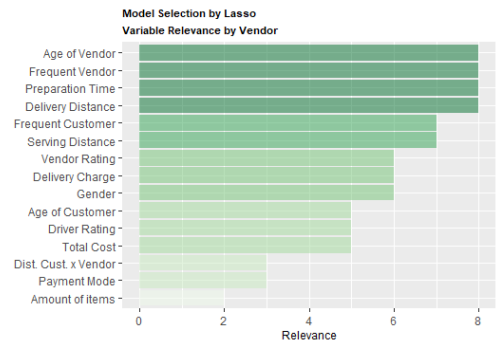

Figure 1.10: Relevance of each variable in Lasso estimation after the initialization phase for a training sample containing 1100 observations and 8 vendors.

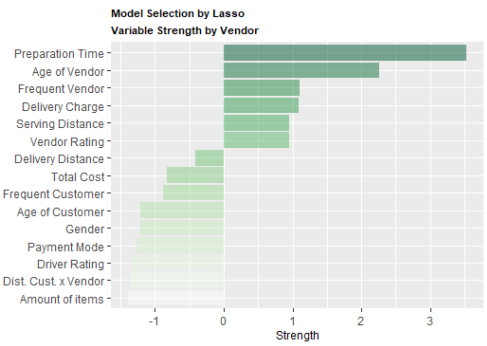

Figure 1.11: Centered (demeaned) strength of each variable in Lasso estimation after the initialization phase for a training sample containing 1100 observations and 8 vendors.

Figure 1.10 exhibits the relevance of each variable, while the right panel is about the demeaned strength in the Lasso estimation. From both panels, it is possible to infer that preferences are strongly influenced by how frequent is the vendor in performing transactions, preparation time and if it is an experienced vendor regarding the time elapsed since it first registered in the app. To see this, take age of vendor register as example. It is selected by Lasso as a relevant variable for all 8 vendors and the absolute sum of its estimated entries across all these 8 vendors is the second largest.

Main results are presented from Figures 1.12 to 1.14. The objective here is to explore the benefits of being conservative in a framework where users might be loyal to a set of specific vendors. For this, we compare the CHD ϵ_t-Greedy with its non-conservative version, using as a benchmark a naive rule that always guess. The later can be understood as a HD algorithm restricted to its exploration only. Figure 1.12 compares the algorithms, while the right panel focus on the post-initialization stage. Notice that the naive benchmark regret grows at an apparently linear rate, while both CHD and HD algorithms learns to choose actions in the long run.

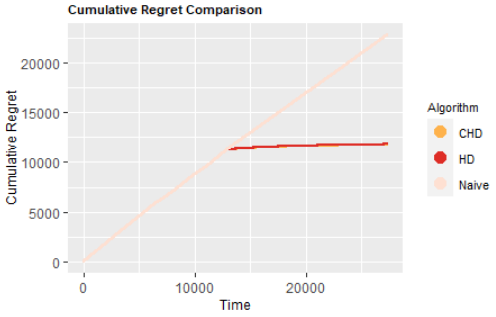

Figure 1.12: Comparison among CHD, HD ϵ_t-Greedy and a Pure Exploitation algorithm.

Figure 1.13: Comparison among CHD, HD ϵ_t-Greedy and a Pure Exploitation algorithm after initialization.

Figure 1.14 focus on the comparison between CHD and HD algorithms in the post-initialization phase. Results indicate that the rules provided in this chapter effectively learn through experience with mild differences between them. Moreover, it is interesting to highlight that the comparative performance between both algorithms is very similar to what was observed in Figure 2.21, when the learning rules were tested in a controlled environment. Another way to qualitatively attest the learning process of the CHD algorithm is looking at Figures 1.15 and 1.16. Jointly, both figures provide a comparison between the frequency of hits of the naive (which never learns) and that of the CHD. A customer would agree with a recommendation made by the naive algorithm 20.01% of the exploration versus exploitation stage, while for the CHD rule, matching achieves 96.38%.

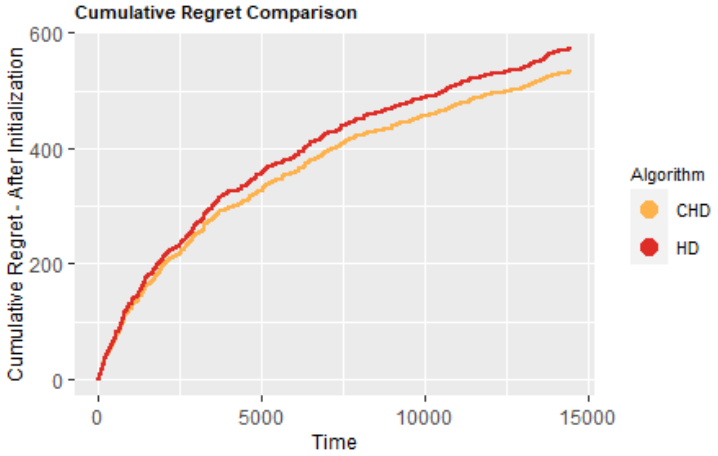

Figure 1.14: Comparison between CHD and HD ϵ_t-Greedy, considering only the exploration x exploitation stage.

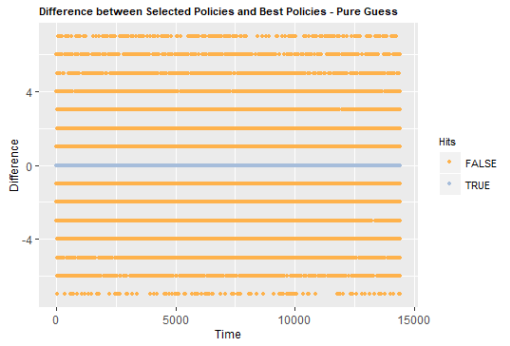

Figure 1.15: Differences between the selected policy and the best policy for the naive algorithm considering only the exploration x exploitation stage.

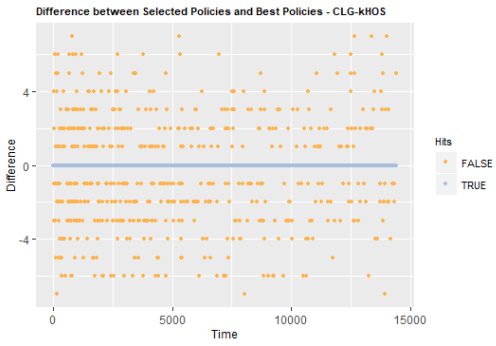

Figure 1.16: Differences between the selected policy and the best policy for the CHD ϵ_t-Greedy algorithm considering only the exploration x exploitation stage.

Since the results so far are conditional on the 8 pre-selected vendors, we perform a simple robustness check that consists of running 30 simulations, where in each of them we select randomly a different set of 8 vendors and, consequently, a different set of customers and their respective transactions. This exercise may be understood as a robustness to different sets of customers's preferences and to different sets of vendors features (which may impact preferences). Figure 1.17 compares the frequency of hits in the exploration versus exploitation phase for the CHD and HD ϵ_t-greedy algorithms. One can see that previous conclusions do not change, regarding the learning capacity and the potential benefits the algorithms would generate to a recommendation system. Actually, the observed frequencies of hits are not too much dispersed and the two respective coefficients of variation are very similar (0.4082 and 0.4080, respectively), indicating that both learning rules operate similar to different sets of inputs.

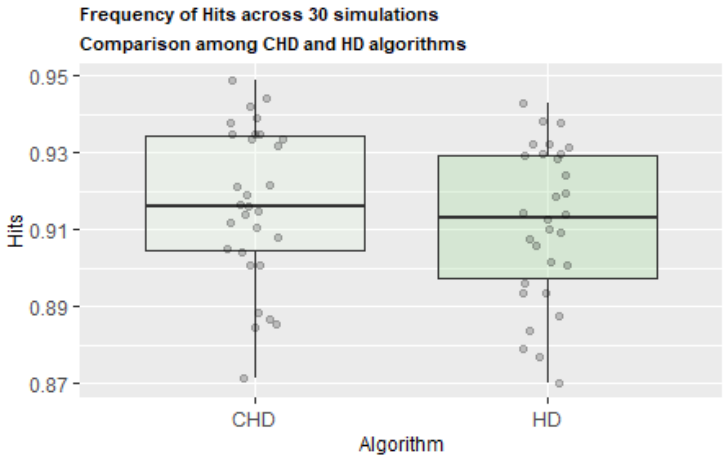

Figure 1.17: Comparison of frequency of hits between among CHD and HD ϵ_t-Greedy algorithms, across 30 simulations.

1.7 Concluding Remarks

In this work, we contribute to augment the basket of online learning solutions related to contextual bandits in high-dimensional scenarios. We extend a popular multiarmed bandit heuristic, the decaying ϵ_t-greedy heuristic, to high-dimensional contexts and we augment it with a conservative exploitation solution. The resulting learning rule can be useful for practical applications where an agent uses the experience and the repeated observation of a large pool of covariates to conservatively learn the best course of action relatively to some reward.

For a decreasing ϵ_t-greedy multiarmed bandit, we find that adding a high-dimensional context to the original setting does not substantially jeopardize the original results, except that in our case, regret not only grows reasonably with time but also depends on the covariate dimensions, as the latter grows with the former in high-dimensional problems. We find an upper bound growing less that $\mathcal{O}\{s_0\sqrt{Tlog(2p)}\}$ which seems to be comparable and, in some cases, even better than similar alternatives in the literature.

Moreover, we show that the consideration of a higher-order statistics searching set as an alternative to random exploration introduces safety to the decision-making process, without deteriorating the regret properties. More specifically, we show that the regret bound when the order statistics searching set is considered is at most equal to but mostly better than the case when random searching is the sole exploration mechanism, provided that the cardinality of the set of actions is sufficient large. Furthermore, we show that the upper bound on the cumulative regret function of the CHD ϵ_t-Greedy algorithm is not affected by the cardinality of the higher-order statistics searching

set, which, *per se*, provides flexibility for end-users facing constraints on the number of viable actions.

In a simulation exercise, we show that the algorithms proposed in this chapter outperform adapted competitors. Also, by employing both algorithms at a recommendation system data base, we confirm their learning trough experience, attesting their potential usefulness for companies that want to leverage their profits with an accurate recommendation engine.

2
Deterministic Treatment Effects Estimation with Unknown Complex Assignment Rules: A Learning Forest Approach

2.1
Introduction

According to Cattaneo et al. (2020):

"*The first step to employ the Regression Discontinuity (RD) design in practice is to learn how to recognize it. There are three fundamental components in the RD design - a score, a cutoff, and a treatment. Without these three basic defining features, RD methodology cannot be employed.*"

In this chapter, we explore situations where the above sentence can be softened. In particular, we present a procedure that can be useful for programs with deterministic rules that are unknown to the researcher and maybe more complex (as defined by Imbens and Zajonc, 2009) than the usual ones in the Regression Discontinuity Designs (RDD) literature.

More specifically, suppose there is a program with a deterministic assignment rule where one of the two occurs: either the researcher, for some reason, does not know the cutoff or she knows a previous version that is not entirely used to assign units to treated and non-treated groups. In the first case, the researcher is prevented to estimate treatment effects using standard sharp or fuzzy procedures since the words of Cattaneo et al. (2020) above-mentioned are binding. One does even not know which unit is eligible or not to the treatment. In the last case, if the researcher tries to estimate effects based on the obsolete cutoff, she may observe some misassignment of units to the treatment groups, compatible to what is observed in fuzzy procedures. At first glance, the respective fuzzy toolkit might appear to be the most suitable econometric procedure to be used, but our work is for cases when one can still discover a sharp rule based on observables, maybe more complex than the previous one. The source of the observed misassignment would be different from the traditional compliance problems characteristic of the fuzzy setup. Example 3 in Section 2.2 explores a real program that, in a first sight, presents some fuzziness, which is overruled since one discovers a sharp hidden complex rule.

The method proposed in this chapter can be useful to both cutoff anomalies described here.

2.1.1 Motivation and Comparison with the Literature

The problem of how to infer about causal relations from observations has challenged social scientists for decades and has led to a huge literature on the theme. In practice there is a multitude of distinct programs being offered to targeted units, each one of them with its own particularities. For instance, it is well-known that for a random assignment of individuals to treatment and non-treatment groups, resulting effects can be easily identified and estimated. However, in the absence of randomness, RDD appears as one of the most credible non-experimental strategies for the analysis of causal effects (Cattaneo et al., 2020).

The idea of using discontinuities in assignment rules to identify local causal effects is not new and can be traced back to works such as Thistlethwaite and Campbell (1960). However, the usage of RDD in economic applications exponentially expanded only after the seminal work in Hahn et al. (2001), that formalized the methodology in a language common to program evaluation. Subsequent works have deepen the understanding of every feature of RDD (Lee and Card, 2008 and Lee and Lemieux, 2010), developed the theory behind estimation (Porter, 2003 and Sun, 2005), provided practical guidance on bandwidth selection for local polynomial regressions (Ludwig and Miller, 2007 and Imbens and Kalyanaraman, 2012), among many others. For comprehensive reviews of RDD please refer to Imbens and Lemieux (2008), Lee and Lemieux (2010) and Cattaneo et al. (2020).

In this work we expand the RDD literature in the following directions:

Unknown Assignment Rules: Real world programs may have intrinsic features that impose some level of non-disclosure of assignment rules. In these cases, the lack of knowledge can be total, when the cutoff has not been published, or partial if a previously announced rule has evolved to a different one at the time of the program's implementation. In both cases, we implicitly consider in this chapter that there is a deterministic (unknown) assignment working behind the scenes and we show in this chapter that a tree-based methodology can be useful to provide consistent estimatives in contexts like these.

For example, in Chay et al. (2005) the authors investigate a Chilean government-based program (P900) designed to assist schools with low fourth-grade test scores. However, at execution, it is possible to recognize that the

government decision was not only based on grades, but other features may have played an important role. This conclusion is not only stated by the authors but also is clearly recognized in some administrative regions of Chile where there exists no possible cutoff based solely on students grades that can exactly segregate units in treated and non-treated groups. Example 3 in Section 2.2 illustrates our point for the Chilean first administrative region, showing that one can still discover a deterministic assignment, but more complex than the previous one based solely in grades.

Other cases could be considered in this context. For example, the work in Van Der Klaauw (2002) evaluates the effect of colleges and universities financial aid offers on student enrollment decisions. The specific financial aid allocation mechanism chosen is linked with college's objectives, such as, for example, those related with total enrollment, quality of accepted students and an appropriate level of diversity. The forcing variable in this case is an underlying index of various individual characteristics and, as commented in Porter and Yu (2015), the respective threshold could not be disclosed to mitigate manipulation by individuals or competition from other schools.

Another example with an unknown discontinuity point is discussed in Card et al. (2008) who analyze the tipping effect in the dynamics of segregation. Specifically, when the minority share in a neighborhood exceeds a "tipping point", all the whites leave. Such a tipping point depends on the strength of white distaste for minority neighbors which is generally unknown.

To the best of our knowledge, the most important paper dealing with unknown assignment rules is the work in Porter and Yu (2015). The authors propose tests to check about selection and the existence of a quantifiable effect from treatment, as well as the theory underlying estimation procedures both in the sharp and in the fuzzy contexts. The main idea is to recover the unknown cutoff, estimating it by the Difference Kernel Estimator proposed in Qiu et al. (1991) and, then, to estimate treatment effects as if the cutoff were known (plugging the estimated one). In our work, we propose a completely different methodology that does not need to uncover the hidden assignment rule. Instead, based on classification trees, it gradually learns treatment effects as the respective forest grows. However, since we know that in some cases the researcher may want to know the assignment rule, we also provide a suggestion of how one could shed light on this.

Complex Assignment Rules: In the classic RDD setup, the probability of treatment changes discontinuously if a scalar characteristic, only one forcing variable, falls above or below a cutoff point. In reality, things can be more complex than that and, in some cases, researchers can be compelled to

build new construct variables to remain in the univariate setup to estimate treatment effects.

For example, as commented in Papay et al. (2011), in public education students often take tests with clear cutoffs in several different subject areas and frequently must pass externally defined multiple criteria to, for example, avoid summer school, to be promoted to the next grade, or to graduate from high school.

Also, the work in Leuven et al. (2007) is about effects in schools that receive two kinds of subsidies. The first scheme gives primary schools with at least 70% disadvantaged minority pupils extra funding for personnel. The second scheme gives primary schools with at least 70% pupils from any disadvantaged group extra funding for computers and software. Eligibility for different types of insurance or entitlement program eligibility may also be driven by several criteria, such as family size or family income.

An important by-product generated in this work is related to the fact that the tree-based approach we engineered in this chapter works properly in a multivariate setup with more intricate assignment rules. This fact allowed us to generalize the traditional univariate cutoff to situations more adherent to practical situations. In this sense we inspired ourselves in Imbens and Zajonc (2009) to characterize what would be a treatment effect in a multivariate cutoff context.

2.1.2 Main Takeaways

Our contributions are twofold:

Primary - Theoretical: The main contribution of this work is to extend the work in Porter and Yu (2015) in directions when one can still discover a hidden deterministic assignment rule, based on observables, but more complex than the usual ones. A central product of this work relates to the theoretical investigation we provide for a tree-based methodology that does not estimate an unknown cutoff, but is capable to learn treatment effects from every piece of border between cells with distinct classification, as better clarified in Section 2.3. Basically, the procedure translates itself to a forest of classification trees, where a sequential learning procedure, similar to a bandit problem, is used to guarantee that, with high probability, the proposed empirical trees asymptotically approach those that correctly identify treatment effects. At the end, it estimates an implicit, from some assignment rule that remains mostly unknown, learned treatment effects that we prove to be consistent.

Subsidiary - Practical: We provide an algorithm for estimation of

treatment effects, in a sharp RDD sense, that takes into account specific forms of complex unknown assignment rules. In general terms, at each tree, several RDDs are evaluated and a tree-average treatment effect is computed. As the forest grows, this procedure is repeated in the other base learners and a forest-average treatment effect is computed (better explained in Section 2.3). Also, although not theoretically required, we provide an exploratory analysis that may be useful in cases that the knowledge of the assignment rule is valued. We also provide a robustness check on some properties of the algorithm with respect to parameters imputed by end-users.

We employ the algorithm proposed to revisit part of the P900 program investigated in Chay et al. (2005). We find the P900 suitable for our work, since the true assignment rule used by the government is not known in all Chilean administrative regions in the sample. There is, however, some documented evidence of discretion based, for example, on somewhat counter-intuitive allocations of units to the treatment groups. Despite of this, we find that our procedure delivers higher classification rates than those obtained in Chay et al. (2005). Moreover, focusing on a specific administrative region, we recover a complex assignment rule that probably has been overlooked by the literature so far, which not only provides a better classification for the units (schools) but also provides insights on a possible heterogeneous results among regions, a topic that, as far as we know, has not been touched in any academic work on the P900.

2.1.3 Organization of this chapter

The rest of this chapter is structured as follows. Section 2.2 gives more details and provides few examples of the problem we are investigating in this chapter, as it establishes the main notation and framework for the classification trees in our forest. Section 2.3 presents considerations about treatment effects identification and estimation, and provides an algorithm for practical use. Section 2.4 exhibits the theoretical and main results of this chapter, while Section 2.5 works on a simulation of the algorithm presented in Section 2.3 and some robustness checks. Section 2.6 employs the proposed methodology in this chapter to revisit part of the Chilean P900 program. Finally, Section 2.7 concludes this work. All proofs are relegated to the Appendix.

2.1.4 Notation

Regarding the notation used in this chapter, we provide in this subsection general guidelines. Definitions and particularities are presented throughout this chapter. Bold capital letters $\boldsymbol{X}$ represent matrices, small bold letters $\boldsymbol{x}$ represent vectors and small standard letters x represent scalars. Matrices or vectors followed by subscript or superscript parentheses denote specific elements. For example, $\mathbf{X}^{(j)}$ is the j-th line or column of the matrix $\mathbf{X}$ depending on the context, while $\boldsymbol{x}^{(j)}$ is the j-th scalar element of the vector $\boldsymbol{x}$. Id is the identity matrix. Let M be an arbitrary vector space. The symbol $\|\cdot\|$ is the usual vector norm on M, while $\mathcal{B}(\boldsymbol{x_0}, \tau)$ is the ball defined in M around a point $\boldsymbol{x}_0$, the set $\{\boldsymbol{x} \in M | d(\boldsymbol{x}, \boldsymbol{x}_0) \leq \tau\}$. Let Y be an arbitrary set. Then, $\#Y$ is used to represent the cardinality of Y, while $\mathbb{1}\{y \in Y\}$ is the traditional indicator function that takes a value of 1 when $y \in Y$. Finally, $\hat{h}$ indicates that it is a quantity generated based on a training sample. Depending on the context, it may refer to an estimative of h, a parameter of interest, or, for example, to some variable associated to an empirical tree, that relates asymptotically to some quantity h in a theoretical tree. For a concrete example, $\hat{\tau}$ is used in this chapter for the estimative of treatment effect, while $\hat{h}$ is used for a split generated by an empirical tree.

2.2 General Setup and Problem Formulation

Consider a sample composed by n individual units, all of them candidates for a generic treatment. For each $i \in \{1, ..., n\}$ we observe individual characteristics $\boldsymbol{x}_i$ taking values on $[0, 1]^p$, $p > 0$, that are i.i.d realizations from the same distribution $\mathbb{P}$. We also observe a scalar response variable y_i taking values on $\mathbb{R}$ and a binary variable $d_i \in \{0, 1\}$ representing treatment status, that is, $d_i = 0$ indicates that the unit i have not been treated and $d_i = 1$ represents the opposite. The well-known fundamental problem in causality refers to the impossibility to simultaneously observe an arbitrary unit's outcomes in treatment and non-treatment groups. To identify treatment effects and infer about causal effects, one can rely upon a variety of different approaches, conditional on assumptions made and the environment surrounding the experiment. In this paper, we study local treatment effects in a similar fashion (but with important additions) of the classical sharp RDD.

The vast majority of papers that use RDD to estimate treatment effects deals with one single forcing variable and a simple assignment rule. That is, considering that there is a direct effect of $\boldsymbol{x}_i$ on y_i, as well as an effect on

d_i, units are automatically (abstracting from compliance problems) assigned to a treatment once their relevant observable characteristic exceeds a fixed threshold c. Formally, $d_i = \mathbb{1}\{\boldsymbol{x}_i^{(j)} \geq c\}$, where $\boldsymbol{x}^{(j)}$ is the forcing variable associated to the treatment.

As better exposed in Section 2.1 we study cases where the cutoff is unknown to the researcher. These cases comprise situations that can vary from the total lack of knowledge of the cutoff value, perhaps driven by ethical reasons or to avoid manipulation, to situations where a threshold is previously published by the entity who runs the program but, when assigning units to treatment, different forces come into play leading to divergences between the expected constitution of treatment groups, based on the known cutoff, and what is observed. This last situation is exemplified later in this section through the lens of the P900 program that ended up generating confusing (based on what was mainly understood about the cutoff) treatment groups.

It happens that, in some cases, unknown cutoffs may be more complex than those that rely on a single forcing variable. The work in Imbens and Zajonc (2009) points out that, conceptually, RDD setups derived from multivariate cutoffs are similar to the scalar case, except that the cutoff at the discontinuity becomes a boundary. According to their definition, any point $\boldsymbol{x}$ is defined to belong to a boundary $\mathcal{A}$, labeled by $\boldsymbol{x}^{\mathcal{A}}$, if and only if every neighborhood around $\boldsymbol{x}^{\mathcal{A}}$ contains points both in the treatment and in control group. It is a generalization of the usual cutoff to the multivariate case. In these cases, local treatment effects can be associated to the limits of conditional expectation functions, taken by shrinking a ball centered in any point $\boldsymbol{x}^{\mathcal{A}}$, $\mathcal{B}(\boldsymbol{x}^{\mathcal{A}}, \varepsilon)$, towards its center ($\varepsilon \to 0$). On the limit, units belonging to treated and untreated groups would be so similar to each other that assignment to any group could be considered as random.

Example 1 *To set ideas, consider a bivariate case, assuming a somewhat more complex assignment rule* $d_i = \mathbb{1}\{min\{\boldsymbol{x}_i^{(1)}, \boldsymbol{x}_i^{(2)}\} \geq c\}$*, that is, an unit is treated provided that at least one of its features assumes a value greater than or equal* c*. In this case, a point is on the boundary if* $\boldsymbol{x}^{\mathcal{A}} = (c, r)$ *or* $\boldsymbol{x}^{\mathcal{A}} = (r, c)$*, with* $r \geq c$*. This reveals that a boundary might be formed by "heterogeneous" units, since a unit* i *with features* $(c, c+k)$*,* $k > 0$ *and a unit* i' *with* $(c+k, c)$ *are both treated and may present different responses to the program due to their intrinsic differences. Figure 2.1 illustrates the above-mentioned boundary.*

In practice, the choice of a boundary point $\boldsymbol{x}^{\mathcal{A}}$ in Example 1, leads to a conditional (to this choice) estimated treatment effect. In this paper we follow the suggestion in Imbens and Zajonc (2009) and we summarize treatment

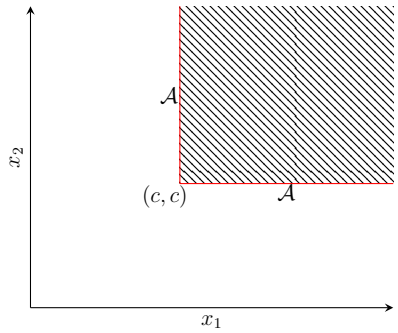

Figure 2.1: Illustration of an heterogeneous boundary in R^2.

effects along the boundary, in line with what they called the sharp integrated treatment effect:

$$\tau \equiv \mathbb{E}[y_i(1) - y_i(0) | \boldsymbol{x}_i \in \mathcal{A}]$$

where the expectation is taken over the boundary. Assumption 5, though, restricts our problem to specifics forms of boundaries that tuns out to be very general to most practical situations. In the sequence, Example 2 provides an hypothetical program that fits to our case and introduces our approach in this paper to deal with the lack of knowledge of the true boundary.

Assumption 5 (Identification of Treatment Boundary) *Consider a program that assigns units to treatment according to a deterministic complex rule $a : [0, 1]^p \to \{0, 1\}$, in the sense that $\forall i$, $a(\boldsymbol{x}_i) = d_i$. Then, there is at least one classification tree $\mathcal{T}$ in a set $\mathcal{D}_a \equiv \{\mathcal{T}_m\}_{m=1,...,M_a}$ that correctly identifies the associated boundary $\mathcal{A}_a$.*

Example 2 *As a simple example of a set $\mathcal{D}$, consider a hypothetical program, maybe a job relocation program designed for poor, but reasonably-educated people that are currently underemployed. In this case, let $\boldsymbol{x}_i^{(1)}$ and $\boldsymbol{x}_i^{(2)}$ be the wealth and the level of education of the person i, and let c_1, c_2 be the respective thresholds. This program fits in assumption 5: it has a complex assignment rule, with an associated boundary $\mathcal{A} = \{\boldsymbol{x}_i | \boldsymbol{x}_i^{(1)} = c_1; \boldsymbol{x}_i^{(2)} \geq c_2\} \cup \{\boldsymbol{x}_i | \boldsymbol{x}_i^{(1)} \leq c_1; \boldsymbol{x}_i^{(2)} = c_2\}$ that can be described by a "tree structure". Figures 2.2 and 2.3 depict two simple trees $\mathcal{T}_1$ and $\mathcal{T}_2$ that correctly identify $\mathcal{A}$, being the only difference between them, the order of splitting: while $\mathcal{T}_1$ uses $\boldsymbol{x}^{(1)}$ as its first direction to split, $\mathcal{T}_2$ uses $\boldsymbol{x}^{(2)}$.*

Firstly, notice how general is the hypothetical program described in Example 2. We could have easily replaced this example for other programs designed to subjects such as: education (students must pass two tests), inequality (fixed criteria on wealth and neighborhood, for example), gender, race, among

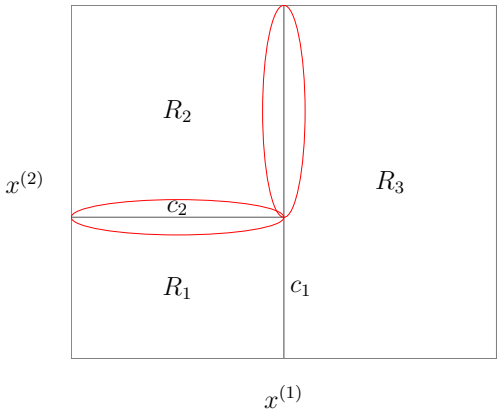

Figure 2.2: $\mathcal{T}_1$

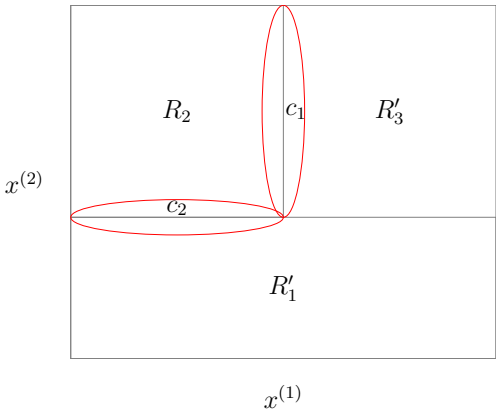

Figure 2.3: $\mathcal{T}_2$

many others. It turns out that the "rectangular shape" formed by the boundaries considered in this paper are not only suitable for a tree-based methodology, but also intuitively associated to most treatments of interest.

Also, building on Example 2, as we work in this paper with unknown assignment rules, consider that for some reason, one does not know $\{c_1, c_2\}$. According to the assignment rule considered, we would expect region R_2 in Figure 2.2 to be the only one containing treated units. Label the edge between R_2 and R_3 as E_{2-3} and the one between R_2 and R_1 as E_{1-2} (red ellipses in Figure 2.2). Notice that both E_{1-2} and E_{2-3} could be considered as cutoffs to a regression discontinuity design where, in each one of them, only one variable plays the role as the forcing one. In fact, as discussed before, units sufficiently close to these borders are so similar among each other with respect to the driving variable, that allocation or not to the treatment could be considered random. Therefore, to estimate integrated treatment effects over this boundary, we consider τ as a composition of two treatment effects, τ_1 and τ_2, whether $\boldsymbol{x}^{(1)}$ or $\boldsymbol{x}^{(2)}$ is the relevant forcing variable. Our problem, then, is to build a procedure that guarantees, with high probability, that our empirical trees approach any theoretical one in $\mathcal{D}$. In Section 2.3 we present our suggestion.

Example 3 uses the P900 treatment as a real example of a program that could be analyzed under the tree-based scope proposed in this paper. In very general terms, the P900 was implemented by the Chilean government to assist schools with poor performance. Back in 1990, the program identified approximately 900 schools presenting low mean fourth-grade test scores that, once treated, should be supported with infrastructure improvements, instructional materials, teacher training, and tutoring for low-achieving students. We provide more details on this program in Section 2.6 and for a thorough exposition please refer to Chay et al. (2005).

Example 3 *Based on what was mainly known about the P900 assignment rule,*

a Chilean school would get assistance from the government based on the grades of its students. Figure 2.4 presents what is observed in terms of the of treated and non-treated schools for the first (out of thirteen) administrative region in Chile. In this figure, the average grade is the mean between the mathematics and language scores obtained by students in the 1988 test in a given school and d_i is the treatment indicator, zero for untreated units and one for treated schools. A simple observation of Figure 2.4 shows that any cutoff based solely on grades that one could possibly consider induces treated schools that should not be treated and non-treated schools that should be. There is a clear and strong overlap that could be associated to a level of misassignment usually observed in a fuzzy setup with possible problems of compliance.

However, a deeper investigation into the data set reveals the hidden threshold: for urban schools, an average grade less than or equal to 51.85 and frequency of students greater than or equal to 14 is sufficient to exactly split both treatment groups without overlapping. That is, a hidden complex rule that lead us back to the sharp setup and that precisely fits to the tree-based methodology of Example 2.

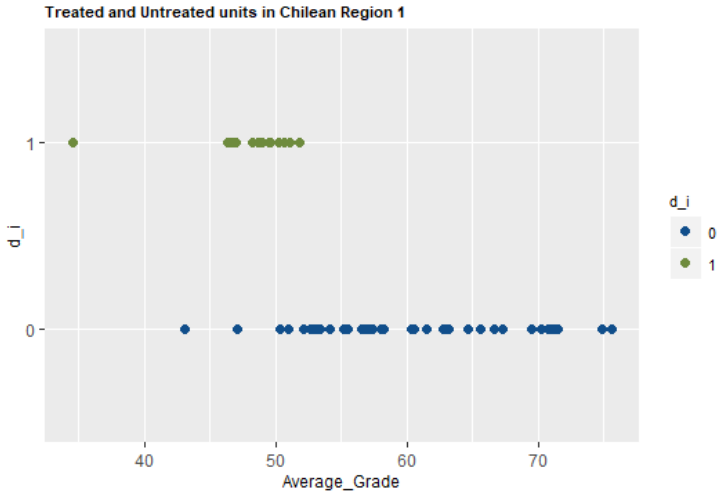

Figure 2.4: Treated and untreated schools in Chilean Administrative Region 1, where we present the average score of each school in mathematics and language in 1988 and d_i is the treatment indicator, zero for untreated units and one for treated schools.

In order to explore the treatment discontinuity to identify the correspondent effects, we impose Assumption 6, which is standard in the regression discontinuity literature (e.g. Imbens and Zajonc, 2009, Hahn et al., 1999, Hahn et al., 2001, Imbens and Lemieux, 2008).

Assumption 6 (Continuity of Conditional Expectation Functions) $\mathbb{E}[y_i(0)|\boldsymbol{x}_i \in [0,1]^p]$ *and* $\mathbb{E}[y_i(1)|\boldsymbol{x}_i \in [0,1]^p]$ *are both continuous functions.*

Remark 4 *Theoretical concepts underlying the RDD methodology require conditional expectation function to be continuous only at the cutoff (boundary). In this paper we extend it to the range of* $\boldsymbol{x}$*, following the comment in chapter 21 of Wooldridge (2001) that says that: "Technically, they are continuous at* $x = c$*, but it is hard to imagine how we could ensure that they are without assuming continuity over the range of* x*". Actually, there is another reason we need Assumption 6. In our procedure, at every tree in the forest different variables may play the role of a forcing variable and different cutoff values can be used as well. Therefore, the conditional expectation function should be required to be continuous at a larger set of different points of the random vector* $\boldsymbol{x}$*. Extending this premise to the entire support of the covariates is sufficient for our purposes.*

Finally, for the sake of completeness, it is important to highlight that, implicitly in this framework, we are considering the SUTVA (Rubin, 1986) assumption. So, units are not affected by the treatment of their peers. As already mentioned, compliance to the treatment is also required.

2.2.1 The Forest Setup

In the last subsection we introduced the idea that a classification tree could be a suitable methodology to be considered when cutoffs are unknown. However, in many real applications, decision trees such as that depicted in Example 2 present high variance (James et al. (2013)). That is, if the training sample is randomly divided into two parts, the results of decision trees fitted to both halves can be quite different. This fact is one of the main drivers that led the evolution of single trees to procedures such as bagged trees and random forests. All of them reduce the variance of a estimative, when compared to the single-tree case and, because of this, have become more and more popular in the last decade. In this paper, we work with what we call learning forests, a modification from the original random forest concept introduced in Breiman (2001), better described in Section 2.3.

Extending definitions already made in this section, we work with a training sample $\mathcal{S}_n = \{(\boldsymbol{x}_i, d_i)\}_{i=1,\dots,n}$ of $[0,1]^p \times \{0,1\}$-valued independent random variables and we build a forest comprising a sequence $\{\mathcal{T}_b\}_{b=1,\dots,B}$ of B classification trees in the forest. Each node of a generic tree represents a set (cell) in the space $[0,1]^p \times \{0,1\}$. The first cell at the top of a tree $\mathcal{T}_b$ is its root and comprises all available units. We represent it by $A_{b,0}$, while $A_{b,k}$ is reserved for other cells as $\mathcal{T}_b$ grows. We work with the case that each cell has

exactly two children, resulting from a splitting rule, or none at all. In the first case we have an internal node and, in the latter, a terminal node or leaf.

Inspired by one of the most famous splitting rule, the Classification and Regression Trees (CART) in Breiman et al. (1984), we impose that, at each node of each tree, the best cut is selected as the one that maximizes the variation in the impurity Gini index. Define $n(A_{b,k}) \equiv \#\{\boldsymbol{x}_i \in A_{b,k}\}$ as the number of units in the k-th cell of the b-th tree and $n_d(A_{b,k}) \equiv \#\{\boldsymbol{x}_i \in A_{b,k} | d_i = d\}$, $d \in \{0,1\}$, as the amount of treated and non-treated units in cell $A_{b,k}$. Then, the impurity Gini ($G(A_{b,k})$) is defined as:

$$G(A_{b,k}) = \phi_0(A_{b,k})(1 - \phi_0(A_{b,k})) + \phi_1(A_{b,k})(1 - \phi_1(A_{b,k}))$$

where $\phi_d(A_{b,k}) \equiv \frac{n_d(A_{b,k})}{n(A_{b,k})}$ is the frequency of units belonging to class $d \in \{0,1\}$ in cell $A_{b,k}$. The purest case occurs when $\phi_0(\cdot) = 1$ or $\phi_0(\cdot) = 0$ leading, in both situations, to $G(\cdot) = 0$.

Let W_b be an uniformly draw subset from $\{1, ..., p\}$ with cardinality $w, \forall b \in \{1, ..., B\}$. The nodes of a tree $\mathcal{T}_b$ choose the best split based on the variation of the Gini index, but restricted to the fact that the chosen direction belongs to W_b. Also, only a random portion of the sample points, with size $s \in \{1, ..., n\}$, is allowed to be used as an input into the root of any tree $\mathcal{T}_b$.

In our notation, a split $h_{b,k} \equiv (j_{b,k}, \zeta_{b,k})$ is a pair that divides the parent cell $A_{b,k-1}$ in its two children, where $j_{b,k} \in W_b$ is the best direction selected for splitting and $\zeta_{b,k}$ is its respective value, that is, a particular realization of $\boldsymbol{x}^{(j_{b,k})}$ used to divide the cell based on the purity gain. Formally, the left child of $A_{b,k-1}$ is $A_{b,k-1,-} \equiv \{(\boldsymbol{x}_i, d_i, y_i) \in A_{b,k-1} | \boldsymbol{x}_i^{(j_{b,k})} \leq \zeta_{b,k}\}$ and the right child is $A_{b,k-1,+} \equiv \{(\boldsymbol{x}_i, d_i, y_i) \in A_{b,k-1} | \boldsymbol{x}_i^{(j_{b,k})} > \zeta_{b,k}\}$. In addition, define $\mathcal{H}_{b,k-1} = \{h_{b,1}, h_{b,2}, ..., h_{b,k-1}\}$ as the set of all previous splits, from the root, used to generate $A_{b,k-1}$, $\mathcal{C} = \{(j, \zeta) | j \in \{1, ..., p\}, \zeta \in \{\boldsymbol{x}_1^{(1)}, ..., \boldsymbol{x}_n^{(p)}\}\}$ as the set of all possible splits derived from $\mathcal{S}_n$, and $\mathcal{C}_{b,k} \subset \mathcal{C}$ as the set of all viable splits $h_{b,k}$ given $\mathcal{H}_{b,k-1}$.

We halt the growing of $\mathcal{T}_b$ based on a stopping criteria Δ, which is compared to the variation of the impurity index after a split. That is, a cell $A_{b,k-1}$ is not further split if:

$$\Delta > \Gamma(A_{b,k-1}, h_{b,k}) \equiv G(A_{b,k-1}) - \Omega(A_{b,k-1}, h_{b,k})$$

where:

$$\begin{aligned}\Omega(A_{b,k-1}, h_{b,k}) \equiv & \phi_-(A_{b,k-1}, h_{b,k}) G(A_{b,k-1,-}, h_{b,k}) - \\ & \phi_+(A_{b,k-1}, h_{b,k}) G(A_{b,k-1,+}, h_{b,k})\end{aligned}$$

is the weighted impurity index for both children of $A_{b,k-1}$, while the weight

$\phi_-(A_{b,k-1}, h_{b,k}) \equiv n(A_{b,k-1,-})/n(A_{b,k-1})$ is the proportion of units in the parent cell assigned to the left child after a split. The same definition is extended to $\phi_+(A_{b,k-1}, h_{b,k})$ regarding the cell $A_{b,k-1,+}$.

Define $\mathcal{L}_b = \{A_{b,k} | A_{b,k,-} = A_{b,k,+} = \emptyset\}$ as the set of K_b leaves in tree $\mathcal{T}_b$. We associate a class for each leaf following the well-known majority vote rule for classification trees. That is, provided that $A_{b,k} \in \mathcal{L}_b$:

$$g(A_{b,k}) = \begin{cases} 1, & \textbf{if} \quad \sum_{i=1}^{n_{A_{b,k}}} d_{i,b,k} > \sum_{i=1}^{n_{A_{b,k}}} (1 - d_{i,b,k}) \\ 0, & \textbf{otherwise} \end{cases} \tag{2-1}$$

where $g(\cdot)$ is the majority vote classifier.

Also, let $\mathcal{F}_b = \{(A_{b,k_i}, A_{b,k_j}) \in \mathcal{L}_b | \exists h \in \mathcal{H}_{b,k_i} \cap \mathcal{H}_{b,k_j}$ and $g(A_{b,k_i}) \neq g(A_{b,k_j})\}$ to be the set of pairs of adjacent leaves with different classifications, with cardinality f_b. This set comprises pairs of leaves that, not only disagree about the classification they assign to their units, but also share a split ($h \in \mathcal{H}_{b,k_i} \cap \mathcal{H}_{b,k_j}$) in some step of the tree growing process. For instance, using the tree $\mathcal{T}_1$ depicted in Figure 2.2, we observe that $\mathcal{L}_1 = \{R_1, R_2, R_3\}$ and $\mathcal{F}_1 = \{(R_1, R_2), (R_2, R_3)\}$.

2.3 Considerations About the Identification of Treatment Effects and Estimation Procedure

Recall from Section 2.2 that our main objective is to estimate the sharp integrated treatment effects τ, based on a true boundary that we do not know in advance. Assumption 5, however, suggests that the boundaries considered in this work are associated with "rectangular regions" of treated groups and, as a consequence, τ could be understood as a composition of what we call border treatment effects (see the discussion in Example 2). We aggregate treatment effects estimatives as: the average of all border treatment effects estimatives after a tree is fully-grown is called the tree treatment effect estimative; and the average of all tree treatment effects estimatives is called the forest treatment effect estimative, which we use to estimate τ. Formally:

Definition 4 (Aggregation of Treatment Effects) *A forest treatment effect estimative $\hat{\tau}$ is:*

$$\hat{\tau} = \frac{1}{B - \mu\psi Q} \sum_{b=\mu\psi Q}^{B} \hat{\tau}_b^{(t)}$$

where μ, Q are defined in this section under the context of the learning methodology proposed, ψ is defined in Lemma 14 and the elements of the sequence $\{\hat{\tau}_1^{(t)}, \hat{\tau}_2^{(t)}, ..., \hat{\tau}_B^{(t)}\}$ are estimatives of the tree treatment effects, calculated by:

$$\hat{\tau}_b^{(t)} = \frac{1}{f_b} \sum_{l=1}^{f_b} \hat{\tau}_{b,l}^{(f)}$$

where, for a tree $\mathcal{T}_b$, the elements of the sequence $\{\hat{\tau}^{(f)}_{b,1}, \hat{\tau}^{(f)}_{b,2}, ..., \hat{\tau}^{(f)}_{b,f_b}\}$ are estimatives of the border treatment effect evaluated on the f_b pairs of leaves $(A_{b,k_i}, A_{b,k_j}) \in \mathcal{F}_b$.

The next section investigates convergence properties of empirical splits and the respective border treatment effects estimatives to their theoretical (population) counterpart. However, as indicated in Assumption 5, maybe more than one theoretical tree correctly identifies the boundary induced by the assignment rule (the one relevant for causal effects) and, on the other hand, certainly there are trees (those that do not belong to $\mathcal{D}$) that identify wrong borders. In this paper we use a sequential learning procedure to guarantee that, with high probability, the borders generated in our operational procedure converge to the right one.

In very general terms, a sequential learning rule with good theoretical properties is capable to effective learn how to take actions, from some alternatives, in order to maximize a reward or some variable of interest. It is not our intention to review these kind of problems in details, since we already work with sequential learning in Chapter 1. For the purposes of this chapter, good references are: Auer et al. (2002), Slivkins (2019) and Charpentier et al. (2021).

Consider the set $\mathcal{Q}$ to contain all possible combinations of w variables, from the p originally considered. That is, a set of sets, with cardinality $Q \equiv \frac{p!}{p!(p-w)!}$, in the sense that an element of $\mathcal{Q}$, $\mathcal{Q}_q$, $q \in \{1, ..., Q\}$, is a set containing a random draw from $\{1, ..., p\}$ with cardinality w. Recall from the framework described in Section 2.2, that set $\mathcal{Q}$ contains all possibilities for sets W_b, splitting direction to be used by the nodes of the b-th tree. In addition to this, we know that any tree in $\mathcal{D}$ correctly recognizes the true boundary and, as so, its leaves attain maximum purity ($G(\cdot) = 0$, leading to correct classification of units in treatment groups). Our problem, then, is to learn how to give our empirical trees the best substrate to work, in order for them to resemble their counterparts in $\mathcal{D}$. In other words, we want empirical trees to attain the maximum purity possible, before a fixed stopping criteria is applied to all of them. In this setting, we consider $\mathcal{Q}$ as the set of possible actions that can be chosen by a learning rule in order to maximize the reward, the sum of the negative Gini index (purity) in the leaves of our empirical trees.

Definition 5 (ϵ_b-Greedy Heuristic) *Let $\mu > 0$ and define the sequence $\epsilon_b \in (0, 1]$, $b \in \{1, ..., B\}$, by $\epsilon_b \equiv \min\left\{1, \frac{\mu Q}{b}\right\}$. Define the action function $I : \mathbb{N}^+ \to \mathcal{Q}$, such that for each tree $\mathcal{T}_b$, $I(b) = \mathcal{Q}_q$ represents that action q was selected by the learning rule. Then, the ϵ_b-greedy algorithm is*

Algorithm 2: ϵ_b-Greedy Heuristic

input parameters: μ, B, Q
for $b \in \{1, ..., B\}$ **do**
  $\epsilon_b \longleftarrow \min\left\{1, \frac{\mu Q}{b}\right\}$;
  $q_b \longleftarrow \mathsf{U}(0, 1)$;
  if $q_b \leq \epsilon_b$ **then**
    $a_b \longleftarrow \mathsf{U}(0, Q)$;
    $I(b) \longleftarrow \mathcal{Q}_{a_b}$;
  else
    $c_b \longleftarrow \underset{q \in \{1,...,Q\}}{\arg\max} \frac{1}{B} \sum_{b=1}^{B} G_{bq}(\cdot)$;
    $I(b) \longleftarrow \mathcal{Q}_{c_b}$;
  end
end

We use one of the most simple and popular learning rule: the decaying ϵ_b-greedy, described in Auer et al. (2002) and repeated in Definition 5 for the reader's convenience. The rule works as: with probability $1 - \epsilon_b$, it selects the action the leads to the best empirical average reward until b; with probability ϵ_b it adopts a random action in trying to find good ones that have not been tested so far.

Given a tree $\mathcal{T}_b$ and a pair of cells in $\mathcal{F}_b$ sharing the l-th border, $l \in \{1, ..., f_b\}$, we estimate the border treatment effect $\hat{\tau}_{b,l}^{(f)}$ using local polynomial regression, a topic that has been around in the literature for a long time. Comprehensive studies on this can be found in Fan (1992), Fan (1993), Fan and Gijbels (1992) and Ruppert and Wand (1994), among many others. Consider the knowledge of the true cutoff ($\zeta_{b,l}^0$) at the particular border considered. Since each border counts with a single forcing variable (see Section 2.2), other covariates are used as control in the estimation procedure. We follow the suggestion in Calonico et al. and estimate $\hat{\tau}_{b,l}^{(f)}(\zeta_{b,l}^0)$ (considering the true cutoff) by:

$$y_{i,b,l} = \alpha_{b,l} + \tau_{b,l}^{(f)} d_{i,b,l} + \beta_{b,l-}(x_{i,b,l} - \zeta_{b,l}^0) + \beta_{b,l+} d_{i,b,l}(x_{i,b,l} - \zeta_{b,l}^0) + \boldsymbol{\gamma}_{b,l}' \boldsymbol{z}_{i,b,l} + e_{i,b,l} \tag{2-2}$$

where the subscripts $\{i, b, l\}$ refers to the i-th unit that lies near the l-th border of the b-th tree, considering a bandwidth choice. Moreover, x is the forcing variable at the l-th border and $\boldsymbol{z}$ is the rest of elements in $\boldsymbol{x}$ used as control, considering the partition $\boldsymbol{x} = [x\ \boldsymbol{z}]'$. Following the notation in Section 2.2, subscripts "-" and "+" refer to cells at each side of the border, regions with treated and non-treated units, since they belong to $\mathcal{F}_b$. We also restrict our problem to cases covered by Assumption 7.

Assumption 7 (Problem Setup) *For each tree $\mathcal{T}_b$ and each pair of cells $(A_{b,k_i}, A_{b,k_j}) \in \mathcal{F}_b$ sharing the l-th border, $l \in \{1, ..., f_b\}$:*

i. $\exists \delta_\alpha, \delta_\beta > 0$, *such that* $|\alpha_{b,l}| \leq \delta_\alpha$ *and* $\forall j \in \{1, ..., p\}$, $|\beta^{(j)}_{b,l-}| < \delta_\beta$ *and* $|\beta^{(j)}_{b,l+}| < \delta_\beta$.

ii. The sequence $\{e_{i,b,l}\}$ *is formed by independent centered random variables with* $E[e^2_{i,b,l}] \leq \sigma^2$, $E[\boldsymbol{z}_{i,b,l} e_{i,b,l}] = \mathbf{0}$ *and* $E[e_{i,b,l} x_{i,b,l}] = 0$.

In practice, researchers often use a weighting scheme according to a kernel function, in order to give relative importance to units whose scores lie within a preselected bandwidth around the cutoff[1]. The most popular choices in the RDD literature are the uniform kernel, that applies an equal weight to observations, and the triangular kernel, that linearly downweights observations far from the threshold (Lee and Lemieux, 2010). In this work, we avoid exploring subjects such as optimal bandwidth choices or different types of kernel. Although important in practical applications, they are not central to the main results of this paper and considerably complicate notation. For more details on these subjects, please refer to the above-mentioned works besides those presented in Fan and Gijbels (1996) and Imbens and Kalyanaraman (2012).

In cases where the true cutoff is known, the authors in Calonico et al. (2014) have proven that $\hat{\tau}^{(f)}_{b,l}(\zeta^0_{b,l})$ is consistent to the true effect $\tau_{b,l}$, plus an additional term that depends on the RD treatment effect on the covariates. A sufficient condition to get rid of this term is described in Assumption 8, imposed in this paper only to reduce the mathematical burden on the proofs. As also described in Calonico et al. (2014), this assumption is weaker than requiring that the marginal distributions of $\boldsymbol{z}$ for treated units are equal those for non-treated units near the cutoff, which is the usual definition of predetermined covariates in randomized experiments. Under Assumption 8, τ is also identified as the difference between intercepts in regressions performed at each side of the discontinuity.

Assumption 8 (Treatment and Additional Covariates) *For each tree $\mathcal{T}_b$ and for each pair of adjacent cells $(A_{b,k_i}, A_{b,k_j}) \in \mathcal{F}_b$ sharing the l-th border:*

$$E[\boldsymbol{z}_{i,b,l} | x = \zeta^0_{b,l}, d_{i,b,l} = 1] = E[\boldsymbol{z}_{i,b,l} | x = \zeta^0_{b,l}, d_{i,b,l} = 0]$$

[1]Since local linear inference relies on regression fits using only a portion of covariates near a threshold.

In equation (2-2), the estimative of interest is $\hat{\tau}_{b,l}^{(f)}$ and, for the rest of this paper, we emphasize its dependence on the type of the cutoff considered. That is, $\hat{\tau}_{b,l}^{(f)}(\zeta_{b,l}^{0})$ refers to the case that the researcher knows the value of the true cutoff. Since in this paper we consider cases that the assignment rule is unknown, we investigate the properties of $\hat{\tau}_{b,l}^{(f)}(\hat{\zeta}_{b,l})$, where $\hat{\zeta}_{b,l}$ is a split empirically determined by the b-th tree in our learning forest.

For practical implementation, Definition 6 presents the RDF algorithm, that comprises the steps discussed in this section. In summary, for the algorithm to be started, one should provide the number of trees in the forest (B), the quantity of pre-selected splitting variables (w), the amount of sample points to be considered as input for the root of each tree (s), a stopping criteria (Δ) and the parameter μ for the learning rule (See Definition 5). Then, at the first tree ($\mathcal{T}_1$), run the ϵ_b-Greedy algorithm to select the set of variables to be considered as splitting directions (W_1), with cardinality w. After that, the feature space is partitioned by selecting best splits according to the procedure in Section 2.2. Once the growing procedure has come to an end, according to the stopping criteria (Δ), identify the set of leaves ($\mathcal{L}_1$) and classify units inside them following the majority vote rule (equation (2-1)). Form the set $\mathcal{F}_1$ by all pairs of adjacent leaves with divergent classifications. Compute all border treatment effects estimatives ($\{\hat{\tau}_{1,1}^{(f)}, ...\hat{\tau}_{1,f_1}^{(f)}\}$) and average them to the first tree treatment effect estimative ($\hat{\tau}_1^{(t)}$). Storage this value and repeat this procedure for all trees in the forest. In the end, $\hat{\tau}$ is computed following Definition 4.

In the resumed way that Definition 6 describes the algorithm, it does not provide a way to uncover the unknown treatment assignment rule. However, in Section 2.5 we propose an exploratory analysis that we call as the first stage of the RDF algorithm that can be useful to shed light on hidden important variables (and their respective cutoffs) that might have been used in some extent to assign units to the treatment groups.

Definition 6 (RDF Algorithm) *Consider a training sample $\mathcal{S}_n$ and specify the total number of trees $B \in \mathbb{N}^+$, the number of pre-selected splitting variables $w \in \{1, ..., p\}$, the amount of sample points $s \in \{1, ..., n\}$ to be randomly selected, the stopping criteria $\Delta \in (0, 0.5)$ and $\mu > 0$. Then, the RDF algorithm is:*

Algorithm 3: RDF Algorithm

input parameters: B, w, s, Δ, μ

for $b \in \{1, ..., B\}$ **do**

Select s units uniformly in $\mathcal{S}_n$;

Select w splitting variables according to the ϵ_b-Greedy rule in Definition 5;

$\Gamma \leftarrow 0$;

while $\Gamma < \Delta$ **do**

Select best splits according to the methodology described in Section 2.2

end

Compute $\mathcal{L}_b$, $\mathcal{F}_b$ and $\{\hat{\tau}_{b,l}^{(f)}\}$, $l \in \{1, ..., f_b\}$;

Compute and storage $\hat{\tau}_b^{(t)}$;

end

Compute $\hat{\tau}$

2.4 Theoretical Properties of the Estimators

In this section we establish the main theoretical properties for the class of estimators proposed in this chapter. As a first step, Lemma 1 exhibits the relationship between the stopping criteria (Δ) and the number of data points inside an arbitrary cell. This is a well-known link in the decision tree literature and very intuitive as well, since an early-stopped tree would have larger leaves, probably leading to a higher bias, considering some tree-based statistics. Since RDD estimation can be severely impacted when there are few points to be used, our procedure is especially useful when a large sample is available. In this case, Lemma 1 implies that, when $n \to \infty$, the number of units inside an arbitrary cell is guaranteed to also grow indefinitely, although at a lower rate.

Lemma 1 (Stopping Criteria and the Number of Units in a Cell) *For any tree $\mathcal{T}_b$, $b = 1, ..., B$, provided that an arbitrary cell $A_{b,k-1}$ has $n(A_{b,k-1}) > 2$ then, any cell $A_{b,k}$, child of $A_{b,k-1}$ and formed by a sequence of splits $\mathcal{H}_{b,k}$ has the property:*

$$n(A_{b,k}) \geq \frac{\Delta n(A_{b,k-1})^2}{2(n(A_{b,k-1}) - 2)} \geq n\left(\frac{\Delta}{2}\right)^k$$

The rest of this section is devoted to establish the asymptotic properties of the estimators in Definition 4. There are not many available papers investigating theoretical properties of forests of classification trees, as recognized by Scornet et al. (2015), but, in our opinion, good references on the theme are the works in Biau et al. (2008), Scornet et al. (2015) and Wager and Athey (2018), among few others. Linking our theoretical effort to what have already been

done in this field, it is possible to identify similarities between our approach and that on Ishwaram (2015), but differences arise from a distinct framework. The proof strategy we follow is: Theorem 3 uses the results of the M-estimation theory to prove consistency of an arbitrary empirical split, while Theorem 4, the main result, extends the results in Calonico et al. (2018) to cases where the true assignment rule is unknown, providing an alternative to the methodology in Porter and Yu (2015).

Recall from Section 2.2 that, for any tree, node splits are found by the maximization of the gain in purity comparing the parent cell with their children, $\Gamma(\cdot)$. This is equivalent to minimize $\Omega(\cdot)$, the weighted impurity of both children. Assumption 9 is a necessary restriction for our proof to work and states that for any cell of any theoretical tree, the function $\Omega(\cdot)$ achieves a unique minimum over the set of viable splits candidates.

Assumption 9 (Unique Global Minimum) *Consider a deterministic complex assignment rule a and the associated set $\mathcal{D}_a$, comprising trees that correctly identify the boundary $\mathcal{A}_a$. Take arbitrarily the $(k-1)$-th cell of the m-th tree in $\mathcal{D}_a$, where $m \in \{1, ..., M_a\}$ and consider the set $\mathcal{C}_{m,k}$ of viable splits given the previous sequence $\mathcal{H}_{m,k}$. Then:*

$$h^0_{m,k} = \underset{h_{m,k} \in \mathcal{C}_{m,k}}{argmin}\ \Omega(A_{m,k-1}, h_{m,k})$$

Our proofs also use the fact that, in the RDF algorithm, once a direction (j) is chosen to split a node, its associated value lies on a discrete set of options, those that belong to the sample accepted by the root. That is, for a split $h_{m,k}$ that divides the $(k-1)$-th cell of tree $\mathcal{T}_m$, $\zeta_{m,k} \in \{\boldsymbol{x}_1^{(j_{m,k})}, ..., \boldsymbol{x}_s^{(j_{m,k})}\}$. Formally, we employ the definition of a discrete in itself bounded down set provided in Burgin (2010) and reproduced in Definition 7 for the reader's convenience.

Definition 7 (Discrete Sets - Burgin (2010)) *An arbitrary set O is discrete in itself if there is no sequence $l_O = \{o_i \in O; i = 1, 2, 3...\}$ such that $\omega = \lim l_O$ for some $\omega \in O$. Moreover, in a bounded down discrete set, all distances between any two adjacent points in O are larger than some number $k_O > 0$.*

Example 4 *A simple example of a discrete set in itself is $O = \{1/(2^i); i = 0, 1, 2...\}$. Notice that this set is not discrete in $\mathbb{R}$ since $\lim_{i\to\infty} 1/(2^i) = 0 \in \mathbb{R}$.*

Lemma 2 shows that $\hat{\Omega}(\cdot)$, the weighted impurity for child cells in a empirical tree, converges uniformly in probability to $\Omega(\cdot)$ over the set of viable

splits. We use this result in the proof of convergence of empirical splits in Theorem 3.

Lemma 2 (Uniform Convergence in Probability of $\hat{\Omega}(\cdot)$) *For any tree $\mathcal{T}_b$, $b = 1, ..., B$ and any cell $A_{b,k-1}$ it is true that for any $h_{b,k} \in \mathcal{C}_{b,k}$, uniformly:*

$$\lim_{n\to\infty} \mathbb{P} \sup_{h_{b,k}\in\mathcal{C}_{b,k}} \left|\hat{\Omega}(A_{b,k-1}, h_{b,k}) - \Omega(A_{b,k-1}, h_{b,k})\right| \xrightarrow{p} 0$$

Theorem 3 (Consistency of empirical splits) *For any tree $\mathcal{T}_b$, $b = 1, ..., B$ and any cell $A_{b,k-1}$, provided that Assumption 9 holds, an empirical split $\hat{h}_{b,k} \in \mathcal{C}_{b,k}$ is consistent in the sense that $\hat{h}_{b,k} - h^0_{b,k}$ is $o_p(1)$.*

Theorem 4 is the main result of this paper. It states that with high probability (for large B), the forest treatment effect estimative is consistent to the one induced by assignment rules of the type considered in Assumption 5. This claim is based on the following arguments, formalized in the proof of Theorem 4: from Theorem 3, empirical splits (those that minimize $\hat{\Omega}(\cdot)$)) are consistent to their population counterparts (those that minimize $\Omega(\cdot)$)); for large B the ϵ_b-Greedy rule learns which set of observables is the best, among viable alternatives, in order to maximize the reward considered (minimization of impurity $\hat{\Omega}(\cdot)$); Assumption 5 assures that there is an unknown complex assignment rule based on observables that yields boundaries perfectly identifiable by trees (in which cases, the impurity is minimum, $\Omega(\cdot) = 0$). In conclusion, with high probability, splits generated from the ϵ_b rule are the best, conditional to observables available, in order to attain maximum purity. From Theorem 3 and Assumption 5 these splits are consistent to those that correctly identify the true boundary induced by the unknown assignment rule. The forest consistency follows because it is an average learner derived from consistent base learners.

Assumption 10 builds on the general continuity properties of projections and requires that the estimative of τ in equation (2-2) to also be a smooth application for different $h_{b,k} \in \mathcal{C}_{b,k}$. In the proof of Theorem 4 we relax a bit this assumption. For any tree $\mathcal{T}_b$, $b = 1, ..., B$ and for any pair of adjacent cells $(A_{b,k_i}, A_{b,k_j}) \in \mathcal{F}_b$ sharing the l-th border, take the regression in the left side of the cutoff (referring to units inside A_{b,k_i}) as example and define $A_{b,k_i|\kappa_l} \equiv \{\boldsymbol{x}_i \in A_{b,k_i} | \zeta_{b,l} - \kappa_{b,l} \leq \boldsymbol{x}_i^{j_{b,l}} \leq \zeta_{b,l}\}$ to be a subset of the cell A_{b,k_i} comprising units to the left of an arbitrary threshold value $\zeta_{b,l}$, but within a distance $\kappa_{b,l}$. Let $\mathbf{Z}_{b,k_i|\kappa_l}$ be a $n(A_{b,k_i|\kappa_l}) \times p$ matrix and $\boldsymbol{\beta}_{b,k_i} \in \mathbb{R}^p$, both defined as:

$$\mathbf{Z}_{b,k_i|\kappa_l} = \begin{bmatrix} x_{1,b,l} - \zeta_{b,l} & \mathbf{z}^{(1)}_{1,b,l} & \cdots & \mathbf{z}^{(p-1)}_{1,b,l} \\ x_{2,b,l} - \zeta_{b,l} & \mathbf{z}^{(1)}_{2,b,l} & \cdots & \mathbf{z}^{(p-1)}_{2,b,l} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n(A_{b,k_i|\kappa_l}),b,l} - \zeta_{b,l} & \mathbf{z}^{(1)}_{n(A_{b,k_i|\kappa_l}),b,l} & \cdots & \mathbf{z}^{(p-1)}_{n(A_{b,k_i|\kappa_l}),b,l} \end{bmatrix} \quad \boldsymbol{\beta}_{b,k_i} = \begin{bmatrix} \beta_{b,l-} \\ \boldsymbol{\gamma}^{(1)}_{b,l} \\ \vdots \\ \boldsymbol{\gamma}^{(p-1)}_{b,l} \end{bmatrix}$$

where the covariates $[x\ \boldsymbol{z}]$ in $\mathbf{Z}_{b,k_i|\kappa_l}$ refer to units in $A_{b,k_i|\kappa_l}$.

Assumption 10 (Smoothness of $\hat{\boldsymbol{\beta}}(\zeta)$) *For each $b \in \{1, ..., B\}$, $l \in \{1, ..., f_b\}$, define $\mathcal{M}_{n(A_{b,k_i|\kappa_l}),p}([0,1])$ as vector spaces of $(n(A_{b,k_i|\kappa_l}), p)$-matrices with entries in the interval $[0,1]$. For any realization of the random vector $[x\ \boldsymbol{z}]'$, the functions $t_{b,k_i|\kappa_l} : \mathcal{E}_{b,k_i|\kappa_l} \to \mathcal{G}_{b,k_i|\kappa_l}$, $\mathcal{E}_{b,k_i|\kappa_l} \subset [0,1]$, $\mathcal{G}_{b,k_i|\kappa_l} \subset \mathcal{M}_{n(A_{b,k_i|\kappa_l}),p}([0,1])$, given by $t_{b,k_i|\kappa_l}(\zeta) = (\mathbf{Z}'_{b,k_i|\kappa_l}(\zeta)(\mathbf{Id} - \mathbf{P}_\iota)\mathbf{Z}_{b,k_i|\kappa_l}(\zeta))^{-1}$ are continuous, where $P_\iota = \boldsymbol{\iota}(\boldsymbol{\iota}'\boldsymbol{\iota})^{-1}\boldsymbol{\iota}'$, $\boldsymbol{\iota} = \begin{bmatrix} 1 & 1 & \cdots & 1 \end{bmatrix}'$, $n(A_{b,k_i|\kappa_{b,l}})$ dimensional. An identical assumption is made for the regression to the right side of the generic cutoff $\zeta_{b,l}$, regarding units inside $A_{b,k_j|\kappa_l}$.*

Theorem 4 (Consistency of forest treatment effect $\hat{\tau}$) *Provided that Assumptions 5 to 10 hold, with probability at least $1 - \mathbb{P}_\epsilon$, for $b \geq \mu\psi Q$, the sequence of trees $\{\mathcal{T}_b\}$, trained on random subsamples of $\mathcal{S}_n$, asymptotically identifies the boundary A_a induced by a deterministic complex assignment rule a and, for $p > 0$ finite and fixed (not growing with sample size), $\hat{\tau} - \tau$ is $o_p(1)$, where $\mathbb{P}_\epsilon$ and ψ are defined in Lemma 14.*

2.5 Simulations

In this section we assess the RDF algorithm in Definition 6, considering a treatment with an unknown complex assignment rule. As a first stage, we suggest an exploratory analysis that can be useful to shed light on the hidden cutoff. Thereafter, we estimate treatment effects (second stage) and investigate the sensitivity of the algorithm with respect to the parameters imputed by end-users.

General Setup: We work with a random sample of $n = 5000$ individual units and corresponding features uniformly draw from $[0,1]^p$, $p = 10$. The forest comprises $B = 10000$ trees. Each one of them receives the following inputs: a set W_b with $w = 5$ (50% of p) features determined by the ϵ_b-Greedy rule and a set with $s = 3750$ (75% of n) randomly draw units. Regarding the model in equation 2-2, $\forall b$ and $\forall l \in \{1, ..., f_b\}$, we assume that $e_{i,b,l} \sim N(0, 0.05)$. Also, both $\beta_{b,l-}$ and $\beta_{b,l+}$ are uniformly draw from $[-1,1]^p$ leading to $\delta_\beta = 1$ in Assumption 7. Also $\forall b, l$, $\alpha_{b,l+} = \delta_\alpha = 4$ and $\alpha_{b,l} = 0.5$, which implies a treatment effect of $\tau = \tau^{(f)}_{b,l} = \alpha_{b,l+} - \alpha_{b,l} = 3.5$.

Regarding the local linear regression estimation, each border bandwidth is selected to be MSE-optimum following the procedure established in Calonico et al. (2014) and we use an Epanechnikov kernel local to the cutoff. We halt the growing process of every tree if any cell has less than $0.02n$ points, which is related to a stopping criteria Δ by Lemma 1. That is, for $n = 5000$ and a tree $\mathcal{T}_b$, a leaf $A_{b,k}$ having at least 100 data points is associated to an implicit $\Delta \leq 2 * 0.02^{1/k}$.

The ϵ_b-Greedy rule in Definition 5 chooses actions (sets with $w = 5$ candidates to be used as splitting variables) in the set $\mathcal{Q}$ with cardinality $Q = \binom{p}{w} = 252$, using $\mu = 1$.

Example 5 (Revisiting Example 2) *Adding to the general setup we suppose that there is a treatment employing a complex unknown assignment rule a, identical to Example 2, with* $c_1 = 0.4$ *and* $c_2 = 0.6$. *That is, the true hidden boundary is* $\mathcal{A} = \{\boldsymbol{x}_i | \boldsymbol{x}_i^{(1)} = 0.4; \boldsymbol{x}_i^{(2)} \geq 0.6\} \cup \{\boldsymbol{x}_i | \boldsymbol{x}_i^{(1)} \leq 0.4; \boldsymbol{x}_i^{(2)} = 0.6\}$.

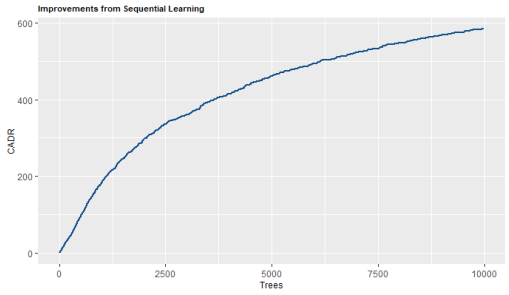

Figure 2.5: Cumulative Average Difference Reward (CADR) computed from the application of ϵ_b-Greedy rule using: $\mu = 1$, $Q = 252$ and a forest with $B = 10000$ trees.

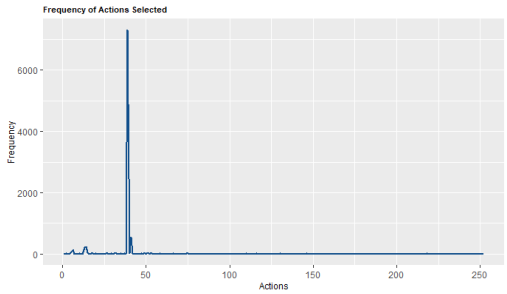

Figure 2.6: Frequency of selected actions by ϵ_b-Greedy rule using: $\mu = 1$, $Q = 252$ and a forest with $B = 10000$ trees.

First Stage of RDF Algorithm: Figure 2.5 presents the Cumulative Average Difference Reward (CADR), computed as the cumulative sum of improvements (differences) in the average reward of selected actions by the ϵ_b rule[2]. Notice that the CADR is strictly increasing at a decreasing rate. The increasing part is a confirmation that the ϵ_b rule selects actions in a way that, for every two sequential trees, the difference between average rewards observed are positive and constitute to real improvements from the learning. The decreasing rate part indicates that the learning process is near to its end and has already elected a temporary best action among the alternatives

[2]For example: when $b = 1000$, the selected action is W_{1000} and we compute its average (among every time W_{1000} was selected before) reward. Then we take the difference between this average reward and the previous one, calculated when $b = 999$ and the selected action was W_{999}, possibly different from W_{1000}. Then, we cumulatively sum all these differences.

considered. Figure 2.6 emphasizes this point by presenting the frequency that each action is selected in the forest.

Figure 2.7 presents the sum of all Gini improvements ($\Gamma(\cdot)$) when a particular variable is used to split a parent cell, considering all splits in the forest. Notice that in these cases that we do not know the assignment rule, we should suspect that it comprises variables 1 and 2 in some way. In fact, the sum of the cumulative Gini improvement generated by the first and the second variables together (5095.07) is more than 193 times the sum of the cumulative Gini improvement generated by the other eight variables (26.38). This is what the ϵ_b-Greedy rule tries to learn, that whenever the first and/or the second variables are selected as candidates for splitting, there is more purity gain compared to cases when the other variables are used.

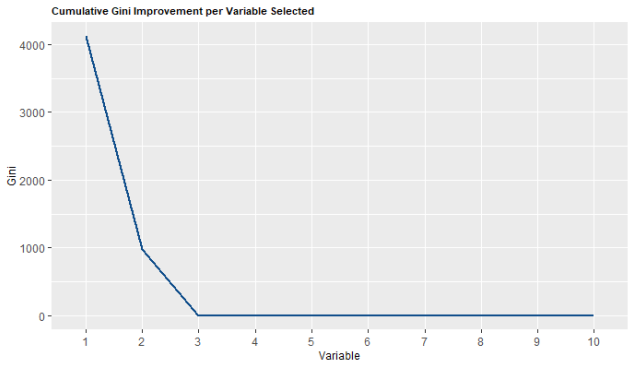

Figure 2.7: Cumulative Gini improvement per variable considering every split in the forest with $n = 5000$ units, $B = 10000$ trees, $w = 0.5p$ candidates for splitting variable and $s = 0.75n$ units randomly selected for each tree.

Each of the ten panels from Figure 2.8 to 2.17 presents the distributions of the cutoff value for each of the ten variables, considering all splits in the forest where that particular variable was selected by the algorithm to split a parent cell. Recall the assignment rule used in this example and, notice in Panel 2.8, that the first variable is selected around a cutoff value of 0.4 with a much more pronounced frequency than it is selected at every other possible value. The same happens with the second variable in Panel 2.9, but at the cutoff value of 0.6. However, a different behavior is observed in the other eight variables. Take the third variable as an example in Panel 2.10 and notice that its selection frequency is not concentrated around any possible cutoff value, indicating that the algorithm is unable to decide which value is the best to split cells among the possible alternatives. Combining this result with the relatively small Gini improvements when the third variable is used, in Figure 2.7, the conclusion points to a variable that is not important to locate the true hidden boundary. And the same happens from the fourth to the tenth variable.

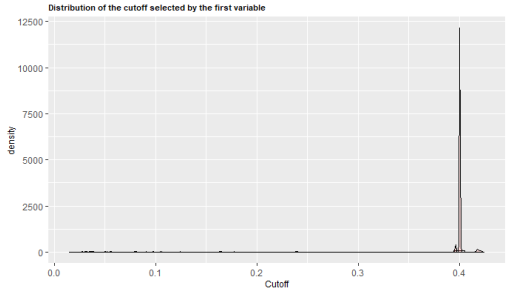

Figure 2.8: Distribution of the cutoff selected by the first variable.

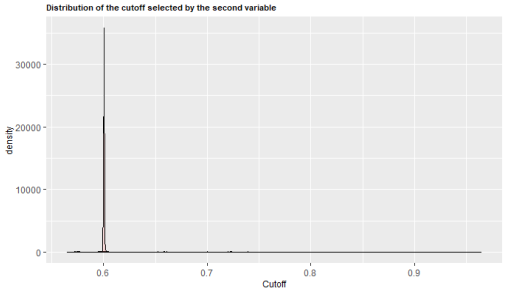

Figure 2.9: Distribution of the cutoff selected by the second variable.

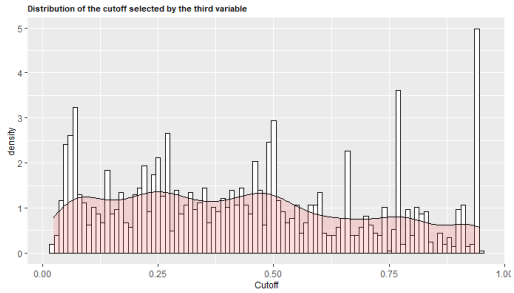

Figure 2.10: Distribution of the cutoff selected by the third variable.

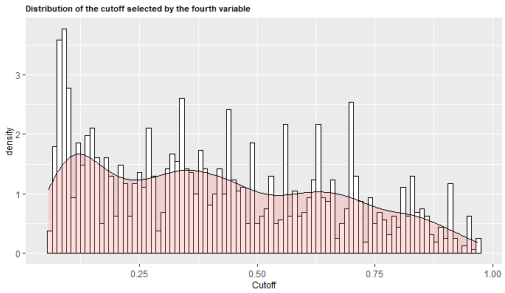

Figure 2.11: Distribution of the cutoff selected by the fourth variable.

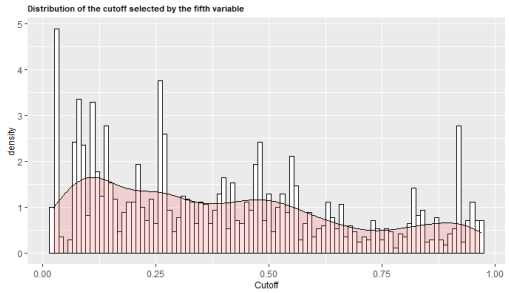

Figure 2.12: Distribution of the cutoff selected by the fifth variable.

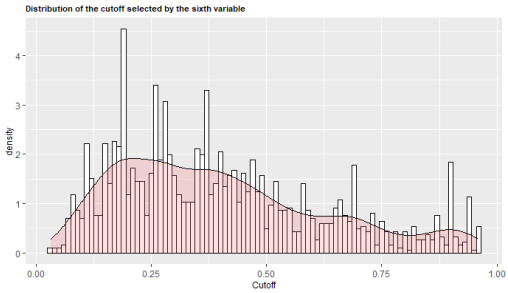

Figure 2.13: Distribution of the cutoff selected by the sixth variable.

Figure 2.18 presents the misclassification rate observed in trees when different types of combination of pre-selected variables occur, considering the same setup as before. More specifically, the type 1×1 in Figure 2.18 denotes that in a particular tree, both the first and the second variables belong to the set of selected variables by the ϵ_b-Greedy rule to be considered as splitting candidates. The combination 1×0 means that the first variable is pre-selected but the second is not and the other types follow the same rationale. The misclassification rate of the type 1×1 is at maximum 0.19% while the type 0×0 accounts for a maximum misclassification of 16.08%.

To sum up, even in a unknown cutoff scenario, an exploratory analysis like the proposed first stage, from Figure 2.5 to Figure 2.18 (which clearly can be expanded), can be helpful to shed light on the rules used in the program. For instance, in our example, we would be pretty confident to state that the

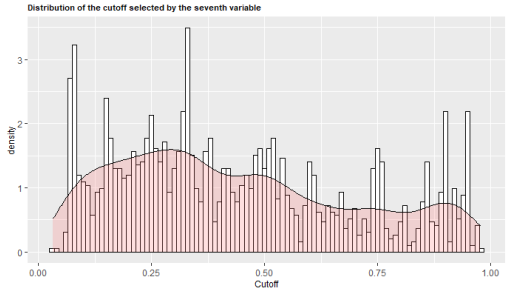

Figure 2.14: Distribution of the cutoff selected by the seventh variable.

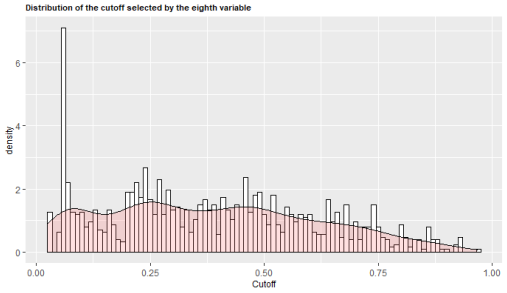

Figure 2.15: Distribution of the cutoff selected by the eighth variable.

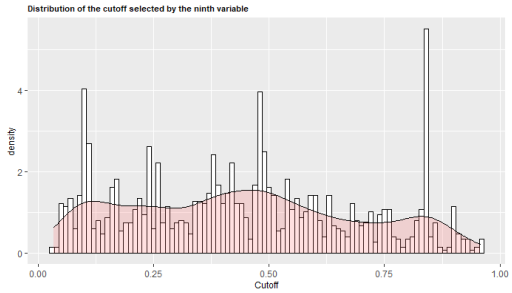

Figure 2.16: Distribution of the cutoff selected by the ninth variable.

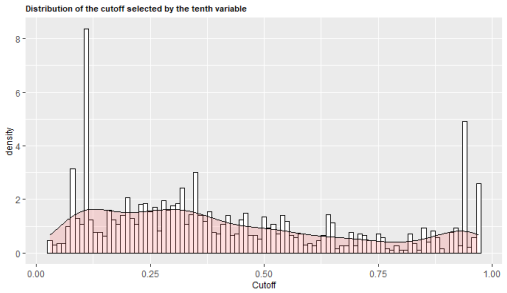

Figure 2.17: Distribution of the cutoff selected by the tenth variable.

first and the second variable both play important roles as assignment features and the values of 0.4 and 0.6, respectively, seem to be important cutoffs to be considered.

Second Stage of RDF Algorithm: Estimation of treatment effects could be carried out jointly with the learning process in the first stage, or in a separate shorter step, that considers information gathered in the first one. We pursue the second option[3] and perform a single simulation considering the same setup described in this section, but with $B = 5000$ trees. The reduced forest is possible in this case since we take into consideration the important role played by the first and second variables, by forcing then to always be provided to the roots of the trees, that are still trained on random subsamples. Figure 2.19 presents the distribution of tree treatment effect estimatives.

Figures 2.20 and 2.21 provide a preliminary sensitivity analysis of the RDF estimatives of tree treatment effects estimatives with respect to parameters imputed by end-users: the size of the forest B (left panel) and the size of the subsample provided to each tree s (right panel). We use the same general setup used in Figure 2.19 and consider $B \in \{100, 500, ..., 10000\}$ for Figure

[3]In a separate exercise we estimate treatment effects using the first option and we get very similar results: Estimatives of tree treatment effects reasonably surrounds the true one (3.5). We observe that 94.26% of the estimated $\hat{\tau}^{(t)}$ are inside the interval $[3, 4]$. Outliers are frequent in the initial stage of the learning process or when the ϵ_b-Greedy rule decides to randomly explore a new action.

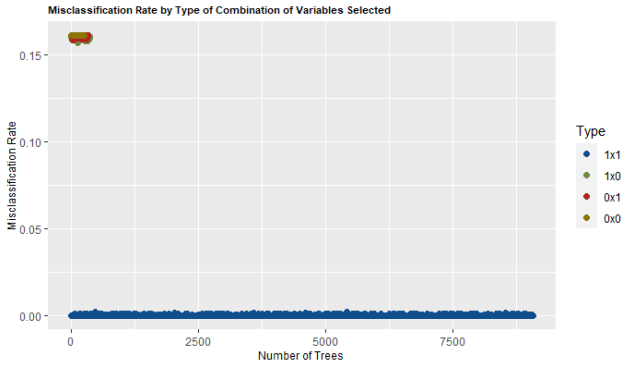

Figure 2.18: Misclassification rate for different types of combination of pre-selected variables occur. Simulations consider $n = 5000$ units, $B = 10000$ trees, $w = 0.5p$ candidates for splitting variable and $s = 0.75n$ units randomly selected for each tree.

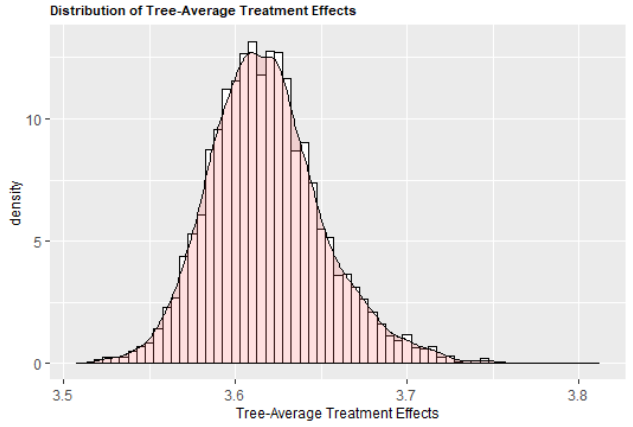

Figure 2.19: Distribution of tree-average treatment effect estimatives for a forest with $n = 5000$ units, $B = 5000$ trees, $w = 0.5p$ candidates for splitting variable conditional to the fact that the first and the second are always selected and $s = 0.75n$ units randomly selected for each tree.

2.20 and $s \in \{0.5n, 0.6n, ..., n\}$ for Figure 2.21. We observe that there is a very mild dependency between tree-average treatment effects and the size of the forest, which seems to be an issue restricted to relatively small forests. After some point, both average and dispersion of estimatives do not vary considerably with B. Also, Figure 2.21 depicts a very intuitive result that the smaller the subsample provided the larger the dispersion in results, but with virtually no effect on the average estimative.

2.6 Revisiting the P900 - A Chilean Government Assistance to Low Performing Schools

In this section we apply the methods described in this paper in part of a real public program, the P900. First, we provide a brief overview of its main

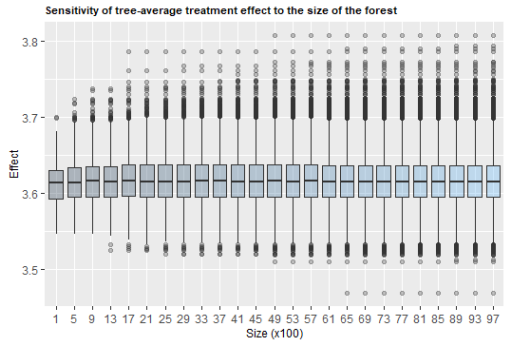

Figure 2.20: Sensitivity of tree-average treatment effect estimatives with respect to the size of forest. Simulations use $n = 5000$ units, $w = 0.5p$ candidates for splitting variable conditional to the fact that the first and the second are always selected,$s_b = 0.75n$ units randomly selected for each tree and $B \in \{100, 500, ..., 10000\}$.

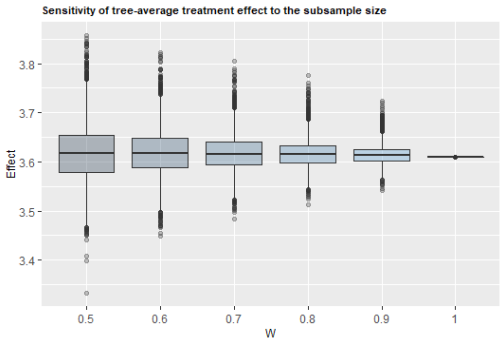

Figure 2.21: Sensitivity of tree-average treatment effect estimatives with respect to the size of the subsample admitted to each tree in the forest. Simulations use $n = 5000$ units, $w = 0.5p$ candidates for splitting variable conditional to the fact that the first and the second are always selected, $s \in \{0.5n, 0.6n, ..., n\}$ units randomly selected for each tree and $B = 5000$.

characteristics and the reader can get a richer set of details in several other papers that have been studying the P900, such as in Chay et al. (2005) and references therein. In the sequence, we implement the first and second stages of the RDF algorithm described in Section 2.5 to a selected subset of the program just to underline how our procedure could be useful in real contexts with unknown and possibly complex assignment rules.

2.6.1 Brief Overview of P900

Back in 1990, the Chilean government implemented a program to assist low fourth-grade performing, publicly funded schools, the so called P900, since it initially identified approximately 900 schools to be treated. After proper selecting schools in thirteen administrative regions in Chile, the program basically consisted of four waves of support, where in each of them, schools received improvements in their infrastructure, a variety of instructional materials, training workshops for related teachers and after school tutoring for low-performing students. The first two years of the program (1990 and 1991) focused on the two first waves, providing the basic means for schools and only in 1992 the third and fourth waves started.

Even though the P900 main objective was to enhance educational learning and overall results in the Chilean administrative regions, the program's assignment rule not always selected schools based solely on their students' grades. Actually, according to Chay et al. (2005) selection occurred essentially in two stages where only the first one relied on the results of tests applied to

fourth-graders in 1988, basically the average between language and mathematics scores. Later in a second stage, regional teams of officials reviewed each list of preselected schools and changed them according to other criteria.

Our data set is similar to that used in Chay et al. (2005) and is composed of the following variables: A unique identifier for each school, its administrative region, a dummy variable indicating if the school is located inside an urban or in a rural area, the average grades per school both in language and in mathematics in 1988, 1990 and in 1992, the number of students that took the test per school in 1988, 1990 and in 1992 and a socioeconomic index per school in 1990 and in 1992, where the higher the index the high the vulnerability (poverty). Moreover, we create the following variables: the average between the language and mathematics grades per school in 1988 (Avg), the difference between grades in language and in mathematics from 1988 to 1992 (Glan8892 and Gmat8892), the difference in the socioeconomic index from 1990 to 1992 (Gsei9092) and the square and cubic of the average between the language and mathematics grades per school in 1988 (Savg and Cavg). Although we find both the original and created variables self-instructive, since we use exactly the same variables as in Chay et al. (2005), any question about them can be settled by visiting their work. Table 2.1 provides descriptive statistics for selected variables per administrative region.

Notice that both the average grades in 1988 and their variabilities are relatively similar among regions, as they are the differences in average grades from 1988 to 1992. Also, there is a weak negative relation between average grade (Avg (88)) and percentage of participation in the program (%P900), indicating that grades are, indeed, important to assign units to treatment. However, there are some inconsistencies, as commented in Chay et al. (2005) and further elaborated later in this section. To briefly illustrate the point, compare the thirteenth region with the first one. In a very simple analysis, we should expect, on average, a higher rate of participation in the former since it presented a lower average grade than the latter. However, the participation rate in the first region was more than twice that on the thirteenth region.

The deterministic nature of the program suggests that an RDD should be employed to recover potential impacts of the government aid in treated schools. However, Example 3 in Section 2.2 highlights two intrinsic characteristics of the P900 that pose themselves as difficulties to the standard application of the sharp RDD. Firstly, the Chilean government did not make public the specific grade values that should be used in assigning units to treatment groups. Secondly, even if a researcher tries to uncover the implicit cutoff used by the government, she will conclude that there is no simple cutoff based solely in

Table 2.1: *Descriptive statistics of selected observables per region. Avg.(88) is the average grade in language and in mathematics in 1988. Math (88-92) and Lang (88-92) are the differences between grades in language and in mathematics from 1988 to 1992, SEI (90) is the socioeconomic index in 1990, %P900 is the percentage of schools that received the benefits of P900 and Size is the number of schools per region. All quantities except Size are expressed as average per region. The left parenthesis is the minimum value per region, the centered parenthesis is the standard deviation and the right is the maximum value.*

Region	Avg. (88)	Math (88-92)	Lang (88-92)	SEI (90)	%P900	Size
1	56.8 (34.5),(8.6),(75.6)	14.6 (-5.7),(8.2),(43.9)	10.1 (-5.7),(7.1),(31.9)	30.3 (2.0),(20.7),(82.8)	25.4	59
2	55.2 (39.4),(8.4),(77.7)	14.6 (-8.9),(8.3),(50.4)	10.6 (-9.6),(8.0),(47.4)	39.1 (8.5),(15.4),(90.1)	12.2	74
3	56.7 (32.7),(9.4),(74.6)	14.6 (-2.9),(7.5),(33.2)	10.3 (-6.6),(7.2),(25.5)	33.1 (1.4),(19.0),(84.4)	18.2	55
4	51.3 (24.8),(10.8),(82.5)	13.3 (-29.7),(10.6),(51.9)	10.8 (-30.0),(9.6),(40.6)	51.8 (0.8),(24.6),(94.7)	32.3	192
5	53.9 (32.3),(8.9),(79.5)	12.7 (-21.8),(8.7),(44.0)	9.8 (-33.7),(8.9),(39.2)	35.4 (0.2),(22.7),(87.1)	15.2	433
6	49.5 (29.6),(10.2),(86.8)	14.5 (-34.0),(11.5),(46.3)	13.3 (-30.2),(10.6),(40.2)	56.3 (0.5),(24.5),(97.7)	18.1	288
7	47.5 (18.7),(10.4),(81.6)	13.8 (-31.7),(10.7),(47.3)	12.8 (-29.5),(10.2),(46.6)	60.2 (0.0),(26.0),(100.0)	21.5	382
8	48.5 (23.1),(11.0),(80.9)	13.3 (-37.2),(10.5),(54.1)	11.3 (-29.0),(9.4),(46.9)	49.9 (0.7),(26.0),(100.0)	26.4	629
9	47.2 (21.3),(10.2),(81.3)	10.9 (-37.3),(9.8),(40.8)	10.8 (-14.7),(8.7),(34.9)	46.8 (2.4),(24.9),(99.2)	44.3	282
10	48.7 (24.3),(9.7),(83.5)	14.3 (-22.7),(10.4),(45.3)	12.3 (-33.6),(9.9),(49.1)	56.6 (0.5),(25.3),(100.0)	41.3	346
11	55.3 (40.8),(8.8),(75.5)	16.1 (-0.8),(8.8),(39.6)	12.0 (-6.8),(9.2),(32.8)	44.0 (4.4),(23.6),(88.4)	35.0	20
12	60.9 (46.5),(7.8),(82.7)	16.1 (0.8),(8.4),(37.8)	11.4 (-4.1),(7.0),(27.6)	26.0 (6.5),(16.1),(69.6)	9.7	31
13	53.1 (29.0),(9.4),(83.4)	13.2 (-16.4),(7.9),(44.5)	11.3 (-15.9),(7.5),(40.1)	30.3 (0.1),(19.5),(88.4)	11.2	1087

grades that can possibly be adopted in order to segregate units. Actually, although Example 3 refers to the first administrative region in Chile, the absence of a solely grade-based cutoff can also be verified in all other twelve administrative regions. The first problem can be adressed by the methodology proposed in this paper, but the second one is a little bit more intricate. While in some administrative regions (such as region 1) we are able to find a hidden complex rule that justifies the sharp methodology, there are other regions where it is impossible to recover an exact assignment rule, considering the whole set of observed covariates. This is the case where unobservables and/or discretion may have played some role in the assignment rule, conflicting with Assumption 5, since no tree-based methodology can be completely accurate, based on the sample collected and the available set of covariates. In this respect, we follow Chay et al. (2005) and accept a minimum level of misclassification by efficiently choosing the cutoff.

To illustrate, consider the example of administrative region 2 restricted to urban schools. As an exploratory exercise, consider fitting the highest possible tree to the data, in the sense that we allow it to grow to its most extent until, perhaps, one single unit remains in a given leaf. Although it is not a feasible tree for our purposes of estimating treatment effects, it provides insights of

what may be going on behind the scenes. In this exercise, we find that the most important variables, in terms of purity gain in the splitting process, seem to be the average grade and enrollment in the 1988 test. Actually, 66.7% of treated units present average grade less than 47.62 and attendance to the test grater than 58 students, which is consistent with described in Chay et al. (2005) that small schools did not actually participate in the program in order to reduce costs. However, there is a treated school with average grade equal to 47.54, enrollment equal to 48 and a non-treated school with 44.92 for the average grade and where 57 students took the test. Since both schools presented very similar socioeconomic indicators (48.55 and 41.15), we cannot find a plausible justification based on available observables, for a relatively larger (enrollment) and worse (grades) school to not be treated. As a consequence, the highest tree above-mentioned still misclassify one school.

First Stage of RDF Algorithm:

To illustrate the potential of the first stage of the RDF algorithm, we apply it to the ninth administrative region (south of Chile), which was also used in Figure 3 of Chay et al. (2005) as an example to highlight the difficulties intrinsic to the P900 assignment rule. Since a cutoff based on grades is unknown, the authors estimate it based on two different definitions. The first one places the cutoff at the rounded-up value of the highest (average) score observed among all treated schools in the region and the result is that only 55.1% and 69.3% of units are correctly classified, considering either all schools in the sample or its restriction to the urban larger schools (more than 15 students enrolled in the test), respectively. The second one defines the cutoff as the score that maximizes the percentage of schools correctly classified across all thirteen regions. In this case the authors achieve 75.5% and 98.0% of correct classification considering all schools or only the urban larger units.

We use the same specification as in Section 2.5 but with 5000 trees, where each one of them receives a random subsample of the ninth region comprising 75% of the total number of units in this region. We consider five observables for cell splitting, resulting in a ϵ_b-Greedy rule running on a set of actions $\mathcal{Q}$ with cardinality $Q = \binom{5}{3} = 10$. They are: the grade in language, the attendance to the test, the grade in mathematics, the socioeconomic index and the mean grade in 1988. The other variables collected refer to years after 1988, unknown at the time of the program implementation and, because of this were not used in this analysis. To better understand which variables are important to assignment, we let trees to grow to their most extent, until at most one school remains at final leaves. Figures 2.22 and 2.23 present information from the application of the ϵ_b-Greedy learning rule, while Figure 2.24 presents the

cumulative Gini improvement per variable, considering every split in the forest.

Figures 2.22 and 2.23 present similar results to what was observed through simulation in Figures 2.5 and 2.6. In the left panel, the CADR (defined in Section 2.5) increases at a decreasing rate denoting the benefits from learning, while in the right panel the algorithm selects the best action to be the set $\{Fre, Sei, Avg\}$. Figure 2.27 exhibits the cumulative Gini improvement per variable considering every split in the forest, in which we confirm that average grade is the most powerful variable to increase purity, followed by the enrollment in the test. In fact, the cumulative Gini improvement generated by the average grade is more than 30 times the one resulted from the participation in the test, which is not so different from the other variables.

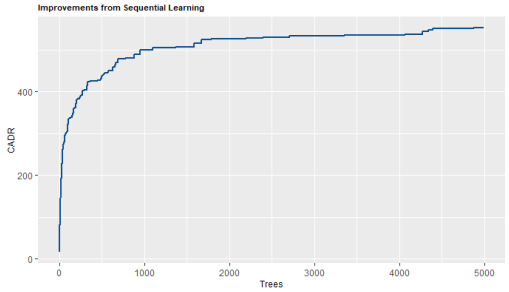

Figure 2.22: Cumulative Average Difference Reward (CADR) computed from the application of ϵ_b-Greedy rule using: $\mu = 1$, $Q = 20$ and a forest with $B = 5000$ trees.

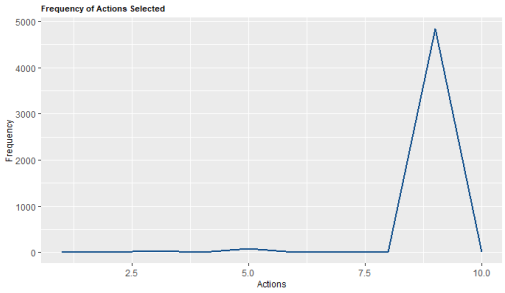

Figure 2.23: Frequency of selected actions by ϵ_b-Greedy rule using: $\mu = 1$, $Q = 20$ and a forest with $B = 5000$ trees.

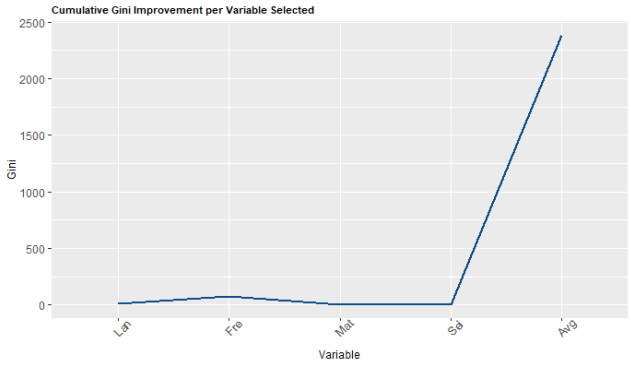

Figure 2.24: Cumulative Gini improvement per variable considering every split in the forest with $B = 5000$ trees, $s = 75\%$ of units of the ninth region randomly selected for each tree, $W = \{Lan, Fre, Mat, Sei, Avg\}$ is the set of candidates for splitting variable and $w = \lceil 0.5\#W \rceil$ is the amount of variables in W selected by the ϵ_b-Greedy rule, where each variable's label in W follows the above-defined explanation.

Moreover, Figure 2.25 exhibits the frequency of selected cutoff values when the average grade is the splitting variable. Notice that the frequency at which the value of 47.45 is selected is about four times the second largest

frequency, associated to the value of 47.44 (both very close numbers). Figure 2.26 presents the same analysis but based on the frequency of selected cutoff values when the attendance to the test is the splitting variable. In this case it appears that a probable cutoff value would be 69.

To sum up, recall our discussion about the important role that unobservables may play in the assignment rule of P900. Following Assumption 5 it is difficult for a tree-based methodology based on a limited set of observables to recover the exact boundary induced by the program. However, if precision is not mandatory, one should recognize that a once completely unknown assignment rule is, in fact, something in the neighborhood of a complex rule composed by an average grade of 47.45 and, marginally, by frequency to the test close to 69.

Figure 2.27 depicts the percentage of correct classification considering administrative region 9. On average, our procedure achieves 99.5% of schools correctly classified, which is better than the results already commented in this subsection obtained by Chay et al. (2005). The worst result refers to 91.8% of correct classification that occurs in a sole event when the learning rule explores a bad action. On the other hand, in 25.9% of the trees the ϵ_b-Greedy rule chooses actions that generates 100% of correct classified schools.

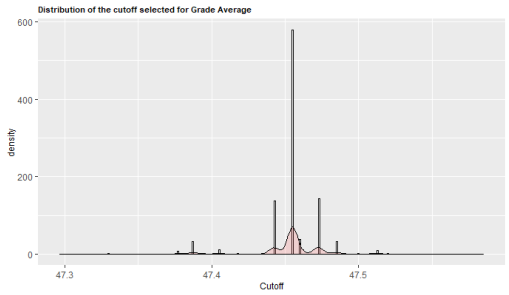

Figure 2.25: Distribution of the cutoff selected by the test grade average in 1988.

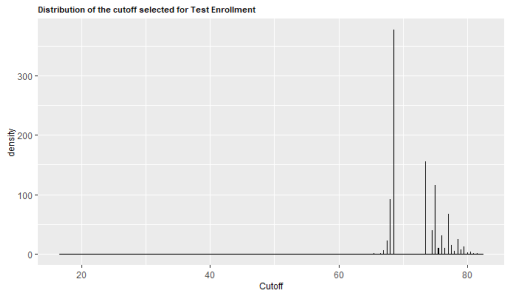

Figure 2.26: Distribution of the cutoff selected by the test attendance in 1988.

The RDF first stage procedure to explore unknown boundaries is not directly comparable to the two cutoffs definitions in Chay et al. (2005). Different from their paper, our procedure considers that cutoffs might be multivariate. Also, the authors use cutoffs that are globally best, for all thirteen regions as a whole. Nevertheless, we extend the analysis made in Figure 2.27 to the other twelve regions in Chile and Table 2.2 compares the percentage of correctly classified units between our procedure and those based on the two cutoff definitions cited in Chay et al. (2005). We consider urban larger schools, but its extension to the whole sample yields similar results. Since for each administrative region we have a forest composed of different trees, the percentage of correct classification per region reported in Table 2.2 (Tree-

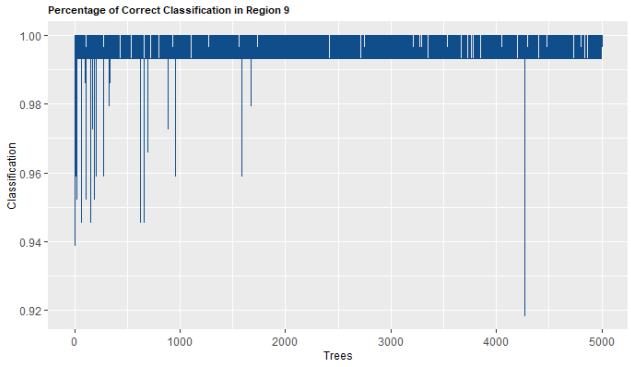

Figure 2.27: Percentage of correct classification considering administrative region 9. We use $B = 5000$ trees, $s = 75\%$ of units of the ninth region randomly selected for each tree, $W = \{Lan, Fre, Mat, Sei, Avg\}$ is the set of candidates for splitting variable and $w = \lceil 0.5\#W \rceil$ is the amount of variables in W selected by the ϵ_b-Greedy rule, where each variable's label in W follows the above-defined explanation.

Cutoff) is the respective forest average. Also, for each region, we present the stopping criteria adopted as the average minimum number of schools in any treated leaf (Leaf Size). In all regions, we achieve higher rates of classification.

Second Stage of RDF Algorithm:

In this part we apply the second stage of RDF Algorithm to the ninth administrative region, concluding the analysis started in the previous paragraphs. One of the reasons that led us to investigate this region is the fact that several regions in our data set do not possess enough data points for a reliable estimation. For instance, Table 2.2 shows that regions 1, 2, 3, 6, 7, 11 and 12 present, each one of them, on average, less than thirteen treated units in a given leaf.

In the case of the ninth region, there are 196 urban larger schools in the sample, from which 84 received the benefits of the program. We use the same set of covariates employed to build Table 3 in Chay et al. (2005), that is: the participation in the program indicator, the average grade per school in 1988, the socioeconomic index in 1990 as well as its evolution from 1990 to 1992, and the cubic average grade in 1988. The set with viable actions for the ϵ_b-Greedy has cardinality $Q = \binom{4}{2} = 6$ and we use the same specification as before: $B = 5000$, $w = 2$ and $s = 147$ (75% of 196). Figure 2.28 illustrates that the learning procedure quickly focus on the best perceived action: the set $\{Sei, Avg\}$.

Figures 2.29 to 2.32 present information of estimated Border Treatment Effects (BTE) for grades in mathematics. At each identified border (see Section 2.3), we estimate BTE using equation (2-2) and Figure 2.29 exhibits their values, per border, while Figure 2.30 presents their frequencies. Figures 2.31

Table 2.2: *Comparison between the percentage of correctly classified schools resulting from the stage 1 of RDF algorithm and those generated under the scope of the cutoffs definition 1 and 2 described in Chay et al. (2005) for the thirteen administrative regions in Chile. We consider urban larger schools in the sample and leaf size is the minimum number of schools in any treated leaf.*

Region	Definition 1	Definition 2	Tree-Cutoff	Leaf Size
1	100.0	98.0	100.0	9
2	91.4	91.4	97.2	4
3	95.7	95.7	97.9	7
4	87.4	94.7	95.8	17
5	74.5	89.8	93.2	33
6	75.8	97.6	98.4	6
7	79.6	97.5	97.6	8
8	69.5	97.1	97.7	29
9	69.3	98.0	99.5	19
10	57.2	91.3	91.7	34
11	87.5	87.5	89.0	5
12	92.9	89.3	93.1	5
13	32.0	86.4	95.4	49

and 2.32 show a more focused information than in both previous figures, after eliminating some extreme values and inserting a 95% confidence bound for each estimative. Figures 2.33 to 2.36 follow the same rationale but investigate possible effects in language grades, also from 1988 to 1992.

Firstly, notice from Figures 2.29 and 2.33 that extreme values of BTE estimatives occur when the ϵ_b-Greedy is in its early stage or when it randomly explores a new action. Since the set of actions is not large, the learning methodology quickly decides which action is the best, leading to 96.3% of BTE estimatives concentrated inside the interval $[-0.2, 0.2]$ when considering possible effects in mathematics grades, for example. A similar pattern occurs in the case of language grades and we avoid dealing with this small portion of extreme values.

Also, Figures 2.31 and 2.35 indicate that P900 did not produce benefits to the ninth region. Actually, considering the possible effects in mathematics grades (Figure 2.31), after removing extreme values, 71.5% of all 95%-confidence intervals would comprise true negative treatment effects for the respective borders. In this case, the average of BTE estimatives is -0.064. When it comes to language grades (Figure 2.35), 76.6% of all confidence intervals are composed of strictly negative values, while the average of BTE estimatives is -0.047. Only 10.8% and 8.6% of confidence intervals in mathematics and

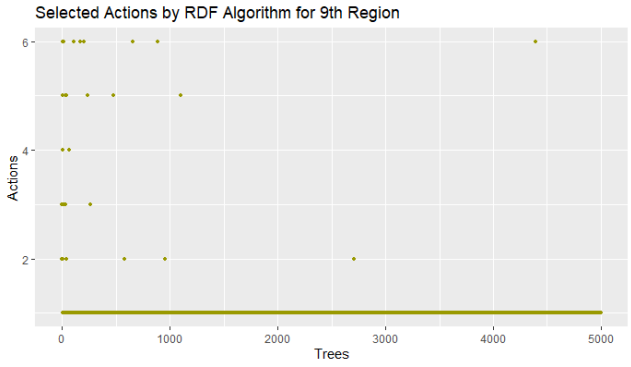

Figure 2.28: Selected actions by ϵ_b-Greedy rule using: $\mu = 1$, $Q = 6$ and a forest with $B = 5000$ trees, considering the sample of urban larger schools of the ninth administrative region. We allow $s = 75\%$ of units to be randomly selected and imputed to each tree and $W = \{Sei, Avg, Gsei9092, Cavg\}$ is the set of candidates for splitting variable, with $w = 2$.

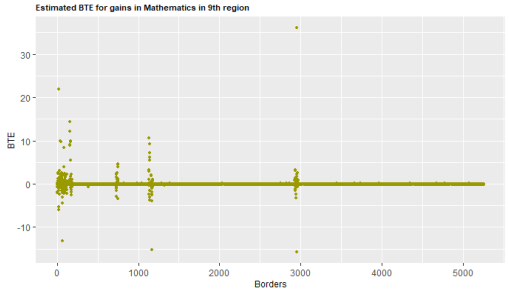

Figure 2.29: BTE estimatives per border, considering the observed change in mathematics grades, from 1988 to 1992, in the ninth administrative region. We use $B = 5000$, $w = 2$, $s = 147$ and $W = \{Sei, Avg, Gsei9092, Cavg\}$. For the ϵ_b-Greedy, we adopt $\mu = 1$ and $Q = 6$.

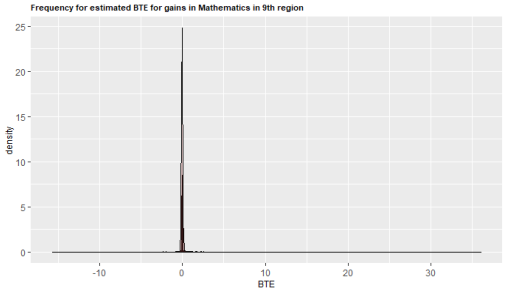

Figure 2.30: Histogram of BTE estimatives, considering the observed change in mathematics grades, from 1988 to 1992, in the ninth administrative region. We use $B = 5000$, $w = 2$, $s = 147$ and $W = \{Sei, Avg, Gsei9092, Cavg\}$. For the ϵ_b-Greedy, we adopt $\mu = 1$ and $Q = 6$.

language, respectively, are in the positive region, with respective averages of 0.044 and 0.031.

We relate this finding with a possible heterogeneity across regions. In fact, in table 3 of Chay et al. (2005) the authors found a positive statistically significant overall effect, considering all regions and urban larger schools in the sample. That is, after determining the best cutoff for each region, the authors computed the difference from the average grade in each urban larger school to the respective region's cutoff. And they used all units in all regions to find a positive effect of the program both in mathematics and in language from 1988 to 1992. Once more, we highlight that the procedure proposed in this paper is not directly comparable to the one in Chay et al. (2005), given the possibility for complex assignment rules. Nevertheless, there are not enough evidences in the results presented from Figure 2.29 to 2.36 to support a positive effect of the program in the ninth region, whether in mathematics or in language.

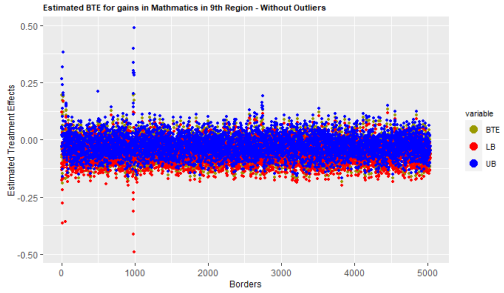

Figure 2.31: BTE estimatives per border and 95% confidence bounds after eliminating extreme values, considering the observed change in mathematics grades, from 1988 to 1992, in the ninth administrative region. We use $B = 5000$, $w = 2$, $s = 147$ and $W = \{Sei, Avg, Gsei9092, Cavg\}$. For the ϵ_b-Greedy, we adopt $\mu = 1$ and $Q = 6$.

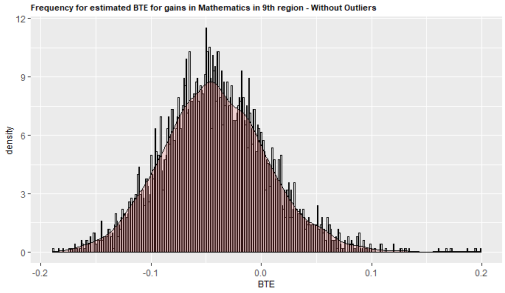

Figure 2.32: Histogram of BTE estimatives after eliminating extreme values, considering the observed change in mathematics grades, from 1988 to 1992, in the ninth administrative region. We use $B = 5000$, $w = 2$, $s = 147$ and $W = \{Sei, Avg, Gsei9092, Cavg\}$. For the ϵ_b-Greedy, we adopt $\mu = 1$ and $Q = 6$.

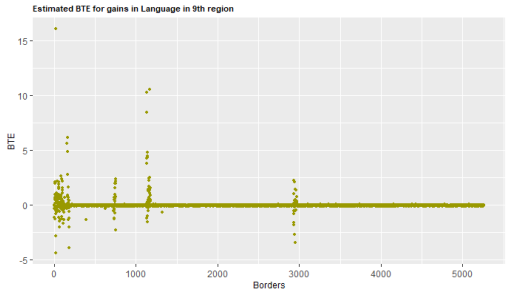

Figure 2.33: BTE estimatives per border, considering the observed change in Language grades, from 1988 to 1992, in the ninth administrative region. We use $B = 5000$, $w = 2$, $s = 147$ and $W = \{Sei, Avg, Gsei9092, Cavg\}$. For the ϵ_b-Greedy, we adopt $\mu = 1$ and $Q = 6$.

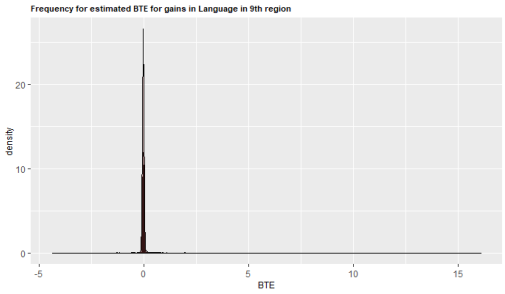

Figure 2.34: Histogram of BTE estimatives, considering the observed change in Language grades, from 1988 to 1992, in the ninth administrative region. We use $B = 5000$, $w = 2$, $s = 147$ and $W = \{Sei, Avg, Gsei9092, Cavg\}$. For the ϵ_b-Greedy, we adopt $\mu = 1$ and $Q = 6$.

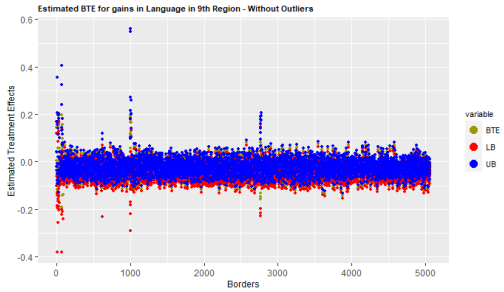

Figure 2.35: BTE estimatives per border and 95% confidence bounds after eliminating extreme values, considering the observed change in Language grades, from 1988 to 1992, in the ninth administrative region. We use $B = 5000$, $w = 2$, $s = 147$ and $W = \{Sei, Avg, Gsei9092, Cavg\}$. For the ϵ_b-Greedy, we adopt $\mu = 1$ and $Q = 6$.

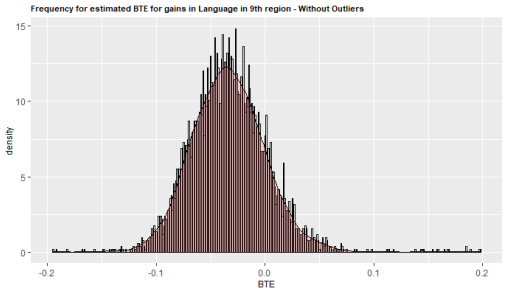

Figure 2.36: Histogram of BTE estimatives after eliminating extreme values, considering the observed change in Language grades, from 1988 to 1992, in the ninth administrative region. We use $B = 5000$, $w = 2$, $s = 147$ and $W = \{Sei, Avg, Gsei9092, Cavg\}$. For the ϵ_b-Greedy, we adopt $\mu = 1$ and $Q = 6$.

2.7 Concluding Remarks

In this paper we propose a new class of estimators that combine tree-based methodologies with sequential leaning, that can be especially useful in situations where a complex unknown deterministic assignment rule is in place. We contextualize the motivation of this particular setup with real word examples stemming from explicit non-disclosure of treatment's rule, perhaps to avoid competition, manipulation or due to ethical reasons, to innocuous disclosed rules that are not followed.

Theoretically speaking, we aid to the treatment effect literature, more specifically to the sharp regression discontinuity design, by providing a consistent way to estimate a program's impact without having to care about the real cutoff. On the other hand, when the knowledge of the real assignment rule is valued in practical applications, we provide an example of an exploratory analysis that can be fruitful in this regard. We also show, in a simple robustness check that parameter imputed by end-users do not greatly impact treatment effects estimatives.

Also, we employ our procedure on part of the P900, a real Chilean program created to assist low-performing schools. Recognizing that cutoffs are not disclosed and may be complex, as exemplified, the procedure proposed in this paper sheds light on the unknown assignment rule. It also reveals a possible level of heterogeneity among regions covered by the program, a topic that, as far as we know, was never discussed by the literature on P900.

3
Conclusions

In this thesis we use sequential learning in two different setups. In the first chapter, we extend one of the most popular learning solutions, the ϵ_t-greedy heuristics, to high-dimensional contexts considering a conservative directive. We do this by allocating part of the time the original rule uses to adopt completely new actions to a more focused search in a restrictive set of promising actions. The resulting rule might be useful for practical applications that still values surprises, although at a decreasing rate, while also has restrictions on the adoption of unusual actions. We find that, with high probability, cumulative regret of a conservative high-dimensional decaying ϵ_t-greedy rule is reasonably bounded. We also provide a lower bound for the cardinality of the set of viable actions that implies in an improved regret bound for the conservative version when compared to its non-conservative counterpart. Additionally, we show that end-users have sufficient flexibility when establishing how much safety they want, since it can be tuned without impacting theoretical properties. We illustrate our proposal both in a simulation exercise and using a real dataset.

In the second chapter we study deterministic treatment effects when the assignment rule is both more complex than traditional ones and unknown to the public perhaps, among many possible causes, due to ethical reasons, to avoid data manipulation or unnecessary competition. We circumvent the lack of knowledge of true cutoffs by employing a forest of classification trees, which also uses the sequential learning rule ϵ-greedy, as in the first chapter, to guarantee that, asymptotically, the true unknown assignment rule is correctly identified. The tree structure also turns out to be suitable if the program's rule is more sophisticated than traditional univariate ones. We show that, with high probability and based on reasonable assumptions, it is possible to consistently estimate treatment effects under this setup. For practical implementation we propose an algorithm that not only sheds light on the previously unknown assignment rule but also is capable to robustly estimate treatment effects regarding different specifications imputed by end-users. Moreover, we exemplify the benefits of our methodology by employing it on part of the Chilean P900 school assistance program, which proves to be suitable for our framework.

Bibliography

[1] ABBASI-YADKORI, Y.; PÁL, D. ; C.SZEPESVÁRI. Improved algorithms for linear stochastic bandits. In: Shawe-Taylor, J.; Zemel, R.; Bartlett, P.; Pereira, F. ; Weinberger, K., editors, ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS 24, p. 2312–2320. 2011.

[2] ABBASI-YADKORI, Y.; PÁL, D. ; SZEPESVÁRI, C.. Online-toconfidence-set conversions and application to sparse stochastic bandits. In: PROCEEDINGS OF THE 15TH INTERNATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE AND STATISTICS, p. 1–9, 2012.

[3] AGRAWAL, S.; GOYAL, N.. Analysis of Thompson sampling for the multi-armed bandit problem. In: PROCEEDINGS OF THE 25TH ANNUAL CONFERENCE ON LEARNING THEORY, p. 39.1–39.26, 2012.

[4] AGARWAL, A.; BASU, S.; SCHNABEL, T. ; JOACHIMS, T.. Effective evaluation using logged bandit feedback from multiple loggers. In: PROCEEDINGS OF THE 23RD ACM SIGKDD INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING, p. 687– –696, 2017.

[5] AMEMIYA, T.. Advanced Econometrics. Harvard University Press, 1985.

[6] AUER, P.; CESA-BIANCHI, N. ; FISCHER, P.. Finite-time analysis of the multiarmed bandit problem. Machine Learning, 47:235–256, 2002.

[7] AUER, P.. Using confidence bounds for exploitation-exploration trade-offs. Journal of Machine Learning Research, 3:397–422, 2003.

[8] BASTANI, H.; BAYATI, M.. Online decision-making with highdimensional covariates. Operations Research, 68:276–294, 2020.

[9] BASTANI, H.; BAYATI, M. ; KHOSRAVI, K.. Mostly exploration-free algorithms for contextual bandits. working paper 1704.09011, arXiv, 2020.

[10] BÜHLMANN, P.; VAN DE GEER, S.. Statistics for High-Dimensional Data: Methods, Theory and Applications. Springer, 2011.

[11] BIAU, G.; DEVROYE, L. ; LUGOSI, G.. Consistency of random forests and other averaging classifiers. Journal of Machine Learning Research, 9:2015–2033, 2008.

[12] DEN BOER, A.. Dynamic pricing and learning: Historical origins, current research, and new directions. Surveys in Operations Research and Management Science, 20:1–18, 2015.

[13] BOUNEFFOUF, D.; RISH, I.; CECCHI, G. ; FÉRAUD, R.. Context attentive bandits: Contextual bandit with restricted context. In: PROCEEDINGS OF THE 26TH INTERNATIONAL JOINT CONFERENCE ON ARTIFICIAL INTELLIGENCE, p. 1468–1475, 2017.

[14] BREIMAN, L.; FRIEDMAN, J. H.; OLSHEN, R. A. ; STONE, C. J.. Classification and Regression Trees. Wadsworth and Brooks, 1984.

[15] BREIMAN, L.. Random forests. Machine Learning, 45(1):5–32, 2001.

[16] BURGIN, M.. Continuity in discrete sets, 2010.

[17] CALONICO, S.; CATTANEO, M. D. ; TITIUNIK, R.. Robust nonparametric confidence intervals for regression-discontinuity designs. Econometrica, 82(6):2295–2326, 2014.

[18] CALONICO, S.; CATTANEO, M.; FARRELL, M. ; TITIUNIK, R.. Regression discontinuity designs using covariates. The Review of Economics and Statistics, 101, 07 2018.

[19] CARD, D.; MAS, A. ; ROTHSTEIN, J.. Tipping and the Dynamics of Segregation*. The Quarterly Journal of Economics, 123(1):177–218, 2008.

[20] CARVALHO, C.; MASINI, R. ; MEDEIROS, M.. ArCo: An artificial counterfactual approach for high-dimensional panel time-series data. Journal of Econometrics, 207:352–380, 2018.

[21] CARPENTIER, A.; MUNOS, R.. Bandit theory meets compressed sensing for high dimensional stochastic linear bandit. In: PROCEEDINGS OF THE 15TH INTERNATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE AND STATISTICS, p. 190–198, 2012.

[22] CATTANEO, M. D.; IDROBO, N. ; TITIUNIK, R.. A Practical Introduction to Regression Discontinuity Designs: Foundations. Elements in Quantitative and Computational Methods for the Social Sciences. Cambridge University Press, 2020.

[23] CESA-BIANCHI, N.; GENTILE, C. ; MANSOUR, Y.. Regret minimization for reserve prices in second-price auctions. In: PROCEEDINGS OF THE 24TH ANNUAL ACM-SIAM SYMPOSIUM ON DISCRETE ALGORITHMS, p. 1190–1204, 2013.

[24] CHARPENTIER, A.; ELIE, R. ; REMLINGER, C.. Reinforcement learning in economics and finance. Computational Economics, p. 1–38, 2021.

[25] CHAY, K. Y.; MCEWAN, P. J. ; URQUIOLA, M.. The central role of noise in evaluating interventions that use test scores to rank schools. American Economic Review, 95(4):1237–1258, 2005.

[26] DESHPANDE, Y.; MONTANARI, A.. Linear bandits in high dimension and recommendation systems. In: PROCEEDINGS OF THE 50TH ANNUAL ALLERTON CONFERENCE ON COMMUNICATION, CONTROL, AND COMPUTING, p. 1750–1754, 2012.

[27] FAN, J.; GIJBELS, I.. Variable bandwidth and local linear regression smoothers. The Annals of Statistics, 20:2008–2036, 1992.

[28] FAN, J.. Design-adaptive nonparametric regression. Journal of the American Statistical Association, 87:998–1004, 1992.

[29] FAN, J.. Local linear regression smoothers and their minimax efficiencies. The Annals of Statistics, 21:196–216, 1993.

[30] FAN, J.; GIJBELS, I.. Local polynomial Modelling and its Applications. Número 66 em Monographs on statistics and applied probability series. Chapman & Hall, 1996.

[31] GOLDENSHLUGER, A.; ZEEVI, A.. A linear response bandit problem. Stochastic Systems, 3:230–261, 2013.

[32] HAHN, J.; TODD, P. ; VAN DER KLAAUW, W.. Evaluating the effect of an antidiscrimination law using a regression-discontinuity design. NBER Working Papers 7131, National Bureau of Economic Research, Inc, 1999.

[33] HAHN, J.; TODD, P. ; VAN DER KLAAUW, W.. Identification and estimation of treatment effects with a regression-discontinuity design. Econometrica, 69(1):201–209, 2001.

[34] IMBENS, G. W.; LEMIEUX, T.. Regression discontinuity designs: A guide to practice. Journal of Econometrics, 142(2):615–635, 2008.

[35] IMBENS, G.; ZAJONC, T.. Regression discontinuity design with vector-argument assignment rules. Mimeo, 05 2009.

[36] IMBENS, G. W.; KALYANARAMAN, K.. Optimal bandwidth choice for the regression discontinuity estimator. The Review of Economic Studies, 79(3):933–959, 2012.

[37] ISHWARAM, H.. The effect of splitting on random forests. Machine Learning, 99:75–118, 2015.

[38] JAMES, G.; WITTEN, D.; HASTIE, T. ; TIBSHIRANI, R.. An Introduction to Statistical Learning: with Applications in R. Springer, 2013.

[39] KIM, G.-S.; PAIK, M. C.. Doubly-robust lasso bandit. In: ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS, volumen 32. Curran Associates, Inc., 2019.

[40] KANDASAMY, K.; GONZALEZ, J.; JORDAN, M. ; STOIC, I.. Mechanism design with bandit feedback. working paper 2004.08924, arXiv, 2020.

[41] KOCK, A.; THYRSGAARD, M.. Optimal sequential treatment allocation. working paper 1705.09952, arXiv, 2017.

[42] KOCK, A.; PREINERSTORFER, D. ; VELIYEV, B.. Functional sequential treatment allocation. working paper 1812.09408, arXiv, 2018.

[43] KOCK, A.; PREINERSTORFER, D. ; VELIYEV, B.. Treatment recommendation with distributional targets. working paper 2005.09717, arXiv, 2020.

[44] KRISHNAMURTHY, S.; ATHEY, S.. Survey bandits with regret guarantees. working paper 2002.09814, arXiv, 2020.

[45] LANGFORD, J.; ZHANG, T.. The epoch-greedy algorithm for multiarmed bandits with side information. In: Platt, J.; Koller, D.; Singer, Y. ; Roweis, S., editors, ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS 20, p. 817–824. 2008.

[46] LEE, D. S.; CARD, D.. Regression discontinuity inference with specification error. Journal of Econometrics, 142(2):655 – 674, 2008.

[47] LEE, D. S.; LEMIEUX, T.. Regression discontinuity designs in economics. Journal of Economic Literature, 48(2):281–355, 2010.

[48] LEUVEN, E.; LINDAHL, M.; OOSTERBEEK, H. ; WEBBINK, D.. The effect of extra funding for disadvantaged pupils on achievement. The Review of Economics and Statistics, 89(4):721–736, 2007.

[49] LI, L.; CHU, W.; LANGFORD, J. ; SCHAPIRE, R.. A contextualbandit approach to personalized news article recommendation. In: PROCEEDINGS OF THE 19TH INTERNATIONAL CONFERENCE ON WORLD WIDE WEB, p. 661—-670, 2010.

[50] LI, W.; BARIK, A. ; HONORIO, J.. A simple unified framework for high dimensional bandit problems, 2021.

[51] LIN, Z.; BAI, Z.. Probability Inequalities. Springer-Verlag, 2011.

[52] LUDWIG, J.; MILLER, D. L.. Does Head Start Improve Children's Life Chances? Evidence from a Regression Discontinuity Design*. The Quarterly Journal of Economics, 122(1):159–208, 2007.

[53] PAPAY, J. P.; WILLETT, J. B. ; MURNANE, R. J.. Extending the regression-discontinuity approach to multiple assignment variables. Journal of Econometrics, 161(2):203–207, 2011.

[54] PORTER, J.. Estimation in the regression discontinuity model. Unpublished Manuscript, Department of Economics, University of Wisconsin at Madison, 2003.

[55] PORTER, J.; YU, P.. Regression discontinuity designs with unknown discontinuity points: Testing and estimation. Journal of Econometrics, 189(1):132 – 147, 2015.

[56] QIU, P.; ASANO, C. ; LI, X.. Estimation of jump regression function. Bulletin of Informatics and Cybernetics, 24(197):197 – 212, 1991.

[57] RUBIN, D. B.. Statistics and causal inference: Comment: Which ifs have causal answers. Journal of the American Statistical Association, 81(396):961–962, 1986.

[58] RUPPERT, D.; WAND, M. P.. Multivariate locally weighted least squares regression. The Annals of Statistics, 22:1346–1370, 1994.

[59] RUSSO, D.; VAN ROY, B.. An information-theoretic analysis of Thompson sampling. Journal of Machine Learning Research, 17:2442—- 2471, 2016.

[60] SAURÉ, D.; ZEEVI, A.. Optimal dynamic assortment planning with demand learning. Manufacturing & Service Operations Management, 15:387–404, 2013.

[61] SCORNET, E.; BIAU, G. ; VERT, J.. Consistency of random rorests. Annals of Statistics, 43(4):1716–1741, 2015.

[62] SLIVKINS, A.. Introduction to multi-armed bandits. arXiv preprint arXiv:1904.07272, 2019.

[63] SUN, Y.. Adaptive estimation of the regression discontinuity model. Available at SSRN 739151, 2005.

[64] THISTLETHWAITE, D. L.; CAMPBELL, D. T.. Regressiondiscontinuity analysis: An alternative to the ex-post facto experiment. Journal of Educational Psychology, 51:309–317, 1960.

[65] TRAN-THANH, L.; CHAPMAN, A.; LUNA, J. M. D. C. F. ; JENNINGS, A. R. N.. Epsilon–first policies for budget–limited multi-armed bandits. In: PROCEEDINGS OF THE 24TH AAAI CONFERENCE ON ARTIFICIAL INTELLIGENCE, p. 1211–1216, 2010.

[66] TSYBAKOV, A.. Optimal aggregation of classifiers in statistical learning. Annals of Statistics, 32:135–166, 2004.

[67] VAN DER KLAAUW, W.. Estimating the effect of financial aid offers on college enrollment: A regression–discontinuity approach. International Economic Review, 43(4):1249–1287, 2002.

[68] WAGER, S.; ATHEY, S.. Estimation and inference of heterogeneous treatment effects using random forests. Journal of the American Statistical Association, 113(523):1228–1242, 2018.

[69] WANG, X.; WEI, M. ; YAO, T.. Minimax concave penalized multiarmed bandit model with high-dimensional covariates. In: PROCEEDINGS OF THE 35TH INTERNATIONAL CONFERENCE ON MACHINE LEARNING, volumen 80 de Proceedings of Machine Learning Research, p. 5200–5208, 2018.

[70] WOOLDRIDGE, J. M.. Econometric Analysis of Cross Section and Panel Data, volumen 1 de MIT Press Books. The MIT Press, 2001.

[71] WU, Y.; SHARIFF, R.; LATTIMORE, T. ; SZEPESVÁRI, C.. Conservative bandits, 2016.

A
Appendix to Chapter 1

In this appendix, we provide the proofs of the Theorems proposed in Chapter 1, and respective Auxiliary Lemmas.

A.1
Auxiliary Lemmas

Lemmas 3 and 4 establish the properties for the Lasso estimation.

Lemma 3 (Finite-Sample Properties of $\widehat{\boldsymbol{\beta}}_k$) *For any $\vartheta \in \mathcal{T}$, define:*

$$\mathcal{G}_{k\vartheta} := \left\{ \frac{2}{n_{k\vartheta}} \max_{1\leq j\leq p} \left| \boldsymbol{\epsilon}'_{k\vartheta} \boldsymbol{X}^{(j)}_{k\vartheta} \right| \leq a \right\}$$

Provided that $\lambda_\vartheta \geq 2a$ and that $\frac{32bs_0}{\phi_0^2} \leq 1$, where $b \geq \max_{i,j} |\widehat{\boldsymbol{\Sigma}}_{k\vartheta(i,j)} - \boldsymbol{\Sigma}_{k\vartheta(i,j)}|$ and s_0, ϕ_0 are established in Definition 2 and Assumption 3 respectively, if $\widehat{\boldsymbol{\beta}}_k$ is the solution of (1-2), on $\mathcal{G}_{k\vartheta}$, it is true that:

$$\left\| \widehat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k \right\|_1 \leq \frac{\left\| \widehat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k \right\|^2_{\widehat{\boldsymbol{\Sigma}}_{k\vartheta}}}{\lambda_\vartheta} + \frac{4\lambda_\vartheta s_0}{\phi_0^2},$$

where $\left\| \widehat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k \right\|^2_{\widehat{\boldsymbol{\Sigma}}_{k\vartheta}} \equiv (\widehat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k)' \widehat{\boldsymbol{\Sigma}}_{k\vartheta} (\widehat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k)$ and $\widehat{\boldsymbol{\Sigma}}_{k\vartheta} \equiv \frac{1}{n_{k\vartheta}} \boldsymbol{X}'_{k\vartheta} \boldsymbol{X}_{k\vartheta}$.

Prova. This proof has been already provided in other papers, such as Carvalho et al. (2018). For the sake of completeness, we provide its main steps, even though it is a well-known result.

In equation (1-2), if $\widehat{\boldsymbol{\beta}}_k$ is the minimum of the optimization problem, then, for $\vartheta \in \mathcal{T}$, it is true that

$$\frac{1}{n_{k\vartheta}} \left\| \boldsymbol{y}_{k\vartheta} - \boldsymbol{X}_{k\vartheta} \widehat{\boldsymbol{\beta}}_k \right\|_2^2 + \lambda_\vartheta \left\| \widehat{\boldsymbol{\beta}}_k \right\|_1 \leq \frac{1}{n_{k\vartheta}} \left\| \boldsymbol{y}_{k\vartheta} - \boldsymbol{X}_{k\vartheta} \boldsymbol{\beta}_k \right\|_2^2 + \lambda_\vartheta \left\| \boldsymbol{\beta}_k \right\|_1 .$$

Using Assumption 1, we can replace $\boldsymbol{y}_{k\vartheta}$ in the above expression to obtain the basic inequality (see Buhlmann and van de Geer, 2011, page 103):

$$\begin{aligned} &\frac{1}{n_{k\vartheta}} \left\| \boldsymbol{X}_{k\vartheta} (\boldsymbol{\beta}_k - \widehat{\boldsymbol{\beta}}_k) + \boldsymbol{\epsilon}_{k\vartheta} \right\|_2^2 + \lambda_\vartheta \left\| \widehat{\boldsymbol{\beta}}_k \right\|_1 \leq \frac{1}{n_{k\vartheta}} \left\| \boldsymbol{\epsilon}_{k\vartheta} \right\|_2^2 + \lambda_\vartheta \left\| \boldsymbol{\beta}_k \right\|_1 \iff \\ &\frac{1}{n_{k\vartheta}} \left\| \boldsymbol{X}_{k\vartheta} (\widehat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k) \right\|_2^2 + \lambda_\vartheta \left\| \widehat{\boldsymbol{\beta}}_k \right\|_1 \leq \frac{2}{n_{k\vartheta}} \boldsymbol{\epsilon}'_{k\vartheta} \boldsymbol{X}_{k\vartheta} (\widehat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k) + \lambda_\vartheta \left\| \boldsymbol{\beta}_k \right\|_1 \end{aligned} \tag{A-1}$$

Define $\left\|\widehat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k\right\|^2_{\widehat{\boldsymbol{\Sigma}}_{k\vartheta}} \equiv (\widehat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k)'\widehat{\boldsymbol{\Sigma}}_{k\vartheta}(\widehat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k)$, and the same for $\left\|\widehat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k\right\|^2_{\boldsymbol{\Sigma}_{k\vartheta}}$ replacing $\widehat{\boldsymbol{\Sigma}}_{k\vartheta}$ for $\boldsymbol{\Sigma}_{k\vartheta}$, where $\boldsymbol{\Sigma}_{k\vartheta} := \mathbb{E}[\boldsymbol{X}'_{k\vartheta}\boldsymbol{X}_{k\vartheta}]$ and $\widehat{\boldsymbol{\Sigma}}_{k\vartheta} := \frac{1}{n_{k\vartheta}}\boldsymbol{X}'_{k\vartheta}\boldsymbol{X}_{k\vartheta}$.

The first term on the right side of (A-1) can be bounded in absolute terms as:

$$\frac{2}{n_{k\vartheta}}|\boldsymbol{\epsilon}'_{k\vartheta}\boldsymbol{X}_{k\vartheta}(\widehat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k)| \leq \left(\frac{2}{n_{k\vartheta}}\max_{1\leq j\leq p}|\boldsymbol{\epsilon}'_{k\vartheta}\boldsymbol{X}^{(j)}_{k\vartheta}|\right)\left\|\widehat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k\right\|_1.$$

On $\mathcal{G}_{k\vartheta}$, we have that

$$\left\|\widehat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k\right\|^2_{\widehat{\boldsymbol{\Sigma}}_{k\vartheta}} + \lambda_\vartheta\left\|\widehat{\boldsymbol{\beta}}_k\right\|_1 \leq a\left\|\widehat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k\right\|_1 + \lambda_\vartheta\left\|\boldsymbol{\beta}_k\right\|_1 \tag{A-2}$$

Using our previous definitions (see Section 1.2) for $\boldsymbol{\beta}_k[S_0]$ and $\boldsymbol{\beta}_k[S_0^c]$ and the respective counterparts for the estimators, by the triangle inequality of the left-hand side of equation (A-2), we have that:

$$\left\|\widehat{\boldsymbol{\beta}}_k\right\|_1 = \left\|\widehat{\boldsymbol{\beta}}_k[S_0]\right\|_1 + \left\|\widehat{\boldsymbol{\beta}}_k[S_0^c]\right\|_1 \geq \left\|\boldsymbol{\beta}_k[S_0]\right\|_1 - \left\|(\widehat{\boldsymbol{\beta}}_k[S_0] - \boldsymbol{\beta}_k[S_0])\right\|_1 + \left\|\widehat{\boldsymbol{\beta}}_k[S_0^c]\right\|_1$$

Using this result in (A-2) and the fact that $\left\|\widehat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k\right\|_1 = \left\|\widehat{\boldsymbol{\beta}}_k[S_0] - \boldsymbol{\beta}_k[S_0]\right\|_1 + \left\|\widehat{\boldsymbol{\beta}}_k[S_0^c]\right\|_1$:

$$\begin{aligned}
&\left\|\widehat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k\right\|^2_{\widehat{\boldsymbol{\Sigma}}_{k\vartheta}} + \lambda_\vartheta\Big(\left\|\boldsymbol{\beta}_k[S_0]\right\|_1 - \left\|(\widehat{\boldsymbol{\beta}}_k[S_0] - \boldsymbol{\beta}_k[S_0])\right\|_1 + \left\|\widehat{\boldsymbol{\beta}}_k[S_0^c]\right\|_1\Big) \leq \\
&a\Big(\left\|\widehat{\boldsymbol{\beta}}_k[S_0] - \boldsymbol{\beta}_k[S_0]\right\|_1 + \left\|\widehat{\boldsymbol{\beta}}_k[S_0^c]\right\|_1\Big) + \lambda_\vartheta\left\|\boldsymbol{\beta}_k\right\|_1 \iff \\
&\left\|\widehat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k\right\|^2_{\widehat{\boldsymbol{\Sigma}}_{k\vartheta}} + (\lambda_\vartheta - a)\left\|\widehat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k\right\|_1 \leq 2\lambda_\vartheta\left\|\widehat{\boldsymbol{\beta}}_k[S_0] - \boldsymbol{\beta}_k[S_0]\right\|_1.
\end{aligned}$$

By Assumption 3, we have that:

$$\left\|\widehat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k\right\|^2_{\widehat{\boldsymbol{\Sigma}}_{k\vartheta}} + (\lambda_\vartheta - a)\left\|\widehat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k\right\|_1 \leq \frac{2\lambda_\vartheta\sqrt{s_0}}{\phi_0}\left\|\widehat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k\right\|_{\boldsymbol{\Sigma}_{k\vartheta}} \tag{A-3}$$

Recall that Assumption 3 also requires that $\max_{i,j}|\widehat{\boldsymbol{\Sigma}}_{k\vartheta(i,j)} - \boldsymbol{\Sigma}_{k\vartheta(i,j)}| \leq b$. Then, using Lemma 8, provided that $\frac{32bs_0}{\phi_0^2} \leq 1$, we have that $\left\|\widehat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k\right\|_{\boldsymbol{\Sigma}_{k\vartheta}} \leq \sqrt{2}\left\|\widehat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k\right\|_{\widehat{\boldsymbol{\Sigma}}_{k\vartheta}}$. Substituting in (A-3):

$$\left\|\widehat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k\right\|^2_{\widehat{\boldsymbol{\Sigma}}_{k\vartheta}} + (\lambda_\vartheta - a)\left\|\widehat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k\right\|_1 \leq \frac{2\sqrt{2}\lambda_\vartheta\sqrt{s_0}}{\phi_0}\left\|\widehat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k\right\|_{\widehat{\boldsymbol{\Sigma}}_{k\vartheta}}$$

Since for any $\vartheta \in \mathcal{T}$, $\lambda_\vartheta \geq 2a$, $a > 0$, multiplying the last expression by 2 and using the fact that $4vu \leq u^2 + 4v^2$, we have:

$$\left\|\widehat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k\right\|_1 \leq \frac{\left\|\widehat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k\right\|^2_{\widehat{\boldsymbol{\Sigma}}_{k\vartheta}}}{\lambda_\vartheta} + \frac{4\lambda_\vartheta s_0}{\phi_0^2} \tag{A-4}$$

■

Lemma 4 (Finite-Sample Properties of $\widehat{\boldsymbol{\beta}}_k$ - Continuation) *Given that Assumptions 1 and 3 and the conditions of Lemma 3 are satisfied, then, for any $\vartheta \in \mathcal{T}$:*

$$\mathbb{P}\left(\left\|\widehat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k\right\|_1 > \frac{4s_0\lambda_\vartheta}{\phi_0^2}\right) \leq \frac{\log(2p)}{n_{k\vartheta}}\left\{\frac{C_1}{\lambda_\vartheta^2} + C_2 + C_3\left[\frac{\log(2p)}{n_{k\vartheta}}\right]^{-1/2}\right\} =: P_{\boldsymbol{\beta}_\vartheta}, \text{ where}$$

$$C_1 := C_1(\sigma, \theta_x) = 128\sigma^2\theta_x^2,\ C_2 := C_2(b, \theta_x) = \frac{\theta_x^2}{b}, \text{ and } C_3 := C_3(b, \theta_x) = \sqrt{2}C_2.$$

Prova. Provided that $\lambda_\vartheta \geq 2a$, on $\mathcal{G}_{k\vartheta}$, that $\frac{32bs_0}{\phi_0^2} \leq 1$, where $b \geq \max_{i,j}|\widehat{\boldsymbol{\Sigma}}_{k\vartheta(i,j)} - \boldsymbol{\Sigma}_{k\vartheta(i,j)}|$, Lemma 3 indicates that $\left\|\widehat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k\right\|_1 \leq \frac{4s_0\lambda_\vartheta}{\phi_0^2}$. Then,

$$\begin{aligned}
\mathbb{P}\left(\left\|\widehat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k\right\|_1 > \frac{4s_0\lambda_\vartheta}{\phi_0^2}\right) &= \mathbb{P}\left[\left(\mathcal{G}_{k\vartheta} \cap \max_{i,j}|\widehat{\boldsymbol{\Sigma}}_{k\vartheta(i,j)} - \boldsymbol{\Sigma}_{k\vartheta(i,j)}| \leq b\right)^c\right] \\
&= \mathbb{P}(\mathcal{G}_{k\vartheta}^c \cup \max_{i,j}|\widehat{\boldsymbol{\Sigma}}_{k\vartheta(i,j)} - \boldsymbol{\Sigma}_{k\vartheta(i,j)}| > b) \\
&\leq \mathbb{P}(\mathcal{G}_{k\vartheta}^c) + \mathbb{P}(\max_{i,j}|\widehat{\boldsymbol{\Sigma}}_{k\vartheta(i,j)} - \boldsymbol{\Sigma}_{k\vartheta(i,j)}| > b) \qquad \text{(A-5)} \\
&= \mathbb{P}\left(\frac{2}{n_{k\vartheta}}\max_{1\leq j\leq p}|\boldsymbol{\epsilon}_{k\vartheta}'\boldsymbol{X}_{k\vartheta}^{(j)}| > \frac{\lambda_\vartheta}{2}\right) \\
&\quad + \mathbb{P}(\max_{i,j}|\widehat{\boldsymbol{\Sigma}}_{k\vartheta(i,j)} - \boldsymbol{\Sigma}_{k\vartheta(i,j)}| > b)
\end{aligned}$$

where the second equality is De Morgan's law and the first inequality is an application of the union bound.

For the first term of (A-5), given that $\max_{1\leq j\leq p}|\boldsymbol{\epsilon}_{k\vartheta}'\boldsymbol{X}_{k\vartheta}^{(j)}|$, $j = 1, \ldots, p$ is a positive random variable, for $l > 0$, we employ the Markov inequality to obtain:

$$\begin{aligned}
\mathbb{P}\left(\frac{2}{n_{k\vartheta}}\max_{1\leq j\leq p}\left|\boldsymbol{\epsilon}_{k\vartheta}'\boldsymbol{X}_{k\vartheta}^{(j)}\right| > \frac{\lambda_\vartheta}{2}\right) &\leq 4^l\frac{\mathbb{E}\left(\max_{1\leq j\leq p}|\boldsymbol{\epsilon}_{k\vartheta}'\boldsymbol{X}_{k\vartheta}^{(j)}|^l\right)}{(n_{k\vartheta}\lambda_\vartheta)^l} \\
&= 4^l\frac{\mathbb{E}\left(\max_{1\leq j\leq p}\left|\sum_{i=1}^{n_{k\vartheta}}\boldsymbol{\epsilon}_{k\vartheta(i)}\boldsymbol{X}_{k\vartheta(i,j)}/n_{k\vartheta}\right|^l\right)}{\lambda_\vartheta^l}.
\end{aligned} \qquad \text{(A-6)}$$

Since (A-6) holds for any value of $l > 0$, take $l = 2$. Therefore, by Lemma 9:

$$16\frac{\mathbb{E}\left(\max_{1\leq j\leq p}\left|\sum_{i=1}^{n_{k\vartheta}}\boldsymbol{\epsilon}_{k\vartheta(i)}\boldsymbol{X}_{k\vartheta(i,j)}/n_{k\vartheta}\right|^2\right)}{\lambda_\vartheta^2} \leq \frac{128}{n_{k\vartheta}^2\lambda_\vartheta^2}\sigma^2\log(2p)\sum_{i=1}^{n_{k\vartheta}}\left(\max_{1\leq j\leq p}|\boldsymbol{X}_{k\vartheta(i,j)}|\right)^2 \leq \frac{128}{n_{k\vartheta}\lambda_\vartheta^2}\sigma^2\log(2p)\theta_x^2 \tag{A-7}$$

For the second term in (A-5), we also have that $\max_{i,j}|\widehat{\boldsymbol{\Sigma}}_{k\vartheta(i,j)} - \boldsymbol{\Sigma}_{k\vartheta(i,j)}|$ is a positive random variable. Then, by the Markov inequality, for $l = 1$:

$$\mathbb{P}\left(\max_{i,j}\left|\widehat{\boldsymbol{\Sigma}}_{k\vartheta(i,j)} - \boldsymbol{\Sigma}_{k\vartheta(i,j)}\right| > b\right) \leq \frac{1}{b}\mathbb{E}\left(\max_{i,j}\left|\widehat{\boldsymbol{\Sigma}}_{k\vartheta(i,j)} - \boldsymbol{\Sigma}_{k\vartheta(i,j)}\right|\right) \tag{A-8}$$

Recall that $\widehat{\boldsymbol{\Sigma}}_{k\vartheta} := \frac{1}{n_{k\vartheta}}\boldsymbol{X}'_{k\vartheta}\boldsymbol{X}_{k\vartheta}$ and, therefore, its elements are given by $\widehat{\boldsymbol{\Sigma}}_{k\vartheta(i,j)} = \frac{1}{n_{k\vartheta}}\sum_{m=1}^{n_{k\vartheta}}\boldsymbol{X}_{k\vartheta(m,i)}\boldsymbol{X}_{k\vartheta(m,j)}$

Define the function $\gamma(\cdot)$, such that for bounded random variables $\boldsymbol{X}_{k\vartheta(m,i)}, \boldsymbol{X}_{k\vartheta(m,j)}$ taking values in a subset of $\mathbb{R}$, $m \in \{1, ..., n_{k\vartheta}\}$, $i, j \in \{1, ..., p\}$:

$$\gamma(\boldsymbol{X}_{k\vartheta(m,i)}, \boldsymbol{X}_{k\vartheta(m,j)}) = \frac{\boldsymbol{X}_{k\vartheta(m,i)}\boldsymbol{X}_{k\vartheta(m,j)} - \mathbb{E}(\boldsymbol{X}_{k\vartheta(m,i)}\boldsymbol{X}_{k\vartheta(m,j)})}{\theta_x^2}$$

where θ_x is defined in Assumption 1.i.

Then, equation (A-8) can be rewritten as:

$$\frac{1}{b}\mathbb{E}\left(\max_{i,j}\left|\widehat{\boldsymbol{\Sigma}}_{k\vartheta(i,j)} - \boldsymbol{\Sigma}_{k\vartheta(i,j)}\right|\right) = \frac{1}{b}\mathbb{E}\left[\max_{i,j}\left|\frac{1}{n_{k\vartheta}}\sum_{m=1}^{n_{k\vartheta}}\theta_x^2\gamma(\boldsymbol{X}_{k\vartheta(m,i)}, \boldsymbol{X}_{k\vartheta(m,j)})\right|\right]. \tag{A-9}$$

Now, notice that $\mathbb{E}\left[\gamma\left(\boldsymbol{X}_{k\vartheta(m,i)}, \boldsymbol{X}_{k\vartheta(m,j)}\right)\right] = 0$ and that for $\mu = 2, 3, 4, \ldots$, such that $\mu \leq 1 + \log(p)$:

$$\frac{1}{n_{k\vartheta}}\sum_{m=1}^{n_{k\vartheta}}\mathbb{E}\left[\left|\gamma\left(\boldsymbol{X}_{k\vartheta(m,i)}, \boldsymbol{X}_{k\vartheta(m,j)}\right)\right|^\mu\right] =$$
$$\frac{1}{n_{k\vartheta}\theta_x^{2\mu}}\sum_{m=1}^{n_{k\vartheta}}\mathbb{E}\left[\left|\boldsymbol{X}_{k\vartheta(m,i)}\boldsymbol{X}_{k\vartheta(m,j)} - \mathbb{E}(\boldsymbol{X}_{k\vartheta(m,i)}\boldsymbol{X}_{k\vartheta(m,j)})\right|^\mu\right] \leq \frac{\theta_x^{2\mu}}{\theta_x^{2\mu}} = 1.$$

Then, the conditions of Lemma 10 are satisfied, and we can apply it to (A-9) to find that:

$$\frac{1}{b}\mathbb{E}\left(\max_{i,j}\left|\frac{1}{n_{k\vartheta}}\sum_{m=1}^{n_{k\vartheta}}\theta_x^2\gamma(\boldsymbol{X}_{k\vartheta(m,i)}, \boldsymbol{X}_{k\vartheta(m,j)})\right|\right) \leq \frac{\theta_x^2}{b}\left[\frac{\log(2p)}{n_{k\vartheta}} + \sqrt{\frac{2\log(2p)}{n_{k\vartheta}}}\right]. \tag{A-10}$$

Merging (A-7) and (A-10), we have:

$$\mathbb{P}\left(\left\|\widehat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k\right\|_1 > \frac{4s_0\lambda_\vartheta}{\phi_0^2}\right) \leq \frac{128}{n_{k\vartheta}\lambda_\vartheta^2}\sigma^2\log(2p)\theta_x^2 + \frac{\theta_x^2}{b}\left[\frac{\log(2p)}{n_{k\vartheta}} + \sqrt{\frac{2\log(2p)}{n_{k\vartheta}}}\right]$$
$$= \frac{\log(2p)}{n_{k\vartheta}}\left\{\frac{C_1}{\lambda_\vartheta^2} + C_2 + C_3\left[\frac{\log(2p)}{n_{k\vartheta}}\right]^{-1/2}\right\} =: P_{\boldsymbol{\beta}_\vartheta},$$

where $C_1 = 128\sigma^2\theta_x^2$, $C_2 = \frac{\theta_x^2}{b}$ and $C_3 = \sqrt{2}C_2$. ■

Lemma 5 presents the cumulative regret for the initialization phase ($0 \leq t \leq vw$), which is common to both HD ϵ_t-Greedy and CHD ϵ_t-Greedy algorithms. On the other hand, Lemmas 6 and 7 exhibit results for their instantaneous regret for $t > vw$.

Lemma 5 (Initialization Regret) *Given the duration vw for the initialization phase and provided that Assumptions 1 and 2 are satisfied, the cumulative regret for both HD ϵ_t-Greedy and CHD ϵ_t-Greedy algorithms in the initialization phase (R^I) is bounded as:*

$$R^I \leq vw\theta_x h\tau_{\mathcal{W}}.$$

where $\tau_{\mathcal{W}} \equiv \underset{k_1,k_2\in\{0,\ldots,w-1\}}{max} \|\boldsymbol{\omega}_{k1} - \boldsymbol{\omega}_{k2}\|_1$.

Prova. Let the sequence $\{\psi_t\}_{0\leq t\leq vw}$ comprises the indexes for the actions in $\mathcal{W}$ that lead to best rewards for each $t \leq vw$, that is, each $\psi_t \in \{0,\ldots,w-1\}$, such that $\psi_t \equiv \underset{j\in\{0,\ldots,w-1\}}{\arg\max}\, y_{jt}$. Considering Definition 1, the regret for the initialization phase is $R^I = \sum_{t=1}^{vw}\mathbb{E}(y_{\psi_t t} - y_{kt})$ for an arbitrary action $\boldsymbol{\omega}_k$ adopted at t. We can bound R^I by considering the worst case possible: to adopt wrong actions for all $t \leq vw$, in which case, $k \neq \psi_t$. By Assumption 1:

$$R^I = \sum_{t=1}^{vw}\mathbb{E}\left[\boldsymbol{x}_t'(\boldsymbol{\beta}_{\psi_t} - \boldsymbol{\beta}_k)\right]. \tag{A-11}$$

The right-hand side of equation (A-11) can be bounded in absolute terms as:

$$|\boldsymbol{x}_t'(\boldsymbol{\beta}_{\psi_t} - \boldsymbol{\beta}_k)| \leq \max_{1\leq j\leq p}|\boldsymbol{x}_{t(j)}|\left\|\boldsymbol{\beta}_{\psi_t} - \boldsymbol{\beta}_k\right\|_1.$$

For a finite set of actions $\mathcal{W}$, define $\tau_{\mathcal{W}} \equiv \underset{k_1,k_2\in\{0,\ldots,w-1\}}{max} \|\boldsymbol{\omega}_{k1} - \boldsymbol{\omega}_{k2}\|_1$. Then, by Assumptions 1.i and 2, we find that $R^I \leq \sum_{t=1}^{vw}\theta_x h\tau_{\mathcal{W}} \leq vw\theta_x h\tau_{\mathcal{W}}$. ■

Lemma 6 (Instantaneous Regret of the HD ϵ_t-Greedy Algorithm) *For every $\vartheta > vw$, provided that each $\lambda_\vartheta \geq 2a$ on $\mathcal{G}_{k\vartheta}$ defined in Lemma 3,*

that $\frac{32bs_0}{\phi_0^2} \leq 1$*, where* $b \geq \max_{i,j}|\widehat{\boldsymbol{\Sigma}}_{k\vartheta(i,j)} - \boldsymbol{\Sigma}_{k\vartheta(i,j)}|$ *and given that Assumptions 1 to 4 hold, with probability* $1 - P_{\boldsymbol{\beta}_\vartheta}$ *the instantaneous regret of the HD* ϵ_t*-Greedy algorithm (*r_ϑ^{HD}*) is bounded as:*

$$r_\vartheta^{HD} \leq w\theta_x h\tau P_{k\vartheta}^{HD}$$

where

$$P_{k\vartheta}^{HD} \leq \frac{v}{\vartheta} + \left(1 - \frac{vw}{\vartheta}\right)\frac{8C_m\theta_x s_0\lambda_\vartheta}{\phi_0^2}$$

$P_{\boldsymbol{\beta}_\vartheta}$ *and* C_m *are established in Lemma 4 and Assumption 4, respectively.*

Prova. For $\vartheta > vw$, define ψ_t in the same way as in the proof of Lemma 5 and consider the definition of the action function $I(\cdot)$ in Section 1.2. Then, by the law of total expectation, the instantaneous regret r_ϑ^{HD} of the HD ϵ_t-Greedy algorithm is:

$$r_\vartheta^{HD} = \sum_{k=0}^{w-1} \mathbb{E}\left[\boldsymbol{x}'_\vartheta(\boldsymbol{\beta}_{\psi_\vartheta} - \boldsymbol{\beta}_k)|I(\vartheta) = \boldsymbol{\omega}_k\right]\mathbb{P}\left[I(\vartheta) = \boldsymbol{\omega}_k\right]. \tag{A-12}$$

By the learning rule of the HD ϵ_t-Greedy algorithm (see Section 1.3), we have that:

$$\mathbb{P}\left[I(\vartheta) = \boldsymbol{\omega}_k\right] = \frac{\epsilon_\vartheta}{w} + (1-\epsilon_\vartheta)\mathbb{P}\left(\boldsymbol{x}'_\vartheta\widehat{\boldsymbol{\beta}}_k \geq \boldsymbol{x}'_\vartheta\widehat{\boldsymbol{\beta}}_j\right), \quad \forall j \in \{0,\ldots,w-1\}. \tag{A-13}$$

From the properties of the maximum of a sequence of random variables, we have the following fact applied to the last term of (A-13):

$$\begin{aligned}\mathbb{P}\left(\max_{j\in\{0,\ldots,w-1\}} \boldsymbol{x}'_\vartheta\widehat{\boldsymbol{\beta}}_j \leq \boldsymbol{x}'_\vartheta\widehat{\boldsymbol{\beta}}_k\right) &= \mathbb{P}\left(\bigcap_{j=0}^{w-1} \boldsymbol{x}'_\vartheta\widehat{\boldsymbol{\beta}}_j \leq \boldsymbol{x}'_\vartheta\widehat{\boldsymbol{\beta}}_k\right) \\ &\leq \mathbb{P}\left(\boldsymbol{x}'_\vartheta\widehat{\boldsymbol{\beta}}_j \leq \boldsymbol{x}'_\vartheta\widehat{\boldsymbol{\beta}}_k\right) \quad \text{for some} \quad j \in \{0,\ldots,w-1\},\end{aligned}$$

since for any sequence of sets A_i, $i = 1,\ldots,n$, the event $\{\bigcap_{i=1}^n A_i\}$ is a subset of every A_i.

Note that

$$\begin{aligned}\mathbb{P}\left(\boldsymbol{x}'_\vartheta\widehat{\boldsymbol{\beta}}_j \leq \boldsymbol{x}'_\vartheta\widehat{\boldsymbol{\beta}}_k\right) &= \mathbb{P}\left(\boldsymbol{x}'_\vartheta\widehat{\boldsymbol{\beta}}_j - \boldsymbol{x}'_\vartheta\boldsymbol{\beta}_j + \boldsymbol{x}'_\vartheta\boldsymbol{\beta}_j - \boldsymbol{x}'_\vartheta\widehat{\boldsymbol{\beta}}_k + \boldsymbol{x}'_\vartheta\boldsymbol{\beta}_k - \boldsymbol{x}'_\vartheta\boldsymbol{\beta}_k \leq 0\right) \\ &= \mathbb{P}\left[\boldsymbol{x}'_\vartheta(\boldsymbol{\beta}_j - \boldsymbol{\beta}_k) \leq \boldsymbol{x}'_\vartheta(\boldsymbol{\beta}_j - \widehat{\boldsymbol{\beta}}_j) + \boldsymbol{x}'_\vartheta(\widehat{\boldsymbol{\beta}_k} - \boldsymbol{\beta}_k)\right]\end{aligned} \tag{A-14}$$

Bounding the term $\boldsymbol{x}'_\vartheta(\widehat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k) - \boldsymbol{x}'_\vartheta(\widehat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j)$ in absolute value and

using the triangle inequality, we find that:

$$\begin{aligned}|\boldsymbol{x}'_\vartheta(\widehat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k - \widehat{\boldsymbol{\beta}}_j + \boldsymbol{\beta}_j)| &\leq \left(\max_{1\leq j\leq p}|\boldsymbol{x}_{\vartheta(j)}|\right)\left\|\widehat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k - \widehat{\boldsymbol{\beta}}_j + \boldsymbol{\beta}_j\right\|_1 \\ &\leq \left(\max_{1\leq j\leq p}|\boldsymbol{x}_{\vartheta(j)}|\right)\left(\left\|\widehat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k\right\|_1 + \left\|\boldsymbol{\beta}_j - \widehat{\boldsymbol{\beta}}_j\right\|_1\right) \\ &\leq \theta_x\left(\left\|\widehat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k\right\|_1 + \left\|\boldsymbol{\beta}_j - \widehat{\boldsymbol{\beta}}_j\right\|_1\right)\end{aligned}$$

Therefore,

$$\mathbb{P}\left(\boldsymbol{x}'_\vartheta\widehat{\boldsymbol{\beta}}_j \leq \boldsymbol{x}'_\vartheta\widehat{\boldsymbol{\beta}}_k\right) \leq \mathbb{P}\left[x'_\vartheta(\boldsymbol{\beta}_j - \boldsymbol{\beta}_k) \leq \theta_x\left(\left\|\widehat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k\right\|_1 + \left\|\boldsymbol{\beta}_j - \widehat{\boldsymbol{\beta}}_j\right\|_1\right)\right] \tag{A-15}$$

Now consider the set $\mathcal{G}_{k\vartheta}$ defined in Lemma 3. Provided that for every $\vartheta > vw$, $\lambda_\vartheta \geq 2a$ and that $\frac{32bs_0}{\phi_0^2} \leq 1$, where $b \geq \max_{i,j}|\widehat{\boldsymbol{\Sigma}}_{k\vartheta(i,j)} - \boldsymbol{\Sigma}_{k\vartheta(i,j)}|$, results of Lemmas 3 and 4 indicates that with probability $1 - P_{\boldsymbol{\beta}_\vartheta}$, for every $k \in \{0, \ldots, w-1\}$, $\left\|\widehat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k\right\|_1 \leq \frac{4s_0\lambda_\vartheta}{\phi_0^2}$. Using this fact in equation (A-15) and Assumption 4, we find that:

$$\mathbb{P}\left(\boldsymbol{x}'_\vartheta\widehat{\boldsymbol{\beta}}_j \leq \boldsymbol{x}'_\vartheta\widehat{\boldsymbol{\beta}}_k\right) \leq \mathbb{P}\left[x'_\vartheta(\boldsymbol{\beta}_j - \boldsymbol{\beta}_k) \leq \frac{8\theta_x s_0\lambda_\vartheta}{\phi_0^2}\right] \leq \frac{8C_m\theta_x s_0\lambda_\vartheta}{\phi_0^2} \tag{A-16}$$

Inserting the result obtained in equation (A-16) into equation (A-13), we find that:

$$\mathbb{P}\left[I(\vartheta) = \boldsymbol{\omega}_k\right] \leq \frac{\epsilon_\vartheta}{w} + (1-\epsilon_\vartheta)\frac{8C_m\theta_x s_0\lambda_\vartheta}{\phi_0^2} \tag{A-17}$$

As described in Section 1.3, the authors in Auer et al. (2002) suggest $\epsilon_\vartheta = \frac{cw}{d^2\vartheta}$, for $c > 0$, $0 < d < 1$ and $\vartheta \geq \frac{cw}{d^2}$. Since equation (A-17) is valid for $\vartheta > vw$ it suffices to take c, d, such that $c/d^2 = v$. In this case:

$$\mathbb{P}\left[I(\vartheta) = \boldsymbol{\omega}_k\right] \leq \frac{v}{\vartheta} + \left(1 - \frac{vw}{\vartheta}\right)\frac{8C_m\theta_x s_0\lambda_\vartheta}{\phi_0^2} =: P_{k\vartheta}^{HD}$$

Finally, recall the definition of $\tau_{\mathcal{W}}$ made in Lemma 5. Then, with probability $1 - P_{\boldsymbol{\beta}_\vartheta}$, the instantaneous regret of the HD ϵ_t-Greedy algorithm after the initialization phase can be bounded as:

$$r_\vartheta^{HD} \leq \theta_x h\tau_{\mathcal{W}}\sum_{k=0}^{w-1}\mathbb{P}\left[I(\vartheta) = \boldsymbol{\omega}_k\right] \leq w\theta_x h\tau_{\mathcal{W}}P_{k\vartheta}^{HD}.$$

■

Corollary 1 *Suppose that for each* $\vartheta > vw$, $\lambda_\vartheta \equiv C_\lambda\sigma\sqrt{\frac{2log(2p)}{\vartheta}}$, $C_\lambda \in \mathbb{R}$. *Then:*

$$r_\vartheta^{HD} \leq w\theta_x h\tau\left(\frac{v}{\vartheta} + \left(1 - \frac{vw}{\vartheta}\right)\frac{8C_m\theta_x s_0 C_\lambda\sigma\sqrt{2log(2p)}}{\phi_0^2\sqrt{\vartheta}}\right)$$

Prova.

This is a straightforward proof, since one plugs the proposed λ_ϑ into $P_{k\vartheta}^{HD}$ defined in Lemma 6. ■

The suggested time dependency for λ_ϑ in Corollary 1 is adapted from Buhlmann and van de Geer (2011) and very similar to the versions used in other papers in this literature such as: Wang et al. (2018), Bastani and Bayati (2020), Kim and Paik (2019) and Li et al. (2021).

Remark 5 *Recall that in some parts of the paper, as in Assumption 4, we assume the existence of λ_{min} and λ_{max}, such that for each $\vartheta \in \mathcal{T}$, $\lambda_{min} \leq \lambda_\vartheta \leq \lambda_{max}$. The path established for λ_ϑ in Corollary 1 restricts the range of its possible values and sheds light on its respective bounds. For example, from the results of Lemma 4, an increasing sequence $\{\lambda_\vartheta\}_{\vartheta\in\mathcal{T}}$ guarantees that $\mathcal{G}_{k\vartheta}$ occurs at higher probabilities as time passes. This is the most interesting case in our high-dimensional setup and, for this to occur, it is sufficient that, at each incremental time step $\vartheta_2 = \vartheta_1 + 1$, the growth in the problem dimension $p_{\vartheta_2}/p_{\vartheta_1}$ is less than $e^{log(2p_{\vartheta_1})/\vartheta_1}$. In this case, $\{\lambda_\vartheta\}_{\vartheta\in\mathcal{T}}$ is a increasing sequence and its lower bound would be $\lambda_{min} = C_\lambda\sigma\sqrt{2log(2p_0)}$, for p_0 the dimension at the beginning of the problem. The upper bound can be easily established considering that we study learning problems with finite horizon.*

Lemma 7 (Instantaneous Regret of the CHD ϵ_t-Greedy Algorithm) *For every $\vartheta > vw$, provided that each $\lambda_\vartheta \geq 2a$ on $\mathcal{G}_{k\vartheta}$ defined in Lemma 3, that $\frac{32bs_0}{\phi_0^2} \leq 1$, where $b \geq \max_{i,j}|\widehat{\mathbf{\Sigma}}_{k\vartheta(i,j)} - \mathbf{\Sigma}_{k\vartheta(i,j)}|$. Provided that $\mathcal{D}_\vartheta \leq w(1 - P_{\boldsymbol{\beta}_\vartheta})$, where*

$$\mathcal{D}_\vartheta := \frac{4\theta_x s_0 \lambda_\vartheta}{\phi_0^2} + \theta_x h \tau_{\mathcal{W}}.$$

and given that Assumptions 1 to 4 hold, with probability $1 - P_{\boldsymbol{\beta}_\vartheta}$ the instantaneous regret of the CHD ϵ_t-Greedy algorithm (r_ϑ^{CHD}) is bounded as:

$$r_\vartheta^{CHD} \leq w\theta_x h\tau_{\mathcal{W}}\left(P_{k\vartheta}^{CHD} - \frac{\epsilon_\vartheta s_\vartheta}{w} + P_{k\vartheta}^{HD}\right)$$

where

$$P_{k\vartheta}^{CHD} := \epsilon_\vartheta s_\vartheta \exp\left\{-\frac{2}{w}\left[\left(w\left(1 - P_{\boldsymbol{\beta}_\vartheta}\right) - \mathcal{X}_\vartheta\right)^2\right]\right\}. \tag{A-18}$$

$P_{\boldsymbol{\beta}_\vartheta}$ is the result of Lemma 4, $\tau_{\mathcal{W}}$ is defined in Lemma 5 and $P_{k\vartheta}^{HD}$ is provided in Lemma 6.

Prova. For any $\vartheta > vw$, define ψ_t in the same way as in the proof of Lemma 5 and consider the definition of the action function $I(\cdot)$ in Section 1.2. Then, by the law of total expectation, the instantaneous regret r_ϑ^{CHD} of the CHD ϵ_t-Greedy algorithm is:

$$r_{\vartheta}^{CHD} = \sum_{k=0}^{w-1} \mathbb{E}\left[\boldsymbol{x}'_{\vartheta}(\boldsymbol{\beta}_{\psi_{\vartheta}} - \boldsymbol{\beta}_k)|I(\vartheta) = \boldsymbol{\omega}_k\right] \mathbb{P}\left[I(\vartheta) = \boldsymbol{\omega}_k\right] \tag{A-19}$$

By the learning rule of the CHD ϵ_t-Greedy algorithm (see Section 1.3), we have that $\forall k \in \{0, \ldots, w-1\}$:

$$\mathbb{P}\left[I(\vartheta) = \boldsymbol{\omega}_k\right] = \frac{\epsilon_{\vartheta} s_{\vartheta}}{\kappa_{\vartheta}} \mathbb{P}(\boldsymbol{x}'_{\vartheta}\widehat{\boldsymbol{\beta}}_k \in \mathcal{H}_{\vartheta}^{(\kappa_{\vartheta})}) + \frac{1}{w}\left[\epsilon_{\vartheta}(1 - s_{\vartheta})\right] + (1-\epsilon_{\vartheta})\mathbb{P}(\boldsymbol{x}'_{\vartheta}\widehat{\boldsymbol{\beta}}_k \geq \boldsymbol{x}'_{\vartheta}\widehat{\boldsymbol{\beta}}_j). \tag{A-20}$$

The last term of the right side of equation (A-20) is the same as the last term of $\mathbb{P}\left[I(\vartheta) = \boldsymbol{\omega}_k\right]$ in the HD ϵ_t-Greedy algorithm. Regarding the first term of equation (A-20), by the definition of $\mathcal{H}_{\vartheta}^{(\kappa_{\vartheta})}$ (Section 1.3):

$$\mathbb{P}(\boldsymbol{x}'_{\vartheta}\widehat{\boldsymbol{\beta}}_k \in \mathcal{H}_{\vartheta}^{(\kappa_{\vartheta})}) = \mathbb{P}\left(\bigcup_{j=w-\kappa_{\vartheta}}^{w} \{\widehat{y}_{k\vartheta} \geq \widehat{y}_{(j:w)\vartheta}\}\right) \tag{A-21}$$

Now notice that restricted to the set of κ_{ϑ} higher-order statistics, the event $\{\widehat{y}_{k\vartheta} \geq \widehat{y}_{(j:w)\vartheta}\} \subset \{\widehat{y}_{k\vartheta} \geq \widehat{y}_{(w-\kappa_{\vartheta}:w)\vartheta}\}$, for $j \in \{w - \kappa_{\vartheta}, \ldots, w\}$. This implies that $\{\widehat{y}_{(w-\kappa_{\vartheta}:w)\vartheta}\}$ is is the most probable to occur since it is the lowest possible order statistic. Then, employing the union bound, we have that:

$$\mathbb{P}\left(\bigcup_{j=w-\kappa_{\vartheta}}^{w} \{\widehat{y}_{k\vartheta} \geq \widehat{y}_{(j:w)\vartheta}\}\right) \leq \kappa_{\vartheta}\mathbb{P}\left(\widehat{y}_{k\vartheta} \geq \widehat{y}_{(w-\kappa_{\vartheta}:w)\vartheta}\right). \tag{A-22}$$

Using Assumption 1, it is clear from the developments made in Lemmas 5 and 6 that $|\boldsymbol{x}'_{\vartheta}\widehat{\boldsymbol{\beta}}_k| \leq \theta_x \left\|\widehat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k\right\|_1 + \theta_x \left\|\boldsymbol{\beta}_k\right\|_1$. Moreover, using Assumption 2, provided that $\mathbf{0} \in \mathcal{C}$, the parametric space, Lemmas 3 and 4 indicates that, on $\mathcal{G}_{k\vartheta} \cap \max_{i,j}|\widehat{\boldsymbol{\Sigma}}_{k\vartheta(i,j)} - \boldsymbol{\Sigma}_{k\vartheta(i,j)}| \leq b$, with probability $1 - P_{\boldsymbol{\beta}_{\vartheta}}$:

$$|\boldsymbol{x}'_{\vartheta}\widehat{\boldsymbol{\beta}}_k| \leq \frac{4\theta_x s_0 \lambda_{\vartheta}}{\phi_0^2} + \theta_x h \tau_{\mathcal{W}} =: \mathcal{X}_{\vartheta}$$

where $\tau_{\mathcal{W}}$ is defined in Lemma 5.

Then, equation (A-22) leads to:

$$\begin{aligned}\kappa_{\vartheta}\mathbb{P}(\widehat{y}_{k\vartheta} \geq \widehat{y}_{(w-\kappa_{\vartheta}:w)\vartheta}) &\leq \kappa_{\vartheta}\mathbb{P}\Big(\widehat{y}_{(w-\kappa_{\vartheta}:w)\vartheta} \leq \mathcal{X}_{\vartheta}\Big) \\ &\leq \kappa_{\vartheta} \sum_{j=w-\kappa_{\vartheta}}^{w} \binom{w}{j} \left[\mathbb{P}(\widehat{y}_{k\vartheta} \leq \mathcal{X}_{\vartheta})\right]^j \left[1 - \mathbb{P}(\widehat{y}_{k\vartheta} \leq \mathcal{X}_{\vartheta})\right]^{w-j},\end{aligned} \tag{A-23}$$

since, as an intermediate-order statistic, $\widehat{y}_{(w-\kappa_{\vartheta}:w)\vartheta} \sim Bin[w, p_{k\vartheta}(y)]$, for $p_{k\vartheta}(y) \equiv \mathbb{P}(\widehat{y}_{k\vartheta} \leq y)$, which in this case, we can take $y = \mathcal{X}_{\vartheta}$.

If $\mathcal{X}_{\vartheta} \leq w p_{k\vartheta}(\mathcal{X}_{\vartheta})$, we can use Lemma 11 to bound equation (A-23) as:

$$\kappa_{\vartheta}\mathbb{P}\left(\widehat{y}_{(w-\kappa_{\vartheta}:w)\vartheta} \leq \mathcal{X}_{\vartheta}\right) \leq \kappa_{\vartheta} \exp\left[-2\frac{(w p_{k\vartheta}(\mathcal{X}_{\vartheta}) - \mathcal{X}_{\vartheta})^2}{w}\right].$$

However, notice that

$$p_{k\vartheta}(\mathcal{X}_\vartheta) := \mathbb{P}(\boldsymbol{x}'_\vartheta \widehat{\boldsymbol{\beta}}_k \leq \mathcal{X}_\vartheta) \geq \mathbb{P}\left(\left\|\widehat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k\right\|_1 \leq \frac{4s_0\lambda_\vartheta}{\phi_0^2} + h\tau_{\mathcal{W}}\right)$$
$$\geq \mathbb{P}\left(\left\|\widehat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k\right\|_1 \leq \frac{4s_0\lambda_\vartheta}{\phi_0^2}\right) = 1 - P_{\boldsymbol{\beta}_\vartheta},$$

Then,

$$\kappa_\vartheta \exp\left[-2\frac{(wp_{k\vartheta}(\mathcal{X}_\vartheta) - \mathcal{X}_\vartheta)^2}{w}\right] \leq \kappa_\vartheta \exp\left[-2\frac{\left(w(1-P_{\boldsymbol{\beta}_\vartheta}) - \mathcal{X}_\vartheta\right)^2}{w}\right] \quad \text{(A-24)}$$

Therefore, for $\vartheta > vw$, $\mathcal{X}_\vartheta \leq w(1-P_{\boldsymbol{\beta}_\vartheta})$ is sufficient to replace the above requisite of Lemma 11 and we restate it as:

$$\mathbb{P}(\boldsymbol{x}'_\vartheta \widehat{\boldsymbol{\beta}}_k \in \mathcal{H}_\vartheta^{(\kappa_\vartheta)}) \leq \kappa_\vartheta \exp\left\{-\frac{2}{w}\left[w\left(1-P_{\boldsymbol{\beta}_\vartheta}\right) - \mathcal{X}_\vartheta\right]^2\right\}$$

Define $P_{k\vartheta}^{CHD} := \frac{\epsilon_\vartheta s_\vartheta}{\kappa_\vartheta}\mathbb{P}(\boldsymbol{x}'_\vartheta \widehat{\boldsymbol{\beta}}_k \in \mathcal{H}_\vartheta^{(\kappa_\vartheta)})$ and, with probability $1 - P_{\boldsymbol{\beta}_\vartheta}$, the instantaneous regret of the CHD ϵ_t-Greedy algorithm, equation (A-19), can be bounded as:

$$r_\vartheta^{CHD} \leq w\theta_x h\tau_{\mathcal{W}}\left(P_{k\vartheta}^{CHD} - \frac{\epsilon_\vartheta s_\vartheta}{w} + P_{k\vartheta}^{HD}\right)$$

■

Note from Lemmas 5, 6 and 7 that all bounds are increasing with θ_x, $\tau_{\mathcal{W}}$ and w, this last one as a function of the initialization period. The intuition behind this fact is clear since the larger the level of dissimilarity among policies or the larger the number of policies to be tested is, the greater the difficulty for the algorithm to select the right policy. In particular, the initialization phase should also be longer, in order to gather information over a large set of alternatives.

Lemma 8 *Suppose that the $\boldsymbol{\Sigma}_0$-compatibility condition holds for the set S with cardinality s with compatibility constant $\phi_{\boldsymbol{\Sigma}_0}(S)$ and that $\|\boldsymbol{\Sigma}_1 - \boldsymbol{\Sigma}_0\|_\infty \leq \tilde{\lambda}$, where*

$$\frac{32\tilde{\lambda}s}{\phi_{\Sigma_0}^2(S)} \leq 1.$$

Then, for the set S, the $\boldsymbol{\Sigma}_1$-compatibility condition holds as well, with $\phi_{\boldsymbol{\Sigma}_1}^2(S) \geq \phi_{\boldsymbol{\Sigma}_0}^2(S)/2$.

Prova. See Corollary 6.8 in Buhlmann and van de Geer (2011) ■

Lemma 9 *For arbitrary n and p, consider independent centered random variables $\epsilon_1, \ldots, \epsilon_n$, such that $\forall\, i$, there is a σ^2 that bounds the variance as*

$\mathbb{E}(\epsilon_i^2) \leq \sigma^2$. *Moreover, let* $\{x_{i,j} : i = 1, \ldots, n,\ j = 1, \ldots, p\}$ *be such that for* $i = 1, \ldots, n$, *there is a* $K_i := \max_{1 \leq j \leq p} |x_{i,j}|$ *such that*

$$\mathbb{E}\left(\max_{1 \leq j \leq p} \left| \sum_{i=1}^{n} \frac{\epsilon_i x_{i,j}}{n} \right|^2 \right) \leq \sigma^2 \left[\frac{8 \log(2p)}{n} \right] \left(\frac{\sum_{i=1}^{n} K_i^2}{n} \right)$$

Prova. See Lemma 14.24 in Buhlmann and van de Geer (2011) ■

Lemma 10 *Let* $Z_1, \ldots, Z_n$ *be independent random variables and* $\gamma_1, \ldots, \gamma_p$ *be real-valued functions satisfied for* $j = 1, \ldots, p$,

$$\mathbb{E}[\gamma_j(Z_i)] = 0$$
$$\frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[|\gamma_j(Z_i)|^m] \leq \frac{m!}{2} K^{m-2}$$

for $K > 0$ *and* $m \leq 1 + log(p)$ *(easily satisfied for large p). Then,*

$$\mathbb{E}\left[\max_{1 \leq j \leq p} \left| \frac{1}{n} \sum_{i=1}^{n} \gamma_j(Z_i) \right|^m \right] \leq \left[\frac{K \log(2p)}{n} + \sqrt{\frac{2 \log(2p)}{n}} \right]^m .$$

Prova. See Lemma 14.12 in Buhlmann and van de Geer (2011) ■

Lemma 11 *Let* $X \sim Bin(n, p)$. *For* $k \leq np$:

$$\mathbb{P}(X \leq k) \leq \exp\left[-\frac{2(np - k)^2}{n} \right]$$

Prova. This is an application of Höeffding's inequality to random variables that follow a binomial distribution. For more details, see Lemma 7.3 of Lin and Bai (2011) ■

Lemma 12 *If* f *is a monotone decreasing function and* g *is a monotone increasing function, both integrable on the range* $[r - 1, s]$, *then:*

$$\sum_{t=r}^{s} f(t) \leq \int_{r-1}^{s} f(t)dt \quad \text{and} \quad \sum_{t=r}^{s} g(t) \leq \int_{r}^{s} g(t)dt$$

Prova. This is a well-known fact for monotone functions linked to left and right Riemann sums. ■

A.2 Theorems

Theorem 1 (Cumulative Regret of CHD ϵ_t-Greedy algorithms) *Provided that the conditions required by Lemmas 5, 6, 7 in this Appendix are satisfied, at least with probability* $1 - P_{\beta_{max}}$, $P_{\beta_{max}} \equiv \underset{vw<\vartheta<T}{max} P_{\beta_\vartheta}$, *for* $s_\vartheta \equiv s$ *imputed by end-user, the cumulative regret until time* T *of the CHD* ϵ_t*-Greedy learning rule can be bounded as:*

$$R_{T-1}^{CHD} \leq R_{T-1}^{HD} + w\theta_x h\tau_{\mathcal{W}} \left[vs \log\left(\frac{T-1}{vw}\right) \left(w \exp\left\{ -\frac{2}{w} \left[\left[w\left(1 - P_{\beta_\vartheta}\right) - \mathcal{X}_\vartheta \right]^2 \right\} - 1 \right) \right] \right.$$
$$= \mathcal{O}\{s_0\sqrt{Tlog(2p)}\}.$$

where P_{β_ϑ}, $\mathcal{X}_\vartheta$ *and* C_m *are provided in Lemmas 4, 7 and Assumption 4, respectively.*

Prova. For $\vartheta \leq vw$, the cumulative regret of both HD ϵ_t-Greedy and CHD ϵ_t-Greedy algorithms are given by Lemma 5.

For $\vartheta > vw$, Lemmas 6 and 7 indicates that, with probability $1 - P_{\beta_\vartheta}$, instantaneous regret of both algorithms are bounded. Considering the whole period $vw < \vartheta < T$, there exists $P_{\beta_{max}} \equiv \underset{vw<\vartheta<T}{max} P_{\beta_\vartheta}$ such that, at least, with probability $1 - P_{\beta_{max}}$ all instantaneous regret are bounded. Following this line, we first compute the cumulative regret for the HD ϵ_t Greedy algorithm until time T and then, express the regret for the conservative version as a function of the first. Since $1/\vartheta$ is a decreasing function of ϑ, using Lemma 12:

$$R_{T-1,\vartheta>vw}^{HD} \leq w\theta_x h\tau_{\mathcal{W}} \sum_{\vartheta=vw+1}^{T-1} P_{k\vartheta}^{HD} \leq w\theta_x h\tau_{\mathcal{W}} \sum_{\vartheta=vw+1}^{T-1} \frac{v}{\vartheta} +$$
$$\left(1 - \frac{vw}{\vartheta}\right) \frac{8C_m\theta_x s_0 C_\lambda \sigma\sqrt{2log(2p)}}{\phi_0^2\sqrt{\vartheta}} \leq$$
$$w\theta_x h\tau_{\mathcal{W}} \left[v \log\left(\frac{T-1}{vw}\right) + \frac{16C_m\theta_x s_0 C_\lambda \sigma\sqrt{2log(2p)}}{\phi_0^2} \right.$$
$$\left. \left[\left(\sqrt{T-1} - \sqrt{vw} + \frac{1}{\sqrt{T-1}} - \frac{1}{\sqrt{vw}} \right) \right] \right] = \mathcal{O}\{s_0\sqrt{Tlog(2p)}\}.$$

Adding the period before vw, then with probability at least $1 - P_{\beta_{max}}$, the total cumulative regret for the HD ϵ_t-Greedy until time T is bounded as:

$$R_{T-1}^{HD} \leq w\theta_x h\tau_{\mathcal{W}} \left[v + v \log\left(\frac{T-1}{vw}\right) + \right.$$
$$\left. \frac{16C_m\theta_x s_0 C_\lambda \sigma\sqrt{2log(2p)}}{\phi_0^2} \left[\left(\sqrt{T-1} - \sqrt{vw} + \frac{1}{\sqrt{T-1}} - \frac{1}{\sqrt{vw}} \right) \right] \right]$$

For the CHD ϵ_t-Greedy algorithm, from Lemma 7:

$$\begin{aligned} R^{CHD}_{T-1,\vartheta>vw} &\leq w\theta_x h\tau_{\mathcal{W}} \sum_{\vartheta=vw+1}^{T-1} \left(P^{CHD}_{k\vartheta} - \frac{\epsilon_\vartheta s_\vartheta}{w} + P^{HD}_{k\vartheta} \right) \\ &\leq R^{HD}_{T-1,\vartheta>vw} + w\theta_x h\tau_{\mathcal{W}} \sum_{\vartheta=vw+1}^{T-1} \frac{vws_\vartheta}{\vartheta} \\ &\left(\exp\left\{ -\frac{2}{w} \left[w\left(1 - P_{\beta_\vartheta}\right) - \mathcal{X}_\vartheta \right]^2 \right\} \right) - \frac{vs_\vartheta}{\vartheta} \end{aligned}$$

For an $s_t \equiv s$ chosen by the end-user:

$$\begin{aligned} R^{CHD}_{T-1,\vartheta>vw} \leq & R^{HD}_{T-1,\vartheta>vw} + w\theta_x h\tau_{\mathcal{W}} \Big[vws \log\left(\frac{T-1}{vw}\right) \\ & \left(\exp\left\{ -\frac{2}{w} \left[w\left(1 - P_{\beta_\vartheta}\right) - \mathcal{X}_\vartheta \right]^2 \right\} \right) - vs \log\left(\frac{T-1}{vw}\right) \Big] \end{aligned}$$

Finally, also with probability at least $1 - P_{\beta_{max}}$, the total cumulative regret for the CHD ϵ_t-Greedy algorithm until time T is bounded as:

$$\begin{aligned} R^{CHD}_{T-1} &\leq R^{HD}_{T-1} + w\theta_x h\tau_{\mathcal{W}} \left[vs \log\left(\frac{T-1}{vw}\right) \left(w \exp\left\{ -\frac{2}{w} \Big[\left[w\left(1 - P_{\beta_\vartheta}\right) - \mathcal{X}_\vartheta \right]^2 \right\} - 1 \right) \right] \\ &= \mathcal{O}\{s_0 \sqrt{Tlog(2p)}\}. \end{aligned}$$

Provided that $w > (12 + 2\sqrt{2})\mathcal{X}_{max}$, Theorem 2 indicates that:

$$w \exp\left\{ -\frac{2}{w} \left[w\left(1 - P_{\beta_\vartheta}\right) - \mathcal{X}_\vartheta \right]^2 \right\} < 1$$

and R^{CHD}_{T-1} as a whole is $\mathcal{O}\{log(T)\}$ strictly lower than the bound for R^{HD}_{T-1}. ■

Theorem 2 (Flexibility and Dominance of CHD ϵ_t-Greedy algorithm) *Provided that the conditions required by Lemmas 6 and 7 are satisfied, the upper bound for the CHD ϵ_t-Greedy algorithm does not depend on κ_ϑ and, at least with probability $1 - P_{\beta_{max}}$, for an increasing sequence $\{\lambda_\vartheta\}_{\vartheta>vw}$, if $w \geq (12 + 2\sqrt{2})\mathcal{X}_{max}$:*

$$\sup_{\vartheta \in \mathcal{T} \cap \{\vartheta | \vartheta > vw\}} r^{CHD}_\vartheta < \sup_{\vartheta \in \mathcal{T} \cap \{\vartheta | \vartheta > vw\}} r^{HD}_\vartheta,$$

where P_{β_ϑ} is defined in Lemma 4, r^{CHD}_ϑ is provided in Lemma 6, while r^{HD}_ϑ and $\mathcal{X}_{max}$ are defined in Lemma 7, where $\mathcal{X}_{max}$ is the usual $\mathcal{X}_\vartheta$ plugged with λ_{max}.

Prova.

For the second part of the theorem, note that from the results of Lemmas 6 and 7, we know that for every $k \in \{0, ..., w-1\}$ and for $\vartheta > vw$, at least

with probability $1 - P_{\boldsymbol{\beta}_{max}}$ (see Theorem 1):

$$\sup_{\vartheta > vw} r_{\vartheta}^{CHD} \leq w\theta_x h\tau_{\mathcal{W}} \left(P_{k\vartheta}^{CHD} - \frac{\epsilon_\vartheta s_\vartheta}{w} + P_{k\vartheta}^{HD} \right) = \sup_{\vartheta > vw} r_{\vartheta}^{HD} + w\theta_x h\tau_{\mathcal{W}} \left(P_{k\vartheta}^{CHD} - \frac{\epsilon_\vartheta s_\vartheta}{w} \right)$$

For the CHD ϵ_t-Greedy to have a stricter instantaneous bound than its non-conservative version, it is sufficient that:

$$P_{k\vartheta}^{CHD} - \frac{\epsilon_\vartheta s_\vartheta}{w} < 0 \iff \frac{1}{\kappa_\vartheta} \mathbb{P}(\boldsymbol{x}'_\vartheta \widehat{\boldsymbol{\beta}}_k \in \mathcal{H}_\vartheta^{(\kappa_\vartheta)}) < \frac{1}{w}$$

which is guaranteed to occur when:

$$\begin{aligned} w < \frac{2}{w} \left[w \left(1 - P_{\boldsymbol{\beta}_\vartheta} \right) - \mathcal{X}_\vartheta \right]^2 \implies w &< \exp\left\{ \frac{2}{w} \left[w \left(1 - P_{\boldsymbol{\beta}_\vartheta} \right) - \mathcal{X}_\vartheta \right]^2 \right\} \\ &= \frac{\kappa_\vartheta}{\mathbb{P}(\boldsymbol{x}'_\vartheta \widehat{\boldsymbol{\beta}}_k \in \mathcal{H}_\vartheta^{(\kappa_\vartheta)})} \end{aligned} \tag{A-25}$$

The LHS of equation (A-25) unfolds three cases of interest:

Case 1: Consider the set $\mathcal{E}_1 \equiv \{\vartheta > vw | (1 - P_{\boldsymbol{\beta}_\vartheta}) \geq 0.75\}$, where 0.75 is arbitrarily chosen to guarantee that $P_{\boldsymbol{\beta}_\vartheta} < 1 - \sqrt{1/2}$. On $\mathcal{E}_1$, the LHS of equation (A-25) has positive curvature and a discriminant of $\Delta_\vartheta = 8\mathcal{X}_\vartheta^2 > 0$, indicating the existence of two real roots, $w_l < w_r$.

Then, for any $\vartheta \in \mathcal{E}_1$, it is sufficient for our purposes to analyze when w can be greater than or equal to the largest root ($w \geq w_r$). Since w_r is increasing with $P_{\boldsymbol{\beta}_\vartheta}$, we require that:

$$w \geq \frac{2\mathcal{X}_{max}(2(0.75) + \sqrt{2})}{4\left(0.75\right)^2 - 2} = (12 + 2\sqrt{2})\mathcal{X}_{max} \geq w_r = \frac{2\mathcal{X}_\vartheta(2(1 - P_{\boldsymbol{\beta}_\vartheta}) + \sqrt{2})}{4\left(1 - P_{\boldsymbol{\beta}_\vartheta}\right)^2 - 2} \tag{A-26}$$

where $\mathcal{X}_{max} \equiv \mathcal{X}_\vartheta$ for values of $\lambda_\vartheta = \lambda_{max}$.

Refer to the discussion made in Remark 5 for the cases when $\{\lambda_\vartheta\}_{\vartheta > vw}$ is an increasing sequence. Coupled with a decreasing relationship between $P_{\boldsymbol{\beta}_\vartheta}$ and λ_ϑ (see Lemma 4), case 1 ($P_{\boldsymbol{\beta}_\vartheta} < 1 - \sqrt{1/2}$) is the most important situation for the high-dimensional case studied in this work. Nonetheless we also briefly provide below the analysis for the other cases.

Case 2: Consider the set $\mathcal{E}_2 \equiv \{\vartheta > vw | (1 - P_{\boldsymbol{\beta}_\vartheta}) \leq 0.7\}$ where, like case 1, 0.7 is chosen to guarantee a negative curvature for the LHS of equation (A-25). The discriminant is the same as in case 1 ($8\mathcal{X}_\vartheta^2$) but, easy calculations show that $\nexists w > 0$ for our result to hold, since $w_l < w_r < 0$.

Case 3: Consider the set $\mathcal{E}_3 \equiv \{\vartheta > vw | (1 - P_{\boldsymbol{\beta}_\vartheta}) = 1 - \sqrt{1/2}\}$. In this case, $1 < w < \frac{\mathcal{X}_{min}}{\sqrt{2}}$ is sufficient for our result to hold, for $\mathcal{X}_{min}$ the counterpart of the above-defined $\mathcal{X}_{max}$.

For the first part of this theorem, recall that

$$\begin{aligned}\sup_{\vartheta>vw} r_{\vartheta}^{CHD} =& w\theta_x h\tau_{\mathcal{W}}\left(P_{k\vartheta}^{CHD} - \frac{\epsilon_\vartheta s_\vartheta}{w} + P_{k\vartheta}^{HD}\right) = \\ & w\theta_x h\tau_{\mathcal{W}}\epsilon_\vartheta s_\vartheta\left[\frac{\mathbb{P}(\boldsymbol{x}'_\vartheta \in \mathcal{H}_\vartheta^{(\kappa_\vartheta)})}{\kappa_\vartheta} - \frac{1}{w}\right] + w\theta_x h\tau_{\mathcal{W}} P_{k\vartheta}^{HD} \leq \\ & w\theta_x h\tau_{\mathcal{W}}\epsilon_\vartheta s_\vartheta\left(\exp\left\{-\frac{2}{w}\left[w\left(1-P_{\beta_\vartheta}\right) - \mathcal{X}_\vartheta\right]^2\right\} - \frac{1}{w}\right) + w\theta_x h\tau_{\mathcal{W}} P_{k\vartheta}^{HD}\end{aligned} \tag{A-27}$$

None of the terms in inequality (A-27) depend on κ_ϑ, which completes the proof. ■

B
Appendix to Chapter 2

In this appendix, we provide the proofs of the Theorems proposed in Chapter 2, and respective Auxiliary Lemmas.

B.1
Auxiliary Lemmas

Proof of Lemma 1

For any tree $\mathcal{T}_b$, $b = 1, ..., B$, provided that an arbitrary cell $A_{b,k-1}$ has $n(A_{b,k-1}) > 2$ then, any cell $A_{b,k}$, child of $A_{b,k-1}$ and formed by a sequence of splits $\mathcal{H}_{b,k}$ has the property:

$$n(A_{b,k}) \geq \frac{\Delta n(A_{b,k-1})^2}{2(n(A_{b,k-1}) - 2)} \geq n\Big(\frac{\Delta}{2}\Big)^k$$

Prova.

Consider a tree $\mathcal{T}_b$, and an arbitrary cell $A_{b,k-1}$ formed by the sequence of splits $\mathcal{H}_{b,k-1}$. Let $h_{b,k} \in \mathcal{C}_{b,k}$ be a new split and label, without loss of generality, the left child of $A_{b,k-1}$ as $A_{b,k}$ and its right child as $A_{r,b,k-1}$. Since results are valid for any tree, in the rest of this proof we suppress the subscript b to simplify notation.

The purity improvement generated by an arbitrary split h_k is:

$$
\begin{aligned}
\Gamma(A_{k-1}, h_k) =& G(A_{k-1}) - \phi_l(A_{k-1})G(A_k) - \phi_r(A_{k-1})G(A_{r,k-1}) = \\
& 2\phi_0(A_{k-1})(1-\phi_0(A_{k-1})) - 2\phi_l(A_{k-1})\phi_0(A_k)(1-\phi_0(A_k)) - \\
& 2\phi_r(A_{k-1})\phi_0(A_{r,k-1})(1-\phi_0(A_{r,k-1})) = \\
& 2\Big(\frac{n_0(A_{k-1})n_1(A_{k-1})}{n(A_{k-1})^2} - \frac{n_0(A_k)n_1(A_k)}{n(A_{k-1})n(A_k)} - \frac{n_0(A_{r,k-1})n_1(A_{r,k-1})}{n(A_{k-1})n(A_{r,k-1})}\Big) \leq \\
& \frac{2}{n(A_{k-1})^2}\Big(n_0(A_{k-1})n_1(A_{k-1}) - n_0(A_k)n_1(A_k) - n_0(A_{r,k-1})n_1(A_{r,k-1})\Big) = \\
& \frac{2}{n(A_{k-1})^2}\Big(n_0(A_{k-1})n_1(A_{k-1}) - n_0(A_k)n_1(A_k) - \\
& (n_0(A_{k-1}) - n_0(A_k))(n_1(A_{k-1}) - n_1(A_k))\Big) = \\
& \frac{2}{n(A_{k-1})^2}\Big(n_0(A_{k-1})n_1(A_k) + n_1(A_{k-1})n_0(A_k) - 2n_0(A_k)n_1(A_k)\Big) \leq \\
& \frac{2}{n(A_{k-1})^2}((n_0(A_{k-1}) + n_1(A_{k-1}))(n_1(A_k) + n_0(A_k)) - 2(n_0(A_k) + n_1(A_k)) = \\
& \frac{2(n(A_{k-1}) - 2)n(A_k)}{n(A_{k-1})^2}
\end{aligned}
\tag{B-1}
$$

where the first inequality uses the fact the $\forall k$, both $n(A_k) \leq n(A_{k-1})$ and $n(A_{r,k-1}) \leq n(A_{k-1})$, while the second inequality is a simplification that recognizes both that $\forall x, y, z, w \geq 2$, $xy + zw \leq (x+z)(y+w)$ and that $yw \geq y + w$.

Since A_{k-1} was in fact divided to generate two children, by the stopping rule described in Section 2.2, the purity improvement was higher enough to not trigger an early stop in the tree growing process. That is:

$$
\Delta \leq \Gamma(A_{k-1}, h_k) \leq \frac{2(n(A_{k-1}) - 2)n(A_k)}{n(A_{k-1})^2} \iff n(A_k) \geq \frac{\Delta n(A_{k-1})^2}{2(n(A_{k-1}) - 2)} \tag{B-2}
$$

which implies, recursively that:

$$
n(A_k) \geq \frac{\Delta n(A_{k-1})^2}{2(n(A_{k-1}) - 2)} > \frac{\Delta}{2}n(A_{k-1}) > \left(\frac{\Delta}{2}\right)^2 n(A_{k-2}) > \cdots > \left(\frac{\Delta}{2}\right)^k n
$$

■

Proof of Lemma 2

For any tree $\mathcal{T}_b$, $b = 1, ..., B$ and any cell $A_{b,k-1}$ it is true that for any $h_{b,k} \in \mathcal{C}_{b,k}$, uniformly:

$$
\lim_{n\to\infty} \mathbb{P} \sup_{h_{b,k} \in \mathcal{C}_{b,k}} \left|\hat{\Omega}(A_{b,k-1}, h_{b,k}) - \Omega(A_{b,k-1}, h_{b,k})\right| \xrightarrow{p} 0
$$

Prova.

As we did in the proof of Lemma 1, we also remove the index b here, since

all steps of this proof can be equally applied to any tree without distinction. If an arbitrary cell A_{k-1}, formed by the sequence of splits $\hat{\mathcal{H}}_{k-1}$, is going to be split into $A_{l,k-1}$ and $A_{r,k-1}$ then, the procedure to select best splits (see Section 2.2) finds a pair $\hat{h}_k = (\hat{j}_k, \hat{\zeta}_k) \in \mathcal{C}_k$, such that the gain in purity is maximum, which is equivalent to minimize:

$$\hat{\Omega}(A_{k-1}, \hat{h}_k) = \hat{\phi}_l(A_{k-1}, \hat{h}_k)\hat{G}(A_{l,k-1}, \hat{h}_k) + \hat{\phi}_r(A_{k-1}, \hat{h}_k)\hat{G}(A_{r,k-1}, \hat{h}_k) \quad \text{(B-3)}$$

Also, notice that any set $\mathcal{C}_k$ is discrete in itself bounded down, according to the Definition 7. To see that, by contradiction, assume that $\mathcal{C}_k$ is not discrete in itself. Then, by definition, there exists a sequence of splits $\{\hat{h}_i \in \mathcal{C}_k; i = 1, 2, 3...\}$ such that $\hat{h}_i^* = \lim \hat{h}_i$ for some $\hat{h}_i^* \in \mathcal{C}_k$. That is, for each $\epsilon > 0$, $\exists N \in \mathbb{N}$ such that $\forall n \geq N$ implies that $\left\|\hat{h}_i^* - \hat{h}_n\right\|_2 \leq \epsilon$.

Define: $\delta = \underset{i,i'}{min}|\hat{\zeta}_i - \hat{\zeta}_{i'}|$ for all possible cuts $\hat{h}_i$, $\hat{h}_{i'} \in \mathcal{C}_k$ and take $\epsilon = \delta/2$. Then, for all $n \geq N$:

$$\left\|\hat{h}_i^* - \hat{h}_n\right\|_2 \leq \frac{\delta}{2} \iff \sqrt{(\hat{j}_i^* - \hat{j}_n)^2 + (\hat{\zeta}_i^* - \hat{\zeta}_n)^2} \leq \frac{\delta}{2} \implies |\hat{\zeta}_i^* - \hat{\zeta}_n| \leq \frac{\delta}{2}$$

which is impossible, since both $\hat{\zeta}_i^*$ and $\hat{\zeta}_n$, $\forall n \geq N$, belong to $\mathcal{C}_k$.

For the bounded down part, one needs only to recognize that δ as above defined is the lower bound for the distance between any two points in $\mathcal{C}_k$. Therefore, there is a finite amount of possibilities to choose $\hat{h}_k$. Since $\hat{j}_k \in W_b$, with cardinality w, and $\hat{\zeta}_k \in \{\boldsymbol{x}_1^{(\hat{j}_k)}, ..., \boldsymbol{x}_s^{(\hat{j}_k)}\}$, then $\#\mathcal{C}_k = ws$.

In this context, take $\hat{\phi}_l(A_{k-1}, \hat{h}_k)$ in equation (B-3) as example. By definition:

$$\hat{\phi}_l(A_{k-1}, \hat{h}_k) = \frac{1}{n(A_{k-1})} \sum_{i=1}^{n(A_{k-1})} \mathbb{1}\{\boldsymbol{x}_i^{(\hat{j}_k)} < \hat{\zeta}_k\}$$

For every $\hat{h}_k \in \mathcal{C}_k$, the random sequence $\left\{\mathbb{1}\{\boldsymbol{x}_i^{(\hat{j}_k)} < \hat{\zeta}_k\}\right\}$ is uniformly integrable, since $E\left[\mathbb{1}\{\boldsymbol{x}_i^{(\hat{j}_k)} < \hat{\zeta}_k\}\right] = \mathbb{P}(\boldsymbol{x}_i^{(\hat{j}_k)} < \hat{\zeta}_k | \boldsymbol{x}_i \in A_{k-1}) \in [0, 1]$. Then, a Pointwise Law of Large Numbers holds, leading to:

$$\hat{\phi}_l(A_{k-1}, \hat{h}_k) = \frac{1}{n(A_{k-1})} \sum_{i=1}^{n(A_{k-1})} \mathbb{1}\{\boldsymbol{x}_i^{(\hat{j}_k)} < \hat{\zeta}_k\} \xrightarrow{p} \mathbb{P}(\boldsymbol{x}_i^{(\hat{j}_k)} < \hat{\zeta}_k | \boldsymbol{x}_i \in A_{k-1}) \equiv \phi_l(A_{k-1}, \hat{h}_k)$$

In our discrete set context, this is sufficient for the uniform convergence in probability. In fact, notice that for every $\hat{h}_k \in \mathcal{C}_k$:

$$\sup_{\hat{h}_k \in \mathcal{C}_k} |\hat{\phi}_l(A_{k-1}, \hat{h}_k) - \phi_l(A_{k-1}, \hat{h}_k)| \leq \sum_{i=1}^{ws} |\hat{\phi}_l(A_{k-1}, \hat{h}_i) - \phi_l(A_{k-1}, \hat{h}_i)| \xrightarrow{p} 0$$

The same operational steps is applied to $\hat{G}(\cdot)$ in equation (B-3). For the sake of completeness take the sequence $\{\hat{\phi}_0(A_{l,k-1}, \hat{h}_k)\}$. Also, by definition,

$$\hat{\phi}_0(A_{l,k-1}, \hat{h}_k) = \frac{1}{n(A_{l,k-1})} \sum_{i=1}^{n(A_{l,k-1})} \mathbb{1}\{d_i = 0\}\mathbb{1}\{\boldsymbol{x}_i^{(\hat{j}_k)} < \hat{\zeta}_k\}$$

which is also integrable, leading to the uniform convergence of $\hat{\phi}_0(A_{l,k-1}, \hat{h}_k)$ to $\phi_0(A_{l,k-1}, \hat{h}_k)$.

Consequently, by the Uniform Continuous Mapping Theorem:

$$\hat{G}(A_{l,k-1}, \hat{h}_k) = 2\hat{\phi}_0(A_{l,k-1}, \hat{h}_k)(1 - \hat{\phi}_0(A_{l,k-1}, \hat{h}_k)) \xrightarrow{p}$$
$$2\mathbb{P}(d_i = 0|\boldsymbol{x}_i^{(\hat{j}_k)} < \hat{\zeta}_k, \boldsymbol{x}_i \in A_{k-1})(1 - \mathbb{P}(d_i = 0|\boldsymbol{x}_i^{(\hat{j}_k)} < \hat{\zeta}_k, \boldsymbol{x}_i \in A_{k-1})) \equiv$$
$$2\phi_0(A_{l,k-1}, \hat{h}_k)(1 - \phi_0(A_{l,k-1}, \hat{h}_k)) = G(A_{l,k-1}, \hat{h}_k)$$

The same developments can be made for $A_{r,k-1}$, which yields the result. ■

Lemma 13 (Marcinkiewicz-Zygmund-Burkholder Inequality) *Let $r \geq 1$, $\{X_j\}$ a sequence of independent random variables with $E[X_j] = 0$, $j = 1, ..., n$. Then, there are positive constants a_r and b_r such that:*

$$a_r \mathbb{E}\Big(\sum_{j=1}^{n} X_j^2\Big)^{r/2} \leq \mathbb{E}\Big|\sum_{j=1}^{n} X_j\Big|^r \leq b_r \mathbb{E}\Big(\sum_{j=1}^{n} X_j^2\Big)^{r/2}$$

Prova. See Section 9.7 of Lin and Bai (2011) ■

Lemma 14 (ϵ_b-Greedy) *For $Q > 1$, if policy ϵ_b-Greedy is run with input parameter $\mu > 0$ then, the probability that after any number $b \geq \mu\psi Q$ of plays, ϵ_b-Greedy chooses a suboptimal action is at most:*

$$\mathbb{P}_\epsilon \equiv \frac{\mu}{b} + 2\Big(\mu log\Big(\frac{(b-1)e^{1/2}}{\mu Q}\Big)\Big)\Big(\frac{Q}{(b-1)\mu e^{1/2}}\Big)^{\mu/5} + \frac{4e}{\psi^2}\Big(\frac{Q}{(b-1)\mu e^{1/2}}\Big)$$

where $0 < \psi < 1$ and Q is defined in Section 2.3.

Prova. See Theorem 3 in Auer et al. (2002) ■

B.2
Theorems

Proof of Theorem 3

For any tree $\mathcal{T}_b$, $b = 1, ..., B$ and any cell $A_{b,k-1}$, provided that Assumption 9 holds, an empirical split $\hat{h}_{b,k} \in \mathcal{C}_{b,k}$ is consistent in the sense that $\hat{h}_{b,k} - h^0_{b,k}$ is $o_p(1)$.

Prova.

First of all, notice that for every tree $\mathcal{T}_b$, $\hat{\Omega}(A_{k-1}, h_k)$ is uniformly continuous in $\mathcal{C}_k$. If it were not, there should exists an $\epsilon > 0$, such that for all $\delta > 0$, we could find an $h'_k \in \mathcal{C}_k$ such that $\|h_k - h'_k\| < \delta$ and $|\hat{\Omega}(A_{k-1}, h_k) - \hat{\Omega}(A_{k-1}, h'_k)| > \epsilon$. Take the same $\delta = \underset{k,k'}{min}|\zeta_k - \zeta_{k'}|$ defined in the proof of Lemma 2. Although h_k is defined in terms of the pair (j_k, ζ_k), we do not mention j_k and j'_k when choosing δ, since by assumption, $\boldsymbol{x}_i \in [0, 1]^p$ $\forall i$. Therefore, this particular δ guarantees that we are referring to cuts at the same direction, since $\underset{k,k'}{min}|\zeta_k - \zeta_{k'}| < 1$.

Therefore, for two splits at the same direction, $h_k = (j_k, \zeta_k)$ and $h'_k = (j_k, \zeta'_k)$, given the discreteness of $\mathcal{C}_k$ and argued by Burgin (2010), the only possibility for $\|h_k - h'_k\| < \delta$ is when $h_k = h'_k$. This implies that there is not an $\epsilon > 0$, such that $|\hat{\Omega}(A_{k-1}, h_k) - \hat{\Omega}(A_{k-1}, h'_k)| > \epsilon$. Uniform continuity comes from the Heine-Cantor Theorem, since $\mathcal{C}_k$ is a compact set.

Moreover, recall that Lemma 2 shows that $\hat{\Omega}(A_{k-1}, h_k)$ uniformly converges in probability in $h_k \in \mathcal{C}_k$ to its population expectation $\Omega(A_{k-1}, h_k)$. Then, provided that Assumption 9 holds, all conditions of Theorem 4.1.1 of Amemiya (1985) are satisfied, and the result follows. ■

Proof of Theorem 4

Provided that Assumptions 5 to 10 hold, with probability at least $1 - \mathbb{P}_\epsilon$, for $b \geq \mu\psi Q$, the sequence of trees $\{\mathcal{T}_b\}$, trained on random subsamples of $\mathcal{S}_n$, asymptotically identifies the boundary A_a induced by a deterministic complex assignment rule a and, for $p > 0$ finite and fixed (not growing with sample size), $\hat{\tau} - \tau$ is $o_p(1)$, where $\mathbb{P}_\epsilon$ and ψ are defined in Lemma 14.

Prova.

Considering Assumption 8 and the model in equation (2-2), border treatment effects can be estimated by the difference between intercepts at each side of cutoff: $\hat{\tau}^{(f)}_{b,l} = \hat{\alpha}_{b,l+} - \hat{\alpha}_{b,l}$.

Consider a tree $\mathcal{T}_b$ and a pair $(A_{b,k_i}, A_{b,k_j}) \in \mathcal{F}_b$ sharing the l-th border. Without loss of generality, take the regression to the left of the cutoff (units inside A_{b,k_i}) as example and consider the definitions made in Assumption 10. In the following development we eliminate the subscripts "b" and "$k_i|\kappa_l$" but keep in mind that we are investigating the regression using units inside $A_{b,k_i|\kappa_l}$. Write, in matrix representation, $\mathbf{y} = \boldsymbol{\iota}\alpha + \mathbf{Z}\boldsymbol{\beta} + \mathbf{e}$.

Using an empirical split $\hat{\zeta}$ as an approximation for ζ^0, the well known OLS estimator for the intercept is given by $\hat{\alpha}(\hat{\zeta}) = (\boldsymbol{\iota}'\boldsymbol{\iota})^{-1}\boldsymbol{\iota}'(\mathbf{y} - \mathbf{Z}(\hat{\zeta})\hat{\boldsymbol{\beta}}(\hat{\zeta}))$.

Then:

$$\begin{aligned}\hat{\alpha}(\hat{\zeta}) - \hat{\alpha}(\zeta^0) =& \frac{1}{n(A)}\Big[\sum_{i=1}^{n(A)} \boldsymbol{y}^{(i)} - \boldsymbol{Z}^{(i)}(\hat{\zeta})\hat{\boldsymbol{\beta}}(\hat{\zeta}) - \sum_{i=1}^{n(A)} \boldsymbol{y}^{(i)} - \boldsymbol{Z}^{(i)}(\zeta^0)\hat{\boldsymbol{\beta}}(\zeta^0)\Big] = \\ & \frac{1}{n(A)}\sum_{i=1}^{n(A)}\sum_{j=1}^{p}\Big[\boldsymbol{Z}^{(i,j)}(\zeta^0)\hat{\boldsymbol{\beta}}^{(j)}(\zeta^0) - \boldsymbol{Z}^{(i,j)}(\hat{\zeta})\hat{\boldsymbol{\beta}}^{(j)}(\hat{\zeta})\Big] = \\ & \frac{1}{n(A)}\sum_{i=1}^{n(A)}\sum_{j=1}^{p}\Big[\boldsymbol{Z}^{(i,j)}(\zeta^0)(\hat{\boldsymbol{\beta}}^{(j)}(\zeta^0) - \boldsymbol{\beta}^{(j)}) + (\boldsymbol{Z}^{(i,j)}(\zeta^0) - \boldsymbol{Z}^{(i,j)}(\hat{\zeta}))\boldsymbol{\beta}^{(j)} + \\ & \boldsymbol{Z}^{(i,j)}(\hat{\zeta})(\boldsymbol{\beta}^{(j)} - \hat{\boldsymbol{\beta}}^{(j)}(\hat{\zeta}))\Big]\end{aligned}$$

Consequently, for $\epsilon > 0$:

$$\begin{aligned}&\mathbb{P}\Big(|\hat{\alpha}(\hat{\zeta}) - \alpha| > \epsilon\Big) = \mathbb{P}\Big(|\hat{\alpha}(\hat{\zeta}) - \hat{\alpha}(\zeta^0) + \hat{\alpha}(\zeta^0) - \alpha| > \epsilon\Big) \leq \\ &\mathbb{P}\Big(\Big|\frac{1}{n(A)}\sum_{i=1}^{n(A)}\sum_{j=1}^{p}\Big[\boldsymbol{Z}^{(i,j)}(\zeta^0)(\hat{\boldsymbol{\beta}}^{(j)}(\zeta^0) - \boldsymbol{\beta}^{(j)}) + (\boldsymbol{Z}^{(i,j)}(\zeta^0) - \boldsymbol{Z}^{(i,j)}(\hat{\zeta}))\boldsymbol{\beta}^{(j)}\Big| > \epsilon/4\Big) + \\ &\mathbb{P}\Big(\Big|\frac{1}{n(A)}\sum_{i=1}^{n(A)}\sum_{j=1}^{p}\boldsymbol{Z}^{(i,j)}(\hat{\zeta})(\boldsymbol{\beta}^{(j)} - \hat{\boldsymbol{\beta}}^{(j)}(\hat{\zeta}))\Big| > \epsilon/4\Big) + \mathbb{P}\Big(|\hat{\alpha}(\zeta^0) - \alpha| > \epsilon/2\Big) \leq \\ &\underbrace{\mathbb{P}\Big(\frac{1}{n(A)}\sum_{i=1}^{n(A)}\sum_{j=1}^{p}\Big|\boldsymbol{Z}^{(i,j)}(\zeta^0)(\hat{\boldsymbol{\beta}}^{(j)}(\zeta^0) - \boldsymbol{\beta}^{(j)})\Big| > \epsilon/8\Big)}_{A} + \\ &\underbrace{\mathbb{P}\Big(\frac{1}{n(A)}\sum_{i=1}^{n(A)}\sum_{j=1}^{p}\Big|(\boldsymbol{Z}^{(i,j)}(\zeta^0) - \boldsymbol{Z}^{(i,j)}(\hat{\zeta}))\boldsymbol{\beta}^{(j)}\Big| > \epsilon/8\Big)}_{B} + \\ &\underbrace{\mathbb{P}\Big(\frac{1}{n(A)}\sum_{i=1}^{n(A)}\sum_{j=1}^{p}\Big|\boldsymbol{Z}^{(i,j)}(\hat{\zeta})(\boldsymbol{\beta}^{(j)} - \hat{\boldsymbol{\beta}}^{(j)}(\hat{\zeta}))\Big| > \epsilon/4\Big)}_{C} + \underbrace{\mathbb{P}\Big(|\hat{\alpha}(\zeta^0) - \alpha| > \epsilon/2\Big)}_{D}\end{aligned} \tag{B-4}$$

where all inequalities use either the triangle inequality or the proposition 5.1 in Lin and Bai (2011). In our framework, for every $\zeta \in \mathcal{C}$, $\max_{i\in\{1,\ldots,n(A)\},j\in\{1,\ldots,p\}} \boldsymbol{Z}^{(i,j)}(\zeta) \leq 1$. Using this fact, equation B-4 resumes to:

$$\mathbb{P}\Big(|\hat{\alpha}(\hat{\zeta}) - \alpha| > \epsilon\Big) \leq \underbrace{\mathbb{P}\Big(\max_{j\in\{1,...,p\}} \Big|\hat{\boldsymbol{\beta}}^{(j)}(\zeta^0) - \boldsymbol{\beta}^{(j)}\Big| > \epsilon/8p\Big)}_{A'} + B + \underbrace{\mathbb{P}\Big(\max_{j\in\{1,...,p\}} \Big|\boldsymbol{\beta}^{(j)} - \hat{\boldsymbol{\beta}}^{(j)}(\hat{\zeta}))\Big| > \epsilon/4p\Big)}_{C'} + D \tag{B-5}$$

Since $\hat{\alpha}(\zeta^0)$ and $\hat{\boldsymbol{\beta}}(\zeta^0)$ are OLS local parameters estimates based on the knowledge of true splits and considering the exogeneity premise in Assumption 7, for any $p > 0$ fixed and finite (not growing with the sample size) $\lim_{n\to\infty} A' \to 0$ and the same happens with D in equation (B-5).

Regarding the term B, notice that $\forall i \in \{1, ..., n(A)\}$ and $\forall j \in \{1, ..., p\}$:

$$\boldsymbol{Z}^{(i,j)}(\zeta^0) - \boldsymbol{Z}^{(i,j)}(\hat{\zeta}) = \begin{cases} \zeta^0 - \hat{\zeta}, \text{ if } j = 1 \\ 0, \text{ otherwise} \end{cases}$$

Then, using assumption 7.i:

$$\begin{aligned} &\mathbb{P}\Big(\frac{1}{n(A)} \sum_{i=1}^{n(A)} \sum_{j=1}^{p} \Big|(\boldsymbol{Z}^{(i,j)}(\zeta^0) - \boldsymbol{Z}^{(i,j)}(\hat{\zeta}))\boldsymbol{\beta}^{(j)}\Big| > \epsilon/8\Big) \leq \\ &\mathbb{P}\Big(\frac{1}{n(A)} \sum_{i=1}^{n(A)} \Big|(\zeta^0 - \hat{\zeta}) \max_{j\in\{1,...,p\}} \boldsymbol{\beta}^{(j)}\Big| > \epsilon/8\Big) \leq \\ &\mathbb{P}\Big(\Big|\zeta^0 - \hat{\zeta}\Big| > \epsilon/8\delta_\beta\Big) \end{aligned} \tag{B-6}$$

which also goes to zero as $n \to \infty$ by the results of theorem 3.

Finally, regarding the term C':

$$\begin{aligned} &\mathbb{P}\Big(\max_{j\in\{1,...,p\}} \Big|\boldsymbol{\beta}^{(j)} - \hat{\boldsymbol{\beta}}^{(j)}(\hat{\zeta}))\Big| > \epsilon/4p\Big) \leq \\ &\underbrace{\mathbb{P}\Big(\max_{j\in\{1,...,p\}} \Big|\boldsymbol{\beta}^{(j)} - \hat{\boldsymbol{\beta}}^{(j)}(\zeta^0))\Big| > \epsilon/8p\Big)}_{C'_1} + \underbrace{\mathbb{P}\Big(\max_{j\in\{1,...,p\}} \Big|\hat{\boldsymbol{\beta}}^{(j)}(\zeta^0) - \hat{\boldsymbol{\beta}}^{(j)}(\hat{\zeta}))\Big| > \epsilon/8p\Big)}_{C'_2} \end{aligned} \tag{B-7}$$

The term C'_1 in equation (B-7) is $o_p(1)$ by the arguments already discussed in this proof. Regarding the term C'_2, notice that the functions defined as $r(\zeta) = \mathbf{Z}'(\zeta)(\mathbf{Id} - \mathbf{P}_\iota)\mathbf{y}$ are continuous for $\zeta \in [0, 1]$. To see this, consider two points ζ, ζ' in the domain. Then:

$$\begin{aligned} &r(\zeta) - r(\zeta') = \mathbf{Z}'(\zeta)(\mathbf{Id} - \mathbf{P}_\iota)\mathbf{y} - \mathbf{Z}'(\zeta')(\mathbf{Id} - \mathbf{P}_\iota)\mathbf{y} = \\ &\Big[(\overline{y}(1 - n(A)))(\zeta' - \zeta) \quad 0 \quad \cdots \quad 0\Big]' \end{aligned}$$

where $\overline{y}$ is the sample average $\frac{1}{n(A)} \sum_{i=1}^{n(A)} \boldsymbol{y}^{(i)}$.

Then, $\forall \tau > 0$, choose $\delta_n(\tau)$ small enough, such as $\delta_n(\tau) = \tau/n$, and notice that for for any realization of the random matrix $\mathbf{Z}$, $|\zeta - \zeta'| < \delta_n$ implies

that

$$\mathbb{P}\Big(\|r(\zeta) - r(\zeta')\|_2 > \tau \Big) \leq \mathbb{P}\Big(\Big| \sum_{i=1}^{n(A)} y^{(i)} \Big(\frac{1}{n(A)} - 1 \Big) \Big| > \frac{\tau}{\delta_n} \Big) = \mathbb{P}\Big(\Big| \sum_{i=1}^{n(A)} y^{(i)} \Big| > \frac{\tau}{\delta_n \Big| \Big(\frac{1}{n(A)} - 1 \Big) \Big|} \Big) \leq \frac{\Big(\delta_n \Big| \Big(\frac{1}{n(A)} - 1 \Big) \Big| \Big)^r}{\tau^r} \mathbb{E}\Big[\Big| \sum_{i=1}^{n(A)} y^{(i)} \Big|^r \Big] \tag{B-8}$$

where the last inequality is Markov inequality. Since equation (B-8) is valid for any $r > 0$, take $r = 1$ and $b_1 > 0$ to get that

$$\begin{aligned}
&\frac{\Big(\delta_n \Big| \Big(\frac{1}{n(A)} - 1 \Big) \Big| \Big)}{\tau} \mathbb{E}\Big[\Big| \sum_{i=1}^{n(A)} y^{(i)} \Big| \Big] \leq \\
&\frac{\delta_n \Big| \Big(\frac{1}{n(A)} - 1 \Big) \Big|}{\tau} \mathbb{E}\Big[\Big| \sum_{i=1}^{n(A)} \alpha + \sum_{j=1}^{p} \mathbf{Z}^{(i,j)} \boldsymbol{\beta}^{(j)} \Big| + \Big| \sum_{i=1}^{n(A)} \mathbf{e}^{(i)} \Big| \Big] \leq \\
&\frac{\delta_n \Big| \Big(\frac{1}{n(A)} - 1 \Big) \Big|}{\tau} \Big(n(A)\delta_\alpha + n(A)p\delta_\beta + \mathbb{E}\Big[\Big| \sum_{i=1}^{n(A)} \mathbf{e}^{(i)} \Big| \Big] \Big) \leq \\
&\frac{\delta_n \Big| \Big(\frac{1}{n(A)} - 1 \Big) \Big|}{\tau} \Big(n(A)\delta_\alpha + n(A)p\delta_\beta + b_1 \mathbb{E}\Big(\sum_{i=1}^{n(A)} \mathbf{e}^{(i)2} \Big)^{1/2} \Big) \leq \\
&\frac{\delta_n \Big| \Big(\frac{1}{n(A)} - 1 \Big) \Big|}{\tau} (n(A)\delta_\alpha + n(A)p\delta_\beta + b_1 \sigma \sqrt{n(A)}) \leq \\
&\frac{C_l \delta_n}{\tau} (C_1 n(A) + C_2 n(A)^{1/2})
\end{aligned}$$

where the first inequality is triangle inequality, the second uses the bounds in Assumption 7, the third uses Lemma 13, the fourth is Jensen inequality and the Assumption 7, the fifth only recognizes that for $n(A) \in \mathbb{N}^+$, $\Big| \Big(\frac{1}{n(A)} - 1 \Big) \Big| = \frac{n(A)-1}{n(A)} = C_l < 1$, for every border l and C_1, C_2 are $(\delta_\alpha + p\delta_\beta)$, $b_1\sigma$, respectively. Since $n \geq n(A)$, picking the above mentioned suggestion for $\delta_n(\tau)$, for every $\tau > 0$, guarantees that with high probability the functions r are continuous.

Using Assumption 10 and the functions t defined therein, $\hat{\beta}(\zeta) = t(\zeta)q(\zeta)$ is continuous on $[0, 1]$. From the Continuous Mapping Theorem and the results of Theorem 3, we get that $|\hat{\boldsymbol{\beta}}(\hat{\zeta}) - \hat{\boldsymbol{\beta}}(\zeta^0)|$ is $o_p(1)$. The whole term C' also goes to zero, applying the triangle inequality, for p fixed and finite when $n \to \infty$.

Therefore, $\hat{\alpha}(\hat{\zeta})$ is a consistent estimator of α. It rests to argue that RDF algorithm provides a way to select "good" empirical splits $\hat{h}$ so that $\hat{\alpha}(\hat{\zeta})$ is not only consistent, but estimated on empirical boundaries that asymptotically approach those identified by trees in $\mathcal{D}_a$ (causal α's).

Take a tree $\mathcal{T}_b$, $b \geq \mu\psi Q$, and f'_b unique leaves $A_{b,k}$ in $\mathcal{F}_b$, $f'_b > f_b$. From Lemma 14, with probability at least $1 - \mathbb{P}_\epsilon$, the ϵ_b-Greedy rule (superscript (ϵG)) selects $W_b^{\epsilon G} \in \mathcal{Q}$ that, once imputed to the root of $\mathcal{T}_b$, leads to sequences of splits that result in empirical leaves with overall minimum impurity among all possibilities in $\mathcal{Q}$. That is:

$$\sum_{k=1}^{f'_b} \hat{G}(A_{b,k}, \hat{h}_{b,k}^{(\epsilon G)}) \equiv \underset{\{\hat{h}_{b,k}\}\in\mathcal{C}}{min} \sum_{k=1}^{f'_b} \hat{G}(A_{b,k}, \hat{h}_{b,k})$$

where $\mathcal{C}$ is the set of all possible splits, defined in Section 2.2.

Since $\hat{h}_{b,k}^{(\epsilon G)}$ is an empirical split like any other, from Theorem 3, $\hat{h}_{b,k}^{(\epsilon G)} \xrightarrow{p} h_{b,k}^{(\epsilon G)} \equiv \underset{h\in\mathcal{C}_{b,k}}{argmin}\, G(A_{b,k}, h)$, such that:

$$\sum_{k=1}^{f'_b} G(A_{b,k}, h_{b,k}^{(\epsilon G)}) \equiv \underset{\{h_{b,k}\}\in\mathcal{C}}{min} \sum_{k=1}^{f'_b} G(A_{b,k}, h_{b,k})$$

Now we claim that there must be a $m \in \{1, ..., M_a\}$ and a $k' > 0$, such that with high probability and for $b \geq \mu\psi Q$, $h_{b,k}^{(\epsilon G)} = h_{m,k'}$ for some tree $\mathcal{T}_m \in \mathcal{D}_a$ as defined in Assumption 5. Otherwise, $\forall m$ and $\forall b \geq \mu\psi Q$, $\sum_{k'=1}^{f'_b} G(A_{m,k'}, h_{m,k'}) > \sum_{k=1}^{f'_b} G(A_{b,k}, h_{b,k}^{(\epsilon G)})$ indicating that $h_{m,k'}$ does not attain maximum purity, an absurd, since from Assumption 5 there exists a unique boundary correctly identified by trees $\mathcal{T}_m \in \mathcal{D}_a$ using observables $\boldsymbol{x}$.

Since the same developments can be made for the regression to the right of cutoff (using units inside $A_{b,k_j|\kappa_l}$), for a tree $\mathcal{T}_b$, consistency for the l-th border treatment effects follows:

$$|\hat{\tau}_l^{(f)}(\hat{\zeta}_l) - \tau_l^{(f)}| = |\hat{\alpha}_{l+}(\hat{\zeta}_l) - \hat{\alpha}_l(\hat{\zeta}_l) + \alpha_l - \alpha_{l+}| \leq \underbrace{|\hat{\alpha}_{l+}(\hat{\zeta}_l) - \alpha_{l+}|}_{o_p(1)} + \underbrace{|\hat{\alpha}_l(\hat{\zeta}_l) - \alpha_l|}_{o_p(1)}$$

Finally, since the sequence $\{\mathcal{T}_b\}_{b\in\{\mu\psi Q,...,B\}}$ is trained on random subsamples of $\mathcal{S}_n$ and considering the model in equation (2-2):

$$\begin{aligned}
&\hat{\tau} = \frac{1}{B - \mu\psi Q} \sum_{b=\mu\psi Q}^{B} \hat{\tau}_b^{(t)} = \frac{1}{B - \mu\psi Q} \sum_{b=\mu\psi Q}^{B} \frac{1}{f_b} \sum_{l=1}^{f_b} \hat{\tau}_{b,l}^{(f)} = \\
&\frac{1}{B - \mu\psi Q} \sum_{b=\mu\psi Q}^{B} \frac{1}{f_b} \sum_{l=1}^{f_b} (\hat{\alpha}_{b,l+}(\hat{\zeta}_{b,l}) - \hat{\alpha}_{b,l}(\hat{\zeta}_{b,l})) \Longrightarrow \\
&|\hat{\tau} - \tau| \leq \frac{1}{B - \mu\psi Q} \sum_{b=\mu\psi Q}^{B} \frac{1}{f_b} \sum_{l=1}^{f_b} \left[\left| \hat{\alpha}_{b,l+}(\hat{\zeta}_{b,l}) - \alpha_{b,l+} \right| + \left| \alpha_{b,l} - \hat{\alpha}_{b,l}(\hat{\zeta}_{b,l})) \right| \right] \xrightarrow{p} 0
\end{aligned} \tag{B-9}$$

by the convergence preservation through summation.

■