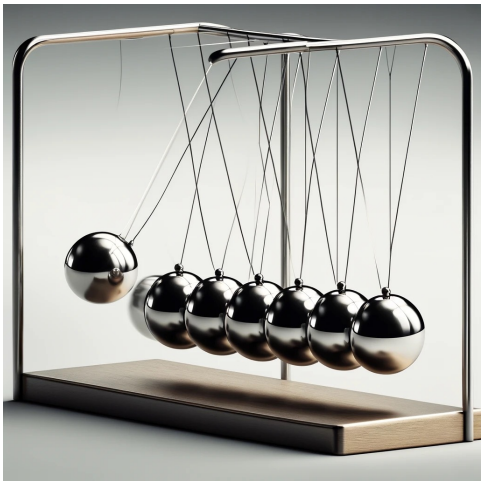
Causal forests and model shift transfer learning to estimate heterogeneous treatment effect

Bérénice-Alexia Jocteur

Institut Camille Jordan
UMR 5208 du CNRS

Causal forests and model shift transfer learning to estimate heterogeneous treatment effect



Bérénice-Alexia Jocteur

Thèse de doctorat

Université Claude Bernard Lyon 1

n°. d'ordre : 2024LYO10119

THESE de DOCTORAT DE L'UNIVERSITE DE LYON

opérée au sein de

Université Claude Bernard Lyon 1

École doctorale InfoMaths (ED 512)
Spécialité : Mathématiques

Thèse soutenue publiquement le 3 juillet 2024 par
Bérénice-Alexia Jocteur

Forêts causales et transfert par changement de modèle pour l'estimation d'effets de traitement hétérogènes

devant le jury composé de :

Antoine Chambaz	Professeur	(Université Paris Cité)	Rapporteur
Josselin Garnier	Professeur	(Ecole Polytechnique)	Rapporteur
Nicolas Bousquet	Professeur associé & Senior Researcher	(Sorbonne Université) & EDF)	Examinateur
Nicolas Brunel	Professeur & Scientific Director	(ENSIIE & Quantmetry)	Examinateur
Gabriela Ciuperca	Professeure	(UCBL)	Examinatrice
Marianne Clausel	Professeure	(Université de Lorraine)	Examinatrice
Véronique Maume-Deschamps	Professeure	(UCBL)	Directrice de thèse
Pierre Ribereau	Maître de conférence	(UCBL)	Co-directeur de thèse
Karim Allouche	Head of Credit and Non-Financial Risks Modelling	(Natixis)	Encadrant
Ibrahima Niang	Credit Quantitative Analyst	(Natixis)	Encadrant

Institut Camille Jordan (ICJ)
Université Claude Bernard Lyon 1
43 boulevard du 11 novembre 1918
F-69622 Villeurbanne Cedex

Natixis
7, promenade Germaine Sablon
75013 Paris
France

Remerciements

En premier lieu je tiens à remercier Véronique Maume-Deschamps et Pierre Ribereau, mes directeurs de thèse, pour le temps qu'ils m'ont accordé durant ces trois années. Les réunions très régulières, vos conseils et vos relectures m'ont été très précieuses.
Cette thèse de doctorat CIFRE a été financée par l'entreprise Natixis. Je remercie Karim Allouche et Yassine Labi pour leur confiance et pour m'avoir permis de travailler dans les meilleures conditions.
Je remercie Ibrahima Niang, mon encadrant en entreprise pour m'avoir transmis ses connaissances sur le risque de crédit en particulier.
Je remercie particulièrement, Antoine Chambaz et Josselin Garnier, de m'avoir fait l'honneur d'être rapporteurs de ma thèse. Je vous adresse toute ma reconnaissance pour le temps consacré à la lecture de mon manuscrit ainsi que pour vos commentaires constructifs ayant contribué à améliorer sa qualité. Je remercie également Nicolas Bousquet, Nicolas Brunel, Gabriela Ciuperca et Marianne Clausel d'avoir accepté d'être membres de mon jury.
Enfin merci à Brayan pour son aide pour l'obtention des données internes, Nadège pour ses réponses à mes questions sur le KECO, Lamyaa pour répondre à mes très nombreuses questions sur le coût du risque et enfin Shane pour ses précieuses relectures.

Forêts causales et transfert par changement de modèle pour l'estimation d'effets de traitement hétérogènes

Résumé : Cette thèse a été réalisée dans le cadre d'un partenariat CIFRE entre l'Université Lyon 1 et Natixis. Elle a pour objectif de développer des méthodes d'apprentissage statistique permettant l'estimation d'effets causaux. Pour ce faire un modèle spécifique de forêt aléatoire a été développé et ses propriétés asymptotiques ont été étudiées. Puis des applications sur des données réelles ont été proposées, notamment sur une quantité d'intérêt pour la direction des risques de Natixis, mais aussi sur une problématique climatique. Enfin une méthode d'apprentissage par transfert sur la forêt précédemment introduite est proposée et des propriétés de convergence ainsi qu'une borne de généralisation sont établies.

Le premier chapitre de cette thèse concerne la construction d'une forêt causale nommée HTERF (Heterogeneous Treatment Effect based Random Forest) qui permet d'estimer la quantité CATE (Conditionnal average treatment effect). Cet estimateur non paramétrique s'inscrit dans la lignée d'autres forêts causales telles que celle introduite par [Athey *et al.* 2019] nommée GRF pour Generalised Random Forest.

La forêt GRF présente des qualités limitées en termes d'interprétabilité et le résultat de consistance et de normalité asymptotique est soumis a un jeu d'hypothèses assez fort. La forêt HTERF qui utilise un critère de partition dédié à l'évaluation des effets causaux permet de pallier ces limitations. D'une part de meilleurs résultats sont obtenus empiriquement sur des simulations que ce soit en termes de qualité de l'estimation de l'effet causal ou en termes d'interprétabilité du modèle. D'autre part un résultat théorique de convergence presque sûre de l'estimateur HTERF est obtenu avec un jeu d'hypothèses plus faibles que pour GRF, un résultat théorique d'interprétabilité est également obtenu.

Une implémentation de HTERF en `Julia` a été créée avec le package `CausalForest`. Une présentation détaillée de ce package est disponible en appendice du Chapitre 2.

Le second chapitre regroupe deux applications de HTERF sur des jeux de données réelles.

Le premier exemple concerne le coût du risque de crédit, une quantité d'intérêt pour la gestion du risque de Natixis. Le backtesting des modèles obtenus est partiellement satisfaisant en termes d'erreur. Cependant les résultats en termes d'interprétabilité sont prometteurs au regard de l'expertise métier.

Le second exemple porte sur le phénomène climatique ENSO (El Niño – Oscillation australe) et plus particulièrement sur l'impact de El Niño sur les précipitations dans l'est australien. Deux stations météorologiques ont été sélectionnées dans des régions différentes d'Australie. Les résultats sont convaincants pour la première station qui met bien en avant l'impact de El Niño et fait ressortir deux variables plus informatives.

Pour la seconde station les données disponibles sont de moins bonne qualité et les résultats obtenus avec HTERF sont moins convaincants.

Le troisième chapitre traite d'apprentissage par transfert dans le cas particulier du *model shift*, lorsque l'on veut estimer un effet causal. La méthode offset introduite par [Wang 2016], propose un algorithme de transfert dans le cadre de la régression, et une borne de généralisation est obtenue.

Nous proposons une adaptation causale de cette méthode offset utilisant l'algorithme HTERF. Un résultat de consistance L^1 est alors obtenu sous des hypothèses, en accord avec les conditions rencontrées en pratique. Une borne de généralisation est également obtenue, elle permet de décomposer cette erreur en un premier terme correspondant à l'erreur propre à HTERF et un second terme correspondant à l'erreur supplémentaire due à la méthode offset.

Des simulations sur des jeux de données synthétiques et semi-synthétiques confirment le bon comportement empirique de cette méthode d'apprentissage par transfert sur les forêts causales.

Mots-clés : forêt causale, inférence causale, effet de traitement hétérogène, résultats contrefactuels, apprentissage par transfert, décalage de modèle, risque de crédit, El Niño

Causal forests and model shift transfer learning to estimate heterogeneous treatment effect

Abstract: This thesis was carried out within the framework of a CIFRE partnership between University Lyon 1 and Natixis. Its aim is to develop statistical learning methods for the estimation of causal effects. To this end, a specific model of random forest was developed, and its asymptotic properties were studied. Subsequent applications on real data were proposed, mainly on data of interest for the risk management department of Natixis, but also on a climate-related issue. Finally, a transfer learning method on the previously introduced forest is proposed, and convergence properties as well as a generalization bound are established. The first chapter of this thesis focuses on the construction of a causal forest named HTERF (Heterogeneous Treatment Effect based Random Forest), which allows the estimation of the CATE (Conditional Average Treatment Effect). This non-parametric estimator follows in the footsteps of other causal forests such as the one introduced by [Athey *et al.* 2019] called GRF for Generalized Random Forest. The GRF forest has limited qualities in terms of interpretability, and the result of consistency and asymptotic normality is subject to a fairly strong set of assumptions. The HTERF forest, which uses a partitioning criterion dedicated to the evaluation of causal effects, addresses these limitations. Better empirical results are obtained in simulations both in terms of the quality of the causal effect estimation and the interpretability of the model. Furthermore, a theoretical result of almost sure convergence of the HTERF estimator is achieved with a set of assumptions that are weaker than those for GRF, and a theoretical result of interpretability is also obtained. An implementation of HTERF in `Julia` has been created using the `CausalForest` package. A detailed presentation of this package is available in the appendix of Chapter 2. The second chapter includes two applications of HTERF on real data sets. The first example concerns the cost of credit risk, an estimate of interest for Natixis's risk management. The backtesting of the obtained models is partially satisfactory in terms of error. However, the results in terms of interpretability are promising from a business expertise standpoint. The second example focuses on the climate phenomenon ENSO (El Niño-Southern Oscillation), specifically on the impact of El Niño on rainfall in Eastern Australia. Two weather stations were selected in different regions of Australia. The results are convincing for the first station, which clearly highlights the impact of El Niño and identifies two particularly informative variables. For the second station, the available data are of lower quality and the results obtained with HTERF are less convincing. The third chapter deals with transfer learning in the specific case of model shift, with the aim of estimating a causal effect. The offset method introduced by [Wang 2016] proposes a transfer algorithm in the context of regression, and a generalization bound is obtained. We propose a causal adaptation

of this offset method using the HTERF algorithm. An L1 consistency result is then obtained under assumptions that align with conditions encountered in practice. A generalization bound is also obtained, which allows for the decomposition of this error into a first term corresponding to the intrinsic error of HTERF and a second term corresponding to the additional error due to the offset method. Simulations on synthetic and semi-synthetic datasets confirm the appropriate empirical behaviour of this transfer learning method on causal forests.

Keywords: causal forest, causal inference, heterogeneous treatment effect, potential outcomes, transfer learning, model shift, credit risk, El Niño

Contents

List of Acronyms

Chapter 1

Introduction

Les travaux présentés dans cette thèse ont été réalisés dans le cadre d'une thèse CIFRE entre la banque Natixis et l'Institut Camille Jordan. L'objectif est de développer des algorithmes d'apprentissage statistique pour évaluer des effets causaux tout en mettant l'accent sur l'interprétabilité de ces modèles. Cette étude est motivée par des applications en risque de crédit.

Nous présentons dans cette introduction le contexte, la notion de causalité ainsi que les différentes méthodes statistiques pour évaluer les effets causaux, puis nous introduisons l'apprentissage par transfert, enfin nous mettons en avant les différentes contributions de cette thèse.

Contents

1.1 Risque de crédit

La réglementation prudentielle vise à garantir la solidité des institutions bancaires, ce qui est crucial pour assurer la stabilité de l'économie, en particulier la sécurité des économies des individus. Cette réglementation établit des normes pour les banques en ce qui concerne leurs fonds propres, la gestion des risques, la diversification des actifs et la communication financière, tout en définissant les mécanismes de supervision.

Le Comité de Bâle joue un rôle primordial dans l'élaboration des règles prudentielles pour les banques. Il est composé de 28 membres représentant les autorités de contrôle nationales des principales puissances économiques mondiales. Ce comité émet des normes et des lignes directrices visant à renforcer la stabilité du système financier dans son ensemble.

Par la suite, l'Union européenne transpose les normes du Comité de Bâle en droit européen à travers des règlements et des directives, applicables à tous les États membres. Dans la zone euro, la Banque centrale européenne (BCE) est chargée d'organiser la supervision des institutions bancaires, en collaboration avec les autorités nationales responsables du contrôle des établissements de crédit, comme l'Autorité de Contrôle Prudentiel et de Résolution (ACPR) en France.

1.1.1 Aperçu historique

En 1974, les gouverneurs des banques centrales du G10 établissent le Comité de Bâle dans le but de renforcer la stabilité des relations bancaires, notamment en harmonisant les réglementations nationales.
En 1988, le Comité de Bâle publie ses premières recommandations, connues sous le nom de Bâle 1, concernant les principes réglementaires d'un ratio de solvabilité : le ratio de Cooke. Cette publication servira de fondement à toutes les réformes bancaires ultérieures. Elle exige des banques qu'elles détiennent un montant de fonds propres proportionnel à leur exposition au risque, selon la relation suivante :

$$\frac{\text{Fonds propres réglementaires}}{\text{Risque de crédit} + \text{Risque de marché}} \geq 8\% \tag{1.1.1}$$

En 2004, le Comité de Bâle présente de nouvelles recommandations, appelées Bâle 2, qui introduisent une définition plus précise du risque de crédit. Ces recommandations prennent désormais en compte la qualité de l'emprunteur à travers un système de notation interne propre à chaque établissement (Internal Rating-Based (IRB)). De plus, un nouveau ratio de solvabilité, le ratio de McDonough, est instauré en conséquence.

$$\frac{\text{Fonds propres réglementaires}}{\text{Risque de crédit} + \text{Risque de marché} + \text{Risque opérationnel}} \geq 8\% \tag{1.1.2}$$

Cette réforme vise principalement à mettre en place un calcul des fonds propres réglementaires qui reflète au mieux le profil de risque de l'établissement bancaire. Pour ce faire, les accords de Bâle II proposent une approche plus nuancée du risque

de crédit et intègrent également le risque opérationnel. Ils encouragent également les établissements à développer des dispositifs de contrôle des risques internes afin de répondre à l'ensemble des critères définis dans les accords de Bâle.
Ce dispositif prudentiel renforcé renforcera également les exigences en matière de communication d'informations aux directions générales, aux autorités de supervision et au marché. Plus qu'une simple norme de solvabilité, Bâle II établit un cadre solide visant à maintenir le niveau global actuel des fonds propres du système bancaire.
Enfin les accords de Bâle III visent à répondre aux limitations de Bale II rencontrées lors de la crise de 2007/2008. Les exigences sont renforcées sur les fonds propres et de nouvelles normes sont introduites :

- La quantité de fonds propres est renforcé pour les risques de marché et pour les banques systémiques.
- Introduction d'un ratio de levier, qui limite la quantité de dettes qu'une banque peut avoir par rapport à ses fonds propres.
- Introduction de deux nouveaux ratios de liquidité : le ratio de liquidité à court terme (LCR) et le ratio de financement stable net (NSFR). Le LCR exige que les banques maintiennent un niveau suffisant de liquidités de haute qualité pour faire face à des périodes de stress à court terme, tandis que le NSFR vise à garantir un financement stable à long terme.
- Des coussins (fonds propres supplémentaires) sont introduits en haut de cycle pour freiner la croissance excessive du crédit.

1.1.2 Typologie des risques

Les différents risques présents dans la banque peuvent être classés suivant la Figure 1.1, selon le document Pilier III de 2023 du groupe BPCE [BPCE 2024].

Seuls les risques de crédit, de marché et opérationnels apparaissent dans les formules susmentionnées, nous nous focaliserons par la suite sur le risque de crédit et nous verrons rapidement comment les autres risques sont pris en compte dans la réglementation bâloise.

1.1.3 Organisation en trois piliers

Le cadre réglementaire de Bâle énonce ses exigences selon trois piliers distincts :

- Pilier I - Méthodes d'évaluation des fonds propres et exigences de qualité : Ce premier pilier établit les méthodes pour évaluer les fonds propres réglementaires en couvrant les encours pondérés du risque de crédit, de marché et opérationnel. Il définit des exigences qualitatives que les banques doivent mettre en œuvre en fonction de la méthode cible envisagée.

Macro-familles de risques	Définitions
Risques de crédit et de contrepartie	
• Risques de crédit	Risque de pertes résultants de l'incapacité des clients, d'émetteurs ou d'autres contreparties à faire face à leurs engagements financiers. Il Inclut le risque de contrepartie afférant aux opérations de marché (risque de remplacement) et aux activités de titrisation. Il peut être aggravé par le risque de concentration.
• Risques de titrisation	Opérations pour lesquelles le risque de crédit inhérent à un ensemble d'expositions est logé dans une structure dédiée (en général un fonds commun de créances ou « conduit ») puis divisé en tranches en vue le plus souvent de leur acquisition par des investisseurs.
Risques financiers	
• Risque de marché	Risque de perte de valeur d'instruments financiers résultants des variations de paramètres de marché, de la volatilité de ces paramètres et des corrélations entre ces paramètres. Les paramètres concernés sont notamment les taux de change, les taux d'intérêt ainsi que les prix des titres (actions, obligations) et des matières premières, des dérivés et de tout autre actif tels que les actifs immobiliers.
• Risque de liquidité	Risque que le groupe ne puisse faire face à ses besoins de trésorerie ou à ses besoins de collatéral au moment où ils sont dus et à un coût raisonnable.
• Risque structurel de taux d'intérêt	Risques de pertes de marge d'intérêt ou de valeur de la position structurelle à taux fixe en cas de variation sur les taux d'intérêt. Les risques structurels de taux d'intérêt sont liés aux activités commerciales et aux opérations de gestion propre
• Risque de spread de crédit	Risque lié à la dégradation de la qualité de la signature d'un émetteur particulier ou d'une catégorie particulière d'émetteurs.
• Risque de change	Risque de pertes de marge d'intérêt ou de valeur de la position structurelle à taux fixe en cas de variation sur le taux d'intérêt de change. Les risques structurels de taux et de change sont liés aux activités commerciales et aux opérations de gestion propre.
Risques non-financiers	
• Risque de non-conformité	Risque de sanction judiciaire, administrative ou disciplinaire, de perte financière significative ou d'atteinte à la réputation, qui naît du non-respect de dispositions propres aux activités bancaires financières, qu'elles soient de nature législative ou règlementaire, nationales ou européennes directement applicables, ou qu'il s'agisse de normes professionnelles et déontologiques, ou d'instructions des dirigeants effectifs prises notamment en application des orientations de l'organe de surveillance.
• Risque opérationnel	Risque de pertes découlant d'une inadéquation ou d'une défaillance des processus, du personnel et des systèmes internes ou d'événements extérieurs, y compris le risque juridique. Le risque opérationnel inclut notamment les risques liés à des événements de faible probabilité d'occurrence mais à fort impact, les risques de fraude interne et externe définis par la réglementation, et les risques liés au modèle.
• Risques de souscription d'assurance	Risque, au-delà de la gestion des risques actifs/passifs (risques de taux, de valorisation, de contrepartie et de change, de tarification des primes du risque de mortalité et des risques structurels liés aux activités d'assurance vie et dommage y compris les pandémies, les accidents et les catastrophes (séismes, ouragans, catastrophes industrielles, actes de terrorismes et conflits militaires).
• Risque de modèle	Risque de modèle est défini comme le risque de conséquences défavorables – perte financière et/ou éventuelle atteinte à la réputation du Groupe – résultant de décisions basées sur des modèles dues à des erreurs dans la conception, la mise en œuvre ou l'utilisation de ces modèles.
• Risque juridique	Risque juridique défini dans la réglementation française comme le risque de tout litige avec une contrepartie, résultant de toute imprécision, lacune ou insuffisance susceptible d'être imputable à l'entreprise au titre de ses opérations.
• Risque de réputation	Risque de réputation est défini comme le risque d'atteinte à la confiance que portent à l'entreprise, ses clients, ses contreparties, ses fournisseurs, ses collaborateurs, ses actionnaires ou tout autre tiers dont la confiance, à quelque titre que ce soit, est une condition nécessaire à la poursuite normale de l'activité.
Risques stratégiques d'activité et d'écosystème	
• Risque de solvabilité	Risque d'incapacité de la société à faire face à ses engagements à long terme et/ou à assurer la continuité des activités ordinaires dans le futur.
• Risque climatique et environnemental	Vulnérabilité directe ou indirecte (*i.e.* via les actifs/passifs détenus) des activités bancaires aux risques liés au climat et à l'environnement, incluant les risques physiques (aléas climatiques, pollution, perte de biodiversité, etc.) et les risques liés à la transition (réglementaire, technologique, attente des clients).

Figure 1.1: Définitions des différents types de risques, selon le document Pilier III du groupe BPCE, [BPCE 2024].

- Pilier II - Processus de contrôle renforcé : Ce deuxième pilier détermine le cadre de gestion des risques non couverts par le Pilier I, tels que le risque de taux sur le portefeuille bancaire, le risque de concentration, et le risque de liquidité. Dans le cadre du Pilier II, les banques doivent calculer des fonds propres économiques pour couvrir ces risques. Le capital économique est soumis à des exigences de notation interne qui peuvent être supérieures aux exigences réglementaires. Ce pilier renforce également la supervision du régulateur.

- Pilier III - Communication financière : Le dernier pilier impose aux banques une plus grande transparence sur leurs risques grâce à un renforcement de la communication financière. Son objectif est de mettre en place une discipline de marché pour éviter d'éventuelles asymétries d'information sur les risques encourus par les banques.

Pilier I

Le régulateur propose différentes méthodes de calcul pour chaque composante au dénominateur du ratio de de McDonough :

Pilier I		
Risque de crédit	**Risque de marché**	**Risque opérationnel**
Approche Standard	Approche standard	Indicateurs de base
IRBA Fondation	Modèles internes	Approche standard
IRBA Avancée		Modèles avancés

Table 1.1: Méthodes de calculs du pilier I.

Le risque de crédit est la possible perte financière qu'encourt une institution bancaire en cas d'incapacité de ses débiteurs de s'acquitter de leurs obligations contractuelles. Trois paramètres réglementaires doivent êtres estimés :

- La probabilité de défaut (PD - "Probability of Default"),

- la perte en cas de défaut (LGD - "Loss Given Default"),

- l'exposition au moment du défaut (EAD - "Exposure At Default").

Le régulateur propose trois méthodes pour calculer ces paramètres:

- Méthode standard : Les actifs sont classés en différentes catégories en fonction de critères tels que la qualité de crédit de l'emprunteur (si notation externe disponible) et la nature de l'actif. Chaque catégorie se voit attribuer une pondération de risque fixe, déterminée par les autorités de régulation. Les pondérations de risque sont appliquées à l'exposition brute des actifs de crédit pour calculer le montant des fonds propres requis. La formule générale pour

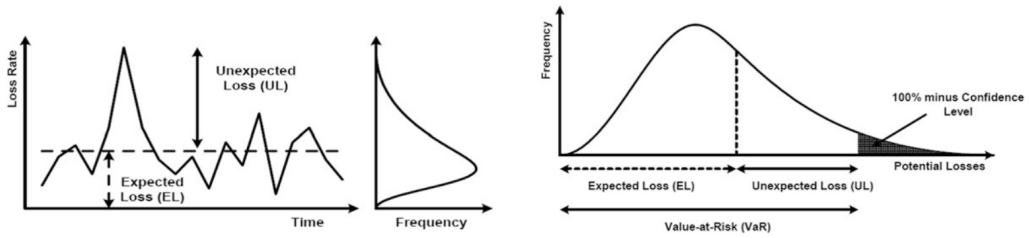
calculer les exigences en fonds propres (Capital Requirement - CR) est la suivante :

$$CR = EAD \times 8\% \times \text{Pondération de risque} \tag{1.1.3}$$

- La méthode IRB-F repose sur l'estimation interne de la Probabilité de Défaut (PD) à l'aide de modèles internes. La Perte en Cas de Défaut (LGD) est fixée forfaitairement à 45%, quelle que soit la contrepartie, tandis que la maturité est prise en compte conformément aux recommandations réglementaires, en se basant sur la maturité résiduelle du contrat en vie.
- La méthode IRB-A estime les paramètres PD, LGD et EAD par des modèles internes et la maturité M est traitée comme précédemment.

Pour déterminer les exigences en fonds propres liées au risque de crédit, il est nécessaire d'estimer le niveau de pertes potentielles sur les portefeuilles de crédit de la banque. Ces pertes sont définies selon deux concepts distincts :

- Pertes Attendues (Expected Loss - EL) : Elles représentent la perte moyenne anticipée sur une exposition donnée. Les banques utilisant les approches IRB doivent fournir des mesures quantitatives des pertes attendues sur leurs expositions. La formule de calcul est la suivante : $EL = PD \times LGD \times EAD$. Ces pertes attendues sont couvertes par des provisions.
- Pertes Inattendues (Unexpected Loss - UL) : Elles correspondent à l'écart-type autour de la moyenne. Dans le cadre des réformes du Comité de Bâle, les pertes inattendues sont calculées à l'aide de formules de pondérations différenciées selon les types d'exposition. La formule de calcul est : $UL = f(PD, LGD, M) \times EAD$, où la fonction f est définie par les régulateurs. Les pertes inattendues sont destinées à être couvertes par les fonds propres.



(a) Exemple de profil de perte d'une banque. (b) Mise en évidence de EL et de UL.

Figure 1.2: Source: BCBS 2005a

Dans le cadre réglementaire de Bâle (*Basel framework*), dans la partie *Calculation of RWA for credit risk*, dans le chapitre 30 au paragraphe 4, des portefeuilles sont définis de la manière suivante : Under the IRB approach, banks must categorise banking-book exposures into broad classes of assets with different underlying risk

characteristics, subject to the definitions set out below. The classes of assets are (a) corporate, (b) sovereign, (c) bank, (d) retail, and (e) equity. Within the corporate asset class, five sub-classes of specialised lending are separately identified. Within the retail asset class, three sub-classes are separately identified. Within the corporate and retail asset classes, a distinct treatment for purchased receivables may also apply provided certain conditions are met. For the equity asset class the IRB approach is not permitted, as outlined further below.

Dans le cadre de l'approche IRB, les paramètres devant être estimés le sont indépendamment sur chacun de ces portefeuilles.

Passons au calcul de UL et du risque de crédit pondéré par les encours (RWA – risk weighted average). Le niveau de fonds propres devrait être suffisant pour couvrir la différence entre les pertes inattendues (UL) et les pertes totales, qui incluent les pertes attendues (EL), étant donné que les pertes attendues sont déjà couvertes par les provisions, comme précédemment mentionné. Les exigences de fonds propres sont alors :

$$CR = f(PD, LGD) \times EAD - PD \times LGD \times EAD. \tag{1.1.4}$$

La fonction f est déterminée selon un modèle dit de Merton. La modélisation dans le cadre de la réglementation bâloise de cette fonction dépend d'une part du risque systémique (c'est à dire le risque porté par l'environnement économique général) et d'autre part le risque idiosyncratique (c'est à dire le risque spécifique de l'actif considéré). La formule suivante est obtenue pour les exigences de fonds propres exprimés en terme de pourcentage de l'EAD :

$$CR = \left(LGD \times \underbrace{N\left(\frac{G(PD)}{\sqrt{1-R}} + \sqrt{\frac{R}{1-R}} G(0.999) \right)}_{\text{Modèle de Merton}} - \underbrace{PD \times LGD}_{\text{EL}} \right) \times \underbrace{\frac{1 + (M - 2.5) b(PD)}{1 - 1.5 b(PD)}}_{\text{Ajustement sur la maturité}}, \tag{1.1.5}$$

où N est la distribution normale standard, G est l'inverse de cette distribution, R est la corrélation des actifs, elle est calculés par classes d'actifs et enfin $b(PD)$ est un ajustement lissé de la maturité M obtenu par régression contre PD : $b(PD) = (0.11852 - 0.05478 \times \log(PD))^2$. On retrouve le quantile 99.9% de l'UL qui apparaît dans la formule de Merton.

Le RWA (Risk-Weighted Assets) pour le risque de crédit représente la totalité des pertes associées au risque de crédit, ajustées en fonction des montants exposés. En d'autres termes, il s'agit de la pondération des pertes liées au risque de crédit en fonction des encours, et s'exprime comme suit :

$$RWA = 12.5 \times CR \times EAD. \tag{1.1.6}$$

12,5 l'inverse de la limite du ratio de McDonough fixé à 8%.

Le ratio de McDonough inclut aussi les risques de marché et opérationnels pour lesquels des RWA spécifiques sont calculées avec d'autres méthodes dédiées.

Pilier II

Le deuxième pilier de la réglementation établie par le Comité de Bâle vise à une identification et une mesure exhaustive des risques du premier pilier (crédit, marché et opérationnel), ainsi que des autres risques bancaires (taux, liquidité, concentration, réputation, stratégique et valeurs résiduelles), voir la Figure 1.3 pour une illustration. Il encourage également les banques à utiliser des indicateurs de risque, par exemple en fixant des seuils. En effet, les banques doivent non seulement être capables de mesurer leurs risques, mais aussi de suivre l'évolution de ces paramètres au fil du temps et d'identifier toute éventuelle dégradation du profil de risque.

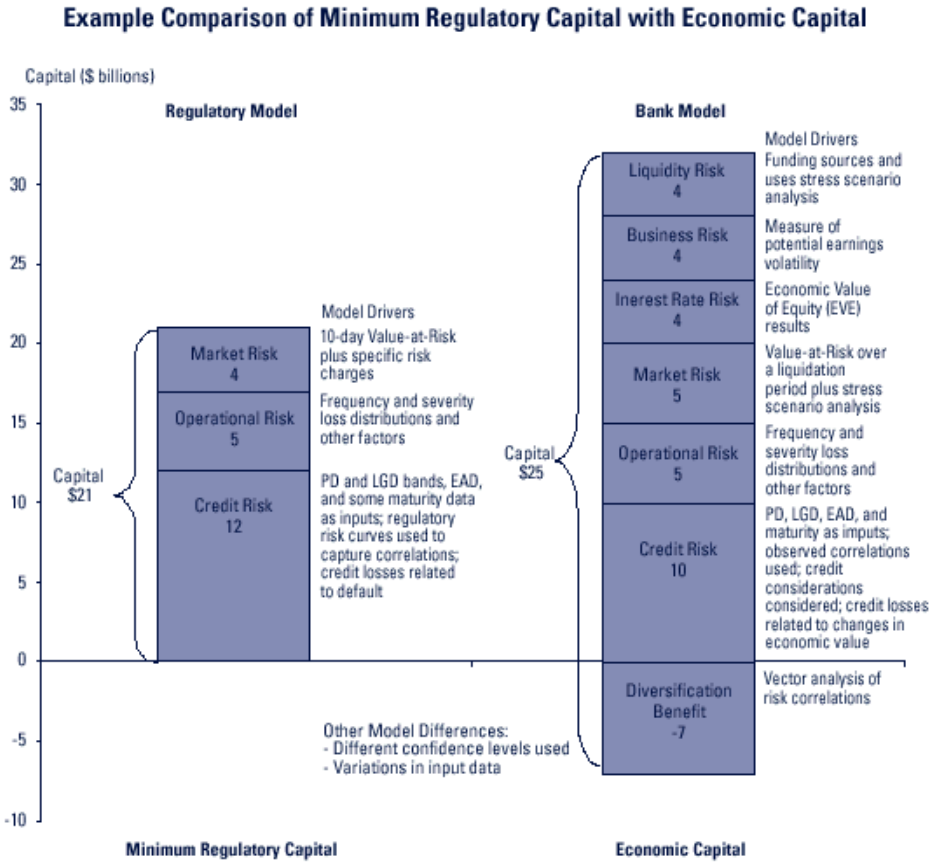


Figure 1.3: Comparaison du capital réglementaire et du capital économique. Image de la FDIC (Federal Deposit Insurance Corporation).

De plus, ce pilier requiert la mise en place de stress tests pour chaque type de risque, permettant notamment d'estimer l'impact potentiel d'une crise ou d'une situation imprévue sur les fonds propres de la banque. Enfin, il encourage les banques à établir un dispositif d'auto-évaluation du capital nécessaire pour couvrir tous les risques (capital économique). Le pilier II ne se limite pas au calcul du

capital économique, mais englobe également le contrôle des méthodologies utilisées dans la gestion des différents risques et la gouvernance globale de la gestion des risques (information de la direction, audit, etc.). Son objectif est de stimuler les établissements à développer leurs techniques de contrôle et de gestion des risques et des fonds propres (ICAAP - Internal Capital Adequacy Assessment Process), et de permettre aux autorités de s'assurer que les banques disposent d'un niveau de fonds propres conforme à leur profil de risque, en demandant éventuellement des mesures correctrices (régies par le processus de revue prudentielle, ou SREP - Supervisory Review Process).

Capital économique

Les évaluations du capital réglementaire et du capital économique visent à déterminer le montant de fonds propres nécessaire pour couvrir les pertes exceptionnelles liées aux activités de la banque pour une probabilité de défaut et à un horizon donnés. Toutefois, il convient de faire une distinction claire entre ces deux concepts. Le capital réglementaire vise à garantir la solvabilité des banques et la pérennité du système financier international dans un scénario de crise standard avec une exigence de rating moyen BBB+. Il est défini par les accords de Bâle (Pilier I). En revanche, le capital économique correspond au montant de fonds propres économiques que l'établissement estime nécessaire pour couvrir ses risques selon son rating cible (voir Figure 1.4). Ainsi, le régulateur évalue la qualité globale du dispositif de gestion des risques de l'établissement et approuve l'ensemble des capitaux réglementaires (calculés selon les méthodes et exigences réglementaires) et des capitaux économiques (calculés par des méthodes internes spécifiques à chaque banque). L'utilisation de stress tests et de scénarios doit donc permettre d'assurer la pertinence et l'efficacité du processus interne d'évaluation du capital économique. Les fonds propres économiques sont ainsi alignés sur les fonds propres réglementaires. Selon le profil de risque de chaque banque, le régulateur peut demander un ajustement de ces derniers.

Pilier III

L'objectif principal de ce troisième pilier est de promouvoir la transparence et la discipline sur le marché financier. Son but est d'encourager les banques à communiquer leurs informations financières de manière transparente, cohérente et appropriée. Il est ainsi recommandé aux établissements de vérifier la pertinence des informations divulguées en fonction du principe de matérialité (c'est-à-dire leur degré de significativité) et de confidentialité (afin de préserver les avantages compétitifs, par exemple), ainsi que de garantir la cohérence entre les informations comptables et financières diffusées.

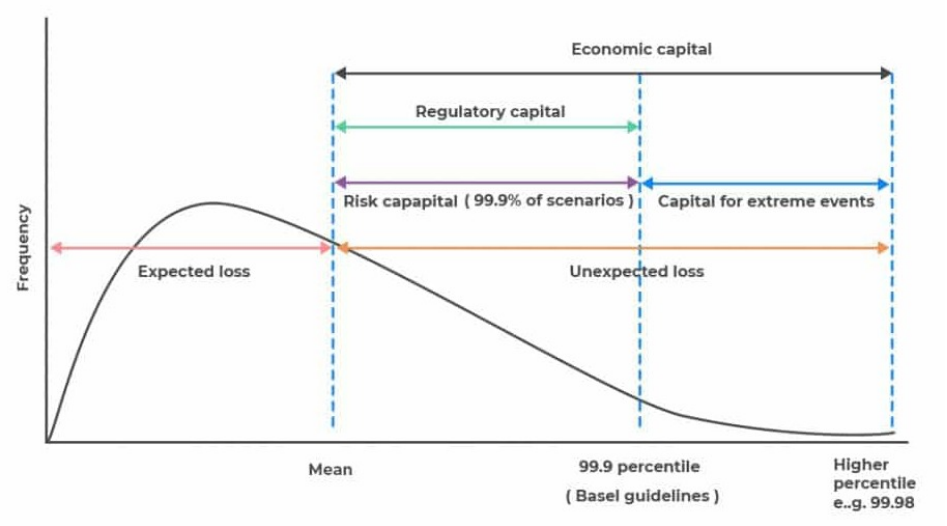


Figure 1.4: Intervalles de confiance utilisés en approche interne pour le calcul du risque de crédit

1.2 Causalité

L'objectif est de présenter la causalité dans un premier temps à partir des premières considérations historiques, puis plus formellement dans un contexte mathématique [Drouet 2012].

Les différentes théories de la causalité peuvent toutes être considérées comme des descendantes de l'analyse de la causalité proposée par Hume. Dans son ouvrage *Traité de la nature humaine* [Hume 1995] il caractérise la causalité ainsi :

- Contiguïté des causes et effets : « En premier lieu, je constate que tous les objets que l'on considère comme causes ou comme effets sont contigus ».
- Antériorité temporelle des causes par rapport aux effets : « La seconde relation dont j'observerai qu'elle est essentielle aux causes et aux effets [...] est celle d'antériorité temporelle de la cause par rapport à l'effet ».
- Régularité de l'enchaînement cause effet : « Des objets semblables ont toujours été placés dans des relations semblables de contiguïté et de succession ».

Nous pouvons alors distinguer deux grandes familles de conceptions de la causalité [Drouet 2012] :

- Des approches physiques qui donnent une interprétation en terme des phénomènes physiques des liens de cause à effet.
- Des approches inférentielles où l'on cherche à quantifier les différences des causes relativement à leurs effets (difference-makers).

Certains auteurs développent des théories « physiques » de la causalité, visant à la définir en référence aux caractéristiques physiques des relations de cause à effet. Ils proposent de penser la causalité soit en termes de processus qui transmettent de l'énergie, transmettent une quantité de mouvement ou plus généralement transmettent ou manifestent une grandeur physique conservée.

Nous nous concentrerons sur la conception inférentielle de la causalité en présentant en détail ses différents courants.

1.2.1 Théorie probabiliste

Les théorie reposant sur l'approche probabiliste partent du principe suivant :

$$A \text{ cause } B \text{ si et seulement si } \mathbb{P}(B|A) > \mathbb{P}(B).$$

Or si cette propriété est vérifiée pour A et B cela implique également $\mathbb{P}(A|B) > \mathbb{P}(A)$. On a alors, que toute cause qui augmente la probabilité de ses effets voit sa propre probabilité augmentée par chacun de ces effets. Or en général les relations de causalité ne sont pas symétriques, des définitions plus poussées ont ainsi été introduites.

[Suppes 1970], introduit la théorie probabiliste suivante de la causalité, A cause B si et seulement si :

- $\mathbb{P}(B|A) > \mathbb{P}(B)$.
- A est antérieure à B.
- Il n'existe pas de propriété C antérieure à A telle que $\mathbb{P}(B|A \wedge C) = \mathbb{P}(B|C)$.

Cependant cette théorie est à nouveau incomplète, et des relations peuvent être identifiées comme causales alors qu'elles ne le sont pas. Le paradoxe le plus connu qui illustre cela est le paradoxe de Simpson. Le paradoxe de Simpson survient lorsque la corrélation observée est positive (ou négative) dans l'ensemble de la population, mais devient négative (ou positive) dans chacune des sous-populations de cette population. Une illustration de ce paradoxe apparaît dans l'étude clinique [Charig *et al.* 1986], où deux traitements sont considérés, on compare leurs performances respectives grâce à leur taux de réussite, deux groupes de patients sont également considérés ceux avec des petits calculs rénaux et ceux avec de gros calculs rénaux. Les résultats obtenus sont résumés dans la Table 1.2.

La théorie de Suppes amène à conclure en comparant les taux de réussites sur la population totale, que le traitement 2 cause une meilleure guérison des patients. Cependant en regardant dans le détail les deux groupes, on voit que le taux de réussite du traitement 1 est plus faible car il est plus proposé aux patients avec des gros calculs qui sont plus difficiles à traiter.

Pour pallier à ces limitations d'autres théories ont été introduites ([Cartwright 2012], [Skyrms 1980]). Cependant toutes ces méthodes sont conceptuelles et n'ont pas été conçues pour faire de l'inférence causale en calculant des effets causaux par exemple. La théorie contrefactuelle permet justement l'inférence causale.

Traitement / Taille des calculs	Traitement 1	Traitement 2
Petits calculs rénaux	93% (81/87)	87% (234/270)
Gros calculs rénaux	73% (192/263)	69% (55/80)
Toutes tailles confondues	**78% (273/350)**	**83% (289/350)**

Table 1.2: Illustration du paradoxe de Simpson, les pourcentages représentent le taux de patient guéris (nombre patients guéris / nombre de patients ayant reçus le traitement).

1.2.2 Théorie contrefactuelle

L'idée de base de la contrefactualité est: "Que serait-il arrivé si, contrairement aux faits, nous avions fait quelque chose d'autre que ce que nous avons effectivement fait ?" Par exemple que serait-il arrivé si un traitement A avait été administré à un individu donné au lieu du traitement B. [Lewis 1973] : "If c and e are two actual events such that e would not have occurred without c, then c is a cause of e."

Soit une comparaison de deux traitements, pour chaque individu nous avons deux résultats contrefactuels (ou "potential outcomes", [Rubin 1974]). Soit Y l'état de santé et A une indicatrice de traitement binaire:

- $Y(1)$: état de santé de l'individu si le traitement $A = 1$ est administré
- $Y(0)$: état de santé de l'individu si le traitement $A = 0$ est administré

Si l'individu a reçu le traitement 1 c'est $Y(1)$ qui est observé. Si il a reçu le traitement 0 c'est $Y(0)$ qui est observé. Dans aucun cas les deux résultats contrefactuels peuvent être observé concomitamment pour un individu donné.

On peut alors évaluer $\mathbb{E}[Y(1)] - \mathbb{E}[Y(0)]$ pour voir si le traitement 1 est meilleur que le traitement 0.

La situation idéale permettant d'estimer cet effet causal serait le cas où les individus qui ont reçu le traitement 1 et ceux qui ont reçu le traitement 0 sont comparables au niveau de leurs résultats contrefactuels. Si tel était le cas les résultats des individus qui ont reçu le traitement 1 seraient similaires aux résultats si la population totale avait reçu le traitement 1, et de même les résultats des individus qui ont reçu le traitement 0 seraient similaires aux résultats si la population total avait reçu le traitement 0. Cependant cette situation n'est que rarement vérifiée en pratique. Un cadre plus large pour pouvoir estimer les effets causaux est alors proposé.

Même si les groupes qui ont reçu le traitement 1 et ceux qui ont reçu le traitement 0 ne sont pas comparables, il est possible qu'à l'intérieur de strates d'autres variables (appelons les C), ceux qui ont reçu le traitement 1 et ceux qui ont reçu le traitement 0 soient comparables. Si c'est le cas, alors les moyennes à l'intérieur de ces strates de C refléteront les résultats contrefactuels moyens pour les strates.

Notons $X \perp\!\!\!\perp Y|Z$ pour signifier que X est indépendant de Y conditionnellement à Z. L'hypothèse d'ignorabilité (unconfoundedness en anglais pour absence de confusion) permet de formaliser ces conditions permettant l'inférence causale. L'effet du traitement A sur le résultat Y est sans confusion pour des covariables C données si pour toutes les valeurs possibles de a : $Y(a) \perp\!\!\!\perp A|C$, id est au sein des strates des variables confondantes, les groupes de traitements sont comparables (elles ont des "potential outcomes" similaires) et nous pouvons alors en tirer des conclusions causales. En effet :
$\mathbb{E}[Y(1)|C = c] = \mathbb{E}[Y(1)|A = 1, C = c] = \mathbb{E}[Y|A = 1, C = c]$
$\mathbb{E}[Y(0)|C = c] = \mathbb{E}[Y(0)|A = 0, C = c] = \mathbb{E}[Y|A = 0, C = c]$
Nous pouvons alors calculer des effets causaux à partir de la donnée observée :
$\mathbb{E}[Y(1)|C = c] - \mathbb{E}[Y(0)|C = c] = \mathbb{E}[Y|A = 1, C = c] - \mathbb{E}[Y|A = 0, C = c]$

Cette définition de la causalité sera celle utilisée dans les travaux qui suivent.

1.2.3 Théorie interventionniste

La théorie interventionniste est introduite telle que Woodward la présente dans [Woodward 2003]. Une cause est définie de la façon suivante : "A necessary and sufficient condition for X to be a direct cause of Y with respect to some variable set V is that there be a possible intervention on X that will change Y (or the probability distribution of Y) when all other variables in V besides X and Y are held fixed at some value by interventions."

L'un des concepts clés de la causalité de Woodward est la notion de graphiques causaux, qui représentent les relations causales entre les variables dans un système. Ces graphiques peuvent être utilisés pour visualiser la structure des mécanismes causaux et pour identifier les voies par lesquelles la causalité opère. Il existe plusieurs types d'interventions selon [Woodward 2003], une illustration de l'intérêt de la notion d'intervention est présentée avec l'exemple d'un baromètre.

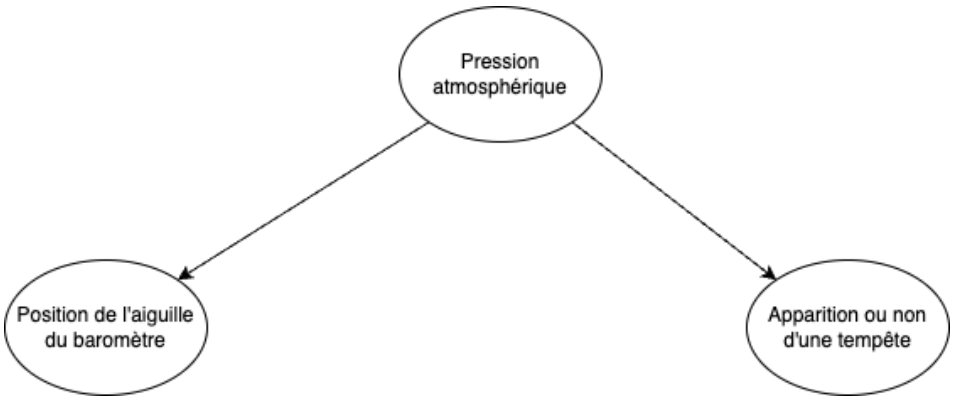


Figure 1.5: Structure causale de l'exemple du baromètre.

Par la simple observation on peut identifier l'association entre position de l'aiguille du baromètre et l'apparition ou non de tempêtes. On pourrait alors être tenté d'établir

un lien de causalité : la position de l'aiguille causerait l'apparition du phénomène météorologique. Une intervention serait de déplacer manuellement l'aiguille du baromètre, même si l'on réplique cette expérience de multiples fois, une tempête ne serait pas déclenchée de ce fait, la position de l'aiguille n'est pas la cause des tempêtes. En effet une cause commune a ces deux événements est la pressions atmosphérique, une intervention sur celle-ci altérerait bien à la fois la position de l'aiguille du baromètre et l'apparition de tempête.

[Pearl 2009], propose également une approche basée sur les intervention pour évaluer les effets causaux, dans ce cas des réseaux bayésiens sont utilisés. Le "do-calculus" est introduit, il permet d'analyser les effets des interventions sur les variables dans un système complexe. Nous n'étudierons pas cette approche dans la suite.

1.3 Méthodes d'inférence causale existantes

Soit (Y_i, X_i, W_i) i.i.d. $\sim \mathcal{P}$, où $Y \in \mathcal{Y}$ est un résultat continu ou binaire d'intérêt, $X \in \mathcal{X} \subset \mathbb{R}^d$ est un vecteur de covariables d-dimensionnelles des possibles confondants et $W_i \in \{0, 1\}$ est une variable de traitement binaire, qui est attribuée selon le score de propension $\pi(x) = P(W = 1|X = x)$, avec une probabilité marginale d'attribution de traitement $p_\pi = P(W = 1)$. Un échantillon $\mathcal{D}_n = \{(Y_i, X_i, W_i)\}_{i=1}^n$ est observé. En utilisant le cadre des résultats potentiels de Neyman-Rubin ([Imbens & Rubin 2015]), notre intérêt principal réside dans l'effet de traitement individualisé : la différence entre les résultats potentiels $Y_i(0)$ si l'individu i ne reçoit pas de traitement $(W_i = 0)$ et $Y_i(1)$ si le traitement est administré $(W_i = 1)$. Cependant, en raison du problème fondamental de l'inférence causale, seulement l'un des résultats potentiels est observé, puisque $Y_i = W_i Y_i(1) + (1 - W_i)Y_i(0)$. Par conséquent, conformément à la majorité de la littérature existante, nous nous concentrons sur l'estimation de l'effet moyen du traitement conditionnel (Conditional Average Treatment Effect (CATE)), $\tau(x) = \mathbb{E}_P[Y(1) - Y(0)|X = x]$, l'effet moyen du traitement attendu pour un individu avec des valeurs de covariables $X = x$.

Il est bien connu que l'identification des effets causaux à partir de données observationnelles repose sur l'imposition d'hypothèses non testables. Ici, nous considérons l'estimation sous les hypothèses standards (consistance, absence de confondants et recouvrement)

- Consistance : Si l'individu i est assigné au traitement w_i, nous observons le résultat potentiel associé $Y_i = Y_i(w_i)$.
- Absence de confondants : il n'y a pas de confondants non observés, tels que $Y(0), Y(1) \perp\!\!\!\perp W|X$.
- Recouvrement : l'attribution du traitement est non déterministe, c'est-à-dire $0 < \pi(x) < 1, \forall x \in X$.

Sous ce jeu d'hypothèses, le CATE peut être écrit comme $\tau(x) = \mu_1(x) - \mu_0(x)$

pour $\mu_w(x) = \mathbb{E}_P[Y|W = w, X = x]$, ce sont les hypothèses standards pour permettre l'identifiabilité de CATE.

1.3.1 Metalearners

Les metalearners sont des estimateurs de CATE, qui combinent des méthodes généralistes d'apprentissage statistique appelées apprenants de base d'une manière spécifique tout en autorisant une liberté totale sur le choix de ces apprenants de base (*base learners* en anglais). Le choix de ces apprenants de base peut ainsi être motivé par une connaissance à priori de la nature de la donnée étudiée. Ces méthodes sont présentées plus ou moins exhaustivement dans [Künzel *et al.* 2019] et [Jacob 2021]. Dans la suite de cette partie sont présentées deux approches les plus simples à savoir T et S-learners, puis le DR-learner est abordé vu comme une amélioration du T-learner, puis le R-learner introduit par [Nie & Wager 2021] et enfin le X-learner de [Künzel *et al.* 2019] est étudié.

T-learner

Le principe du T-learner (abréviation de *two-learner*) est d'estimer séparément les moyennes conditionnelles suivantes : $\mu_1(x) = \mathbb{E}[Y(1)|\mathbf{X} = \mathbf{x}]$ et $\mu_0(\mathbf{x}) = \mathbb{E}[Y(0)|\mathbf{X} = \mathbf{x}]$. La première quantité est estimée avec n'importe quel apprenant de base en utilisant les observation du groupe traité : $\{\mathbf{X}_i, Y_i\}_{W_i=1}$. L'estimateur est noté $\hat{\mu}_1$. De la même façon un estimateur $\hat{\mu}_0$ de μ_0 est construit avec les données du groupe contrôle. Finalement le T-learner est construit ainsi :

$$\hat{\tau}_T(\mathbf{x}) = \hat{\mu}_1(\mathbf{x}) - \hat{\mu}_0(\mathbf{x}). \tag{1.3.1}$$

Des limitations de cette approche sont qu'en séparant le jeu de données d'apprentissage en deux selon le traitement reçu, d'une part un biais de sélection peut apparaître et d'autre part cela donne moins de puissance aux apprenants de base pour détecter de l'information commune au deux groupes, ce qui est par exemple le cas dans un essai randomisé.

S-learner

L'idée du S-learner (abréviation de *single learner*) est de n'utiliser qu'un seul apprenant de base en considérant la variable W comme une simple covariable semblable à $\mathbf{X}$. Ainsi il n'y a plus qu'un unique apprenant de base $\hat{\mu}$ qui cherche à estimer la quantité $\mu(\mathbf{x}, w) = \mathbb{E}[Y(0)|\mathbf{X} = \mathbf{x}, W = w]$. Désormais l'estimateur peut être entraîné sur le jeu de données complet. Finalement le S-learner peut être construit :

$$\hat{\tau}_S(\mathbf{x}) = \hat{\mu}(\mathbf{x}, 1) - \hat{\mu}(\mathbf{x}, 0). \tag{1.3.2}$$

Contrairement au T-learner quand la surface de réponse varie peu en fonction du traitement, le S-learner devrait avoir de meilleures performances car il apprend sur

plus de données. Cependant cette méthode est particulièrement limitée dans le cas où les données ne sont pas réparties équitablement en terme de traitement selon les régions où évolue $\mathbf{X}$ (voir [Hahn *et al.* 2020]), en effet de par la simplicité du modèles les mêmes conditions de régularisation sont imposées aux deux groupes de traitement.

DR-learner

Le DR-learner (abréviation de "Doubly Robust"), est une amélioration du T-learner reposant sur le *inverse probabilty weighting* (IPW), une méthode qui permet de corriger le biais induit sur l'estimateur final par les biais des deux apprenants de base. Le DR-learner est introduit par [Kennedy 2020], nous présentons ici l'algorithme, qui se décompose en trois étapes. Initialement le jeu de données d'entraînement est coupé en deux parties indépendantes de taille n : $\mathcal{D}_{1,n}, \mathcal{D}_{2,n}$, constitués de triplets $(\mathbf{X}_i, W_i, Y_i)$.

- Étape 1 : Plusieurs estimateurs sont construits avec $\mathcal{D}_{1,n}$, ce sont les estimateurs des paramètres de nuisances, à savoir $\hat{\pi}$ estimateur de $\pi(\mathbf{x}) = \mathbb{P}(W = 1|\mathbf{X} = \mathbf{x})$ en utilisant $(\mathbf{X}_i, W_i)_{i\in\mathcal{D}_{1,n}}$, $\hat{\mu}_0$ estimateur de μ_0 en utilisant $(\mathbf{X}_i, Y_i)_{W_i=0, i\in\mathcal{D}_{1,n}}$ et $\hat{\mu}_1$ estimateur de μ_1 en utilisant $(\mathbf{X}_i, Y_i)_{W_i=1, i\in\mathcal{D}_{1,n}}$.

- Étape 2 : Le pseudo-outcome est construit de la manière suivante :

$$\hat{\psi}(W, \mathbf{X}, Y) = \frac{W - \hat{\pi}(\mathbf{X})}{\hat{\pi}(\mathbf{X})(1 - \hat{\pi}(\mathbf{X}))} (Y - \hat{\mu}_W(\mathbf{X})) + \hat{\mu}_1(\mathbf{X}) - \hat{\mu}_0(\mathbf{X}). \tag{1.3.3}$$

  Cette quantité est la fonction d'influence de ATE (voir [Kennedy 2022] pour une introduction à cette notion). L'intuition est que pour estimer ATE il faut calculer la moyenne des $\hat{\psi}$ et ainsi pour estimer CATE il faut régresser $\hat{\psi}$ contre les covariables. Aini en régressant cette quantité contre $\mathbf{X}$ sur $\mathcal{D}_{2,n}$, on construit l'estimateur :

$$\hat{\tau}_{DR}(\mathbf{x}) = \hat{\mathbf{E}}_n \left[\hat{\psi}|\mathbf{X} = \mathbf{x}\right]. \tag{1.3.4}$$

- Étape 3 : Cross-fitting, les rôles de $\mathcal{D}_{1,n}$ et $\mathcal{D}_{2,n}$ sont inversés, à savoir $\mathcal{D}_{2,n}$ est utilisé pour estimer les paramètres de nuisances et $\mathcal{D}_{1,n}$ est utilisé pour estimer $\hat{\tau}_{DR}$. La moyenne des deux estimateurs $\hat{\tau}_{DR}$ ainsi obtenus est utilisée comme estimateur final de CATE.

Une validation croisée à deux blocs a été présentée si dessus mais elle peut facilement être étendue à une validation croisée à K blocs.

Un résultat théorique donne une borne d'erreur pour le DR-learner relativement à un estimateur oracle de CATE, et ce pour des estimateurs arbitraires dans l'étape 1, tant que les estimateurs de l'étape 2 ont une certaine stabilité définie dans [Kennedy 2020]. L'estimateur oracle de CATE régresse les vrais pseudo-outcome contre X.

R-learner

Le R-learner est un metalearner proposé par [Nie & Wager 2021], il consiste en un algorithme en deux étapes. L'idée originelle provient de [Robinson 1988], elle permet de remarquer que sous l'hypothèse d'absence de facteurs confondants (*unconfoundedness*),

$$\mathbb{E}[\varepsilon_i(W_i)|X_i, W_i] = 0, \text{ où } \varepsilon_i(w) = Y_i(w) - (\mu_0(X_i) + w\tau(X_i)). \tag{1.3.5}$$

En posant $m(x) = \mathbb{E}[Y|X = x] = \mu_0(x) + \pi(X)\tau(X)$, en simplifiant l'écriture de $\varepsilon_i(W_i)$ par ε_i, une réécriture est alors possible :

$$Y_i - m(X_i) = (W_i - \pi(X_i))\tau(X_i) + \varepsilon_i. \tag{1.3.6}$$

Cette écriture équivaut à :

$$\tau(.) = \arg\min_{\tau^*} \left\{ \mathbb{E}\left[((Y_i - m(X_i)) - (W_i - \pi(X_i))\tau^*(X_i))^2 \right] \right\}. \tag{1.3.7}$$

Ainsi si les quantités $m(x)$ et $\pi(x)$ sont complètement connues, τ pourrait être estimé par minimisation de la perte empirique

$$\hat{\tau}(.) = \arg\min_{\tau^*} \left(\frac{1}{n} \sum_{i=1}^{n} [(Y_i - m(X_i)) - (W_i - \pi(X_i))\tau^*(X_i)]^2 + \Gamma_n(\tau(.)) \right), \tag{1.3.8}$$

où Γ est un terme de régularisation pour contrôler la complexité de τ (par exemple régularisation L^1 ou L^2, pénalisation de la régularité de splines, dropout pour des réseaux de neurones, etc).

La procédure d'estimation de CATE en deux étapes peut maintenant être décrite :

- Étape 1 : Le jeu de donnée est divisé en Q parties de tailles identiques (stratégie de crossfitting). La fonction qui a chaque indice i du jeu de données associe la partition à laquelle il appartient est notée $q(i)$. Q estimateurs de m et π sont construits en excluant successivement une des partitions. Ces estimateurs sont les apprenants de base du R-learner.

- Étape 2 : Un estimateur plug-in de τ basé sur l'équation 1.3.8 peut être défini. Soit $\hat{\pi}^{-q(i)}$ l'estimateur de π obtenu en excluant la partition définie à l'étape précédente contenant l'indice i. L'estimateur plug-in est donné par :

$$\hat{\tau}(.) = \arg\min_{\tau^*} \left[\hat{L}_n(\tau^*(.)) + \Gamma(\tau^*(.)) \right], \tag{1.3.9}$$

$$\hat{L}_n(\tau^*(.)) = \frac{1}{n} \sum_{i=1}^{n} [(Y_i - m(X_i)) - (W_i - \pi(X_i))\tau^*(X_i)]^2. \tag{1.3.10}$$

  Les apprenants de base proposés par [Nie & Wager 2021] sont : penalized regression, kernel ridge regression and boosting.
  En pratique le problème de minimisation est considéré sur une famille de fonctions issues d'un même modèle.

X-learner

Le X-learner proposé par [Künzel *et al.* 2019] est une extension du T-learner, il permet d'estimer CATE en trois étapes successives.

- Étape 1 : Similairement à ce qui est fait dans le T-learner, nous construisons respectivement les estimateurs μ_0 et μ_1 en les entraînant respectivement sur les individus non traités et traités. Ces estimateurs sont les apprenants de base de la première étape.

- Étape 2 : Les effets de traitements pour respectivement les individus traités et les individus non non traités peuvent être imputés à l'aide des estimateurs de l'étape précédente :

$$\tilde{D}_i^1 = Y_i - \hat{\mu}_0(\mathbf{X}_i) \text{ si } W_i = 1 \text{ et } \tilde{D}_i^0 = \hat{\mu}_1(\mathbf{X}_i) - Y_i \text{ si } W_i = 0. \tag{1.3.11}$$

  Cette étape est une tentative de récupérer l'effet de traitement individuel (ITE pour *individual treatment effect*), $D_i = Y_i(1) - Y_i(0)$, pour chaque individu qu'il soit traité ou non traité en remplaçant le potential outcome non observé par son estimation avec les apprenants de base de la première étape. Contrairement au T-learner ces estimateurs ne sont utilisés que pour estimer un potential outcome à cette étape. L'idée est alors d'utiliser deux nouveaux apprenants de base $\hat{\tau}_1$ et $\hat{\tau}_0$ pour estimer τ sur les groupes des traités et des non traités. Pour construire $\hat{\tau}_1$ on régresse $\tilde{D}^1$ contre $\mathbf{X}$ sur l'ensemble des individus traités. De même on construit $\hat{\tau}_0$ en régressant $\tilde{D}^0$ contre $\mathbf{X}$ sur l'ensemble des individus non-traités. Ce sont les apprenants de base de la seconde étape.

- Étape 3 : L'estimateur final de CATE proposé est :

$$\hat{\tau}_X(\mathbf{x}) = g(x)\hat{\tau}_0(\mathbf{x}) + (1 - g(x))\hat{\tau}_1(\mathbf{x}), \tag{1.3.12}$$

  où g est une fonction poids à valeurs dans $[0; 1]$. Un choix habituel pour la fonction g est un estimateur du score de propensité $g = \hat{\pi}$. Cependant lorsque la taille de population d'individu traités est négligeable face à celle de non traités $g = 0$ est aussi un choix judicieux (de la même façon on choisit $g = 1$ si c'est la population d'individus non traités qui est négligeable face à celle des traités). La logique derrière ces choix est la suivante : le modèle ajusté avec la plus grande population est sûrement le mieux spécifié il convient donc de le surpondérer.

Ainsi le X-learner est une alternative au T-learner particulièrement indiquée lorsque qu'il y a une disproportion entre les nombres d'individus traités et non traités.

1.3.2 Forêts causales

Les forêts causales sont une adaptation des forêts aléatoires de régression pour estimer le CATE. Dans une forêt aléatoire de régression, plusieurs arbres de décision sont construits de manière aléatoire à partir de l'ensemble des données d'entraînement.

L'algorithme CART (Classification and Regression Trees présenté par [Breiman 2001]) est un algorithme basé sur les arbres de décision qui peut être utilisé pour résoudre à la fois des problèmes de classification et de régression en apprentissage statistique. Il fonctionne en partitionnant de manière récursive les données d'entraînement en sous-ensembles plus petits à l'aide de divisions binaires. L'arbre commence à la racine, qui contient toutes les données d'entraînement, et divise de manière récursive les données en sous-ensembles plus petits jusqu'à ce qu'un critère d'arrêt soit atteint.

Pour estimer l'hétérogénéité des effets causaux, une approche basée sur les données et utilisant le CART est proposée dans [Athey *et al.* 2019] pour diviser les données en sous-populations présentant des différences dans l'ampleur de leurs effets de traitement. Des intervalles de confiance valides peuvent également être créés pour CATE. Cette approche se différencie du CART conventionnel sur deux points : d'une part, elle se concentre sur l'estimation des effets de traitement moyens conditionnels plutôt que sur la prédiction directe de Y, comme c'est le cas dans le CART conventionnel, cela provient de l'impossibilité d'observer directement la valeur de CATE sur le jeu d'entraînement. D'autre part, des échantillons distincts sont utilisés pour construire la partition et estimer les effets dans chaque sous-population, ce qui est appelé estimation honnête. En revanche, dans le CART conventionnel, les mêmes échantillons sont utilisés pour les deux tâches.

Une description plus détaillée de cet algorithme est présente dans le Chapitre 2.

1.3.3 Méthodes par réseaux de neurones

Des réseaux de neurones peuvent être construits pour estimer les réseaux de neurones, cette approche peut être vue comme une extension des métalearners présentés précédemment. [Curth & van der Schaar 2021] propose une taxonomie de ces approches. Ces algorithmes se décomposent en deux étapes. Lors de la première étape un réseau de neurones permet d'estimer les quantités μ_0, μ_1 et e déjà définies dans la partie sur les metalearners. Puis plusieurs méthodes sont considérées pour combiner ces estimateurs afin d'estimer CATE, c'est cette partie qui est analogue aux metalearners. Dans cet article l'auteur considère les combinaisons des différentes architectures de réseaux de neurones avec les différentes combinaisons des trois estimateurs pour estimer CATE.

Les réseaux de neurones considérés ont des architectures différentes qui permettent des échanges plus ou moins importants d'information pour permettre l'apprentissage des trois fonctions de nuisance (μ_0, μ_1, π). Les réseaux considérés par la suite sont tous des réseaux de neurones à propagation avant (en anglais feedforward neural network).

Le plus naïf, qui consiste à appliquer directement le metalearner choisi sans partage d'information est le TNet. Il s'agit d'entraîner un apprenant de base de type réseau de neurones pour chacune des fonctions de nuisance.

Les architectures suivantes partagent de l'information via des couches

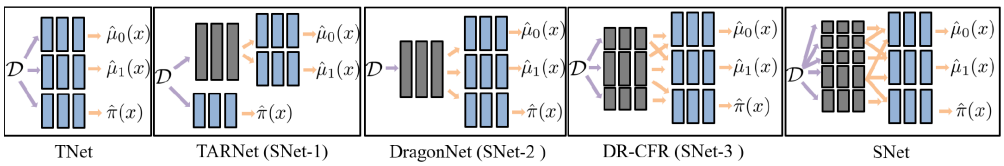


Figure 1.6: Les différentes architectures de réseaux de neurones présentées dans [Curth & van der Schaar 2021].

d'apprentissage communes pour plusieurs fonctions de nuisances. Elles sont désignées par l'appellation SNet. Certaines d'entre elles ont déjà été présentées dans la littératures comme TARNet ([Shalit *et al.* 2017]) et Dragonnet ([Shi *et al.* 2019]). L'apprentissage des représentations est un ensemble de techniques qui permet à un système de découvrir automatiquement les représentations nécessaires à la détection ou à la classification à partir de données brutes. Les réseaux neuronaux multicouches peuvent être utilisés pour de l'apprentissage des représentations, car ils apprennent une représentation de leurs entrées au niveau des couches cachées qui est ensuite utilisée pour la classification ou la régression à la couche de sortie. Le modèle TARNet repose sur une hypothèse de représentativité de μ_0 et μ_1 dans un même espace, pour DragonNet, la fonction de propensité $\pi(.)$ est aussi supposée être représentée dans un même espace commun. Des hypothèses plus complexes sur les espaces de représentation mènent au modèle SNet. Pour chaque hypothèse associée à un réseau de neurones, une fonction de perte adaptée est proposée. Les réseaux de neurones sont ensuite entraînés de manière classique avec la rétropropagation du gradient (*backpropagation* en anglais) sur cette fonction de perte. Finalement plusieurs metalearners sont proposés par [Curth & van der Schaar 2021] afin de combiner les sorties des réseaux de neurones pour obtenir un estimateur de CATE.

1.3.4 Autres méthodes

Enfin d'autres méthodes d'apprentissage statistique adaptées à l'inférence causale sont présentées. Elles reposent sur des arbres mais avec des approches différentes des forêts causales.

BART causal

L'approche des arbres de régression additives bayésiens (BART) est une méthode d'apprentissage automatique qui combine des éléments de l'apprentissage par gradient boosting, de la modélisation bayésienne et de l'échantillonnage par chaîne de Markov Monte Carlo (MCMC). Cette méthode, introduite par [Chipman *et al.* 2010a], vise à estimer des modèles de régression robustes et flexibles tout en prenant en compte l'incertitude dans les prédictions.

L'idée centrale derrière BART est de construire un modèle de régression en agrégeant plusieurs arbres de décision, tout en permettant à chaque arbre d'avoir

une structure et des paramètres spécifiques. Contrairement aux méthodes traditionnelles qui fournissent des estimations ponctuelles, BART produit une distribution postérieure complète pour chaque observation, ce qui permet de quantifier l'incertitude associée à la prédiction.

Pour estimer les paramètres du modèle BART, une approche bayésienne est utilisée, où des prior sont spécifiés pour chaque composante du modèle (profondeur maximale par exemple). Ensuite, l'échantillonnage MCMC est utilisé pour explorer l'espace des paramètres et estimer la distribution postérieure des paramètres.

[Hill 2011] propose d'utiliser BART pour évaluer CATE à la manière d'un S-learner déjà présenté dans la Section 1.3.1. [Hahn *et al.* 2020] propose une extension de cette approche appelée Bayesian Causal Forest (BCF). L'idée d'utiliser une approche bayésienne pour estimer les effets de traitement en exprimant la surface de réponse ainsi :

$$f(X_i, W_i) = \mu(X_i, \hat{\pi}(X_i)) + \tau(X_i)W_i \tag{1.3.13}$$

où $\hat{\pi}(x)$ est le score de propension estimé et les fonctions μ et τ sont des priors de BART indépendants.

Causal boosting

Une alternative à une forêt aléatoire pour la régression des moindres carrés est l'utilisation d'arbres boostés. Le boosting construit une approximation de fonction en ajustant successivement des apprenants faibles aux résidus du modèle à chaque étape. Dans cette section en suivant [Powers *et al.* 2018], le boosting est adapté au problème de l'estimation des effets de traitement hétérogènes.

Algorithm 1 Causal boosting

Input: $\{X_i, W_i, Y_i\}$ l'échantillon de données, $K > 0$ le nombre d'itérations, $\varepsilon > 0$
  $R_i \leftarrow Y_i$
  $\hat{G}_0(x, t) \leftarrow 0$
  for $k = 1$ to K **do**
    Un arbre causal $\hat{g}_k$ est entraîné sur $\{X_i, W_i, R_i\}$.
    $R_i \leftarrow R_i - \varepsilon \hat{g}_k(X_i, W_i)$
    $\hat{G}_k \leftarrow \hat{G}_{k-1} + \varepsilon \hat{g}_k$
  end for
Output: $\hat{G}_K$

L'arbre causal $\hat{g}_k$ estime l'effet causal sur les résidus obtenus lors l'itération précédente.

Comme dans la forêt causale, le problème demeure celui de contrôler le surajustement. En particulier, les méthodes de boosting sont susceptibles de surajuster les données car les arbres ne sont pas construits de manière indépendante. Alors qu'une forêt aléatoire bénéficierait de l'utilisation de plus d'arbres sur lesquels effectuer une moyenne, dans le boosting par gradient, le nombre d'arbres est un

paramètre de réglage important qui doit être contrôlé. En apprentissage supervisé, nous appliquerions idéalement une validation croisée. Dans notre cas, le paramètre d'intérêt est le CATE et nous n'observons pas la vraie valeur pour chaque observation. Par conséquent, la validation croisée ne s'applique pas ici. Au lieu de cela, [Powers *et al.* 2018] propose de faire quelque chose de similaire à l'approche honnête de la forêt causale, en séparant le jeu de données en deux parties disjointes, les arbres successifs sont construits avec le premier jeu de données et en utilisant en second jeu de données pour remplir les feuilles. Une description plus détaillée du procédé est présentée dans [Powers *et al.* 2018].

1.4 Apprentissage par transfert

L'apprentissage par transfert (*transfer learning* en anglais) est une méthode avancée d'apprentissage statistique qui tire parti des connaissances acquises dans un domaine source pour améliorer les performances des modèles dans un domaine cible similaire. Cette approche est particulièrement utile lorsque les données d'apprentissage dans le domaine cible sont rares ou coûteuses à obtenir, mais que des données abondantes sont disponibles dans un domaine source connexe. Le concept fondamental d'apprentissage par transfert réside dans l'idée que les similitudes entre les deux domaines peuvent être exploitées pour transférer des représentations, des modèles ou des connaissances spécifiques d'un domaine à un autre.

Les stratégies d'apprentissage par transfert comprennent selon [Yang *et al.* 2020] le *fine-tuning* des modèles pré-entraînés, l'extraction de caractéristiques, le multi-task learning... Le *fine-tuning* implique l'ajustement d'un modèle pré-entraîné sur des données source avec des données cibles spécifiques. L'extraction de caractéristiques consiste à utiliser les couches cachées d'un modèle de réseau de neurones pré-entraîné comme extracteur de caractéristiques pour les données cibles. L'apprentissage multi-tâche vise à former un modèle à accomplir plusieurs tâches simultanément, tandis que l'adaptation de domaine adapte un modèle appris sur un domaine source à un domaine cible légèrement différent. L'adaptation de domaine est un cas particulier d'apprentissage par transfert, où les domaines sources et cibles vivent dans le même espace (*feature space*) et seules les distributions des variables changent d'un domaine à l'autre. D'autres méthodes sont alors adaptées tels que les algorithmes de recalibration ou la recherche d'un espace de représentation commun. Le premier algorithme tente de repondérer l'échantillon étiqueté source afin qu'il ressemble le plus possible à l'échantillon cible en terme de distribution, la recherche d'espace de représentation commun passe par l'utilisation de RKHS ou est dérivée du problème de transport optimal.

L'apprentissage par transfert est largement utilisé dans des domaines tels que la vision par ordinateur, le traitement du langage naturel, la reconnaissance vocale et la bioinformatique. Par exemple, dans la vision par ordinateur, les modèles pré-entraînés sur des ensembles de données massifs comme ImageNet sont souvent

utilisés comme point de départ pour des tâches spécifiques nécessitant moins de données d'apprentissage. Dans le traitement du langage naturel, les modèles de langage pré-entraînés comme BERT et GPT sont fréquemment utilisés pour améliorer les performances des tâches de classification de texte, de résumé automatique et de traduction. En résumé, l'apprentissage par transfert joue un rôle essentiel en permettant aux modèles d'apprentissage automatique de bénéficier des connaissances acquises dans des domaines connexes, améliorant ainsi leur capacité à généraliser et à résoudre des problèmes dans de nouveaux contextes.

1.5 Présentation des contributions

Le travail qui suit a été structuré en 3 chapitres, le Chapitre 2 introduit la forêt causale HTERF qui est utilisée dans les deux autres chapitres, c'est une version enrichie d'un article coécrit avec Véronique Maume-Deschamps et Pierre Ribereau qui est publié dans le volume 196 de la revue *Computational Statistics & Data Analysis*. Le Chapitre 3 présente des applications de cette forêt sur deux sujets (risque de crédit et données climatiques). Enfin le Chapitre 4 présente des résultats d'apprentissage par transfert à partir de la forêt HTERF.

1.5.1 Chapitre 2 : Estimation de l'effet causal par forêt aléatoire (HTERF)

Ce chapitre est une version enrichie d'un article publié dans le volume 196 de la revue *Computational Statistics & Data Analysis* et intitulé *Heterogeneous Treatment Effect based Random Forest: HTERF*. Une illustration graphique de résultats de simulations, une présentation plus détaillée de la preuve du théorème principal ainsi que d'une présentation détaillée du package associé on été ajoutées dans ce chapitre.

Les méthodes reposant sur des forêts causales pour évaluer CATE sont présentées dans des articles tels que [Athey & Imbens 2016], [Athey *et al.* 2019], [Wager & Athey 2018] et des résultats théoriques de convergences de ces estimateurs sont obtenus. Ces algorithmes sont également largement représentés dans les packages d'apprentissage statistique causal, tels que `EconML` en `Python` ou `grf` en `R`. La méthode intitulée *generalized random forest* (GRF) est la plus aboutie ([Athey *et al.* 2019]), elle présente de bonnes performance en terme d'estimation de CATE et il existe un résultat de convergence en loi sous un certain ensemble d'hypothèses.

La méthode GRF diffère en partie des forêts de régressions classiques de [Breiman 2001], nous présentons ici son fonctionnement.

Une forêt aléatoire est un modèle ensembliste, composé d'un groupe d'arbres de décision. Pendant l'entraînement, plusieurs arbres sont développés sur des sous-échantillons aléatoires de l'ensemble de données. Les arbres individuels sont entraînés selon les étapes suivantes :

- Tout d'abord, un sous-échantillon aléatoire est extrait en échantillonnant sans remplacement à partir de l'ensemble de données complet. Cet échantillon est séparé en deux parties disjointes $\mathcal{I}_1$ et $\mathcal{I}_2$. Seule la première partie est utilisée pour la construction de l'arbre qui suit. Un seul noeud racine est créé contenant cet échantillon aléatoire.

- Le noeud racine est divisé en noeuds enfants, et les noeuds enfants sont divisés de manière récursive pour former un arbre. La procédure s'arrête lorsque aucun noeud ne peut être divisé davantage. Chaque noeud est divisé en utilisant l'algorithme suivant :

  - Un sous-ensemble aléatoire de variables explicatives est sélectionné comme candidats pour la division, nous le nommons $\mathcal{M}_{try}$
  - Pour chacuns de ces variables $X^{(j)}$, toutes ses valeurs possibles z sont examinées pour obtenir la meilleure partition de ce noeud en deux enfants. La qualité d'une division (j, z) est déterminée par l'augmentation de la quantité d'hétérogénéité sur CATE. Certaines divisions ne sont pas considérées, car les noeuds enfants résultants seraient trop petits ou trop différents en taille.
  - Tous les exemples avec des valeurs pour la variable de division $X^{(j)}$ inférieures ou égales à la valeur de partition v sont placés dans un nouveau noeud enfant gauche, et tous les exemples avec des valeurs supérieures à v sont placés dans un noeud enfant droit.
  - Si un noeud n'a pas de divisions valides, ou si la division ne permettra pas d'améliorer l'ajustement, le noeud n'est pas divisé davantage et forme une feuille de l'arbre final.

Optimiser directement le critère d'hétérogénéité est trop coûteux car il requiert l'estimation de deux variances et de deux covariances, à la place, une approximation du gradient de ce critère est utilisée en pratique.

Nous présentons maintenant comment une estimation de CATE est calculée pour un point test donné. Pour chaque arbre, l'exemple de test est "descendu" pour déterminer dans quelle feuille il tombe. Sur la base de cette information, nous créons une liste des exemples d'entraînement voisins issus de $\mathcal{I}_2$, pondérée par le nombre de fois où l'exemple est tombé dans la même feuille que l'exemple de test. Une prédiction est faite en utilisant cette liste pondérée de voisins, l'effet du traitement peut être calculé en utilisant les résultats et le statut du traitement des exemples voisins.

Les principales différences avec les forêts de régression sont :

- L'utilisation de sous-échantillonnage sans remise au lieu de l'échantillonnage bootstrap,

- l'utilisation d'un critère de partitionnement qui cherche à maximiser l'hétérogénéité entre les enfants en termes de CATE,

- le caractère honnête de la forêt : un premier échantillon $\mathcal{I}_1$ est utilisé pour construire les différents noeuds de l'arbre puis un second échantillon $\mathcal{I}_2$ est utilisé pour remplir les feuilles et procéder aux estimations.

Théorème 1.1 ([Athey *et al.* 2019]). *Sous un jeu d'hypothèse bien choisi (voir [Athey et al. 2019]) et en supposant que* $Y = \tau(\mathbf{X})W + \gamma(\mathbf{X}) + \varepsilon$, *l'estimateur* $\hat{\tau}$ *de CATE issu de GRF vérifie :*

$$\frac{\hat{\tau}(\mathbf{x}) - \tau(\mathbf{x})}{\sqrt{Var(\hat{\tau}(\mathbf{x}))}} \longrightarrow N(0,1).$$

Nous nous sommes ensuite penché sur l'interprétabilité des résultats obtenus avec cet algorithme. Pour les forêts aléatoires de régression, [Scornet *et al.* 2015] propose un résultat asymptotique sur la fréquence d'apparition des variables informatives dans les noeuds des arbres. Nous désignons par variables informatives, l'ensemble des variables explicatives dont Y dépend dans un contexte de régression. Ce résultat est également bien vérifié empiriquement sur des exemples simples de régression. Dans le cas de l'estimation de CATE, l'ensemble des variables informatives est l'ensemble des variables explicatives dont τ dépend. Cependant sur des exemples simples causaux, le résultat n'était pas aussi flagrant avec la forêt de [Athey *et al.* 2019], la surreprésentation des variables informatives dans les noeuds n'est pas aussi marquée.

Nous proposons avec un critère de segmentation différent, plus robuste et simple à calculer. Un résultat de convergence presque sûre est également établi sous un jeu d'hypothèses plus permissif que précédemment. Enfin un résultat d'interprétabilité est également obtenu.

Le nouveau critère de partion est optimisé sur un sous-ensemble de variables explicatives $\mathcal{M}try$. Les covariables ont été sélectionnées de manière aléatoire avec une probabilité positive pour chaque covariable d'être sélectionnée. Ensuite, la meilleure partition est celle qui maximise le critère de partition $\Delta(A, j, z)$, où $A = \prod_{i=1}^{d}[a_i, b_i]$ est le noeud actuel, j est choisi dans $\mathcal{M}_{try}$ et $z \in A^j = [a_j, b_j]$.

$$\Delta(A, j, z) = \frac{|A_L||A_R|}{|A|^2}\left(\left(\overline{Y}_{A_{L1}} - \overline{Y}_{A_{L0}}\right) - \left(\overline{Y}_{A_{R1}} - \overline{Y}_{A_{R0}}\right)\right)^2, \tag{1.5.1}$$

où $A_{L1} = \{\mathbf{X}_i \in A | \mathbf{X}_i^{(j)} < z, W_i = 1\}$, $A_{L0} = \{\mathbf{X}_i \in A | \mathbf{X}_i^{(j)} < z, W_i = 0\}$, $A_{R1} = \{\mathbf{X}_i \in A | \mathbf{X}_i^{(j)} \geq z, W_i = 1\}$, $A_{R0} = \{\mathbf{X}_i \in A | \mathbf{X}_i^{(j)} \geq z, W_i = 0\}$, $A_L = A_{L1} \cup A_{L0}$ et $A_R = A_{R1} \cup A_{R0}$. Pour tout ensemble B, on note $\overline{Y}_B = \frac{1}{|B|}\sum_{i \in B} Y_i$.

Pour estimation de τ, nous utilisons une approche qui s'inspire de GRF :

$$\hat{\tau}_{B,n}(\mathbf{x}) = \sum_{i:W_i=1} \alpha_i(\mathbf{x}) Y_i - \sum_{i:W_i=0} \alpha_i'(\mathbf{x}) Y_i, \tag{1.5.2}$$

Le point test $\mathbf{x}$ est à nouveau "descendu" dans chaque arbre, on note $L_b(\mathbf{x})$ l'ensemble des éléments de $\mathcal{I}_2$ qui tombent dans la même feuille que x et tels que $W = 1$. On définit alors :

$$\alpha_{b,i}(\mathbf{x}) = \frac{\mathbb{1}\{\mathbf{X}_i \in L_b(\mathbf{x})\}}{|L_b(\mathbf{x})|} \text{ and } \alpha_i(\mathbf{x}) = \frac{1}{B}\sum_{b=1}^{B}\alpha_{b,i}(\mathbf{x}). \tag{1.5.3}$$

Les coefficients α' définis de façon analogue, en utilisant les observations telles que $W = 0$.

Définition 1.1. *Soit $f : \mathcal{X} \to \mathbb{R}$; elle n'appartient PAS à la classe ♠ s'il existe un rectangle $A = \prod_{j=1}^{d}[a_j, b_j] \subset \mathcal{X}$, avec $a_j \leq b_j$ tel que pour tout $j = 1, \ldots, d$, $z \mapsto \mathbb{E}\left[f\left(z, \mathbf{X}^{-j}\right)\mathbb{1}_{\{\mathbf{X}^{-j} \in A^{-j}\}}\right]$ est constant sur $[a_j, b_j]$, et f n'est pas constante sur A.*

Soit $\tau_1(\mathbf{x}) = \mathbb{E}\left[Y(1)|\mathbf{X} = \mathbf{x}\right]$ et $\tau_0(\mathbf{x}) = \mathbb{E}\left[Y(0)|\mathbf{X} = \mathbf{x}\right]$, et de manière similaire, $\hat{\tau}_1(\mathbf{x}) = \sum_{i:W_i=1}\alpha_i(\mathbf{x})Y_i$ et $\hat{\tau}_0(\mathbf{x}) = \sum_{i:W_i=0}\alpha'_i(\mathbf{x})Y_i$.

On a les hypothèses suivantes:

Hypothèse 1.1.

- $Y = \tau(\mathbf{X})g(\mathbf{W}) + \gamma(\mathbf{X}) + \varepsilon$.
- $\mathbf{X} = (X_1, \ldots, X_d)$ *est un vecteur aléatoire continu avec des coordonnées indépendantes.*
- ε *and* $\mathbf{X}$ *sont indépendants, et* ε *est une variable aléatoire centrée continue à queue légère avec une fonction de répartition croissante. C'est à dire qu'il existe* $0 < \theta < 1$ *tel que pour tout* $D > 0$, $\mathbb{P}\left(|\varepsilon| > D\right) \leq C\theta^D$.
- $\mathbf{X}$ *prend ses valeurs dans* $\mathcal{X}$*, un hyper-rectangle compact de* $\mathbb{R}^d$*:* $\mathcal{X} = \prod_{i=1}^{d}[u_i, v_i]$, $-\infty < u_i \leq v_i < \infty$.
- $\mathbf{x} \mapsto \gamma(\mathbf{x})$, $\mathbf{x} \mapsto \tau_1(\mathbf{x})$ *et* $\mathbf{x} \mapsto \tau_0(\mathbf{x})$ *sont continues. Donc en particulier,* $\mathbf{x} \mapsto \tau(\mathbf{x})$ *est continu.*

On adopte la notation suivante :
$f(n) = \Omega\left(g(n)\right) \iff \exists k > 0, \exists n_0 > 0 \mid \forall n \geqslant n_0 \quad |f(n)| \geqslant k \cdot |g(n)|$

Hypothèse 1.2. *Les hypothèses suivantes sont faites sur B (le nombre d'arbres dans la forêt) et $N_{n,1}(\mathbf{x}; \Theta, \mathcal{D}_n)$ resp. $N_{n,0}(\mathbf{x}; \Theta, \mathcal{D}_n)$ (le nombre d'observation dans une feuille telle que $W = 1$, resp. $W = 0$):*

1. $B = \mathcal{O}\left(n^{\alpha}\right)$, avec $\alpha > 0$.
2. $\forall \mathbf{x} \in \mathcal{X}, \quad \mathbb{E}\left[N_{n,1}(\mathbf{x}; \Theta, \mathcal{D}_n)\right] = \Omega\left(\sqrt{n}\left(\ln(n)\right)^{\beta}\right)$, avec $\beta > 1$.
3. $\forall \mathbf{x} \in \mathcal{X}, \quad \mathbb{E}\left[N_{n,0}(\mathbf{x}; \Theta, \mathcal{D}_n)\right] = \Omega\left(\sqrt{n}\left(\ln(n)\right)^{\beta}\right)$.

Théorème 1.2. *Sous les hypothèses 1.1 et 1.2, avec τ_1 et τ_0 qui appartiennent à la classe ♠; on suppose que pour $\beta > \frac{5}{2}$, $C > 0$ fixés, chaque arbre construit est le plus haut tel que $C\sqrt{n}(\ln n)^\beta \leq N_{n,0}\left(\mathbf{x}; \Theta_\ell, \mathcal{D}_n\right), N_{n,1}\left(\mathbf{x}; \Theta_\ell, \mathcal{D}_n\right)$. Aussi on suppose que $\mathbb{E}[\max \varepsilon_i^2] \leq K(\ln n)^u$ avec $\beta - u > \frac{1}{2}$ et K une constante positive. Alors,*

$$\forall \mathbf{x} \in \mathcal{X}, |\hat{\tau}_{B,n}(\mathbf{x}) - \tau(\mathbf{x})| \xrightarrow[n\to\infty]{p.s.} 0.$$

Des simulations ont été effectuées sur des jeu de données synthétiques et semi-synthétiques. Les performances de HTERF sont meilleures que les forêts causales actuellement utilisées, aussi bien en termes d'erreur sur l'estimation de CATE que en termes d'interprétabilité des forêts avec une meilleure représentativité des variables informatives dans les segmentations des premiers étages de HTERF.

Un package intitulé `CausalForest` a été développé et est disponible sur le répertoire général des bibliothèques `Julia`. Une description détaillée de ce package est disponible en appendice.

1.5.2 Chapitre 3 : Applications de HTERF

Dans ce chapitre deux applications de HTERF sont proposées sur des données empiriques. La première porte sur un sujet de gestion des risques au sein de Natixis et la seconde concerne l'étude du phénomène climatique El Niño.

Application au risque de crédit

Le coût du risque est une mesure du risque futur contenu dans l'encours présent. Il peut être appréhendé comme l'ajustement du stock de provisions entre deux périodes et est composé des dotations nettes et des reprises au dépréciations et provisions pour risque de crédit, et des pertes sur créances irrécouvrables. Le coût du risque est donc un indicateur des pertes attendues et mesure l'effort déployé par une entité sur une période donnée pour se protéger contre les pertes futures estimées sur son portefeuille de prêts.

Ayant connaissance du coût du risque pour chaque contrepartie d'un portefeuille de Natixis à une fréquence trimestrielle, nous proposons d'évaluer l'impact des crises sur ce portefeuille en terme de coût du risque. Pour ce faire, des variables explicatives on été ajoutées, celles-ci avaient des granularités différentes allant de variables mondiales, à des variables propres à une zone géographique ou à un secteur d'activité. Enfin la variable de traitement considérée est la présence ou l'absence de crise économique.

Application climatique

Pour la seconde application, nous considérons ENSO, ou El Niño-Southern Oscillation, un phénomène climatique majeur qui se produit dans l'océan Pacifique tropical. Il se caractérise par des variations périodiques des températures de surface de la mer et

des vents, ce qui a un impact significatif sur les schémas météorologiques mondiaux, notamment en Australie.

L'ENSO comporte trois phases principales : El Niño, La Niña et une phase neutre. Chacune de ces phases a des effets distincts sur le climat australien :

- El Niño, souvent associé à des conditions météorologiques plus chaudes et plus sèches que la normale. Cela peut entraîner des sécheresses, des vagues de chaleur, des feux de brousse et des réductions des précipitations, en particulier dans le sud et l'est du pays.

- La Niña, qui est généralement associée à des précipitations supérieures à la normale, des inondations potentielles et des températures plus fraîches que la normale, en particulier dans le nord et l'est du pays.

- La phase neutre, durant laquelle les températures de surface de la mer et les schémas de circulation atmosphérique sont proches de la moyenne à long terme. Cela peut conduire à des conditions météorologiques relativement normales en Australie, bien que d'autres facteurs régionaux et mondiaux puissent toujours influencer le climat.

En résumé, l'ENSO a un impact significatif sur le climat australien, ces variations peuvent avoir des conséquences importantes sur l'agriculture, les ressources en eau, les écosystèmes naturels et les risques naturels en Australie. Nous proposons d'utiliser HTERF pour étudier l'effet de l'impact de ces phénomène sur la température en deux stations météorologiques situées dans des régions différentes de l'est australien.

1.5.3 Chapitre 4 : Apprentissage par transfert sur des forêts causales

Ce chapitre sera prochainement soumis pour publication, il propose une méthode de transfert utilisant les forêts causales.

Dans ce chapitre le problème de l'estimation de CATE est posé dans un contexte d'apprentissage statistique. Deux distributions sont considérées, d'une part une distribution source qui génère (X^s, Y^s, W^s) et d'autre part une distribution cible qui génère $\left(X^t, Y^t, W^t\right)$. Les variables X^s et X^t (respectivement W^s et W^t) ont les mêmes distributions. L'objectif est de déterminer le comportement de CATE sur le domaine cible pour lequel moins de données sont disponibles, en utilisant de manière efficace les similarités entre les deux domaines.

Dans le cadre d'un problème de régression, lorsque les variables X^s et X^t ont même distribution, un algorithme permettant d'estimer $\mathbb{E}\left(Y^t|X^t\right)$ est proposé par [Wang 2016], il s'agit de l'algorithme offset (Algorithme 2).

Si un modèle de régression ridge à noyau est utilisé pour $\hat{f}^s$ et $\hat{f}^o$, une borne de généralisation pour cet algorithme est proposée, elle dépend des erreurs associées à chacun des deux estimateurs (voir [Wang & Schneider 2015]).

Algorithm 2 Algorithme offset

Input: Un jeu de données source $\{X_i^s, Y_i^s\}$, une jeu de données cible étiqueté $\{X_i^{tL}, Y_i^{tL}\}$ et un jeu de données cible non étiqueté X^{tU}.
Estimer un modèle $\hat{f}^s$ qui régresse $\{Y_i^s\}$ contre $\{X_i^s\}$
Estimer un modèle $\hat{f}^o$ qui régresse $\{\hat{Y}_i^o\} = \{Y_i^{tL} - \hat{f}^s(X_i^{tL})\}$ contre $\{X_i^t\}$
$\{Y_i^{new}\} \leftarrow \{Y_i^s + \hat{f}^o(X_i^s)\}$
Entraîner un modèle M sur $\{X_i^s, Y_i^{new}\} \cup \{X_i^{tL}, Y_i^{tL}\}$.
Output: $\{\hat{Y_i^{tU}}\} \leftarrow \{M(X_i^{tU})\}$

Nous proposons une adaptation causale de cet algorithme en utilisant HTERF comme modèle final M dans l'algorithme (Algorithme 3).

Algorithm 3 Algorithme offset causal avec deux modèles distincts

Input: Un jeu de données source $\{W^s, X^s, Y^s\}$, un jeu de données cible étiqueté $\{W^{tL}, X^{tL}, Y^{tL}\}$ et un jeu de données cible non étiqueté X^{tU}
Estimer un modèle $\hat{f}_0^s$ qui régresse $\{Y_i^s\}_{W_i^s=0}$ contre $\{X_i^s\}_{W_i^s=0}$ et un modèle $\hat{f}_1^s$ qui régresse $\{Y_i^s\}_{W_i^s=1}$ contre $\{X_i^s\}_{W_i^s=1}$
Estimer un modèle $\hat{f}_0^o$ qui régresse $\{Y_i^{tL} - \hat{f}_0^s(X_i^{tL})\}_{W_i^{tL}=0}$ contre $\{X_i^{tL}\}_{W_i^{tL}=0}$ et un modèle $\hat{f}_1^o$ qui régresse $\{Y_i^{tL} - \hat{f}_1^s(X_i^{tL})\}_{W_i^{tL}=1}$ contre $\{X_i^{tL}\}_{W_i^{tL}=1}$
$\{Y_i^{new}\} \leftarrow \{Y_i^s + \hat{f}_{W_i^s}^o(X_i^s)\}$
Entraîner un modèle M sur $\{W_i^s, X_i^s, Y_i^{new}\} \cup \{W_i^{tL}, X_i^{tL}, Y_i^{tL}\}$.
Output: $\{\hat{\tau}^{tU}(X_i^{tU}\} \leftarrow \{M(X_i^{tU})\}$

Sous un jeu d'hypothèses proche de celui utilisé dans le second chapitre pour avoir la convergence de HTERF, et avec des hypothèses raisonnables sur $\hat{f}^s$ et $\hat{f}^o$, nous montrons la convergence de la méthode offset causale. Nous obtenons également une borne de généralisation qui permet de décomposer celle-ci en deux parties. La première partie qui est l'erreur propre à HTERF et une seconde partie qui est ajoutée par l'utilisation de la méthode offset.

En modifiant légèrement le jeu d'hypothèses de HTERF et en ajoutant les nouvelles hypothèses suivantes sur la construction des arbres de HTERF,

Hypothèse 1.3.

1. $Y^t = \tau^t(\mathbf{X}^t)g(\mathbf{W}^t) + \gamma^t(\mathbf{X}^t) + \varepsilon^t$ et $Y^s = \tau^s(\mathbf{X}^s)g(\mathbf{W}^s) + \gamma^s(\mathbf{X}^s) + \varepsilon^s$.
2. $\forall i$ *tel que* $W_i = 1, Y_i^s = f_1^s(X_i^s) + \varepsilon_{1,i}^s, \varepsilon_1^s \perp\!\!\!\perp X^s$ *et* $Y_i^t - f_1^s(X_i^t) = f_1^o(X_i^t) + \varepsilon_{1,i}^t, \varepsilon_1^t \perp\!\!\!\perp X^s, X^t$. ε_1^s *et* ε_1^t *sont des variables aléatoires continues et centrées.*
3. $\forall i$ *tel que* $W_i = 0, Y_i^s = f_0^s(X_i^s) + \varepsilon_{0,i}^s, \varepsilon_0^s \perp\!\!\!\perp X^s$ *et* $Y_i^t - f_0^s(X_i^t) = f_0^o(X_i^t) + \varepsilon_{0,i}^t, \varepsilon_0^t \perp\!\!\!\perp X^s, X^t$. ε_0^s *et* ε_0^t *sont des variables aléatoires continues et centrées.*
4. $B = \mathcal{O}(\sqrt{n})$ et $B = \Theta\left(\frac{\sqrt{n}}{(\ln(n))^\beta}\right)$, avec $\beta > 1$.

5. $\max_{x,\Theta} N_{n,1}(\mathbf{x}; \Theta, \mathcal{D}_n) = o(n)$.

6. $\max_{x,\Theta} N_{n,0}(\mathbf{x}; \Theta, \mathcal{D}_n) = o(n)$.

7. *A chaque étape de la construction d'un arbre, la probabilité que la prochaine division soit faite selon la* $j-eme$ *variable est bornée par en dessous par* π/d *où* $0 < \pi \leq 1$ *pour tout* $j = 1, \ldots, d$.

8. $\mathcal{I}_2$ *véfifie qu'à chaque division une proportion d'au moins* α *de l'échantillon total disponible tel que* $W = 1$ *(resp.* $W = 0$*) va dans chacun des nouveaux noueds issus de la division, où* $0 < \alpha \leq 0.5$.

Nous obtenons alors le résultat suivant de convergence :

Théorème 1.3. *Sous des hypothèses proches de HTERF et 1.3, si de plus pour un* $\beta > \frac{5}{2}$, $C > 0$, *chaque arbre de HTERF est le plus haut tel que* $C\sqrt{n}(\ln n)^{\beta} \leq N_{n,0}(\mathbf{x}; \Theta_\ell, \mathcal{D}_n), N_{n,1}(\mathbf{x}; \Theta_\ell, \mathcal{D}_n)$. *En supposant* $\mathbb{E}\left[\max\left(\varepsilon^s_{1,i}\right)^2\right], \mathbb{E}\left[\max\left(\varepsilon^t_{1,i}\right)^2\right], \mathbb{E}\left[\max\left(\varepsilon^s_{0,i}\right)^2\right], \mathbb{E}\left[\max\left(\varepsilon^t_{0,i}\right)^2\right] \leq K(\ln n)^u$ *avec* $\beta - u > \frac{1}{2}$ *et* K *est une constante positive. En supposant également que* Y *et que le terme d'erreur* E^o *(erreur d'estimation sur* f^o*) sont bornés et que* E^o *converge vers* 0 *dans* L^2. *Alors*

$$\mathbb{E}\left[\left|\hat{\tau}^{new}_{B,n}(\mathbf{X}) - \tau^t(\mathbf{X})\right|\right] \underset{n\to\infty}{\longrightarrow} 0.$$

Enfin, sur des simulations sur des exemples simples avec des jeux de données synthétiques et semi-synthétiques nous avons obtenus des résultats convainquant. La méthode causale offset permet d'obtenir de meilleurs résultats pour l'estimation de CATE que d'apprendre uniquement sur le jeu de données cible de petite taille.

Chapter 2

Heterogeneous Treatment Effect based Random Forest: HTERF

This chapter consists in the article published in the Volume 196 of *Computational Statistics & Data Analysis* [Jocteur *et al.* 2024].

Abstract
Estimates of causal effects are needed to answer what-if questions about shifts in policy, such as new treatments in pharmacology or new pricing strategies for business owners. In this study, a new non-parametric approach is proposed to estimate the heterogeneous treatment effect based on random forests (HTERF). The potential outcome framework with unconfoundedness shows that the HTERF is pointwise almost surely consistent with the true treatment effect. Interpretability results are also presented. A software implementation, `CausalForest` for `Julia` is available on the general repository of `Julia`.

Contents

2.1 Introduction

Automation of decision-making across a wide range of application domains is one of the goals of machine learning. The estimation of the heterogeneous treatment effect or, more specifically, how to determine how an intervention will affect a particular outcome in relation to a variety of observable characteristics of the treated sample present a fundamental challenge in the majority of data-driven personalised decision scenarios. The aim of clinical studies is to evaluate how pharmacological treatments

affect a patient's clinical response in relation to patient variables. This also occurs in empirical research in economics and related fields, where the goal is to determine the impact of realised or hypothetical interventions to assess theories and improve policies.

Two classes of statistical methods can be applied to explore causal inference: metalearners and tree-based methods. The main contribution of this paper is the introduction of a new algorithm for the Conditional Average Treatment Effect (CATE) estimation: the Heterogeneous Treatment Effect based Random Forest (HTERF). This new algorithm uses a random forest with a new splitting criterion specifically designed for binary treatments, which is easier to compute and improved by a preliminary step in the metalearners' spirit (see Section 2.5). Our aim is to emphasise the interpretability of the algorithm. For some simulated examples, we noticed poor representativeness in the Generalized Random forest (GRF) causal forest ([Athey *et al.* 2019]) of the most informative variables. However, for regression random forests, under certain assumptions, the most informative variables appear more often in probabilities in tree construction (see [Scornet *et al.* 2015]). This motivated us to propose a new splitting criterion. In addition, we obtained an almost certain representativity result for the HTERF.

Finally, we compare our approach with previously developed approaches using simulated and semi-synthetic data inspired from those presented in the causal treatment effect literature. We compared the performance of the HTERF algorithm against that of the GRF and found that it dominates both in terms of the CATE RMSE and interpretability in different settings.

The remainder of this paper is organised as follows. Section 2 introduces the potential outcome framework and CATE estimation methods. Section 3 describes the HTERF method, and Section 4 presents the results. Section 5 evaluates the performance of the HTERF. Finally, Section 6 presents our conclusions.

2.2 Inference for treatment effect

This section presents the potential outcome framework and state-of-the-art methods for CATE estimation.

2.2.1 The causal framework

Following the potential outcome framework presented by [Imbens & Rubin 2015], we posit the potential outcomes $Y(1)$ and $Y(0)$ corresponding to the outcome we would have observed had we assigned control or treatment, respectively, to the quantity of interest Y. Assume $Y = Y(W)$, where W is a binary treatment. We also consider a set of covariates $\mathbf{X} \in \mathbb{R}^d$. The conditional average treatment effect (CATE) at $\mathbf{x}$ is defined as follows:

$$\tau(\mathbf{x}) = \mathbb{E}\left[Y(1) - Y(0) | \mathbf{X} = \mathbf{x}\right]. \tag{2.2.1}$$

A standard assumption for the identifiability of CATE is unconfoundedness ([Rosenbaum & Rubin 1983]), meaning that, conditional on $\mathbf{X}$ the treatment assignment W is independent of the potential outcomes for Y :

$$\{Y(1), Y(0)\} \perp\!\!\!\perp W|\mathbf{X}. \tag{2.2.2}$$

We consider n independent and identically distributed training individuals labelled $i = 1, \ldots, n$. Each of them is composed of a feature vector $\mathbf{X}_i \in \mathbb{R}^d$, an outcome $Y_i \in \mathbb{R}$ and a treatment indicator $W_i \in \{0, 1\}$. We denoted the observed data as

$$\mathscr{D}_n = (Y_i, \mathbf{X}_i, W_i)_{1 \leq i \leq n}.$$

The distribution of $\mathscr{D}_n$ is specified by distribution $\mathscr{P}$.

In this study, we are interested in consistent estimators $\hat{\tau}(.)$ of τ. The difficulty in evaluating the function $\tau(.)$ is that we only observe one of the two potential outcomes for a given training example, so we cannot directly train a classical machine learning method on the difference $Y_i(1) - Y_i(0)$.

2.2.2 Methods for causal effect estimation

We categorised the methods used to evaluate the CATE into two groups. On one hand, classical machine learning methods (random forest, boosting, etc.) cannot evaluate CATE directly and are usually called metalearners. On the other hand, there are machine learning methods designed to estimate the CATE directly, such as causal forests or Bayesian regression tree models for causal inference.

A review of metalearners can be found in [Künzel *et al.* 2019]. Metalearner combines base learners in a specific fashion to estimate the CATE. Base learners are supervised learning or regression estimators, but they are not specified in the metalearner.

The T- and S-learners are two basic examples of metalearners. The T-learner estimates Y (1) and Y (0) separately, and the estimated CATE is given by

$$\hat{\tau}_T(\mathbf{x}) = \hat{\mu}_1(\mathbf{x}) - \hat{\mu}_0(\mathbf{x}), \tag{2.2.3}$$

where $\hat{\mu}_1(\mathbf{x})$ (respectively $\hat{\mu}_0(\mathbf{x})$) is an estimator of $\mu_1(\mathbf{x}) = \mathbb{E}[Y(1)|\mathbf{X} = \mathbf{x}]$ (resp. $\mu_0(\mathbf{x}) = \mathbb{E}[Y(0)|\mathbf{X} = \mathbf{x}]$) using the observations in $\{(\mathbf{X}_i, Y_i)\}_{W_i=1}$ (resp. $\{(\mathbf{X}_i, Y_i)\}_{W_i=0}$).

The S-learner uses a single base learner $\hat{\mu}$. It estimates the quantity $\mu(\mathbf{x}, w) = \mathbb{E}[Y|\mathbf{X} = \mathbf{x}, W = w]$ with any base learner on the whole dataset. The CATE estimator is then given by

$$\hat{\tau}_S(\mathbf{x}) = \hat{\mu}(\mathbf{x}, 1) - \hat{\mu}(\mathbf{x}, 0). \tag{2.2.4}$$

These methods allow full control of the estimation method used at each stage. Moreover, they allow cross-validation for additional data-adaptive estimations at

each stage. Hence, they allow the user to perform model selection for both base learners and the final CATE model. However, to obtain good CATE estimations, base learners must match the features of the data and underlying model.

The use of forest-based algorithms to estimate heterogeneous treatment effects has been proposed in the literature. Some of these studies have used the Bayesian Additive Regression Tree (BART) method proposed by [Chipman *et al.* 2010b]; such approaches can be seen in studies by [Hill 2011, Green & Kern 2012, Hill & Su 2013]. Other approaches relying on tree-based methods have been developed that modify the standard random forest algorithm to focus on directly estimating the CATE. These methods are referred to as causal trees or forests. The first approach, using a random forest with a custom splitting criterion, was proposed by [Su *et al.* 2009]. citeathey2016recursive, propose an alternative criterion for causal trees that also allows for the construction of confidence intervals for causal effects. This inspired the causal forest developed by [Wager & Athey 2018]), who introduced the idea of double sampling using one sample to build trees and another for CATE estimation. Finally, GRF causal forests introduced by [Athey *et al.* 2019] are a special case of the previous causal forest.

The GRF approach is a method for non-parametric estimation that applies to a wide variety of quantities of interest, including quantile regression, CATE estimation, and instrumental variable regression. We focus on CATE estimation. We compared our method with the GRF because it improves upon previous methods. The GRF algorithm can be decomposed into two parts: tree growth and quantity of interest estimation. In what follows, the two steps of the GRF causal forest are described in detail.

A random forest, as presented by [Breiman 2001], consists of trees $T_1, \ldots, T_B$. To obtain a prediction for the test point $\mathbf{x}$, this point is pushed down in each tree until it reaches a leaf, and a prediction is associated with each leaf. Let $\hat{\mu}_b$ be the prediction from tree b; then, the random forest prediction for $\mathbf{x}$ is: $\frac{1}{B}\sum_{b=1}^{B}\hat{\mu}_b(\mathbf{x})$.

In GRFs, the strategy is slightly different; the test $\mathbf{x}$ is still pushed down in the trees, but instead of looking for a prediction at each tree, they consider $L_b(\mathbf{x})$ as the set of elements in the training sample that fall into the same leaf as $\mathbf{x}$. For each $i = 1, \ldots, n$, define:

$$\alpha_{b,i}(\mathbf{x}) = \frac{\mathbb{1}\{\mathbf{X}_i \in L_b(\mathbf{x})\}}{|L_b(\mathbf{x})|} \text{ and } \alpha_i(\mathbf{x}) = \frac{1}{B}\sum_{b=1}^{B}\alpha_{b,i}(\mathbf{x}). \tag{2.2.5}$$

The $\alpha_i(\mathbf{x})$ can be seen as a weighting function that indicates how important each training sample is when trying to predict at $\mathbf{x}$. It is straightforward that $\sum_{i=1}^{n}\alpha_i(\mathbf{x}) = 1$.

Once all of the weights have been calculated, CATE is estimated as follows:

$$\hat{\tau}(\mathbf{x}) = \frac{\sum_{i=1}^{n}\alpha_i(\mathbf{x})(W_i - \overline{W_\alpha})(Y_i - \overline{Y_\alpha})}{\sum_{i=1}^{n}\alpha_i(\mathbf{x})(W_i - \overline{W_\alpha})^2}, \tag{2.2.6}$$

where $\overline{W_\alpha} = \sum_{i=1}^{n} \alpha_i(\mathbf{x}) W_i$ and $\overline{Y_\alpha} = \sum_{i=1}^{n} \alpha_i(\mathbf{x}) Y_i$. This estimation step is described in Algorithm 5.

This expression is an empirical version of $\frac{Cov(W,Y|\mathbf{X}=\mathbf{x})}{Var(W|\mathbf{X}=\mathbf{x})}$. A simple computation shows that if W has a linear impact : $Y = \tau(\mathbf{X})W + \gamma(\mathbf{X})$ then $\tau(\mathbf{X}) = \frac{Cov(W,Y|\mathbf{X})}{Var(W|\mathbf{X})}$.

Before studying the splitting criterion used in the GRF, two differences with Breiman random forests can be mentioned: the trees were trained on subsamples of the training data, and a subsampling technique termed honesty was also used. These strategies were used to obtain good theoretical and statistical behaviour. The concept of honesty is to split the training subsample into two subsets before building each tree: the first is used to build the nodes of the tree, and the second is used to fill the tree and will be used to estimate the quantity of interest.

Let P be a parent node, $\mathcal{J}$ be the elements of the sample belonging to P, and C_1 and C_2 be the two child nodes for a given split. A criterion similar to thr CART regression is to minimise:

$$err(C_1, C_2) = \sum_{j=1}^{2} \mathbb{P}(\mathbf{x} \in C_j | \mathbf{x} \in P) \mathbb{E}\left[(\hat{\tau}_{C_j}(\mathcal{J}) - \tau(\mathbf{x}))^2 | \mathbf{x} \in C_j\right], \tag{2.2.7}$$

where $\hat{\tau}_{C_j}(\mathcal{J})$ is the estimation of τ over child nodes C_j.

Unfortunately, the true CATE is unknown and a calculable criterion would be to maximize the following quantity. This favours splits that increase the heterogeneity of the CATE estimates between children. This concept has already been proposed by [Athey & Imbens 2016]:

$$\Delta(C_1, C_2) = \frac{n_{C_1} n_{C_2}}{n_P^2} \left[\hat{\tau}_{C_1}(\mathcal{J}) - \hat{\tau}_{C_2}(\mathcal{J})\right]^2, \tag{2.2.8}$$

where n_{C_1}, n_{C_1} and n_P are the numbers of points that fall into nodes C_1, C_2 and P, respectively.

We present the GRF for the estimation of CATE, but recall that it is applicable to a wide range of quantities of interest. Optimising Equation (2.2.8) over all possible splits means estimating the quantity of interest for both children for each candidate split, which is expensive in terms of complexity in most cases. Instead, gradient-based approximations called pseudo-outcomes were used. For the CATE estimation, the following pseudo-outcomes are computed:

$$\rho_i = A_P^{-1}(W_i - \overline{W}_P)(Y_i - \overline{Y}_P - (W_i - \overline{W}_P)\hat{\beta}_P), \tag{2.2.9}$$

where β_P is the least-squares regression solution of Y_i on W_i and:

$$A_P = \frac{1}{|\{i : X_i \in P\}|} \sum_{i:\mathbf{X}_i \in P} (W_i - \overline{W}_P)^2. \tag{2.2.10}$$

Finally, the chosen split maximises; this is a classic CART regression split over

pseudo-outcomes.

$$\tilde{\Delta}(C_1, C_2) = \frac{1}{n_{C_1}} \left(\sum_{i:\mathbf{X}_i \in C_1} \rho_i \right)^2 + \frac{1}{n_{C_2}} \left(\sum_{i:\mathbf{X}_i \in C_2} \rho_i \right)^2 . \tag{2.2.11}$$

This tree building step is applied in Algorithm 4.

Algorithm 4 Building random forest algorithm

Input: $B > 0$ number of trees, s subsampling rate, $\mathcal{S}$ set of examples
  for $b = 1$ to B **do**
    set of examples $\mathcal{I}_b \leftarrow$ SUBSAMPLE(S, s)
    ▷ Draw a subsample from $\mathcal{S}$ without replacement of size $s|\mathcal{S}|$
    sets of examples $\mathcal{I}_{b,1}, \mathcal{I}_{b,2} \leftarrow$ SPLITSAMPLE$(\mathcal{I}_b)$
    ▷ Randomly divides a set into two evenly sized, non-overlapping halves
    node $P_0 \leftarrow$ CREATENODE$(\mathcal{I}_{b,1})$
    queue $\mathcal{Q} \leftarrow$ INITIALISEQUEUE(P_0)
    while NOTNULL(node $P \leftarrow$ POP$(\mathcal{Q})$) **do**
      vector $R_P \leftarrow$ GETPSEUDOOUTCOMES(P) ▷ Computes (2.2.9)
      split $\Sigma \leftarrow$ MAKECARTSPLIT(P, R_P) ▷ Optimises (2.2.11)
      if SPLITSUCCEEDED **then** ▷ If there is a legal split
        SETCHILDREN$(P,$ GETLEFTCHILD$(\Sigma),$ GETRIGHTCHILD$(\Sigma))$
        ADDTOQUEUE$(\mathcal{Q},$ GETLEFTCHILD$(\Sigma))$
        ADDTOQUEUE$(\mathcal{Q},$ GETRIGHTCHILD$(\Sigma))$
      end if
    end while
    ▷ Tree $\mathcal{T}_b$ has been built
  end for
Output: A causal forest with trees $\mathcal{T}_1, \ldots, \mathcal{T}_B$

In practice, prior centring is applied before running the algorithm, which involves regressing out the effect of feature $\mathbf{X}_i$ on W_i and Y_i separately. This improves performance on finite datasets.

Remark 2.1. *The pseudo-outcomes have been introduced for easier calculations; however, for CATE estimation, we obtain the same algorithmic complexity when computing* $\Delta(C_1, C_2)$ *or* $\tilde{\Delta}(C_1, C_2)$.

In this section, the GRF algorithm is described. This motivated the introduction of a new forest-based method.

2.2.3 Limitations of the GRF approach

Random forests are effective regression algorithms and are quite interpretable under the assumption that Y follows an additive regression model. [Scornet *et al.* 2015]

Algorithm 5 Estimation algorithm

Input: A causal forest with trees $\mathcal{T}_1, \ldots, \mathcal{T}_B$, a test point $\mathbf{x}$, the size of training set n.

weight vector $\alpha \leftarrow \textsc{Zeros}(n)$ ▷ Create a vector of zeros of length n

for $b = 1$ to B **do**

$\mathcal{N} \leftarrow \textsc{Neighbours}(\mathbf{x}, \mathcal{T}_b, \mathcal{I}_{b,2})$

▷ Elements of $\mathcal{I}_{b,2}$ that fall into the same leaf as $\mathbf{x}$ in the tree $\mathcal{T}_b$

for all example $e \in \mathcal{N}$ **do**

$\alpha[e] \mathrel{+}= \frac{1}{|\mathcal{N}|}$

end for

end for

$\alpha = \alpha / B$

Output: $\hat{\tau}(\mathbf{x})$ ▷ Uses (2.2.6)

proved that the algorithm selects splits mostly along informative variables. In simple linear regression examples, only significant variables are present in the first stages of the regression forest. In contrast, when we consider a simple linear causal example, the overrepresentation of informative variables is not as striking(see Table 2.3), which limits the interpretability of GRF causal forests. This is one of the motivations for proposing a new splitting criterion for causal forests.

The GRF approach is a general framework not tailored for causal inference. The HTERF splitting criterion is easier to compute as there are only four means to do so, while GRF needs to compute an OLS solution for each parent node and the calculations of the pseudo-outcomes are more complex to compute and less robust than means. Furthermore, in the GRF, the final splitting criterion uses an approximation using pseudo-outcomes, which results in less precision. Our HTERF splitting criterion is specifically designed to assess the CATE when the treatment is binary and can be adapted for multiple discrete treatments.

2.3 Estimation of causal effect with HTERF

The splitting criterion used in the HTERF is based on the idea of maximising the difference in the treatment effects between child nodes. An empirical version of Equation (2.2.1) is used to define the splitting criterion.

2.3.1 Algorithm

We assume that we are given a training sample $\mathcal{D}_n = (Y_j, \mathbf{X}_j, W_j)_{j=1,\ldots,n}$ of independent random variables distributed as a prototype triple $(Y, \mathbf{X}, W)$, which is a $(d+2)$-dimensional random vector. The purpose is to use the dataset $\mathcal{D}_n$ to construct an estimator $\hat{\tau}_{B,n} : \mathcal{X} \to \mathbb{R}$ of τ.

The tree-building process of the HTERF is as follows. Prior to the construction

of each tree, subsampling and honest splitting were performed, as in the GRF. The splitting criterion was optimised over a subset of features $\mathcal{M}_{try}$. The features were randomly selected with a positive probability for each covariate to be selected, which included uniform selection. Then the best split is the one maximising the splitting criterion $\Delta(A, j, z)$, where $A = \prod_{i=1}^{d}[a_i, b_i]$ is the current node, j is chosen in $\mathcal{M}_{try}$ and $z \in A^j = [a_j, b_j]$.

$$\Delta(A, j, z) = \frac{|A_L||A_R|}{|A|^2}\left(\left(\overline{Y}_{A_{L1}} - \overline{Y}_{A_{L0}}\right) - \left(\overline{Y}_{A_{R1}} - \overline{Y}_{A_{R0}}\right)\right)^2, \tag{2.3.1}$$

where $A_{L1} = \{\mathbf{X}_i \in A | \mathbf{X}_i^{(j)} < z, W_i = 1\}$, $A_{L0} = \{\mathbf{X}_i \in A | \mathbf{X}_i^{(j)} < z, W_i = 0\}$, $A_{R1} = \{\mathbf{X}_i \in A | \mathbf{X}_i^{(j)} \geq z, W_i = 1\}$, $A_{R0} = \{\mathbf{X}_i \in A | \mathbf{X}_i^{(j)} \geq z, W_i = 0\}$, $A_L = A_{L1} \cup A_{L0}$ and $A_R = A_{R1} \cup A_{R0}$. For all sets B, we denote $\overline{Y}_B = \frac{1}{|B|}\sum_{i \in B} Y_i$. This splitting criterion was partially inspired by those used by [Athey & Imbens 2016] and [Athey *et al.* 2019].

For the estimation of τ, we reuse the GRF procedure with Algorithm 5, but the estimation is different, namely:

$$\hat{\tau}_{B,n}(\mathbf{x}) = \sum_{i:W_i=1} \alpha_i(\mathbf{x}) Y_i - \sum_{i:W_i=0} \alpha'_i(\mathbf{x}) Y_i, \tag{2.3.2}$$

where α (resp. α') is the weight vector defined as in Algorithm 5 associated with observations such that $W_i = 1$ (resp. $W_i = 0$).

We use the following notations:

- $\Theta_\ell, \ell = 1, \ldots, B$ are independent random vectors, distributed as a generic random vector $\Theta = \left(\Theta^1, \Theta^2, \Theta^3\right)$ and independent of $\mathcal{D}_n$, and $\left(\Theta^1, \Theta^2\right)$ is independent of Θ^3. Θ^1 contains indices of observations that are used to build each tree. That is, the subsample $\mathcal{I}_1$, Θ^2 contains indices of observations that are used for estimations in each tree; namely, the subsample $\mathcal{I}_2$ and Θ^3 contains indices of splitting candidate variables in each node. We assume that Θ^3 gives a positive probability to each co-variate. We must consider both Θ^1 and Θ^2 because $\mathcal{I}_2$ is the complementary of $\mathcal{I}_1$ in $\mathcal{I}$ which is random itself.

- $\mathcal{D}^\star_{n,1}(\Theta_\ell)$ and $\mathcal{D}^\star_{n,2}(\Theta_\ell)$ are the disjoint subsamples selected prior to tree construction; the first is used to build the tree, and the second allows the building of weights used during the estimation step.

- $A_n(\mathbf{x}; \Theta_\ell, \mathcal{D}_n)$ is the tree cell (subspace of $\mathcal{X}$) containing $\mathbf{x}$.

- $N_{n,1}(\mathbf{x}; \Theta_\ell, \mathcal{D}_n)$ (resp. $N_{n,0}(\mathbf{x}; \Theta_\ell, \mathcal{D}_n)$) is the number of elements of $\mathcal{D}^\star_{n,2}(\Theta_\ell)$ that fall into $A_n(\mathbf{x}; \Theta_\ell, \mathcal{D}_n)$, such that $W_i = 1$ (resp. $W_i = 0$).

We define the weights:

$$\alpha_i(\mathbf{x}) = \frac{1}{B}\sum_{l=1}^{B} \frac{\mathbb{1}_{\mathbf{X}_i \in A_n(\mathbf{x}; \Theta_l, \mathcal{D}_n) \wedge W_i = 1 \wedge i \in \mathcal{D}^\star_{n,2}(\Theta_\ell)}}{N_{n,1}(\mathbf{x}; \Theta_\ell, \mathcal{D}_n)}, \tag{2.3.3}$$

$$\alpha_i'(\mathbf{x}) = \frac{1}{B}\sum_{l=1}^{B}\frac{\mathbb{1}_{\mathbf{X}_i\in A_n(\mathbf{x};\Theta_l,\mathcal{D}_n)\wedge W_i=0\wedge i\in\mathcal{D}^\star_{n,2}(\Theta_\ell)}}{N_{n,0}\left(\mathbf{x};\Theta_\ell,\mathcal{D}_n\right)}. \tag{2.3.4}$$

Remark 2.2. *The output $\hat{\tau}(x)$ can also be seen as an average of estimations obtained by several causal trees (as in a Breiman random forest).*

$$\begin{aligned}\hat{\tau}_{B,n}(\mathbf{x}) = &\sum_{i:W_i=1}\frac{1}{B}\sum_{l=1}^{B}\frac{\mathbb{1}_{\mathbf{X}_i\in A_n(\mathbf{x};\Theta_l,\mathcal{D}_n)\cap i\in\mathcal{D}^\star_{n,2}(\Theta_\ell)}}{N_{n,1}\left(\mathbf{x};\Theta_\ell,\mathcal{D}_n\right)}Y_i\\ &-\sum_{i:W_i=0}\frac{1}{B}\sum_{l=1}^{B}\frac{\mathbb{1}_{\mathbf{X}_i\in A_n(\mathbf{x};\Theta_l,\mathcal{D}_n)\cap i\in\mathcal{D}^\star_{n,2}(\Theta_\ell)}}{N_{n,0}\left(\mathbf{x};\Theta_\ell,\mathcal{D}_n\right)}Y_i\\ =&\frac{1}{B}\sum_{l=1}^{B}\sum_{\substack{i\in\mathcal{D}^\star_{n,2}(\Theta_\ell)\\ W_i=1\\ \mathbf{X}_i\in A_n(\mathbf{x};\Theta_l,\mathcal{D}_n)}}\frac{Y_i}{N_{n,1}\left(\mathbf{x};\Theta_\ell,\mathcal{D}_n\right)}\\ &-\frac{1}{B}\sum_{l=1}^{B}\sum_{\substack{i\in\mathcal{D}^\star_{n,2}(\Theta_\ell)\\ W_i=0\\ \mathbf{X}_i\in A_n(\mathbf{x};\Theta_l,\mathcal{D}_n)}}\frac{Y_i}{N_{n,0}\left(\mathbf{x};\Theta_\ell,\mathcal{D}_n\right)}.\end{aligned}$$

Following the `R` package `grf`, which provides an implementation of the [Athey *et al.* 2019] algorithm, we define the importance of variables. Let $freq$ be the matrix of split depth by feature index; $freq_{i,j}$ is the number of times (over the forest) the split has been performed along X_i at depth j divided by the total number of splits at depth j. This is the frequency of splits for each feature at a given depth. The importance of a feature can be defined as the weighted sum of the number of times the feature is split at each depth in the forest. It depends on two parameters: max_depth, the maximum depth considered to get the $freq$ matrix and *decay*, the decay exponent, which controls the importance of the split depth.

We now define the importance of the ith feature as:

$$Imp_i(max\_depth, decay) = \frac{\sum_{k=1}^{max\_depth} freq_{k,i}k^{-decay}}{\sum_{k=1}^{max\_depth}k^{-decay}}. \tag{2.3.5}$$

In numerical applications, the values used for max_depth and *decay* are 4 and 2, respectively.

2.3.2 Theoretical tree

Similar to what [Scornet *et al.* 2015] did, a random theoretical tree can be defined for the HTERF. The theoretical equivalent of the empirical HTERF splitting criterion for node A is:

$$\begin{aligned}\Delta^\star(A,j,z) =& \mathbb{P}[\mathbf{X}^{(j)} < z|\mathbf{X}\in A]\mathbb{P}[\mathbf{X}^{(j)} \geq z|\mathbf{X}\in A] \\ & (\mathbb{E}[Y(1)-Y(0)|\mathbf{X}^{(j)} < z, \mathbf{X}\in A] \\ & -\mathbb{E}[Y(1)-Y(0)|\mathbf{X}^{(j)} \geq z, \mathbf{X}\in A])^2.\end{aligned} \tag{2.3.6}$$

A theoretical tree is obtained by using the best consecutive cuts $(j^\star, z^\star)$ optimising the previous criterion $\Delta^\star(A,.,.)$.

2.4 Consistency of HTERF

2.4.1 Existing results

The following assumptions are made in order to obtain consistency results on HTERF:

- Unconfoundedness as in Equation (2.2.2)
- Honesty: Two different samples are used to construct splits and predict the labels.
- The resampling is done by subsampling and not by bootstrapping as in Breiman forests.

With Remark 2.2, we can see that $\hat{\tau}_{B,n}$ is a U-statistic and, under additional assumptions below a normality result for $\hat{\tau}_{B,n}$ follows from [Wager & Athey 2018].

Assumption 2.1.

1. $\mathbf{X}$ *is a uniform random vector with independent coordinates:* $\mathbf{X} \sim U([0,1]^d)$.
2. $(\mathbf{X}, Y(u))$ *with* $u \in \{0,1\}$ *verifies* $\mathbf{x} \mapsto \mathbb{E}[Y(u)|\mathbf{X}=\mathbf{x}]$ *and* $\mathbf{x} \mapsto \mathbb{E}[Y(u)^2|\mathbf{X}=\mathbf{x}]$ *are Lipschitz-continuous,* $Var[Y(u)|\mathbf{X}=\mathbf{x}] > 0$ *and* $\mathbb{E}[|Y(u)-\mathbb{E}[Y(u)|\mathbf{X}=\mathbf{x}]|^{2+\delta}|\mathbf{X}=\mathbf{x}] \leq M$ *for some constants* $\delta, M > 0$ *uniformly over all* $\mathbf{x} \in [0,1]^d$.
3. *At every step of the tree building procedure, the probability that the next split is done along the* $j-th$ *feature is bounded below by* π/d *for some* $0 < \pi \leq 1$ *for all* $j = 1, \ldots, d$.
4. *A causal tree is* $\alpha-$*regular at* $\mathbf{x}$ *for some* $\alpha > 0$ *if the sample* $\mathcal{I}_2$ *used for estimation verifies the following: (1) each split leaves at least a fraction* α *of the available training sample on each side of the split, (2) the leaf containing* $\mathbf{x}$ *has at least* k *observations for each treatment group for some* $k \in \mathbb{N}$, *and (3) the leaf containing* $\mathbf{x}$ *has either less than* $2k-1$ *observations with* $W_i = 0$ *or* $2k-1$ *observations with* $W_i = 1$.
5. *The subsample size* s_n *scales as:* $s_n \asymp n^\beta$ *for some* $\beta_{min} := 1 - \left(1 + \frac{d}{\pi}\frac{\log(\alpha^{-1})}{\log((1-\alpha)^{-1})}\right) < \beta < 1$.

Theorem ([Athey *et al.* 2019]). *Under Assumptions 2.1 and considering that* $Y = \tau(\mathbf{X})W + \gamma(\mathbf{X}) + \varepsilon$, *we have the following:*

$$\frac{\hat{\tau}_{B,n}(\mathbf{x}) - \tau(\mathbf{x})}{\sqrt{Var(\hat{\tau}(\mathbf{x}))}} \longrightarrow N(0,1),$$

where the variance of the causal forest can be consistently estimated using the infinitesimal jackknife for random forests (see [Wager & Athey 2018] for more details).

2.4.2 New consistency results

We propose a consistency result under assumptions weaker than Assumptions 2.1 based on [Elie-Dit-Cosaque & Maume-Deschamps 2022]. In what follows, $\mathcal{X}$ is a compact hyper-rectangle of $\mathbb{R}^d$: $\mathcal{X} = \prod_{i=1}^{d}[u_i, v_i]$, $-\infty < u_i \leq v_i < \infty$, and we denote by $\mathcal{A}$ the set of hyper-rectangles in $\mathcal{X}$: $A \in \mathcal{A}$ writes $A = \prod_{i=1}^{d}[a_i, b_i]$ with $u_i \leq a_i \leq b_i \leq v_i$. Additionally, we denote $A^{-j} = \prod_{k\neq j}[a_k, b_k]$ and $A^J = \prod_{k\in J}[a_k, b_k]$ for any $J \subset \{1, \ldots, d\}$. Given that $\mathbf{x} \in \mathbb{R}^d$, $\mathbf{x}^{-j}$ is the vector of $\mathbb{R}^{d-1}$ where the j-th coordinate has been removed and $\mathbf{x}^J$ is the vector of $\mathbb{R}^J$ for which the coordinates are $x^{(j)}$, $j \in J$.

Definition 2.1. *Let* $f : \mathcal{X} \to \mathbb{R}$; *it does NOT belong to the* ♠*-class if there exists a rectangle* $A = \prod_{j=1}^{d}[a_j, b_j] \subset \mathcal{X}$, *with* $a_j \leq b_j$ *such that for all* $j = 1, \ldots, d$, $z \mapsto \mathbb{E}\left[f\left(z, \mathbf{X}^{-j}\right)\mathbb{1}_{\{\mathbf{X}^{-j}\in A^{-j}\}}\right]$ *is constant on* $[a_j, b_j]$, *and* f *is not constant on* A.

Remark 2.3. *The* ♠*-class contains many functions, such as additive and multiplicative functions. A more elaborate list can be found in [Elie-Dit-Cosaque & Maume-Deschamps 2022]. A noteworthy example is the set of linear combinations of Gaussian radial basis functions on* $[0,1]^d$, *with positive weights:*

$$\mathcal{G} = \left\{ \sum_{i=1}^{p} a_i \exp[\sum_{j=1}^{d}(x_j - \mu_j)^2\sigma_j^2],\ a_i \geq 0,\ \sigma_j \geq 0,\ \mu_j \in \mathbb{R} \right\}.$$

It is known that the class $\mathcal{G}$ *is dense in the set of non-negative continuous functions on* $[0,1]^d$ *(see [Park & Sandberg 1991] and [Klusowski 2019] where class* $\mathcal{G}$ *is also considered to study CART).*

Let $\tau_1(\mathbf{x}) = \mathbb{E}[Y(1)|\mathbf{X} = \mathbf{x}]$ and $\tau_0(\mathbf{x}) = \mathbb{E}[Y(0)|\mathbf{X} = \mathbf{x}]$, and, similarly, $\hat{\tau}_1(\mathbf{x}) = \sum_{i:W_i=1}\alpha_i(\mathbf{x})Y_i$ and $\hat{\tau}_0(\mathbf{x}) = \sum_{i:W_i=0}\alpha'_i(\mathbf{x})Y_i$. We make the following assumptions:

Assumption 2.2.

- $Y = \tau(\mathbf{X})g(\mathbf{W}) + \gamma(\mathbf{X}) + \varepsilon$.
- $\mathbf{X} = (X_1, \ldots, X_d)$ *is a continuous random vector with independent coordinates.*
- ε *and* $\mathbf{X}$ *are independent, and* ε *is a continuous, centered random variable with increasing distribution function and light tails. That is, there exists* $0 < \theta < 1$ *such that for any* $D > 0$, $\mathbb{P}(|\varepsilon| > D) \leq C\theta^D$.
- $\mathbf{X}$ *takes its values in* $\mathcal{X}$*, which is assumed to be a compact hyper-rectangle of* $\mathbb{R}^d$*:* $\mathcal{X} = \prod_{i=1}^{d} [u_i, v_i]$, $-\infty < u_i \leq v_i < \infty$.
- $\mathbf{x} \mapsto \gamma(\mathbf{x})$, $\mathbf{x} \mapsto \tau_1(\mathbf{x})$ *and* $\mathbf{x} \mapsto \tau_0(\mathbf{x})$ *are continuous. Thus, in particular,* $\mathbf{x} \mapsto \tau(\mathbf{x})$ *is continuous.*

Remark 2.4. *$g(W) = W$ corresponds to the linear treatment effect considered in [Athey et al. 2019]. As noticed in [Hill 2011], non-linear functions have practical interest.*

We denote the following:
$f(n) = \Omega(g(n)) \iff \exists k > 0, \exists n_0 > 0 \mid \forall n \geqslant n_0 \quad |f(n)| \geqslant k \cdot |g(n)|$

Assumption 2.3. *The following assumptions are made on B (number of trees) and $N_{n,1}(\mathbf{x}; \Theta, \mathcal{D}_n)$ resp. $N_{n,0}(\mathbf{x}; \Theta, \mathcal{D}_n)$ (number of observations in a leaf node such that $W = 1$, resp. $W = 0$):*

1. $B = \mathcal{O}(n^\alpha)$, with $\alpha > 0$.
2. $\mathbb{E}[N_{n,1}(\mathbf{x}; \Theta, \mathcal{D}_n)] = \Omega\left(\sqrt{n}(\ln(n))^\beta\right)$, with $\beta > 1$.
3. $\mathbb{E}[N_{n,0}(\mathbf{x}; \Theta, \mathcal{D}_n)] = \Omega\left(\sqrt{n}(\ln(n))^\beta\right)$.

Remark 2.5. *Items 2 and 3 in Assumption 2.3 are easier to verify than Assumption 2.1 because the number of observations in leaves can be controlled as a standard construction parameter of trees of the forest.*

Theorem 2.1. *Let Assumptions 2.2 and 2.3 be verified, with τ_1 and τ_0 belonging to the ♠-class; assume that for fixed $\beta > \frac{5}{2}$, $C > 0$, each constructed tree is the highest, such that $C\sqrt{n}(\ln n)^\beta \leq N_{n,0}(\mathbf{x}; \Theta_\ell, \mathcal{D}_n), N_{n,1}(\mathbf{x}; \Theta_\ell, \mathcal{D}_n)$. Also assume that $\mathbb{E}[\max \varepsilon_i^2] \leq K(\ln n)^u$ with $\beta - u > \frac{1}{2}$ and K is a positive constant. Then,*

$$\forall \mathbf{x} \in \mathcal{X}, |\hat{\tau}_{B,n}(\mathbf{x}) - \tau(\mathbf{x})| \xrightarrow[n\to\infty]{a.s.} 0.$$

Remark 2.6. *The property on $\mathbb{E}[\max \varepsilon_i^2]$ is verified for subgaussian distributions ([Boucheron et al. 2013]).*

The proof follows the lines of [Elie-Dit-Cosaque & Maume-Deschamps 2022], and the main steps are described in Appendix A.

2.4.3 Interpretability

Using Proposition A.4, we can state an almost surely version of Proposition 1 in [Scornet *et al.* 2015] (which gives an interpretability result in probability). Proposition A.4 gives that for any $h \in \mathbb{N}$ and any empirical tree $\mathcal{T}_e$ satisfying Assumption 2.2 and the upper bounds on $N_{n,0}(\mathbf{x}; \Theta_\ell, \mathcal{D}_n), N_{n,1}(\mathbf{x}; \Theta_\ell, \mathcal{D}_n)$ in Theorem 2.1, there exists a theoretical tree $\mathcal{T}_h$ as close as wanted to $\mathcal{T}_e$ until height h.

We can define the informative variables as the set of variables upon which τ depends. Given that τ belongs to the ♠-class, in the theoretical tree, the splits are only made along informative variables. The theoretical splitting criterion equals zero along non-informative variables. Denoting $inf \subseteq \{1, \dots, d\}$ the set of indices of informative variables, we have $\tau(\mathbf{X}) \perp\!\!\!\perp \mathbf{X}^{-inf}$. Thus, up to height h, empirical cuts are performed along the same coordinates as the theoretical tree $\mathcal{T}_h$. The following result is thus straightforward.

Theorem 2.2. *Assume that Assumption 2.2 is verified and set $|\mathcal{M}_{try}| = d$, let $h \in \mathbb{N}$. Assume that τ belongs to the ♠-class and that τ is non constant in every node up to height h. Assume that for fixed $\beta > \frac{5}{2}$, $C > 0$, each constructed tree is the highest, such that $C\sqrt{n}(\ln n)^\beta \leq N_{n,0}(\mathbf{x}; \Theta_\ell, \mathcal{D}_n), N_{n,1}(\mathbf{x}; \Theta_\ell, \mathcal{D}_n)$. Then for n large enough, all the cuts in an empirical tree up to height h are made along informative variables almost surely.*

In the following section, examples of these good interpretability results are presented in the form of numerical tables and boxplots.

Remark 2.7.

- *By fixing $|\mathcal{M}_{try}| = d$ the only source of randomness in HTERF trees comes from the subsampling step before building each tree.*
- *If $|\mathcal{M}_{try}| = k < d$, considering a probability distribution on Mtry the expected frequency among theoretical trees of splitting along informative variables is the probability of the set inf. Using the proximity result between empirical and theoretical trees Proposition A.4, one can prove that the average frequency of splitting along informative variables in the empirical causal forest, converges to the probability of inf.*

2.5 Simulations results

First, we consider a simulation where $\gamma(W) = W$. Then, we consider a non-linear case. We examine the same examples as [Athey *et al.* 2019]. In a fourth example, a modified version of the HTERF is considered where the term γ is linear. Finally, to assess interpretability, we study an example close to the Ishigami function adapted to a causal perspective.

In practice, a preliminary step called centring is applied. We estimate the quantity $\mu_0(\mathbf{x}) = \mathbb{E}[Y(0)|\mathbf{X} = \mathbf{x}]$ with $\hat{\mu}_0$ on observations such that $W_i = 0$. Then, we consider the quantity $Y_i^e = Y_i - \hat{\mu}_0(\mathbf{X}_i)$. In the `Julia` implementation of the HTERF, we

use a cross-validated regression random forest to get $\hat{\mu}_0$. Cross-validation is used to optimise hyperparameters such as the minimum sample size in nodes and leaves and the value of $|\mathcal{M}_{try}|$. Similar to the fact that any algorithm can be used for metalearners, we can use any algorithm for centring. The building and estimation steps of the HTERF were performed with the centred data Y_i^e. Below is an example of the motivation for prior centring. Let $Y = \tau(\mathbf{X})W + \gamma(\mathbf{X})$ where $\tau(\mathbf{X}) = X^{(1)}$ and $\gamma(\mathbf{X}) = X^{(2)}$. Assume that γ is perfectly known when calculating Y^e. Consider the following data set in Table 2.1.

$X^{(1)}$	$X^{(2)}$	W	Y	Y^e
5	5	1	10	5
5	5	0	5	0
10	5	1	15	10
10	10	1	20	10
10	10	0	10	0

Table 2.1: Motivational example for centring

With no centring, the criterion along $X^{(1)}$ is

$$\frac{2 \times 3}{5^2}\left((10-5) - \left(\frac{15+20}{2} - 10\right)\right)^2 = 1.5. \tag{2.5.1}$$

The criterion along $X^{(2)}$ is

$$\frac{2 \times 3}{5^2}\left(\left(\frac{10+15}{2} - 5\right) - (20-10)\right)^2 = 1.5. \tag{2.5.2}$$

The criterion is equal for both covariates; therefore, none of them seem more informative, which is unfortunate because only $X^{(1)}$ is informative.
However, if we consider the criterion with the centred outcome Y^e, we obtain the following criterion along $X^{(1)}$:

$$\frac{2 \times 3}{5^2}\left((5-0) - \left(\frac{10+10}{2} - 0\right)\right)^2 = 6. \tag{2.5.3}$$

The criterion along $X^{(2)}$ is

$$\frac{2 \times 3}{5^2}\left(\left(\frac{5+10}{2} - 0\right) - (10-0)\right)^2 = 1.5. \tag{2.5.4}$$

With a centred outcome, the criterion is larger when splitting along $X^{(1)}$ as intended.

Remark 2.8. *If the observations are unbalanced regarding the treatment distribution with substantially fewer untreated cases, then the estimator could not be as good as expected. In practice, when more than 55% of the observations are treated, we estimate the quantity $\mu_1(\mathbf{x}) = \mathbb{E}[Y(1)|\mathbf{X} = \mathbf{x}]$ with an estimator $\hat{\mu}_1$ trained on observations such that $W_i = 1$. Then, we define the quantity $Y_i^e = Y_i - \hat{\mu}_1(\mathbf{X}_i)$ and proceed as previously. In this case, the quantity $-\tau$ that is estimated instead of τ.*

In the following examples, the HTERF is compared with the state-of-the-art causal forest method (GRF) using the `R` package `grf` hosted in the CRAN repository, which can be found at https://cran.r-project.org/web/packages/grf/index.html. We also carry out a comparison with two metalearners, namely, the T-learner previously presented and the R-learner introduced by [Nie & Wager 2021] and implemented in the `R` package `rlearner` hosted at the following address: https://github.com/xnie/rlearner. The version using XGBoost as the base learner was retained for these examples. We used 500 trees and the hyperparameters specified by [Nie & Wager 2021]. The concept of interpretability is common to all causal forest methods (GRF and HTERF) but cannot be generalised to metalearners.

2.5.1 First example

We consider the simulated data to be close to the previously studied causal frameworks ([Athey *et al.* 2019]). Let $\mathbf{X}_i \sim U([0,1]^p)$, $W_i \sim Bern(0.5)$, and $Y_i = \tau(\mathbf{X}_i)W_i + \beta\gamma(\mathbf{X}_i)$ where $p = 10, \tau(\mathbf{x}) = \sin(x^{(1)})$ and $\gamma(\mathbf{x}) = \cos(2x^{(2)} + 3x^{(3)})$. The underlying model for Y follows the causal framework presented by [Athey *et al.* 2019], supporting the unconfounding hypothesis. The scalar β allows for consideration of the impact of the magnitude of τ relative to γ.

β	GRF	HTERF	T-learner	R-learner
5	0.278	0.117	15.263	0.425
1	0.121	0.012	1.672	0.169
0.2	0.079	0.004	0.324	0.114

Table 2.2: Mean squared errors of the GRF, HTERF, T-learner, and R-learner methods that estimate the heterogeneous treatment effect. GRF and HTERF causal forests have 500 trees, both the centring forests for the GRF and the forest of the first step in the HTERF have 500 trees, and the same subsampling rate (0.7) is used in both methods. The results are aggregated over 60 simulation replications with 1000 test points each. The mean square errors are multiplied by 1000.

We notice the influence of the order magnitude of the $\beta\gamma$ term relative to τ. The bigger β is, the larger the RMSE is. This is true for both GRF, HTERF, and metalearners. We also notice that for a small β, the gain of the HTERF relative to GRF is more significant in term of RMSE. In all configurations, the metalearners are not competitive.

τ only depends on the variable $X^{(1)}$, so it is expected that for small depths, splits should be performed only for this variable. In addition, the importance of $X^{(1)}$ should be high. These results are clearer for HTERF than GRF (see Table 2.3).

Up to a depth of 3, the split frequency is better overall for the HTERF in all configurations, and there is more volatility for the GRF split frequency (see Figure 2.1). Up to a depth of 5, we have the same conclusion for $\beta = 1$ and $\beta = 0.2$. At a

	GRF				HTERF			
β	dep.3	dep.5	dep.10	imp.	dep.3	dep.5	dep.10	imp.
5	0.870	0.378	0.150	0.852	1	0.498	0.175	0.985
1	0.874	0.526	0.174	0.866	1	0.995	0.282	1
0.2	0.875	0.627	0.2	0.866	1	1	0.603	1

Table 2.3: Frequencies of splitting on $X^{(1)}$ at depths of 3, 5 and 10 and the importance of $X^{(1)}$.

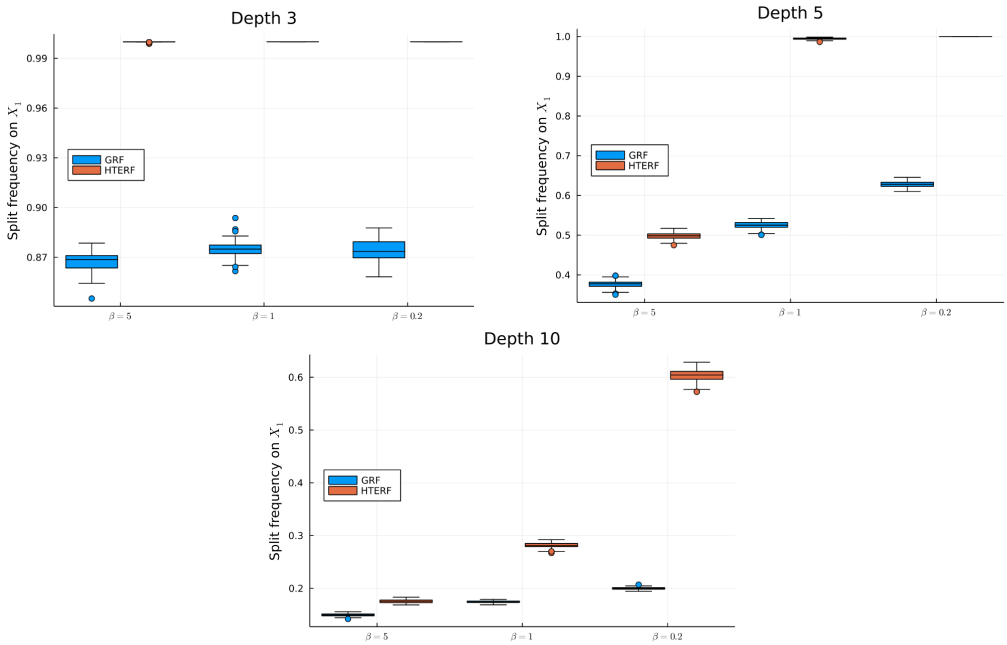


Figure 2.1: Boxplots comparing the split frequency on $X^{(1)}$ in HTERF and GRF trees, for each β, and up to depths of 3, 5 and 10. Each forest consists of 500 trees and, the results are aggregated over 60 simulations, as described above.

high depth and when the γ term has a bigger magnitude than the τ term, the edge of the HTERF over the GRF is less striking.

These observations regarding the relative magnitude of τ highlight the importance of the quality of fit of the model in the first step of the HTERF. To illustrate this, we considered a similar simulation in which the γ term is much simpler to estimate in Section 2.5.4.

2.5.2 Non-linear framework

The BART algorithm ([Hill 2011]) allows for the estimation of the CATE in a more general context where $Y = f(W, X)+\varepsilon$ with ε being normal iid. In [Athey *et al.* 2019], only the case where the relationship between Y and W is linear is considered, which

is not the case in the HTERF. This section aims to illustrate the performance of the HTERF in a nonlinear case.

Let $\mathbf{X} \sim U([0,1]^p)$, $W \sim Bern(0.5)$ and $Y = \sin(X^{(1)})(W+2)^3 + \cos(X^{(2)})$, where $p = 3$. Hence, we have a CATE that satisfies: $\tau(\mathbf{x}) = 19\sin(x^{(1)})$.

Method	RMSE	importance
GRF	0.303	0.792
HTERF	0.099	1
T-learner	0.185	/
R-learner	0.353	/

Table 2.4: Root mean squared errors of the GRF, HTERF, T-learner, and R-learner methods that estimate the heterogeneous treatment effect. GRF and HTERF causal forests have 500 trees, both the centring forests for the GRF and the forest of the first step in the HTERF have 500 trees, and the same subsampling rate (0.7) is used in both methods. The results are aggregated over 60 simulation replications with 1000 test points each. We also consider the importance of $X^{(1)}$.

The HTERF performed better than the GRF in terms of estimation (lower RMSE for the HTERF) and interpretability (Table 2.4). $X^{(1)}$ is the only informative variable; therefore, its importance is expected to be 1. Metalearners do not perform as well as the HTERF.

2.5.3 GRF example

To illustrate the performance of the HTERF, we used a simulation by [Athey *et al.* 2019]. Let $\mathbf{X}_i \sim U([0,1]^p)$, $W_i|\mathbf{X}_i \sim Bern(e(\mathbf{X}_i))$, and $Y_i|\mathbf{X}_i, W_i \sim N(m(\mathbf{X}_i) + (W_i - 0.5)\tau(\mathbf{X}_i), 1)$, where $p = 10$ or $p = 20$ depending on the simulation considered. The authors considered the following three settings:

- No confounding, $m(\mathbf{x}) = 0$, and $e(\mathbf{x}) = 0.5$, but treatment heterogeneity $\tau(\mathbf{x}) = \varsigma(x^{(1)})\varsigma(x^{(2)})$ where $\varsigma(u) = 1 + 1/(1 + e^{-20(u-1/3)})$.
- Confounding, $e(\mathbf{x}) = \frac{1}{4}(1 + \beta_{2,4}(x^{(3)}))$, and $m(\mathbf{x}) = 2x^{(3)} - 1$, where $\beta_{a,b}$ is beta distribution with shape parameters a and b but no treatment heterogeneity, $\tau(\mathbf{x}) = 0$.
- Both confounding $e(\mathbf{x}) = \frac{1}{4}(1 + \beta_{2,4}(x^{(3)}))$ and $m(\mathbf{x}) = 2x^{(3)} - 1$, and treatment heterogeneity, $\tau(\mathbf{x}) = \varsigma(x^{(1)})\varsigma(x^{(2)})$.

The results in Table 2.5 show that, under the three configurations, the HTERF has a similar or better performance than the GRF. When there is no confounding, HTERF performs better. In all considered configurations, the metalearners do not perform well.

In terms of interpretability in the first setting with treatment heterogeneity and no confounding, we expect high and similar importance for $X^{(1)}$ and $X^{(2)}$. These

Conf.	Heterog.	p	n	GRF	HTERF	T-learner	R-learner
no	yes	10	800	0.97	0.83	1.78	1.78
no	yes	10	1600	0.61	0.54	1.29	1.23
no	yes	20	800	1.08	0.93	1.86	2.07
no	yes	20	1600	0.63	0.54	1.31	1.43
yes	no	10	800	0.15	0.17	1.54	0.84
yes	no	10	1600	0.10	0.10	1.16	0.59
yes	no	20	800	0.11	0.11	1.56	0.73
yes	no	20	1600	0.06	0.05	1.16	0.57
yes	yes	10	800	1.03	1.01	1.92	1.95
yes	yes	10	1600	0.68	0.65	1.47	1.58
yes	yes	20	800	1.21	1.13	1.95	2.01
yes	yes	20	1600	0.76	0.71	1.49	1.66

Table 2.5: Mean squared errors of the GRF, HTERF, T-learner, and R-learner methods that estimate the heterogeneous treatment effect. GRF and HTERF causal forests have 500 trees, both the centring forests for the GRF and the forest of the first step in the HTERF have 500 trees, and the same subsampling rate (0.7) is used in both methods. The results are aggregated over 60 simulation replications with 1000 test points each. The mean square errors are multiplied by 10.

		GRF			HTERF		
p	n	$X^{(1)}$	$X^{(2)}$	$\sum$	$X^{(1)}$	$X^{(2)}$	$\sum$
10	800	0.416	0.410	0.826	0.446	0.416	0.862
10	1600	0.413	0.431	0.844	0.440	0.447	0.887
20	800	0.398	0.413	0.811	0.410	0.427	0.837
20	1600	0.416	0.420	0.836	0.434	0.433	0.867

Table 2.6: The importance of $X^{(1)}, X^{(2)}$ and their sum ($\sum$) for the previous setting with treatment heterogeneity and unconfoundedness.

		GRF				HTERF			
p	n	$X^{(1)}$	$X^{(2)}$	$\sum$	$X^{(3)}$	$X^{(1)}$	$X^{(2)}$	$\sum$	$X^{(3)}$
10	800	0.401	0.411	0.812	0.023	0.375	0.486	0.861	0.018
10	1600	0.403	0.435	0.838	0.021	0.381	0.510	0.891	0.014
20	800	0.404	0.395	0.799	0.012	0.359	0.478	0.837	0.010
20	1600	0.385	0.443	0.828	0.010	0.334	0.537	0.871	0.007

Table 2.7: The importance of $X^{(1)}, X^{(2)}$, their sum ($\sum$), and $X^{(3)}$ for the previous setting with treatment heterogeneity and confounding.

are the only informative covariates, and their contributions are quite symmetrical.

In Table 2.6, we can see good results for both methods, with an improvement for the HTERF method. In the third setting, with confounding variables, we expect the same results and no significant importance for the confounding variable $X^{(3)}$. In Table 2.7, for both methods, we find no significant importance for $X^{(3)}$ (for example, for the first line of the GRF, the remaining importance for non-informative variables is 0.188; therefore, we can expect an importance of 0.024 for each non-informative variable, which is very close to the observed importance of 0.023). We see an improvement in terms of the sum of the importance of the informative variables with the HTERF.

2.5.4 Linear γ function

We perform a simulation study to show the importance of an accurate estimation of μ_0 in the pre-processing step.

Let $\mathbf{X}_i \sim U([0,1]^p)$, $W_i \sim Bern(0.5)$, and $Y_i = \tau(\mathbf{X}_i)W_i + \gamma(\mathbf{X}_i)$ where $p = 10, \tau(\mathbf{x}) = \sin(x^{(1)})$ and $\gamma(\mathbf{x}) = 2x^{(2)} + 3x^{(3)}$. We consider a new estimator HTERF-OLS where μ_0 is a linear regression instead of a random forest. Because γ is a simple linear function, μ_0 will fit γ better, and we can expect better results in CATE estimation.

Method	RMSE	Depth 3	Depth 5	Depth 10	Importance
GRF	11.60	0.875	0.514	0.171	0.865
HTERF	9.81	1	0.501	0.163	0.993
HTERF-OLS	1.49	1	1	0.944	1
T-learner	44.26	/	/	/	/
R-learner	11.82	/	/	/	/

Table 2.8: Root mean squared errors of the GRF, HTERF, HTERF-OLS, T-learner, and R-learner methods that estimate the heterogeneous treatment effect. All causal forests have 500 trees, both the centring forests for the GRF and the forest of the first step in the HTERF have 500 trees, and the same subsampling rate (0.7) is used in both methods. The results are aggregated over 60 simulation replications with 1000 test points each. The RMSEs are multiplied by 1000. We also consider the frequency of split on $X^{(1)}$ at depths of 3, 5, and 10 and the importance of this variable.

HTERF-OLS has the best results in terms of quality of fit and interpretability, especially at deeper splits, such as a depth of 10, in comparison with the GRF and HTERF. A low quality of the μ_0 estimator is a flaw for the overall HTERF algorithm. We propose the use of cross-validated random forests in general, but with external knowledge regarding the nature of γ, better choices can be made. In this example, the metalearners perform poorly compared to HTERF-OLS. However, the R-learner has an RMSE close to that of the GRF.

2.5.5 Ishigami-like model

Another example is based on the Ishigami functions, which are often used in sensitivity analyses ([Ishigami & Homma 1990]). Let $\mathbf{X}_i \sim U([-\pi, \pi]^3)$, $W_i \sim Bern(0.5)$ and $Y_i = \tau(\mathbf{X}_i)W_i + \gamma(\mathbf{X}_i)$ where $\tau(\mathbf{x}) = 0.3(x^{(3)})^4 \sin(x^{(1)})$ and $\gamma(\mathbf{x}) = \sin(x^{(1)}) + 7\sin(x^{(2)})^2$.

Method	RMSE	importance $X^{(1)}$	importance $X^{(2)}$	importance $X^{(3)}$
GRF	0.985	0.654	0.076	0.270
HTERF	0.767	0.763	0	0.237
T-learner	0.713	/	/	/
R-learner	0.833	/	/	/

Table 2.9: Root mean squared errors of the GRF, HTERF, T-learner, and R-learner methods that estimate the heterogeneous treatment effect. All causal forests have 500 trees, both the centring forests for the GRF and the forest of the first step in the HTERF have 500 trees, and the same subsampling rate (0.7) is used in both methods. The results are aggregated over 60 simulation replications with 1000 test points each. For the GRF and HTERF, we also consider the importances of the three variables $X^{(1)}, X^{(2)}$ and $X^{(3)}$.

Once again, the HTERF is a better estimator of the CATE in terms of RMSE. Moreover, because $\mathbf{X}^{(2)}$ does not appear in the expression of τ, we expect null importance for this variable. The HTERF had more consistent results in terms of interpretability. In this example, the T-learner performs better than the causal forest algorithm.

2.5.6 Simulation based on real dataset: IHDP

The last example uses the IDHP dataset introduced by [Hill 2011]. It is a semi-synthetic dataset based on the Infant Health and Development Program (IHDP), which is often used in causal literature to benchmark algorithms ([Im *et al.* 2021]; [Louizos *et al.* 2017]; [Shalit *et al.* 2017]; [Johansson *et al.* 2016]). The covariates of the dataset were obtained from a randomised experiment which studied the effects of specialist home visits on children's future cognitive test scores. A subset of the treated population was removed to create an imbalance between the treated and control groups. Simple comparisons of the outcomes would have led to biased estimates of the treatment effect. The dataset contained 747 individuals (139 treated and 608 control) and 25 covariates measuring the aspects of children, their mothers, and pregnancy. Following [Hill 2011], the response surface "A" is used to simulate the potential outcomes (10 such datasets[1] were used), and the noiseless outcome is used to compute the true effect. The models are trained on a randomly selected

[1]The data used comes from : https://github.com/AMLab-Amsterdam/CEVAE/blob/master/datasets/IHDP/csv

training subset of size 7000 and tested on the remaining 470 units; this procedure is replicated 60 times.

Method	RMSE
GRF	10.802
HTERF	8.856
T-learner	8.870
R-learner	14,431

Table 2.10: Root mean squared errors of the GRF, HTERF, T-learner, and R-learner methods that estimate the heterogeneous treatment effect. GRF and HTERF causal forests have 500 trees, both the centring forests for the GRF and the forest of the first step in the HTERF have 500 trees, and the same subsampling rate (0.7) is used in both methods. The results are aggregated over 60 simulation replications with 470 test points each.

In this example, the HTERF performs better than the GRF. The T-learner is the best metalearner, and its performance is slightly lower than that of the HTERF.

2.6 Discussion

In this study, we propose a novel causal forest-based algorithm, the HTERF, to estimate the CATE with a binary treatment. We have shown empirically that the HTERF is more efficient than the GRF in terms of the quality of estimation of the CATE as well as interpretability. We also demonstrate an almost consistent result for the HTERF model under realistic assumptions. The performance of the HTERF is better than or similar to that of the considered metalearners. Additional work could be done on the choice of the μ_0 estimator used in the centring process. When there are clues as to the nature of γ, a well-chosen estimator can drastically improve the performances of the HTERF.

A Proof of consistency

We follow an approach similar to that of [Elie-Dit-Cosaque & Maume-Deschamps 2022]. Here, C denotes any positive constant, allowing us to write $C + C = C$ or $uC = C$ where $u > 0$.

We consider an intermediate result before proving Theorem 2.1.

Assumption A.1. *For all $\ell \in [1, B]$, we assume that the variation of the CATE function within any cell goes to 0:*

$$\forall \mathbf{x} \in \mathcal{X}, \quad \sup_{\mathbf{z} \in A_n(\mathbf{x}; \Theta_\ell, \mathcal{D}_n)} |\tau(\mathbf{z}) - \tau(\mathbf{x})| \xrightarrow[n \to \infty]{a.s.} 0 \ .$$

Assumption A.1 is verified if there is a bounded probability of splitting with regard to each variable, even non-informative variables, as in [Athey *et al.* 2019]. In the following, we prove that Assumption A.1 is satisfied under a hypothesis closer to the random forest practice.

Theorem A.1. *Let Y satisfy Assumption 2.2, with τ belonging to the ♠-class, let $\beta > \frac{5}{2}$, $C > 0$, and let the constructed trees be the highest such that $C\sqrt{n}(\ln n)^{\beta} \leq N_{n,0}(\Theta_\ell, \mathcal{D}_n), N_{n,1}(\Theta_\ell, \mathcal{D}_n)$. Then, Assumption A.1 is verified.*

Lemma A.2. *Assume that Assumption 2.2 is satisfied with the function τ in the ♠-class. Let $S^\infty = (s_j, j = 1, \ldots)$ with $s_j \in \{L, R\}$, it describes an infinite path in a binary tree. Let $S^h = (s_j, j = 1, \ldots, h)$; this describes a path in a binary tree of height h. Let $A_h(S^h, \Theta)$ be the corresponding leaf in a theoretical tree. Then the variation of $\tau(\cdot)$ on $A_h(S^h, \Theta)$ goes to 0 almost surely as h goes to infinity.*

Proof.

The proof is similar to the proof of Lemma 5.3 in [Elie-Dit-Cosaque & Maume-Deschamps 2022]. Let $A_\infty(S^\infty, \Theta) = \bigcap_{h \geq 1} A_h(S^h, \Theta)$, consider a sequence $(j^p, z^p) \in \operatorname{argmax} L^\star_{A_{h_p}(S^{h_p})}(j, z)$ and (j^∞, z^∞) any limit point of the sequence (j^p, z^p). It is noteworthy that

$$
\begin{aligned}
&\Delta^\star(A, j, z) = 0 \Leftrightarrow \\
&\mathbb{E}\left[Y(1) - Y(0) | \, X_{j^\infty} < z, \mathbf{X}^J \in A^J\right] - \mathbb{E}\left[Y(1) - Y(0) | \, X_{j^\infty} \geqslant z, \mathbf{X}^J \in A^J\right] = 0 \\
&\Leftrightarrow \mathbb{P}(\mathbf{X} \in A_\infty)\mathbb{E}\left[(Y(1) - Y(0))\mathbb{1}_{\{X_i \leq z,\ \mathbf{X} \in A_\infty\}}\right] \\
&= \mathbb{P}(X_i \leq z, \mathbf{X} \in A_\infty)\mathbb{E}\left[(Y(1) - Y(0))\mathbb{1}_{\{\mathbf{X} \in A_\infty\}}\right].
\end{aligned}
$$

By differentiating with respect to z, we can observe that this is equivalent to $z \mapsto \mathbb{E}\left[\tau\left(z, \mathbf{X}^{-i}\right)\mathbb{1}_{\{\mathbf{X}^{-i} \in A_\infty^{-i}\}}\right]$ is constant for all $i = 1, \ldots, d$. As we assumed that τ belongs to the ♠-class, either τ is constant on $A_\infty(S^\infty)$ or the diameter of $A_\infty(S^\infty)$ is zero. In both cases, we conclude that the variation of $\tau(\cdot)$ on $A_h(S^h, \Theta)$ goes to 0 as h goes to infinity, as in [Elie-Dit-Cosaque & Maume-Deschamps 2022]. □

Proposition A.3. *Let Assumption 2.2 be satisfied. Let $\beta > \frac{5}{2}$, and let A be a rectangle in $\mathcal{X}$, we shall say that $(A, j, z) \in \mathcal{A}^n$ if $|A_{L0}|$,$|A_{L1}|$,$|A_{R0}|$ and $|A_{R1}|$ are greater than $C\sqrt{n}(\ln n)^{\beta}$ (so this bound is also true for $|A_L|$ and $|A_R|$). We have*

$$
\sup_{(A,j,z) \in \mathcal{A}^n} |\Delta^\star(A, j, z) - \Delta(A, j, z)| \xrightarrow{a.s.} 0 \text{ as } n \to \infty.
$$

Proof.

The proof closely follows that of Proposition 5.3. in [Elie-Dit-Cosaque & Maume-Deschamps 2022] and makes use of the following

decomposition:

$$
\begin{aligned}
&|\Delta^\star(A,j,z) - \Delta(A,j,z)| = |T_1 + T_2| \\
&=: \frac{|A_L||A_R|}{|A|^2}\big[\big(\overline{Y}_{A_{L1}} - \overline{Y}_{A_{L0}} - \overline{Y}_{A_{R1}} + \overline{Y}_{A_{R0}}\big)^2 \\
&- \big(\mathbb{E}[Y(1) - Y(0)|\mathbf{X}^{(j)} < z, \mathbf{X} \in A] - \mathbb{E}[Y(1) - Y(0)|\mathbf{X}^{(j)} \geq z, \mathbf{X} \in A]\big)^2\big] \\
&+ \big(\mathbb{E}[Y(1) - Y(0)|\mathbf{X}^{(j)} < z, \mathbf{X} \in A] - \mathbb{E}[Y(1) - Y(0)|\mathbf{X}^{(j)} \geq z, \mathbf{X} \in A]\big)^2 \\
&\Big(\frac{|A_L||A_R|}{|A|^2} - \mathbb{P}(\mathbf{X}^{(j)} < z|\mathbf{X} \in A)\mathbb{P}(\mathbf{X}^{(j)} \geq z|\mathbf{X} \in A)\Big).
\end{aligned}
$$

To prove this proposition, we prove that $\sup_{A,j,z} T_1$ and $\sup_{A,j,z} T_2$ go to 0 a.s. Using Vapnik-Chervonenkis theory on rectangles in $\mathcal{A}$ we have:

$$
\mathbb{P}\left(\sup_{B\in\mathcal{A}}\left|\frac{|B|}{n} - \mathbb{P}(\mathbf{X} \in B)\right| > \kappa\right) \leq 8(n+1)^{2d}e^{-n\kappa^2/32}. \tag{A.1}
$$

We can write $T_2 = m_2 R_2$, where m_2 is squared difference of expectations and R_2 is the difference with empirical frequencies and conditional probabilities. T_2 decomposes into

$$
|T_2| \leq m_2\left(\frac{|A_L|}{|A|}\left|\frac{|A_R|}{|A|} - \mathbb{P}(\mathbf{X}^{(j)} \geq z|\mathbf{X} \in A)\right| + \mathbb{P}(\mathbf{X}^{(j)} \geq z|\mathbf{X} \in A)\left|\frac{|A_L|}{|A|} - \mathbb{P}(\mathbf{X}^{(j)} < z|\mathbf{X} \in A)\right|\right).
$$

Remark that for $B \in \mathcal{A}$, if

$$
\left|\frac{|B|}{n} - \mathbb{P}(\mathbf{X} \in B)\right| \leq \frac{C}{2}\frac{(\ln n)^\beta}{\sqrt{n}} \tag{A.2}
$$

and $|B| \geq C\sqrt{n}(\ln n)^\beta$, then $\mathbb{P}(\mathbf{X} \in B) \geq \frac{C}{2}\frac{(\ln n)^\beta}{\sqrt{n}}$. So that, for $(A,j,z) \in \mathcal{A}^n$, we have, provided that (A.2) holds for A,

$$
\begin{aligned}
\mathbb{E}(Y(1) - Y(0)|\mathbf{X} \in A) =& \frac{1}{\mathbb{P}(\mathbf{X} \in A)}[\mathbb{E}((Y(1) - Y(0))\mathbb{1}_{\{\mathbf{X}\in A\}}\mathbb{1}_{\{\tau(\mathbf{X})\leq D\}}) \\
&+ \mathbb{E}((Y(1) - Y(0))\mathbb{1}_{\{\mathbf{X}\in A\}}\mathbb{1}_{\{\tau(\mathbf{X})> D\}})] \qquad \text{(A.3)} \\
\leq& D + \mathbb{E}(\tau^p(\mathbf{X}))^{\frac{1}{p}}\frac{\mathbb{P}(\tau(\mathbf{X}) > D)^{\frac{1}{q}}}{\mathbb{P}(\mathbf{X} \in A)^{1-\frac{1}{r}}} \\
&\text{the second term is obtained using Hölder inequality} \\
\leq& (\ln n)^\delta + C\frac{n^{\frac{1}{2}(1-\frac{1}{r})}}{(\ln n)^{\beta(1-\frac{1}{r})}} \\
\leq& C(\ln n)^\delta, \qquad \text{(A.4)}
\end{aligned}
$$

by taking $D = (\ln n)^\delta$, $\delta > 1$, $p,q,r > 0$ with $\frac{1}{p} + \frac{1}{q} + \frac{1}{r} = 1$. The hypothesis

"(A.2) verified for A, A_L, A_R" is noted Ξ.

$$\begin{aligned}
\mathbb{P}(|T_2| > \kappa) =& \mathbb{P}(|T_2| > \kappa, \Xi) + \mathbb{P}(|T_2| > \kappa, \neg\Xi) \\
\leq& \mathbb{P}(\neg((A.2) \text{ verified for } A, A_L, A_R)) + \mathbb{P}(|T_2| > \kappa, \Xi) \\
\leq& 3 \times 8(n+1)^{2d} e^{\frac{-nC^2(\ln n)^{2\beta}}{32\times 4n}} + \mathbb{P}(|T_2| > \kappa, \Xi) \\
\leq& 24(n+1)^{2d} e^{\frac{-C(\ln n)^{2\beta}}{128}} + \mathbb{P}\left(R_2 > \frac{\kappa}{4C(\ln n)^{2\delta}}, \Xi\right) \\
\leq& 24(n+1)^{2d} e^{\frac{-C(\ln n)^{2\beta}}{128}} \\
&+ \mathbb{P}\left(\left[\frac{|A_L|}{|A|}\left|\frac{|A_R|}{|A|} - \mathbb{P}(\mathbf{X}^{(j)} \geq z | \mathbf{X} \in A)\right|\right] > \frac{\kappa}{8C(\ln n)^{2\delta}}, \Xi\right) \\
&+ \mathbb{P}\left(\left[\mathbb{P}(\mathbf{X}^{(j)} \geq z | \mathbf{X} \in A)\left|\frac{|A_L|}{|A|} - \mathbb{P}(\mathbf{X}^{(j)} < z | \mathbf{X} \in A)\right|\right] > \frac{\kappa}{8C(\ln n)^{2\delta}}, \Xi\right) \\
\mathbb{P}(|T_2| > \kappa) \leq& 24(n+1)^{2d} e^{\frac{-C(\ln n)^{2\beta}}{128}} + T_{2,1} + T_{2,2}. \qquad \text{(A.5)}
\end{aligned}$$

Notice that:

$$\begin{aligned}
&\frac{|A_L|}{|A|}\left|\frac{|A_R|}{|A|} - \mathbb{P}(\mathbf{X}^{(j)} \geq z | \mathbf{X} \in A)\right| \\
\leq& \frac{|A_L|}{|A|}\left[\frac{n}{|A|}\left|\frac{|A_R|}{n} - \mathbb{P}(\mathbf{X}^{(j)} \geq z, \mathbf{X} \in A)\right| + \left|\frac{1}{\mathbb{P}(\mathbf{X} \in A)} - \frac{n}{|A|}\right| \mathbb{P}(\mathbf{X}^{(j)} \geq z, \mathbf{X} \in A)\right].
\end{aligned}$$

So $T_{2,1}$ decomposes into:

$$\begin{aligned}
T_{2,1} \leq& \mathbb{P}\left(\frac{|A_L|n}{|A|^2}\left|\frac{|A_R|}{n} - \mathbb{P}(\mathbf{X}^{(j)} \geq z, \mathbf{X} \in A)\right| > \frac{\kappa}{16C(\ln n)^{2\delta}}, \Xi\right) \\
&+ \mathbb{P}\left(\frac{|A_L|}{|A|}\mathbb{P}(\mathbf{X}^{(j)} \geq z, \mathbf{X} \in A)\left|\frac{1}{\mathbb{P}(\mathbf{X} \in A)} - \frac{n}{|A|}\right| > \frac{\kappa}{16C(\ln n)^{2\delta}}, \Xi\right) \\
\leq& T_{2,1,1} + T_{2,1,1}.
\end{aligned}$$

Notice that:

$$\frac{|A_L|n}{|A|^2} \leq \frac{|A_L|n}{|A_L|^2} \leq \frac{n}{|A_L|} \leq \frac{n}{C\sqrt{n}(\ln n)^{\beta}} \leq \frac{\sqrt{n}}{C(\ln n)^{\beta}}.$$

An upper bound for $T_{2,1,1}$ can be determined:

$$\begin{aligned}
T_{2,1,1} \leq& \mathbb{P}\left(\left|\frac{|A_R|}{n} - \mathbb{P}(\mathbf{X}^{(j)} \geq z, \mathbf{X} \in A)\right| > \frac{\kappa}{16C(\ln n)^{2\delta-\beta}\sqrt{n}}, \Xi\right) \\
\leq& \mathbb{P}\left(\left|\frac{|A_R|}{n} - \mathbb{P}(\mathbf{X}^{(j)} \geq z, \mathbf{X} \in A)\right| > \frac{\kappa}{16C(\ln n)^{2\delta-\beta}\sqrt{n}}\right) \\
\leq& 8(n+1)^{2d} e^{\frac{-C\kappa^2}{8192(\ln n)^{4\delta-2\beta}}}.
\end{aligned}$$

Notice that:

$$\begin{aligned}\frac{|A_L|}{|A|}\mathbb{P}(\mathbf{X}^{(j)} \geq z, \mathbf{X} \in A)\left|\frac{1}{\mathbb{P}(\mathbf{X} \in A)} - \frac{n}{|A|}\right| &= \frac{|A_L|}{|A|}\frac{\mathbb{P}(\mathbf{X}^{(j)} \geq z, \mathbf{X} \in A)}{\mathbb{P}(\mathbf{X} \in A)}\frac{n}{|A|}\left|\frac{|A|}{n} - \mathbb{P}(\mathbf{X} \in A)\right| \\ &\leq 1 \times 1 \times \frac{n}{|A|}\left|\frac{|A|}{n} - \mathbb{P}(\mathbf{X} \in A)\right| \\ &\leq \frac{n}{C(\ln n)^{\beta}\sqrt{n}}\left|\frac{|A|}{n} - \mathbb{P}(\mathbf{X} \in A)\right| \\ &\leq \frac{\sqrt{n}}{C(\ln n)^{\beta}}\left|\frac{|A|}{n} - \mathbb{P}(\mathbf{X} \in A)\right|.\end{aligned}$$

An upper bound for $T_{2,1,2}$ can be determined:

$$\begin{aligned}T_{2,1,2} \leq & \mathbb{P}\left(\left|\frac{|A|}{n} - \mathbb{P}(\mathbf{X} \in A)\right| > \frac{\kappa}{16C(\ln n)^{2\delta-\beta}\sqrt{n}}\right) \\ \leq & 8(n+1)^{2d}e^{\frac{-C\kappa^2}{8192(\ln n)^{4\delta-2\beta}}}.\end{aligned}$$

So:

$$T_{2,1} \leq 16(n+1)^{2d}e^{\frac{-C\kappa^2}{8192(\ln n)^{4\delta-2\beta}}}. \tag{A.6}$$

The term $T_{2,2}$ can be treated in a similar way:

$$\begin{aligned}T_{2,2} \leq & \mathbb{P}\left(\frac{n}{|A|}\mathbb{P}(\mathbf{X}^{(j)} \geq z|\mathbf{X} \in A)\left|\frac{|A_L|}{n} - \mathbb{P}(\mathbf{X}^{(j)} < z, \mathbf{X} \in A)\right| > \frac{\kappa}{16C(\ln n)^{2\delta}}, \Theta\right) \\ & + \mathbb{P}\left(\mathbb{P}(\mathbf{X}^{(j)} \geq z|\mathbf{X} \in A)\mathbb{P}(\mathbf{X}^{(j)} < z, \mathbf{X} \in A)\left|\frac{1}{\mathbb{P}(\mathbf{X} \in A)} - \frac{n}{|A|}\right| > \frac{\kappa}{16C(\ln n)^{2\delta}}, \Theta\right) \\ \leq & T_{2,2,1} + T_{2,2,1}.\end{aligned}$$

Notice:

$$\frac{n}{|A|}\mathbb{P}(\mathbf{X}^{(j)} \geq z|\mathbf{X} \in A) \leq \frac{n}{|A|} \leq \frac{n}{|A_R|} \leq \frac{\sqrt{n}}{C(\ln n)^{\beta}}.$$

An upper bound for $T_{2,2,1}$ can be determined:

$$\begin{aligned}T_{2,2,1} \leq & \mathbb{P}\left(\left|\frac{|A_L|}{n} - \mathbb{P}(\mathbf{X}^{(j)} < z, \mathbf{X} \in A)\right| > \frac{\kappa}{16C(\ln n)^{2\delta-\beta}\sqrt{n}}\right) \\ \leq & 8(n+1)^{2d}e^{\frac{-C\kappa^2}{8192(\ln n)^{4\delta-2\beta}}}.\end{aligned}$$

Notice:

$$\begin{aligned}
&\mathbb{P}(\mathbf{X}^{(j)} \geq z|\mathbf{X} \in A)\mathbb{P}(\mathbf{X}^{(j)} < z, \mathbf{X} \in A)\left|\frac{1}{\mathbb{P}(\mathbf{X} \in A)} - \frac{n}{|A|}\right| \\
=&\mathbb{P}(\mathbf{X}^{(j)} \geq z|\mathbf{X} \in A)\frac{\mathbb{P}(\mathbf{X}^{(j)} < z, \mathbf{X} \in A)}{\mathbb{P}(\mathbf{X} \in A)}\frac{n}{|A|}\left|\frac{|A|}{n} - \mathbb{P}(\mathbf{X} \in A)\right| \\
\leq&1 \times 1 \times \frac{n}{|A|}\left|\frac{|A|}{n} - \mathbb{P}(\mathbf{X} \in A)\right| \\
\leq&\frac{n}{C(\ln n)^{\beta}\sqrt{n}}\left|\frac{|A|}{n} - \mathbb{P}(\mathbf{X} \in A)\right| \\
\leq&\frac{\sqrt{n}}{C(\ln n)^{\beta}}\left|\frac{|A|}{n} - \mathbb{P}(\mathbf{X} \in A)\right|.
\end{aligned}$$

An upper bound for $T_{2,2,2}$ can be determined:

$$\begin{aligned}
T_{2,2,2} \leq&\mathbb{P}\left(\left|\frac{|A|}{n} - \mathbb{P}(\mathbf{X} \in A)\right| > \frac{\kappa}{16C(\ln n)^{2\delta-\beta}\sqrt{n}}\right) \\
\leq&8(n+1)^{2d}e^{\frac{-C\kappa^2}{8192(\ln n)^{4\delta-2\beta}}}.
\end{aligned}$$

Finally:

$$T_2 \leq 24(n+1)^{2d}e^{\frac{-C(\ln n)^{2\beta}}{128}} + 32(n+1)^{2d}e^{\frac{-C\kappa^2}{8192(\ln n)^{4\delta-2\beta}}}. \tag{A.7}$$

Then, Borel-Cantelli Lemma gives that $\sup_{(A,j,z)\in\mathcal{A}^n} T_2$ goes to 0 a.s. provided that $2\beta - 4\delta > 1$.
We can now consider the T_1 term:

$$\begin{aligned}
T_1 =&\frac{|A_L||A_R|}{|A|^2}\big[\big(\overline{Y}_{A_{L1}} - \overline{Y}_{A_{L0}} - \overline{Y}_{A_{R1}} + \overline{Y}_{A_{R0}}\big)^2 \\
&- \big(\mathbb{E}[Y(1) - Y(0)|\mathbf{X}^{(j)} < z, \mathbf{X} \in A] - \mathbb{E}[Y(1) - Y(0)|\mathbf{X}^{(j)} \geq z, \mathbf{X} \in A]\big)^2\big] \\
=&\frac{|A_L||A_R|}{|A|^2}\big(\overline{Y}_{A_{L1}} - \mathbb{E}[Y(1) - \mu_0(\mathbf{X})|\mathbf{X}^{(j)} < z, \mathbf{X} \in A] \\
&- \overline{Y}_{A_{L0}} + \mathbb{E}[Y(0)|\mathbf{X}^{(j)} < z, \mathbf{X} \in A] \\
&- \overline{Y}_{A_{R1}} + \mathbb{E}[Y(1)|\mathbf{X}^{(j)} \geq z, \mathbf{X} \in A] \\
&+ \overline{Y}_{A_{R0}} - \mathbb{E}[Y(0)|\mathbf{X}^{(j)} \geq z, \mathbf{X} \in A]\big) \\
&\big(\overline{Y}_{A_{L1}} + \mathbb{E}[Y(1)|\mathbf{X}^{(j)} < z, \mathbf{X} \in A] \\
&- \overline{Y}_{A_{L0}} - \mathbb{E}[Y(0)|\mathbf{X}^{(j)} < z, \mathbf{X} \in A] \\
&- \overline{Y}_{A_{R1}} - \mathbb{E}[Y(1)|\mathbf{X}^{(j)} \geq z, \mathbf{X} \in A] \\
&+ \overline{Y}_{A_{R0}} + \mathbb{E}[Y(0)|\mathbf{X}^{(j)} \geq z, \mathbf{X} \in A]\big) \\
=&\frac{|A_L||A_R|}{|A|^2}T_{1,1}T_{1,2} \\
|T_1| \leq&|T_{1,1}||T_{1,2}|.
\end{aligned}$$

We note Ξ the hypothesis "(A.2) verified for $A_{L1}, A_{L0}, A_{R1}, A_{R0}$", for any $D > 0$ we have:

$$\begin{aligned}\mathbb{P}(|T_1| > \kappa) \leq & \mathbb{P}(|T_{1,2}| > \kappa/D, |T_{1,1}| \leq D, \Xi) + \mathbb{P}(|T_{1,2}| > \kappa, |T_{1,1}| > D, \Xi) + \mathbb{P}(\neg\Xi) \\ \leq & \mathbb{P}(|T_{1,2}| > \kappa/D, \Xi) + \mathbb{P}(|T_{1,1}| > D, \Xi) + 4 \times 8(n+1)^{2d} e^{\frac{-nC^2(\ln n)^{2\beta}}{32\times 4n}}.\end{aligned}$$

$\mathbb{P}(|T_{1,1}| > D, \Xi)$ is treated first:

$$\begin{aligned}\mathbb{P}(|T_{1,1}| > D, \Xi) \leq & \mathbb{P}\left(\left|\overline{Y}_{A_{L1}} - \mathbb{E}[Y(1)|\mathbf{X}^{(j)} < z, \mathbf{X} \in A]\right| > \frac{D}{4}, \Xi\right) \\ & + \mathbb{P}\left(\left|\overline{Y}_{A_{L0}} - \mathbb{E}[Y(0)|\mathbf{X}^{(j)} < z, \mathbf{X} \in A]\right| > \frac{D}{4}, \Xi\right) \\ & + \mathbb{P}\left(\left|\overline{Y}_{A_{R1}} - \mathbb{E}[Y(1)|\mathbf{X}^{(j)} \geq z, \mathbf{X} \in A]\right| > \frac{D}{4}, \Xi\right) \\ & + \mathbb{P}\left(\left|\overline{Y}_{A_{R0}} - \mathbb{E}[Y(0)|\mathbf{X}^{(j)} \geq z, \mathbf{X} \in A]\right| > \frac{D}{4}, \Xi\right) \\ \leq & T_{1,1,1} + T_{1,1,2} + T_{1,1,3} + T_{1,1,4}.\end{aligned}$$

Notice that:

$$\begin{aligned}& \left|\overline{Y}_{A_{L1}} - \mathbb{E}[Y(1)|\mathbf{X}^{(j)} < z, \mathbf{X} \in A]\right| \\ = & \left|\overline{Y}_{A_{L1}} - \mathbb{E}[Y(1)|\mathbf{X} \in A_L]\right| \\ = & \left|\overline{Y}_{A_{L1}} - \mathbb{E}[Y(1)|\mathbf{X} \in A_{L1}]\right| \text{ unconfoundedness} \\ \leq & \frac{n}{|A_{L1}|}\left|\frac{1}{n}\sum_{i=1}^{n} Y_i \mathbb{1}_{\mathbf{X}_i \in A_{L1}} - \mathbb{E}[Y\mathbb{1}_{\mathbf{X}\in A_{L1}}]\right| + |\mathbb{E}[Y\mathbb{1}_{\mathbf{X}\in A_{L1}}]|\left|\frac{n}{|A_{L1}|} - \frac{1}{\mathbb{P}(\mathbf{X} \in A_{L1})}\right|.\end{aligned}$$

$$\begin{aligned}T_{1,1,1} \leq & \mathbb{P}\left(\frac{n}{|A_{L1}|}\left|\frac{1}{n}\sum_{i=1}^{n} Y_i \mathbb{1}_{\mathbf{X}_i \in A_{L1}} - \mathbb{E}[Y\mathbb{1}_{\mathbf{X}\in A_{L1}}]\right| > \frac{D}{8}\right) \\ & + \mathbb{P}\left(|\mathbb{E}[Y\mathbb{1}_{\mathbf{X}\in A_{L1}}]|\left|\frac{n}{|A_{L1}|} - \frac{1}{\mathbb{P}(\mathbf{X} \in A_{L1})}\right| > \frac{D}{8}\right) \\ \leq & \mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} Y_i \mathbb{1}_{\mathbf{X}_i \in A_{L1}} - \mathbb{E}[Y\mathbb{1}_{\mathbf{X}\in A_{L1}}]\right| > \frac{DC(\ln n)^{\beta}}{8\sqrt{n}}\right) \\ & + \mathbb{P}\left(|\mathbb{E}[Y\mathbb{1}_{\mathbf{X}\in A_{L1}}]|\left|\frac{n}{|A_{L1}|} - \frac{1}{\mathbb{P}(\mathbf{X} \in A_{L1})}\right| > \frac{D}{8}\right) \\ \leq & T_{1,1,1,1} + T_{1,1,1,2}.\end{aligned}$$

Using Theorem 9.6 in [Györfi *et al.* 2002] and Lemma A.2 in

[Elie-Dit-Cosaque & Maume-Deschamps 2022], lead for any $L > 0$ and $\frac{1}{p} + \frac{1}{q} = 1$:

$$\begin{aligned}
&\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} Y_i \mathbb{1}_{\mathbf{X}_i \in A_L, W=1} - \mathbb{E}[Y\mathbb{1}_{\mathbf{X}\in A_{L1}}]\right| > \kappa\right) \\
\leq & 24\left(\frac{32eL}{\kappa}\log\left(\frac{48eL}{\kappa}\right)\right)^{2d} \exp\left(\frac{-n\kappa^2}{512L^2}\right) + C\frac{\mathbb{E}[Y^p]^{\frac{1}{p}}\mathbb{P}(Y > L)^{\frac{1}{q}}}{\kappa}.
\end{aligned}$$

And we take $L = (\ln n)^\delta$ as before.

$$T_{1,1,1,1} \leq 24\left(\frac{Ce(\ln n)^\delta\sqrt{n}}{D(\ln n)^\beta}\log\left(\frac{Ce(\ln n)^\delta\sqrt{n}}{D(\ln n)^\beta}\right)\right)^{2d} \exp\left(\frac{-CD^2(\ln n)^{2\beta}}{(\ln n)^{2\delta}}\right) + C\frac{\sqrt{n}e^{(\ln n)^\delta \frac{\ln\theta}{q}}}{D(\ln n)^\beta}$$

Notice that:

$$\begin{aligned}
&|\mathbb{E}[Y\mathbb{1}_{\mathbf{X}\in A_{L1}}]|\left|\frac{n}{|A_{L1}|} - \frac{1}{\mathbb{P}(\mathbf{X}\in A_{L1})}\right| \\
\leq & \frac{|\mathbb{E}[Y\mathbb{1}_{\mathbf{X}\in A_{L1}}]|}{\mathbb{P}(\mathbf{X}\in A_{L1})}\frac{n}{|A_{L1}|}\left|\mathbb{P}(\mathbf{X}\in A_{L1}) - \frac{|A_{L1}|}{n}\right| \\
\leq & |\mathbb{E}[Y|\mathbf{X}\in A_{L1}]|\frac{\sqrt{n}}{C(\ln n)^\beta}\left|\mathbb{P}(\mathbf{X}\in A_{L1}) - \frac{|A_{L1}|}{n}\right|.
\end{aligned}$$

The expected value is treated as previously in A.3 with Hölder under the assumption that Ξ is verified:

$$\begin{aligned}
\mathbb{E}[Y|\mathbf{X}\in A_{L1}] &\leq \frac{1}{\mathbb{P}(\mathbf{X}\in A_{L1})}\left[\mathbb{E}\left(Y\mathbb{1}_{\mathbf{X}\in A_{L1}}\mathbb{1}_{Y\leq D}\right) + \mathbb{E}\left(Y\mathbb{1}_{\mathbf{X}\in A_{L1}}\mathbb{1}_{Y>D}\right)\right] \\
&\leq D + \mathbb{E}[Y^p]^{\frac{1}{p}}\frac{\mathbb{P}(Y>D)^{\frac{1}{q}}}{\mathbb{P}(\mathbf{X}\in A_{L1})^{1-\frac{1}{r}}} \\
&\leq (\ln n)^\delta + Ce^{(\ln n)^\delta\frac{\ln\theta}{q}}\frac{n^{\frac{1}{2}\left(1-\frac{1}{r}\right)}}{(\ln n)^{\beta\left(1-\frac{1}{r}\right)}} \\
&\leq C(\ln n)^\delta.
\end{aligned}$$

Then the quantity is bounded in the same way that $T_{2,1,1}$:

$$\begin{aligned}
T_{1,1,1,2} &\leq \mathbb{P}\left(\left|\mathbb{P}(\mathbf{X}\in A_{L1}) - \frac{|A_{L1}|}{n}\right| > \frac{DC(\ln n)^{\beta-\delta}}{8\sqrt{n}}\right) \\
&\leq 8(n+1)^{2d}e^{-\frac{D^2C(\ln n)^{2\beta-2\delta}}{2048}}.
\end{aligned}$$

Now we treat $T_{1,2}$:

$$\begin{aligned}\mathbb{P}(|T_{1,2}| > \kappa/D, \Xi) \leq & \mathbb{P}\left(\left|\overline{Y}_{A_{L1}} + \mathbb{E}[Y(1)|\mathbf{X}^{(j)} < z, \mathbf{X} \in A]\right| > \frac{\kappa}{D4}, \Xi\right) \\ & + \mathbb{P}\left(\left|\overline{Y}_{A_{L0}} + \mathbb{E}[Y(0)|\mathbf{X}^{(j)} < z, \mathbf{X} \in A]\right| > \frac{\kappa}{D4}, \Xi\right) \\ & + \mathbb{P}\left(\left|\overline{Y}_{A_{R1}} + \mathbb{E}[Y(1)|\mathbf{X}^{(j)} \geq z, \mathbf{X} \in A]\right| > \frac{\kappa}{D4}, \Xi\right) \\ & + \mathbb{P}\left(\left|\overline{Y}_{A_{R0}} + \mathbb{E}[Y(0)|\mathbf{X}^{(j)} \geq z, \mathbf{X} \in A]\right| > \frac{\kappa}{D4}, \Xi\right) \\ \leq & T_{1,2,1} + T_{1,2,2} + T_{1,2,3} + T_{1,2,4}.\end{aligned}$$

$$\begin{aligned}T_{1,2,1} \leq & \mathbb{P}\left(\left|\overline{Y}_{A_{L1}} - \mathbb{E}[Y(1)|\mathbf{X}^{(j)} < z, \mathbf{X} \in A]\right| > \frac{\kappa}{D8}, \Xi\right) \\ & + \mathbb{P}\left(\left|\mathbb{E}[Y(1)|\mathbf{X}^{(j)} < z, \mathbf{X} \in A]\right| > \frac{\kappa}{D16}, \Xi\right).\end{aligned}$$

The first term has already been treated and second term is bounded using Hölder inequality as before. $T_{1,2,2}$, $T_{1,2,3}$ and $T_{1,2,4}$ are treated similarly.

Finally, Borel-Cantelli Lemma gives that $\sup_{(A,j,z)\in\mathcal{A}^n} T_{1,1}$ goes to 0 a.s. and $T_{1,2}$ is bounded, so finally $\sup_{(A,j,z)\in\mathcal{A}^n} T_1$ goes to 0 a.s.. □

Proposition A.4. *Let Assumption 2.2 be satisfied. Assume that for $\beta > \frac{5}{2}$, $N_{n,0}(\Theta_\ell, \mathcal{D}_n), N_{n,1}(\Theta_\ell, \mathcal{D}_n) \geq C\sqrt{n}(\ln n)^\beta$. For $h \in \mathbb{N}$, let $S \in \{L, R\}^h$ describe a path of length h in a binary tree. Let $A^n(S)$ and $A(S)$ be corresponding nodes in empirical and theoretical trees. Denote*

$$A(S) = \prod_{j=1}^{d}[a_j, b_j] \text{ and } A^n(S) = \prod_{j=1}^{d}[a_j^n, b_j^n].$$

Denote $\mathcal{T}_h$ the set of theoretical trees of height h; then,

$$\inf_{\mathcal{T}_h} \max_{j=1,\dots,d} \max\left(|a_j - a_j^n|, |b_j - b_j^n|\right) \longrightarrow 0 \text{ a.s. as } n \to \infty. \tag{A.8}$$

Proof.

The proof is the same as in Proposition 5.4 from [Elie-Dit-Cosaque & Maume-Deschamps 2022]. □

Proof of Theorem A.1.

This theorem can be proved as Theorem 5.1 in [Elie-Dit-Cosaque & Maume-Deschamps 2022]. □

Following the lines of the proof of Theorem A.1, the same result for τ_1 and τ_0 can be obtained.

Assumption A.2. *For all $\ell \in [1, B]$, we assume that the variation of τ_1 and τ_0 within any cell goes to 0:*

$$\forall \mathbf{x} \in \mathcal{X}, \quad \sup_{\mathbf{z} \in A_n(\mathbf{x}; \Theta_\ell, \mathcal{D}_n)} |\tau_1(\mathbf{z}) - \tau_1(\mathbf{x})| \xrightarrow[n\to\infty]{a.s.} 0 \ .$$

$$\forall \mathbf{x} \in \mathcal{X}, \quad \sup_{\mathbf{z} \in A_n(\mathbf{x}; \Theta_\ell, \mathcal{D}_n)} |\tau_0(\mathbf{z}) - \tau_0(\mathbf{x})| \xrightarrow[n\to\infty]{a.s.} 0 \ .$$

Theorem A.5. *Let Y satisfy Assumption 2.2, with τ_0 and τ_1 belonging to the ♠-class, let $\beta > \frac{5}{2}$, $C > 0$, and let the constructed trees be the highest such that $C\sqrt{n}(\ln n)^\beta \leq N_{n,0}(\Theta_\ell, \mathcal{D}_n), N_{n,1}(\Theta_\ell, \mathcal{D}_n)$. Then, Assumption A.2 is verified.*

Theorem 2.1 is a direct consequence of Theorem A.6.

Theorem A.6. *Consider a random forest which satisfies Assumptions A.2, 2.3, and the hypotheses for Theorem 2.1. Then,*

$$\forall \mathbf{x} \in \mathcal{X}, |\hat{\tau}_{B,n}(\mathbf{x}) - \tau(\mathbf{x})| \xrightarrow[n\to\infty]{a.s.} 0 \ .$$

Proof of Theorem A.6.

The proof follows the ideas of [Elie-Dit-Cosaque & Maume-Deschamps 2022]; however, some differences arise because of honest subsampling rather than bootstrapping. We provide this in detail for completeness.

The main component of the proof is the use of a second sample, $\mathcal{D}_n^\diamond$ to address the data-dependent aspect. Thus, we first define a dummy estimator based on two samples, $\mathcal{D}_n$ and $\mathcal{D}_n^\diamond$, which will be used as follows. The trees are grown using $\mathcal{D}_n$, but we consider another sample $\mathcal{D}_n^\diamond$ (independent of $\mathcal{D}_n$ and Θ) which is used to define a dummy estimator

$$\begin{aligned}
&\tau_{B,n}^\diamond(\mathbf{x}; \Theta_1, \ldots, \Theta_k, \mathcal{D}_n^\diamond, \mathcal{D}_n) \\
=&\sum_{j=1}^n \alpha_{n,j}^\diamond(\mathbf{x}; \Theta_1, \ldots, \Theta_k, \mathbf{X}_1^\diamond, \ldots, \mathbf{X}_n^\diamond, W_1^\diamond, \ldots, W_n^\diamond, \mathcal{D}_n) Y_j^{e\diamond} \\
&- \sum_{j=1}^n \alpha_{n,j}^{\prime\diamond}(\mathbf{x}; \Theta_1, \ldots, \Theta_k, \mathbf{X}_1^\diamond, \ldots, \mathbf{X}_n^\diamond, W_1^\diamond, \ldots, W_n^\diamond, \mathcal{D}_n) Y_j^{e\diamond},
\end{aligned}$$

where the weights are

$$\begin{aligned}
&\alpha_{n,j}^\diamond(\mathbf{x}; \Theta_1, \ldots, \Theta_k, \mathbf{X}_1^\diamond, \ldots, \mathbf{X}_n^\diamond, W_1^\diamond, \ldots, W_n^\diamond, \mathcal{D}_n) \\
=&\frac{1}{B}\sum_{\ell=1}^B \frac{\mathbb{1}_{\{\mathbf{X}_j^\diamond \in A_n(\mathbf{x}; \Theta_\ell, \mathcal{D}_n)\} \wedge W_j^\diamond = 1}}{N_{n,1}^\diamond(\mathbf{x}; \Theta_\ell, \mathbf{X}_1^\diamond, \ldots, \mathbf{X}_n^\diamond, W_1^\diamond, \ldots, W_n^\diamond, \mathcal{D}_n)}, \ j = 1, \ldots, n.
\end{aligned}$$

with $N_{n,1}^\diamond(\mathbf{x}; \Theta_\ell, \mathbf{X}_1^\diamond, \ldots, \mathbf{X}_n^\diamond, W_1^\diamond, \ldots, W_n^\diamond, \mathcal{D}_n)$, the number of elements of $\mathcal{D}_n^\diamond$ that fall into $A_n(\mathbf{x}; \Theta_\ell, \mathcal{D}_n)$ such that $W^\diamond = 1$. Throughout this section, we shall use the convention $\frac{0}{0} = 0$ in case $N_{n,1}^\diamond(\mathbf{x}; \Theta_\ell, \mathbf{X}_1^\diamond, \ldots, \mathbf{X}_n^\diamond, W_1^\diamond, \ldots, W_n^\diamond, \mathcal{D}_n) = 0$ and thus

$\mathbb{1}_{\left\{\mathbf{X}_j^\diamond \in A_n(\mathbf{x};\Theta_\ell,\mathcal{D}_n)\right\}\wedge W_j^\diamond=1} = 0$ for $j = 1, \dots n$.
Similarly we have:

$$\begin{aligned}
&\alpha'^{\diamond}_{n,j}\left(\mathbf{x};\Theta_1,\dots,\Theta_k,\mathbf{X}_1^\diamond,\dots,\mathbf{X}_n^\diamond,W_1^\diamond,\dots,W_n^\diamond,\mathcal{D}_n\right)\\
=&\frac{1}{B}\sum_{\ell=1}^{B}\frac{\mathbb{1}_{\left\{\mathbf{X}_j^\diamond \in A_n(\mathbf{x};\Theta_\ell,\mathcal{D}_n)\right\}\wedge W_j^\diamond=0}}{N^\diamond_{n,0}\left(\mathbf{x};\Theta_\ell,\mathbf{X}_1^\diamond,\dots,\mathbf{X}_n^\diamond,W_1^\diamond,\dots,W_n^\diamond,\mathcal{D}_n\right)},\ j=1,\dots,n.
\end{aligned}$$

To lighten the notation in the sequel, we simply write $\tau^\diamond_{B,n}(\mathbf{x}) = \sum_{j=1}^{n}\alpha_j^\diamond(\mathbf{x})\,Y_j^\diamond - \sum_{j=1}^{n}\alpha'^{\diamond}_j(\mathbf{x})\,Y_j^\diamond$.

Let $\mathbf{x}\in\mathcal{X}$, we have:

$$\begin{aligned}
\left|\hat{\tau}(\mathbf{x})-\tau(\mathbf{x})\right| \leq& \left|\hat{\tau}(\mathbf{x})-\tau^\diamond(\mathbf{x})\right|\\
&+\left|\tau^\diamond(\mathbf{x})-\tau(\mathbf{x})\right|.
\end{aligned}$$

Let $\mathbf{x}$ in $\mathcal{X}$: $|\tau^\diamond(\mathbf{x})-\tau(\mathbf{x})| \leq |\tau_1^\diamond(\mathbf{x})-\tau_1(\mathbf{x})| + |\tau_0^\diamond(\mathbf{x})-\tau_0(\mathbf{x})|$ Each of the two terms will be treated the same way.

$$\begin{aligned}
|\tau_1^\diamond(\mathbf{x})-\tau_1(\mathbf{x})| \leq& \left|\sum_{\substack{i=1\\W_i^\diamond=1}}^{n}\alpha_i^\diamond(\mathbf{x})\left[(Y_i^\diamond)-\mathbb{E}\left[Y(1)|\mathbf{X}_i^\diamond\right]\right]\right|\\
&+\left|\sum_{\substack{i=1\\W_i^\diamond=1}}^{n}\alpha_i^\diamond(\mathbf{x})\left[\mathbb{E}\left[Y(1)|\mathbf{X}_i^\diamond\right]-\mathbb{E}\left[Y(1)|\mathbf{X}=\mathbf{x}\right]\right]\right|\\
\leq& U_n+V_n.
\end{aligned}$$

The last term tends towards 0 with Theorem A.5:

$$\begin{aligned}
V_n =& \left|\sum_{\substack{i=1\\W_i^\diamond=1}}^{n}\alpha_i^\diamond(\mathbf{x})\left[\mathbb{E}\left[Y(1)|\mathbf{X}_i^\diamond\right]-\mathbb{E}\left[Y(1)|\mathbf{X}=\mathbf{x}\right]\right]\right|\\
=& \left|\sum_{\substack{i=1\\W_i^\diamond=1\\\exists l|\mathbf{X}_i^\diamond\in A_n(\mathbf{x},\Theta_l)}}^{n}\alpha_i^\diamond(\mathbf{x})\left[\mathbb{E}\left[Y(1)|\mathbf{X}_i^\diamond\right]-\mathbb{E}\left[Y(1)|\mathbf{X}=\mathbf{x}\right]\right]\right|\\
\leq& \sum_{\substack{i=1\\W_i^\diamond=1\\\exists l|\mathbf{X}_i^\diamond\in A_n(\mathbf{x},\Theta_l)}}^{n}\left|\alpha_i^\diamond(\mathbf{x})\left[\mathbb{E}\left[Y(1)|\mathbf{X}_i^\diamond\right]-\mathbb{E}\left[Y(1)|\mathbf{X}=\mathbf{x}\right]\right]\right|\\
\leq& \sup_{\mathbf{z}\in A_n(\mathbf{x})}|\tau_1(\mathbf{z})-\tau_1(\mathbf{x})| \xrightarrow[n\to+\infty]{} 0.
\end{aligned}$$

For the first term, we have

$$U_n = \left| \sum_{\substack{i=1 \\ W_i^\diamond = 1}}^{n} \alpha_i^\diamond \varepsilon_i^\diamond \right|.$$

The following Lemma is useful:

Lemma A.7. *Let $u \in \{0,1\}$, as before, $N_{n,u}(A_n(\Theta)) = N_{n,u}(\mathbf{x}; \Theta, \mathcal{D}_n)$ is the number of observations of $\mathcal{D}_n$ such that $W = u$ that fall into in $A_n(\Theta) = A_n(\mathbf{x}; \Theta, \mathcal{D}_n)$ and $N_{n,u}^\diamond(A_n(\Theta)) = N_{n,u}^\diamond\left(\mathbf{x}; \Theta, \mathbf{X}^{\diamond 1}, \ldots, \mathbf{X}^{\diamond n}, \mathcal{D}_n\right)$, the number of observations of $\mathcal{D}_n^\diamond$ such that $W = u$ that fall into $A_n(\Theta)$. Then,*

$$\forall \varepsilon > 0, \quad \mathbb{P}\left(\left|N_{n,u}(A_n(\Theta)) - N_{n,u}^\diamond(A_n(\Theta))\right| > \varepsilon\right) \leqslant 16(n+1)^{2d} e^{-\varepsilon^2/128n}.$$

Proof.

The proof is similar to Lemma 6.3 in [Elie-Dit-Cosaque & Maume-Deschamps 2022] without bootstrap considerations. □

So:

$$\begin{aligned}
\mathbb{E}\left[(U_n)^2\right] &= \mathbb{E}\left[\left(\sum_{j=1}^{n} \alpha_j^\diamond \varepsilon_j^\diamond\right)^2\right] \\
&= \sum_{j=1}^{n}\sum_{m=1}^{n} \mathbb{E}\left[\alpha_j^\diamond \alpha_m^\diamond \varepsilon_j^\diamond \varepsilon_m^\diamond\right] \\
&= \sum_{j=1}^{n} \mathbb{E}\left[\alpha_j^{\diamond 2} \varepsilon_j^{\diamond 2}\right] + \sum_{\substack{1 \leqslant j,m \leqslant n \\ j \neq m}} \mathbb{E}\left[\alpha_j^\diamond \alpha_m^\diamond \varepsilon_j^\diamond \varepsilon_m^\diamond\right] \\
&\overset{\text{def}}{=} I_n + J_n.
\end{aligned}$$

For I_n:

$$\begin{aligned}
I_n &= \mathbb{E}\left[\sum_{j=1}^{n} \alpha_j^{\diamond 2} \varepsilon_j^{\diamond 2}\right] \\
&\leq \mathbb{E}\left[\max \alpha_j^\diamond \sum_{j=1}^{n} \alpha_j^\diamond \varepsilon_j^{\diamond 2}\right] \\
&\leq \mathbb{E}\left[\max \alpha_j^\diamond \max \varepsilon_j^{\diamond 2}\right] \\
&\leq \mathbb{E}\left[\max \alpha_j^\diamond\right] \mathbb{E}\left[\max \varepsilon_j^{\diamond 2}\right].
\end{aligned}$$

Let $\lambda = \frac{\mathbb{E}[N_{n,1}(A_n(\Theta))]}{2}$:

$$\begin{aligned}
\mathbb{E}\left[\max \alpha_j^{\diamond}\right] &= \mathbb{E}\left[\max \frac{1}{B}\sum_{l=1}^{B} \frac{\mathbb{1}_{\mathbf{X}_i^{\diamond} \in A_n(\mathbf{x};\Theta_l,\mathcal{D}_n) \wedge W_i^{\diamond}=1}}{N_{n,1}^{\diamond}(\mathbf{x};\Theta_\ell,\mathcal{D}_n)}\right] \\
&\leq \mathbb{E}\left[\frac{1}{\min N_{n,1}^{\diamond}(\mathbf{x};\Theta_\ell,\mathcal{D}_n)}\right] \\
&\leq \mathbb{E}\left[\frac{\mathbb{1}_{\{\forall \ell, N_{n,1}^{\diamond}(\mathbf{x};\Theta_\ell,\mathcal{D}_n) \geq \lambda\}}}{\min N_{n,1}^{\diamond}(\mathbf{x};\Theta_\ell,\mathcal{D}_n)}\right] + \mathbb{E}\left[\frac{\mathbb{1}_{\{\exists \ell, N_{n,1}^{\diamond}(\mathbf{x};\Theta_\ell,\mathcal{D}_n) < \lambda\}}}{\min N_{n,1}^{\diamond}(\mathbf{x};\Theta_\ell,\mathcal{D}_n)}\right] \\
&\leq \frac{1}{\lambda} + \mathbb{P}(\exists \ell | N_{n,1}^{\diamond}(\Theta_\ell) < \lambda) \\
&\leq \frac{1}{\lambda} + B\mathbb{P}(N_{n,1}^{\diamond}(\Theta) < \lambda).
\end{aligned}$$

Noticing that

$$\left\{N_{n,1}^{\diamond}(A_n(\Theta)) < \frac{\mathbb{E}[N_{n,1}(A_n(\Theta))]}{2}\right\} \subset \left\{\left|N_{n,1}(A_n(\Theta)) - N_{n,1}^{\diamond}(A_n(\Theta))\right| > \frac{\mathbb{E}[N_{n,1}(A_n(\Theta))]}{2}\right\},$$

we have

$$\mathbb{E}\left[\max \alpha_j^{\diamond}\right] \leq \frac{1}{\lambda} + B\mathbb{P}(\left|N_{n,1}(A_n(\Theta)) - N_{n,1}^{\diamond}(A_n(\Theta))\right| > \lambda). \tag{A.9}$$

Using Assumption 2.3 and Lemma A.7, we have C, K and M positive constants such that:

$$\begin{aligned}
\mathbb{E}\left[\max \alpha_j^{\diamond}\right] \leq& \frac{2}{\mathbb{E}[N_{n,1}(A_n(\Theta))]} + B\mathbb{P}(\left|N_{n,1}(A_n(\Theta)) - N_{n,1}^{\diamond}(A_n(\Theta))\right| > \lambda) \\
\leq& \frac{4}{K\sqrt{n}(\ln n)^{\beta}} + 16Cn^{\alpha}(n+1)^{2d}e^{-K^2(\ln n)^{2\beta}/512}.
\end{aligned}$$

Finally:

$$I_n \leq \frac{4}{K\sqrt{n}(\ln n)^{\beta-u}} + 16Cn^{\alpha}(n+1)^{2d}(\ln n)^{u}e^{-K^2(\ln n)^{2\beta}/512}. \tag{A.10}$$

For J_n:

$$\begin{aligned}
J_n &= \sum_{\substack{1\leqslant j,m\leqslant n\\ j\neq m}} \mathbb{E}\left[\alpha_j^\diamond \alpha_m^\diamond \varepsilon_j^\diamond \varepsilon_m^\diamond\right] \\
&= \sum_{\substack{1\leqslant j,m\leqslant n\\ j\neq m}} \mathbb{E}\left[\mathbb{E}\left[\alpha_j^\diamond \alpha_m^\diamond \varepsilon_j^\diamond \varepsilon_m^\diamond | \mathcal{D}_n, \mathbf{X}_1^\diamond, \ldots, \mathbf{X}_n^\diamond, \varepsilon_j^\diamond\right]\right] \\
&= \sum_{\substack{1\leqslant j,m\leqslant n\\ j\neq m}} \mathbb{E}\left[\alpha_j^\diamond \alpha_m^\diamond \mathbb{E}\left[\varepsilon_j^\diamond \varepsilon_m^\diamond | \mathcal{D}_n, \mathbf{X}_1^\diamond, \ldots, \mathbf{X}_n^\diamond, \varepsilon_j^\diamond\right]\right] \\
&\quad \text{because } \alpha_j^\diamond, \alpha_m^\diamond \text{ are } \mathcal{D}_n, \mathbf{X}_1^\diamond, \ldots, \mathbf{X}_n^\diamond \text{ measurable.} \\
&= \sum_{\substack{1\leqslant j,m\leqslant n\\ j\neq m}} \mathbb{E}\left[\alpha_j^\diamond \alpha_m^\diamond \varepsilon_j^\diamond \mathbb{E}\left[\varepsilon_m^\diamond | \mathcal{D}_n, \mathbf{X}_1^\diamond, \ldots, \mathbf{X}_n^\diamond, \varepsilon_j^\diamond\right]\right] \\
&= \sum_{\substack{1\leqslant j,m\leqslant n\\ j\neq m}} \mathbb{E}\left[\alpha_j^\diamond \alpha_m^\diamond \varepsilon_j^\diamond \mathbb{E}\left[\varepsilon_m^\diamond\right]\right] \\
&= 0.
\end{aligned}$$

Using Bienaimé-Tchebychev's inequality, we have:

$$\forall \varepsilon > 0, \mathbb{P}\left(|U_n| \geq \varepsilon\right) \leq \frac{I_n}{\varepsilon^2}. \tag{A.11}$$

Because $\sum_{n\geq 1} I_{n^2} < \infty$, with the Borel-Cantelli lemma:

$$\forall \varepsilon > 0, \mathbb{P}\left(\overline{\lim_{n\to+\infty}} \{|U_{n^2}| \geq \varepsilon\}\right) = 0. \tag{A.12}$$

Thus, $U_{n^2} \to 0$.
We now show that $(U_n)_{n\geq 1}$ converges almost surely to 0.

$$\begin{aligned}
\mathbb{P}\left(\overline{\lim_{n\to+\infty}} \left\{\left|U_n - U_{\lfloor\sqrt{n}\rfloor^2}\right| \geq \varepsilon\right\}\right) &= \mathbb{P}\left(\overline{\lim_{n\to+\infty}} \left\{\sum_{i=\lfloor\sqrt{n}\rfloor^2+1}^{n} |\alpha_i^\diamond \varepsilon_i^\diamond| \geq \varepsilon\right\}\right) \\
&= \mathbb{P}\left(\forall n, \exists N_0 > n, \left\{\sum_{i=\lfloor\sqrt{N_0}\rfloor^2+1}^{N_0} |\alpha_i^\diamond \varepsilon_i^\diamond| \geq \varepsilon\right\}\right) \\
&= \lim_{n\to+\infty} \mathbb{P}\left(\exists N_0 > n, \left\{\sum_{i=\lfloor\sqrt{N_0}\rfloor^2+1}^{N_0} |\alpha_i^\diamond \varepsilon_i^\diamond| \geq \varepsilon\right\}\right) \\
&= \lim_{n\to+\infty} \sum_{N_0>n} \mathbb{P}\left(\sum_{i=\lfloor\sqrt{N_0}\rfloor^2+1}^{N_0} |\alpha_i^\diamond \varepsilon_i^\diamond| \geq \varepsilon\right).
\end{aligned}$$

For a given N_0, let $D(N_0) = (\ln N_0)^\gamma$:

$$\begin{aligned}\mathbb{P}\left(\sum_{i=\lfloor\sqrt{N_0}\rfloor^2+1}^{N_0}|\alpha_i^\diamond\varepsilon_i^\diamond| \geq \varepsilon\right) &\leq \mathbb{P}\left(\exists i \in [\![\lfloor\sqrt{N_0}\rfloor^2+1, N_0]\!], |\varepsilon_i^\diamond| > D(N_0)\right) \\ &\quad + \mathbb{P}\left(D(N_0)\sum_{i=\lfloor\sqrt{N_0}\rfloor^2+1}^{N_0}|\alpha_i^\diamond| \geq \varepsilon\right) \\ &\leq \mathbb{P}\left(\exists i \in [\![\lfloor\sqrt{N_0}\rfloor^2+1, N_0]\!], |\varepsilon_i^\diamond| > D(N_0)\right) \\ &\quad + \mathbb{P}\left(D(N_0)\frac{2\sqrt{N_0}}{\min N_{n,1}^\diamond(\mathbf{x};\Theta_l,\mathcal{D}_n)} \geq \varepsilon\right).\end{aligned}$$

Let us treat the second term:

$$\begin{aligned}\mathbb{P}\left(D(N_0)\frac{2\sqrt{N_0}}{\min N_{n,1}^\diamond(\mathbf{x};\Theta_l,\mathcal{D}_n)} \geq \varepsilon\right) &\leq \mathbb{P}\left(D(N_0)\frac{2\sqrt{N_0}}{\min N_{n,1}^\diamond(\mathbf{x};\Theta_l,\mathcal{D}_n)} \geq \varepsilon, \forall \ell | N^\diamond(A_n(\Theta_\ell)) \geq \lambda\right) \\ &\quad + \mathbb{P}\left(D(N_0)\frac{2\sqrt{N_0}}{\min N_{n,1}^\diamond(\mathbf{x};\Theta_l,\mathcal{D}_n)} \geq \varepsilon, \exists \ell | N^\diamond(A_n(\Theta_\ell)) < \lambda\right).\end{aligned}$$

The first term is zero since $\mathbb{E}\left[N_{n,1}\left(A_n(\Theta)\right)\right] \geq \frac{8\sqrt{N_0}D(N_0)}{\varepsilon}$ for N_0 large enough according to Assumption 2.3. Once again, using that

$$\left\{N_{n,1}^\diamond\left(A_n(\Theta)\right) < \frac{\mathbb{E}\left[N_{n,1}\left(A_n(\Theta)\right)\right]}{2}\right\} \subset \left\{\left|N_{n,1}\left(A_n(\Theta)\right) - N_{n,1}^\diamond\left(A_n(\Theta)\right)\right| > \frac{\mathbb{E}\left[N_{n,1}\left(A_n(\Theta)\right)\right]}{2}\right\},$$

we get:

$$\begin{aligned}\mathbb{P}\left(D(N_0)\frac{2\sqrt{N_0}}{\min N_{n,1}^\diamond(\mathbf{x};\Theta_l,\mathcal{D}_n)} \geq \varepsilon\right) &\leq \mathbb{P}\left(\exists \ell | N^\diamond(A_n(\Theta_\ell)) < \lambda\right) \\ &\leq B\mathbb{P}\left(N^\diamond(A_n(\Theta)) < \lambda\right) \\ &\leq B\mathbb{P}(\left|N_{n,1}\left(A_n(\Theta)\right) - N_{n,1}^\diamond\left(A_n(\Theta)\right)\right| > \lambda) \\ &\leq 16Cn^\alpha(n+1)^{2d}e^{-K^2(\ln n)^{2\beta}/512}.\end{aligned}$$

Finally we have:

$$\mathbb{P}\left(\sum_{i=\lfloor\sqrt{N_0}\rfloor^2+1}^{N_0}|\alpha_i^\diamond\varepsilon_i^\diamond| \geq \varepsilon\right) \leq C(N_0 - \lfloor\sqrt{N_0}\rfloor^2)\theta^{D(N_0)} + 16Cn^\alpha(n+1)^{2d}e^{-K^2(\ln n)^{2\beta}/2048}.$$

Using the Borel-Cantelli lemma:

$$\forall \varepsilon > 0, \mathbb{P}\left(\overline{\lim_{n\to+\infty}}\left\{\left|U_n - U_{\lfloor\sqrt{n}\rfloor^2}\right| \geq \varepsilon\right\}\right) = 0. \tag{A.13}$$

Finally we have that $(U_n)_{n\geq 1}$ goes to 0 almost surely. Therefore, $|\tau^{\diamond}(\mathbf{x}) - \tau(\mathbf{x})|$ goes to 0.

The quantity $|\hat{\tau}(\mathbf{x}) - \tau^{\diamond}(\mathbf{x})|$ is now treated. We use the same decomposition and consider separately but in a similar fashion $|\hat{\tau}_1(\mathbf{x}) - \tau_1^{\diamond}(\mathbf{x})|$ and $|\hat{\tau}_0(\mathbf{x}) - \tau_0^{\diamond}(\mathbf{x})|$:

$$|\hat{\tau}_1(\mathbf{x}) - \tau_1^{\diamond}(\mathbf{x})| = \left| \frac{1}{B} \sum_{l=1}^{B} \sum_{j=1}^{n} \frac{\mathbb{1}_{\mathbf{X}_j \in A_n(l)} \mathbb{1}_{W_j=1}}{N_{n,1}(\mathbf{x}; \Theta_l, \mathcal{D}_n)} Y_j - \frac{\mathbb{1}_{\mathbf{X}_j^{\diamond} \in A_n(l)} \mathbb{1}_{W_j^{\diamond}=1}}{N_{n,1}^{\diamond}(\mathbf{x}; \Theta_l, \mathcal{D}_n)} Y_j^{\diamond} \right|.$$

We decompose as follows:

$$\begin{aligned}
&\left| \sum_{j=1}^{n} \frac{\mathbb{1}_{\mathbf{X}_j \in A_n(l)} \mathbb{1}_{W_j=1}}{N_{n,1}(\mathbf{x}; \Theta_l, \mathcal{D}_n)} Y_j - \frac{\mathbb{1}_{\mathbf{X}_j^{\diamond} \in A_n(l)} \mathbb{1}_{W_j^{\diamond}=1}}{N_{n,1}^{\diamond}(\mathbf{x}; \Theta_l, \mathcal{D}_n)} Y_j^{\diamond} \right| \\
\leq &\left| \sum_{j=1}^{n} \frac{\mathbb{1}_{\mathbf{X}_j \in A_n(l)} \mathbb{1}_{W_j=1}}{N_{n,1}(\mathbf{x}; \Theta_l, \mathcal{D}_n)} Y_j - \mathbb{E}\left[Y | \mathbf{X} \in A_n(\Theta), W = 1\right] \right| \\
&+ \left| \sum_{j=1}^{n} \frac{\mathbb{1}_{\mathbf{X}_j^{\diamond} \in A_n(l)} \mathbb{1}_{W_j^{\diamond}=1}}{N_{n,1}^{\diamond}(\mathbf{x}; \Theta_l, \mathcal{D}_n)} Y_j^{\diamond} - \mathbb{E}\left[Y | \mathbf{X} \in A_n(\Theta), W = 1\right] \right|.
\end{aligned}$$

Let Λ be the statement "(A.2) verified for $A' = \{\mathbf{X}_i \in A_n(\Theta) | W_i = 1\}$", for any $\varepsilon > 0$ we have:

$$\begin{aligned}
&\mathbb{P}\left(\left| \sum_{j=1}^{n} \frac{\mathbb{1}_{\mathbf{X}_j \in A_n(l)} \mathbb{1}_{W_j=1}}{N_{n,1}(\mathbf{x}; \Theta_l, \mathcal{D}_n)} Y_j - \frac{\mathbb{1}_{\mathbf{X}_j^{\diamond} \in A_n(l)} \mathbb{1}_{W_j^{\diamond}=1}}{N_{n,1}^{\diamond}(\mathbf{x}; \Theta_l, \mathcal{D}_n)} Y_j^{\diamond} \right| > \varepsilon \right) \\
\leq &\mathbb{P}\left(\left| \sum_{j=1}^{n} \frac{\mathbb{1}_{\mathbf{X}_j \in A_n(l)} \mathbb{1}_{W_j=1}}{N_{n,1}(\mathbf{x}; \Theta_l, \mathcal{D}_n)} Y_j - \mathbb{E}\left[Y | \mathbf{X} \in A_n(\Theta), W = 1\right] \right| > \frac{\varepsilon}{2}, \Lambda \right) \\
&+ \mathbb{P}\left(\left| \sum_{j=1}^{n} \frac{\mathbb{1}_{\mathbf{X}_j^{\diamond} \in A_n(l)} \mathbb{1}_{W_j^{\diamond}=1}}{N_{n,1}^{\diamond}(\mathbf{x}; \Theta_l, \mathcal{D}_n)} Y_j^{\diamond} - \mathbb{E}\left[Y | \mathbf{X} \in A_n(\Theta), W = 1\right] \right| > \frac{\varepsilon}{2}, \Lambda \right) \\
&+ 8(n+1)^{2d} e^{\frac{-C(\ln n)^{2\beta}}{128}} \left(\text{i.e. } \mathbb{P}\left(\Lambda^C\right) \right).
\end{aligned}$$

Note that:

$$\begin{aligned}
&\mathbb{P}\left(\left| \sum_{j=1}^{n} \frac{\mathbb{1}_{\mathbf{X}_j \in A_n(l)} \mathbb{1}_{W_j=1}}{N_{n,1}(\mathbf{x}; \Theta_l, \mathcal{D}_n)} Y_j - \mathbb{E}\left[Y | \mathbf{X} \in A_n(\Theta), W = 1\right] \right| > \frac{\varepsilon}{2}, \Lambda \right) \\
\leq &\mathbb{P}\left(\frac{n}{N_{n,1}(\mathbf{x}; \Theta_l, \mathcal{D}_n)} \left| \frac{1}{n} \sum_{j=1}^{n} Y_j \mathbb{1}_{\mathbf{X}_j \in A_n(l)} \mathbb{1}_{W_j=1} - \mathbb{E}[Y \mathbb{1}_{\mathbf{X} \in A_n(\Theta)} \mathbb{1}_{W=1}] \right| > \frac{\varepsilon}{4}, \Lambda \right) \\
&+ \mathbb{P}\left(\left| \mathbb{E}[Y \mathbb{1}_{\mathbf{X} \in A_n(\Theta)} \mathbb{1}_{W=1}] \right| \left| \frac{n}{N_{n,1}(\mathbf{x}; \Theta_l, \mathcal{D}_n)} - \frac{1}{\mathbb{P}(\mathbf{X} \in A')} \right| > \frac{\varepsilon}{4}, \Lambda \right).
\end{aligned}$$

We can treat this term the same way as $T_{1,1,1}$.

$$\begin{aligned}
&\mathbb{P}\left(\left|\sum_{j=1}^{n}\frac{\mathbb{1}_{\mathbf{X}_j\in A_n(l)}\mathbb{1}_{W_j=1}}{N_{n,1}(\mathbf{x};\Theta_l,\mathcal{D}_n)}Y_j-\mathbb{E}\left[Y|\mathbf{X}\in A_n(\Theta),W=1\right]\right|>\frac{\varepsilon}{2},\Lambda\right)\\
\leq&24\left(\frac{Ce(\ln n)^\delta\sqrt{n}}{\varepsilon(\ln n)^\beta}\log\left(\frac{Ce(\ln n)^\delta\sqrt{n}}{\varepsilon(\ln n)^\beta}\right)\right)^{2d}\exp\left(\frac{-C\varepsilon^2(\ln n)^{2\beta}}{(\ln n)^{2\delta}}\right)+C\frac{\sqrt{n}e^{(\ln n)^\delta\frac{\ln\theta}{q}}}{\varepsilon(\ln n)^\beta}\\
&+8(n+1)^{2d}e^{-\frac{\varepsilon^2C(\ln n)^{2\beta-2\delta}}{2048}}.
\end{aligned}$$

Second term is treated with the same idea but needs additional work:

$$\begin{aligned}
&\mathbb{P}\left(\left|\sum_{j=1}^{n}\frac{\mathbb{1}_{\mathbf{X}_j^\diamond\in A_n(l)}\mathbb{1}_{W_j^\diamond=1}}{N_{n,1}^\diamond(\mathbf{x};\Theta_l,\mathcal{D}_n)}Y_j^\diamond-\mathbb{E}\left[Y|\mathbf{X}\in A_n(\Theta),W=1\right]\right|>\frac{\varepsilon}{2},\Lambda\right)\\
\leq&\mathbb{P}\left(\left|\sum_{j=1}^{n}\frac{\mathbb{1}_{\mathbf{X}_j^\diamond\in A_n(l)}\mathbb{1}_{W_j^\diamond=1}}{N_{n,1}^\diamond(\mathbf{x};\Theta_l,\mathcal{D}_n)}Y_j^\diamond-\mathbb{E}\left[Y|\mathbf{X}\in A_n(\Theta),W=1\right]\right|>\frac{\varepsilon}{2},N_{n,1}^\diamond(\mathbf{x};\Theta_l,\mathcal{D}_n)\geq\lambda,\Lambda\right)\\
&+\mathbb{P}\left(N_{n,1}^\diamond(\mathbf{x};\Theta_l,\mathcal{D}_n)<\lambda\right).
\end{aligned}$$

The term:

$$\mathbb{P}\left(\left|\sum_{j=1}^{n}\frac{\mathbb{1}_{\mathbf{X}_j^\diamond\in A_n(l)}\mathbb{1}_{W_j^\diamond=1}}{N_{n,1}^\diamond(\mathbf{x};\Theta_l,\mathcal{D}_n)}Y_j-\mathbb{E}\left[Y|\mathbf{X}\in A_n(\Theta),W=1\right]\right|>\frac{\varepsilon}{2},N_{n,1}^\diamond(\mathbf{x};\Theta_l,\mathcal{D}_n)\geq\lambda,\Lambda\right) \tag{A.14}$$

is treated as previously. The last term is close to an expression already bounded:

$$\begin{aligned}
&\mathbb{P}\left(N_{n,1}^\diamond(\mathbf{x};\Theta_l,\mathcal{D}_n)<\lambda\right)\\
\leq&\mathbb{P}(\left|N_{n,1}\left(A_n(\Theta)\right)-N_{n,1}^\diamond\left(A_n(\Theta)\right)\right|>\lambda)\\
\leq&16C(n+1)^{2d}e^{-K^2(\ln n)^{2\beta}/512}.
\end{aligned}$$

As per to Borel-Cantelli, provided that $2\beta-2\delta>1$, we conclude that $|\hat{\tau}_1(\mathbf{x})-\tau_1^\diamond(\mathbf{x})|$ goes to 0.

Finally we have $|\hat{\tau}(\mathbf{x})-\tau(\mathbf{x})|$ goes to 0.

□

B Graphical illustrations

Here we consider a slightly modified version of the first simulation example to compare graphically the GRF and HTERF results. Let $\mathbf{X}_i \sim U([0,5]^p)$, $W_i \sim Bern(0.5)$ and $Y_i = \tau(\mathbf{X}_i)W_i + \beta\gamma(\mathbf{X}_i)$. Where $p = 10, \tau(\mathbf{x}) = \sin(x^{(1)})$ and $\gamma(\mathbf{x}) = \cos(2x^{(2)} + 3x^{(3)})$. We consider the same values of β as previously.

A test set of X is generated uniformly over $[0,5]^p$. Since τ only depends on variable $X^{(1)}$, we plot the estimated and the true CATE against this variable.

The variance of GRF estimator appears to be higher than HTERF. Also the HTERF estimation is smoother especially where the curvature of the true effect curve is the most important: for the three configurations, GRF struggles at the top of the sinus curve.

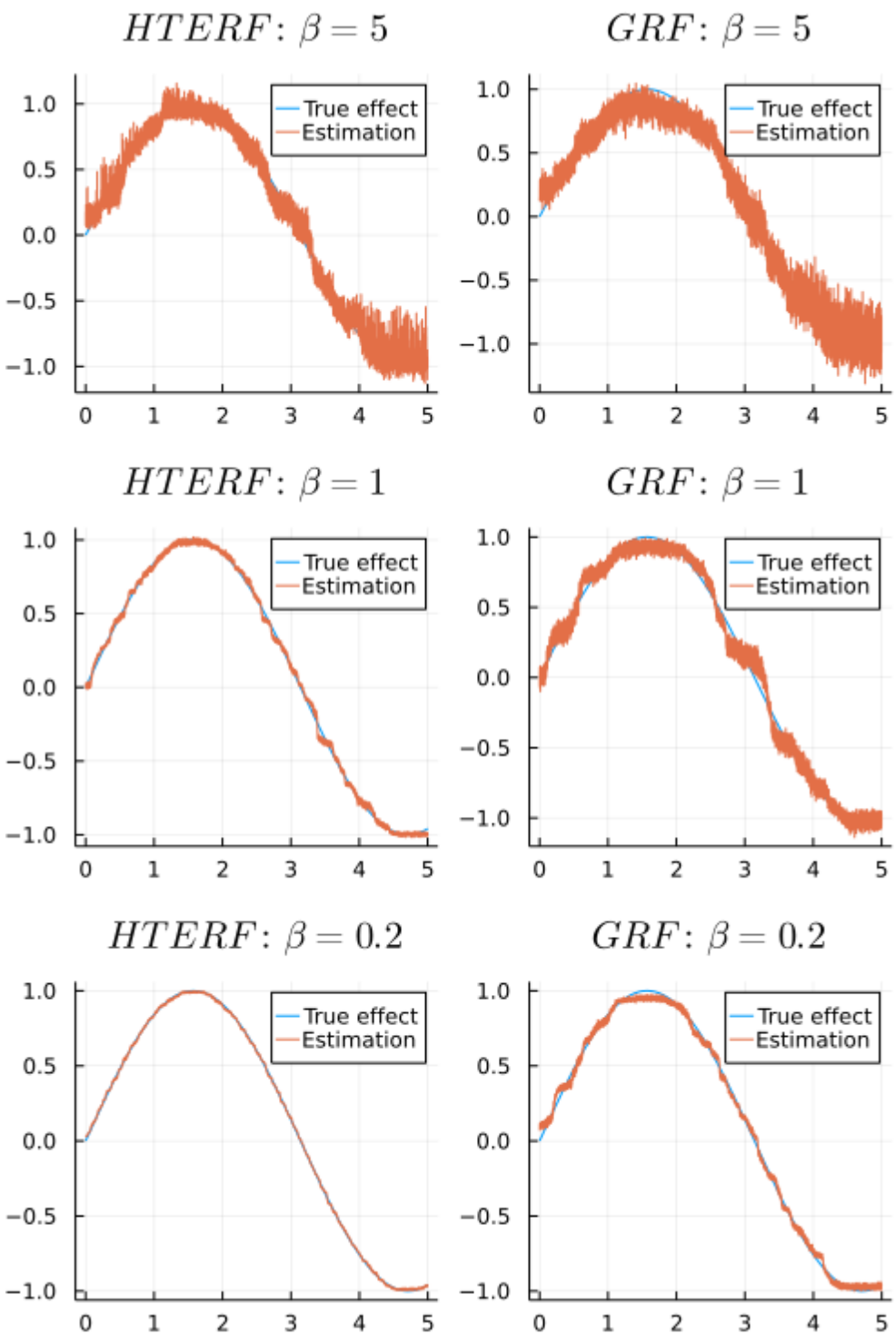


Figure 2.2: Comparison of CATE estimation using HTERF and GRF.

C The Julia package

The designing of the package associated implementing the HTERF algorithm has been a substantial part of the work on this subject. In this section the motivation to create the `Julia` package `CausalForest` is exposed, then the main functions are described and are used to show how some previous results have been obtained. Finally the perspectives of this work and the possible improvements are discussed.

C.1 Introduction

Here are the current packages that allow to build causal forest with the GRF splitting criterion:

- **grf - R**:
  The grf[2] package in R provides functions for fitting generalized random forests, including methods for estimating causal effects.
- **econml - Python**:
  The econml[3] package in Python, developed by Microsoft, includes the CausalForest class for fitting causal forests to estimate heterogeneous treatment effects.
- **causalml - Python**:
  The causalml[4] package in Python includes the CausalForest class, which provides an implementation of causal forest algorithms for estimating treatment effects.

The design of the package `CausalForest` has been heavily inspired by the package `DecisionTree`, that implements classification and regression random forest in `Julia`. Note that there exists a package named `CausalInference` for `Julia`, but it does not concern the CATE estimation, it is dedicated to causal discovery i.e. estimating the causal links between the available variables and estimating causal effect using tools like regressions. Another noteworthy `Julia` package is `CausalELM`, that enables estimation of causal effects extreme learning machines (feed forward neural networks that determine their weights analytically instead of using backpropagation as in the standard neural networks.

`CausalForest` allows the exploration of options that do not fit in the theoretical background exposed in Section 2.4, for example bootstrapping can be used instead of subsampling or the assumption about honesty can be relaxed. Also the package has been designed in such a way that the splitting criterion can be easily modified. One of the motivation to develop this new package, was that is was not straightforward and easy to modify the splitting criterion in the `Python` packages, while the structure of the `Julia` package `DecisionTree` is much more accessible.

[2]Package's CRAN page: https://cran.r-project.org/web/packages/grf/index.html
[3]Package's GitHub page: https://github.com/py-why/EconML
[4]Package's GitHub page: https://github.com/uber/causalml

C.2 The CausalForest package

The `CausalForest` package is available in the General registery of `Julia`, the repository keeping tracks of all available packages and can be obtained in `Julia` using the following commands:

```
using Pkg
Pkg.add("CausalForest")
```

Development of `CausalForest` takes place on GitHub (https://github.com/BereniceAlexiaJocteur/CausalForest.jl), the source files are in the folder "src".

The `src` folder contains five main folders:

1. `causal`: This folder contains the implementation of HTERF, there is two files inside. The first one named `tree.jl`, is dedicated to the construction of one causal tree, it contains especially the implementation of the splitting criterion. The second file named `main.jl`, contains the code to build causal tree objects and causal forest objects, this is where the functions `build_forest` and `apply_forest` are. These two functions are the ones used by the user to fit an HTERF and to apply the causal forest on a test point.

2. `causal_old`: This folder as the same structure as the previous one. The only difference is that the prior centering is done with a linear regression instead of a random forest. It is the code used in Section 2.5.4.

3. `classification`: This folder as the same structure as the previous ones. It contains the code for out of bag (OOB) classification random forests.

4. `regression`: Same as above for OOB regression random forests. Those random forests are used for the centering in the HTERF.

5. `sensitivity`: This is the code for the notion of importance described in Section 2.4.3.

The inputs of the function `build_forest` are:

- `bootstrap`: If `true` we sample for each tree via bootstrap else we use subsampling.
- `honest`: If `true` we use 2 samples one too build splits and the other one to fill leaves otherwise we use the whole sample for the two steps.
- `labels`: A vector containing the values of Y.
- `treatment`: A vector containing the values of W.
- `features`: A matrix containing the values of X.

- `const_mtry`: If `true` we use a constant $\mathcal{M}_{try}$ = `m_pois` otherwise we use a random $\mathcal{M}_{try}$ following $\min(\max(Poisson(\texttt{m\_pois}), 1), \texttt{number\_of\_features})$, where $Poisson(\texttt{m\_pois})$ draws one observation from a Poisson distribution of parameter `m_pois` and `number_of_features` is the total number of covariates (i.e. the number of columns of X); (default=true).

- `m_pois`: If `m_pois` $= -1$, it is set to $sqrt(\texttt{number\_of\_features})$ else it uses the inputted value according to the description in previous point, (default=-1).

- `n_trees`: Number of causal trees to train (default=10).

- `n_trees_centering`: Number of trees in the random forest used for prior centering (default=100).

- `optimisation`: If `true` we use cross validation to tune the hyper parameters of the centering random forest (default=true).

- `partial_sampling`: If subsampling is used this is the subsampling rate (default=0.7).

- `honest_proportion`: If honest trees are built, this is the proportion of the sample used to build the tree and the complementary is used to fill leaves (default=0.5).

- `max_depth`: Maximum depth of the causal trees (default=-1 i.e. no maximum).

- `min_samples_leaf`: The minimum number of samples each leaf needs to have (default=5).

- `min_samples_split`: The minimum number of samples needed for a split (default=10).

- `rng`: The random number generator or the seed to use (default=Random.GLOBAL_RNG).

The output of the function `build_forest` is an object of class `EnsembleCausal`, that contains the following fields:

- `trees`: A vector of objects of class TreeCausal, the details of this class will not be explained here to stay concise.

- `bootstrap`: A boolean indicating if bootstrapping hase been used (`true`) or subsampling (`false`).

- `honest`: A boolean indicating if a honest forest has been built or all the data has been used to build the splits and fill the leaves.

- `X`: The matrix X.

- `Y`: The vector Y.
- `T`: The vector W.
- `model_Y`: The object of the OOB regression random forest used to center Y.
- `model_T`: The object of the OOB classification random forest used to center W.
- `Y_center`: The vector Y after centering.
- `T_center`: The vector W after centering.

The function `apply_forest` is simpler, its inputs are:

- `forest`: An object of class `EnsembleCausal`, this is the causal forest on which we want to make the prediction.
- `x`: The matrix where which line is a test point.

It returns an array that contains the predictions for all the test points.

Finally, the function `importance`, takes as argument an object of class `EnsembleCausal`, and returns a list of the importance of each covariate in their order of apparition in the matrix X.

C.3 Examples

As an example of the use of the `CausalForest` package, the code used to obtain the results in Section 2.5.2 is presented.

```
using CausalForest
using RCall
using StatsBase
using Statistics
using Random
using Distributions
```

Aside from the `CausalForest` package, the other packages are loaded to compare results with `grf` package (`RCall`), compute RMSE and means (`StatsBase) and Statistics`, generate random numbers (`Random`) and generate data according to a certain distribution in order to generate the data (W, X, Y) (`Distributions`). For the CATE estimation using HTERF, the package `CausalForest` is standalone.

```
Random.seed!(123);
n, m = 10^4, 3;
@rlibrary grf
R"""
```

```
set.seed(123)
res <- data.frame()
"""
errors_grf = zeros(1)
errors_hterf = zeros(1)
hterf1 = zeros(1)
grf1 = zeros(1)
for j in 1:1
    u = Uniform(-pi,pi);
    features = rand(u, (n, m));
    X = features;
    b = Bernoulli();
    T = convert(Vector{Int64},rand(b, n));
    Y = cos.(features[:,2])+sin.(features[:,1]).*(T.+2).^3
    Xtest = rand(u, (n, m));
    tau = 19*sin.(Xtest[:,1])
    @rput X T Y Xtest tau
    R"""
    cf <- grf::causal_forest(X, Y, T, num.trees=500,
        tune.num.trees=500, sample.fraction=0.7, ci.group.size=1)
    tau.hat <- predict(cf, Xtest)$predictions
    mse = sqrt(mean((tau.hat - tau)^2))
    g_1 = grf::variable_importance(cf)[1]
    """
    @rget mse g_1
    cf = build_forest(false, true, Y, T, X, true, m, 500, 500)
    tauhat = apply_forest(cf, Xtest)
    hterf1[j] = importance(cf)[1]
    errors_hterf[j] = rmsd(tau, tauhat)
    grf1[j] = g_1
    errors_grf[j] = mse
end
err_hterf = mean(errors_hterf)
err_grf = mean(errors_grf)
grf_1 = mean(grf1)
hterf_1 = mean(hterf1)
@rput err_grf err_hterf grf_1 hterf_1
R"""
dfgrf = data.frame(method = "GRF", RMSE = err_grf, imp1 = grf_1)
dfhterf = data.frame(method = "HTERF", RMSE = err_hterf, imp1 = hterf_1)
res = rbind(res, dfgrf, dfhterf)

"""
```

```
@rget res
print(res)
```

In this example the three main functions of the package have been used in context. Other examples can be found on the GitHub repository (https://github.com/BereniceAlexiaJocteur/CausalForest.jl) in the folder named "Notebooks".

C.4 Discussion

The `CausalForest` is a first package in `Julia` offering to estimate CATE with causal forests. It also offers an interpretability tool, by estimating the importance of variables in the forest. Although the code is already parallelized, it could still be optimized for faster execution.

The emphasis has been placed on the modularity of the package, making it easy to subsequently adapt the code to modify the splitting criterion or even tailor it to estimate other quantities of interest than CATE.

Chapter 3

Applications of HTERF

In this chapter two applications of HTERF on real datasets are proposed. The first application has been developed with data from Natixis, as part of the CIFRE convention. The second application concerns the study of the phenomenons El Niño and La Niña on the Australian climate.

Contents

3.1 Credit risk application

The cost of risk is a measure of future risk inherent in the current portfolio. It can be understood as the adjustment of the provision stock between two periods and is composed of net provisions and reversals of depreciations and provisions for credit risk, as well as losses on irrecoverable receivables. Therefore, the cost of risk is an indicator of expected losses and measures the effort made by an entity over a given period to protect itself against estimated future losses on its loan portfolio. We consider an application of the HTERF on modelisation of Cost of Risk against macroeconomic variables on Natixis data.

3.1.1 Introduction

Cost of Risk (CoR) in euros is considered on counterpart level Non Performing Loans (NPL). Note that this is an intermediate balance excluding interests on doubtful debts.

$$
\begin{aligned}
CoR = &\text{ New Allowances} \\
&- \text{ Reversal of Allowances} \\
&+ \text{ Write-Offs} \\
&- \text{ Recoveries (of previously written-off amounts)} \\
&+ \text{ Conversion Gaps (from exposures in foreign currencies)}
\end{aligned}
$$

Interest-related costs are excluded from CoR calculation.

We consider quarterly data from Q2 2007 to Q4 2020, for counterparts in the corporate scope.

Following the framework previously presented, we consider:

- Y as the quarterly CoR for each counterpart,
- T as a binary variable, $T = 1$ if we are in a crisis period, $T = 0$ otherwise,
- X a set of macroeconomic variables associated to each counterpart at a given quarter.

Thus CATE represents the impact of a crisis on CoR conditionally on the macroeconomic environment.
The crisis "treatment" variable has been set to one for the following years:

- 2009: Subprime mortgage crisis
- 2013: European sovereign debt crisis
- 2019, 2020: COVID-19 crisis

The following macro economic variables have beer retained:

- A stock market index variable: it consists of the daily returns of the following indices depending on the country of the counterpart. For France we consider CAC40, for European countries excluding France we take EuroStoxx50, for north American countries we use S&P500 and for all others countries we take a combination 80% of S&P and 20% of Euro Stoxx.
- VIX returns, VIX is an index that measures the stocks market's expectation of volatility based on S&P500 index options.
- Gold returns.
- Indices relative to European corporate bonds: we consider individually AAA rated bonds and bonds rated down BBB, we also consider the spread between these two quantities (using iBoxx corporate indices).

- Government bonds returns: we consider the three following for 10 years duration, France, Germany and United States, we also consider the spread between France and Germany.

- Returns of Euribor for the following maturities: one month, three months, six months and twelve months. Euribor is a daily reference rate, based on the average interest rate at which Eurozone banks borrow unsecured funds in the euro interbank market.

- Quarterly variations of GDP and unemployment rate for each country, we also considered one year lagged versions of these two variables, indeed the economical research department of Natixis can provide predictions on these two variables.

3.1.2 HTERF results

We fitted two HTERF causal forests, the first one with non lagged version of GDP and unemployment rate variables and the second one with one year lagged version of these variables. Causal forests have 500 trees and the centering forests also have 500 trees.
The most informative variables (with importances) in the first forest are:

- Unemployment rate (48%)

- VIX (16%)

- Euribor 12 months (7%)

With lagged GDP and unemployment rate we have:

- Unemployment rate (46%)

- GDP (26%)

- Stock market index (9%)

The highlighted variables in the second model are consistent with the main risk drivers used to model CoR in the bank.

We propose to backtest the results on the period crisis, assuming that the CoR if there were no crisis would have be the same as the previous years. On the crisis periods we observe $Y(1)$ and we assume that $Y(0)$ unobserved for these crisis years is close to the CoR observed on the previous years. Thus we can propose a true value of the crisis effect against which the results of HTERF will be challenged. We define the error rate as follows:

$$\frac{\text{estimated effect - true effect}}{\text{true effect}} \tag{3.1.1}$$

Year	Non-lagged forest	Lagged model
2009	-165%	-62%
2013	-19%	-46%
2019	-3%	-33%
2020	2%	-60%

Table 3.1: Error rates of HTERF forests on CoR data.

The error rate is considered instead of the CATE RMSE previously used to ensure the privacy of the COR, which is a sensitive information.

According to Table 3.1, the first HTERF model gives the best results except for year 2009. We assumed that the poor performance for the subprime crisis might be due to the lack of macro economics variables explaining the amplitude of its impact.

3.1.3 Inclusion of new variables

We added new explanatory variables from Refinitiv (a global provider of financial market data, subsidiary of London Stock Exchange Group), namely:

- Core inflation index in OECD. Core inflation is the change in the costs of goods and services, but it does not include those from the food and energy sectors. This measure of inflation excludes these items because their prices are much more volatile.
- Trade in goods and services, which is defined as the transactions in goods and services between residents and non-residents. It is measured in million USD at 2015 constant prices, all OECD countries compile their data.
- Total Domestic Expenditure, it refers to the total amount of money spent within OECD on goods and services during each quarter. It includes expenditures by households, businesses, and the government on goods and services produced domestically.
- Spot price of Brent. The spot price is the current price in the marketplace at which a barril of brent can be bought or sold for immediate delivery.

The most informative variables (with importances) in the first forest are:

- Unemployment rate (53%)
- Brent (16%)
- GDP (7%)

With lagged GDP and unemployment rate we have:

- Unemployment rate (29%)

- Brent (17%)
- Domestic expenditure (15%)
- GDP (14%)

Once again GDP and unemployment rate are among the most informative variables. The new variable Brent spot price is also one of the most informative in both settings.

Year	Non-lagged forest	Lagged model
2009	-85%	-76%
2013	-5%	-19%
2019	-26%	-44%
2020	-29%	-9%

Table 3.2: Error rates of HTERF forests on CoR data with new OECD variables.

This time as shown in Table 3.2, it is not clear if the better choice is to use the non-lagged version of GPD and unemployment rate. For the non lagged version, which had the best performance without the introduction of the new variables, a smaller error is observed on the 2009 and 2013 crisis, but the estimations are worse during the COVID crisis.

The granularity of the data for the outcome variable Y is on the counterpart level for each quarter. However most of the variables only depends on time (VIX, Gold, bonds indices, Government bonds returns, Euribor and all the variables from Refinitiv), the others depend on the combination between the quarter considered and the country of the counterpart. Thus the size and the diversity of the dataset cannot be fully exploited due to these limitations.

The 'Full international and global accounts for research in input-output Analysis' (FIGARO) tables represent the EU inter-country supply, use, and input-output tables. They can be used to create a new variable "added value" which is defined at the scale of each combination of country and sector of the counterpart. However this new data is only available between 20011 and 2020, it reduces the size of the available data and remove the subprime crisis which is the most problematic in previous uses of HTERF. Also the results were worse with this new variable in term of error rate on past crisis. Also this time the most informative variable is by far the stock market index. We could conclude that the intensity of the post 2008 crisis are henceforth explained by financial indices and no more by macro economic considerations.

3.1.4 Discussion

The introduction of new explanatory variables which describe better the specificities of each counterparts could lead to better results. The introduction of sectoral value added per country was a first unsuccessful try. Other variables could be considered

in the future to enhance the performance of HTERF.
The use of the importance in these causal forests lead to two interesting findings:

- The GDP and unemployment rate variables, which are the most important in the ECL models, are predominant in the first four HTERF models considered.
- A shift has been noticed between the 2008 crisis and the followings. The financial indices tend to gain more importance to the detriment of macro economic variables, to explain the impact of crisis on cost of risk.

These observations are consistent with the analysis of the Natixis experts.

3.2 Climatic application

This section is dedicated to a study of the impact of El Niño and La Niña phenomena on east Australia rainfall.

3.2.1 Introduction

El Niño and La Niña [Autralian Government - Bureau of Meteorology 2012] are opposite phases of the El Niño-Southern Oscillation (ENSO) climate pattern, which occurs in the tropical Pacific Ocean. These phenomena have significant impacts on weather patterns around the world.

1. El Niño:
    - El Niño is characterized by warmer-than-average sea surface temperatures in the central and eastern tropical Pacific Ocean.
    - It typically occurs irregularly every 2 to 7 years, lasting for about 9 to 12 months.
    - El Niño events can lead to a variety of weather disruptions globally, including:
      - Increased rainfall and flooding in the eastern Pacific (Peru and Ecuador) and drought conditions in the western Pacific (Australia and Indonesia).
      - Warmer temperatures in the United States during the winter, with potential impacts on precipitation patterns.
      - Changes in atmospheric circulation that can influence weather patterns, such as the jet stream, leading to extreme weather events like storms and hurricanes.
2. La Niña:
    - La Niña is the opposite phase of El Niño, characterized by cooler-than-average sea surface temperatures in the central and eastern tropical Pacific Ocean.

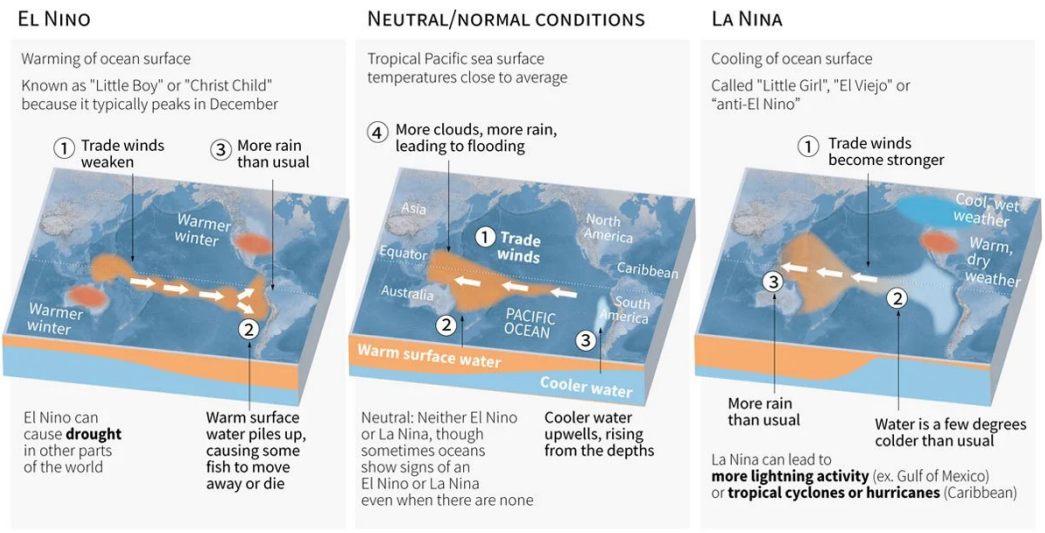


Figure 3.1: Illustration of El Niño and La Niña from [AFP 2023].

- Like El Niño, La Niña occurs irregularly every 2 to 7 years, typically lasting for about 9 to 12 months, a La Niña usually follows 1 or 2 years after an El Niño.
- La Niña events often lead to contrasting weather patterns compared to El Niño, including:
  - Increased rainfall and flooding in the western Pacific (Australia and Indonesia) and drier conditions in the eastern Pacific (Peru and Ecuador).
  - Cooler temperatures in the United States during the winter, with potential impacts on precipitation patterns.
  - Shifts in atmospheric circulation that can influence weather patterns, including increased hurricane activity in the Atlantic basin.

Overall, El Niño and La Niña events have significant impacts on global weather patterns, affecting temperature, precipitation, and atmospheric circulation in various regions of the world. Understanding and monitoring these phenomena are essential for weather forecasting, agriculture, water resource management, and disaster preparedness.
Two indices can be used to monitor these phenomena:

- SOI (Southern Oscillation Index): The Southern Oscillation Index is a measure of the atmospheric pressure differences between Tahiti and Darwin, Australia. It is used as an indicator of the state of the ENSO climate pattern. A negative SOI typically indicates the presence of El Niño conditions, while a positive SOI often signifies La Niña conditions.
- ONI (Oceanic Niño Index): The Oceanic Niño Index is another measure used to monitor the state of the ENSO climate pattern. It is based on sea surface

temperature anomalies in the central and eastern equatorial Pacific Ocean. Positive ONI values indicate El Niño conditions, while negative values indicate La Niña conditions. The ONI is calculated as a three-month running mean of sea surface temperature anomalies.

In what follows, we will focus on the impact of ENSO on the weather in Australia.

3.2.2 Impact on Australian weather

The ENSO has significant impacts on the climate of Australia. These impacts can vary depending on whether the ENSO phase is El Niño, La Niña, or neutral. Here's how each phase generally affects Australia's climate:

El Niño

- Drier and warmer conditions: El Niño typically brings reduced rainfall to many parts of Australia, particularly in the eastern and southeastern regions, see Figure 3.2. This can lead to drought conditions, decreased agricultural productivity, and increased risk of bushfires.
- Increased temperatures: El Niño events are often associated with higher temperatures across much of Australia, see Figure 3.3.
- Stronger winds: Some areas may experience stronger winds during El Niño events, exacerbating fire risk.

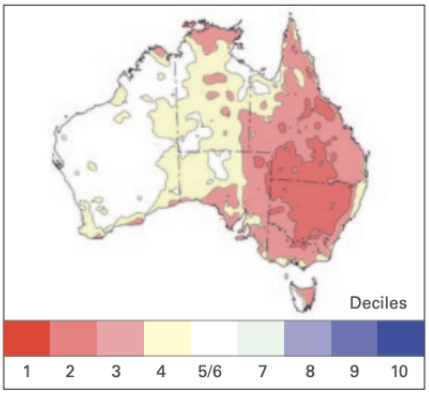
La Niña

- Wetter conditions: La Niña tends to bring above-average rainfall to many parts of Australia, especially in eastern and northern regions, see Figure 3.2. This can lead to flooding, particularly in river catchments.
- Cooler temperatures: La Niña events can bring cooler temperatures to some parts of Australia, particularly in northern and eastern regions, see Figure 3.3.
- Increased tropical cyclone activity: La Niña events can enhance the likelihood of tropical cyclones in the Australian region, particularly in the northern parts of the country.

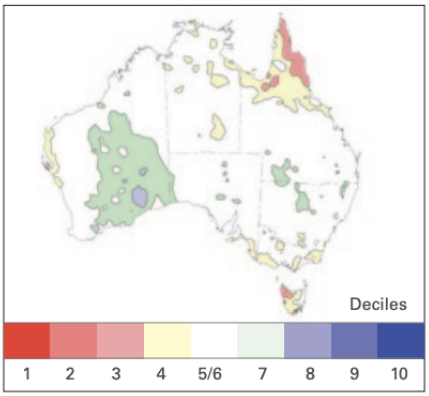
Two meteorological stations have been retained to study the impact of ENSO phases in Australia. Both have good data quality (few missing data for rainfall). Also they are not both in the same region, station 39000 is close to the coral see, while station 44026 is further inland, see Figure 3.4. Data about temperatures were missing for station 39000, so we retrieved temperature data from station 39089 which is 90km away from station 39000, the temperature can be assumed to be similar.

El Niño

El Niño is typically associated with reduced rainfall in northern and eastern Australia



***Winter/spring rainfall** – below average across eastern Australia*

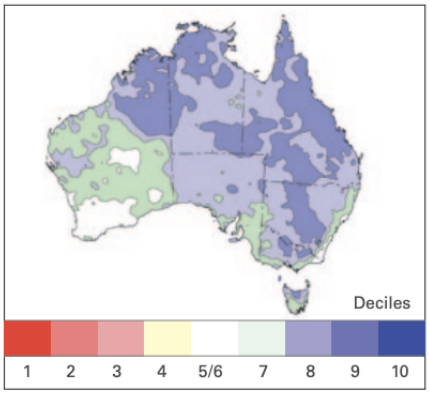


***Summer rainfall** – mostly near average*

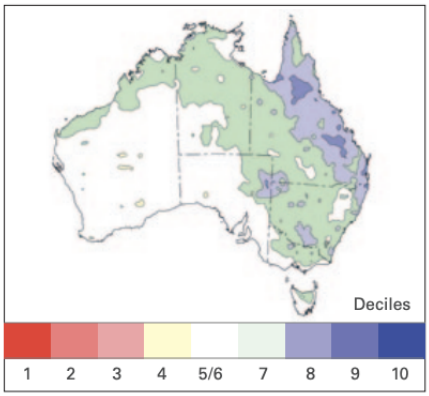
The onset years for the 13 strongest 'classic' El Niño events used are 1905, 1914, 1940, 1941, 1946, 1965, 1972, 1977, 1982, 1991, 1994, 1997 and 2002.

La Niña

La Niña is typically associated with increased rainfall in northern and eastern Australia



***Winter/spring rainfall** – above average across most of eastern and northern Australia*



***Summer rainfall** – above average in eastern and northern Australia*

The onset years for the 13 strongest 'classic' La Niña events used are 1906, 1910, 1916, 1917, 1950, 1955, 1956, 1971, 1973, 1975, 1988, 1998 and 2010.

Figure 3.2: Rainfall patterns during ENSO events in Australia. Each map shows mean rainfall deciles, warm tones indicate below average rainfall while cold tones indicate above average rainfall totals. (Images from [Autralian Government - Bureau of Meteorology 2012])

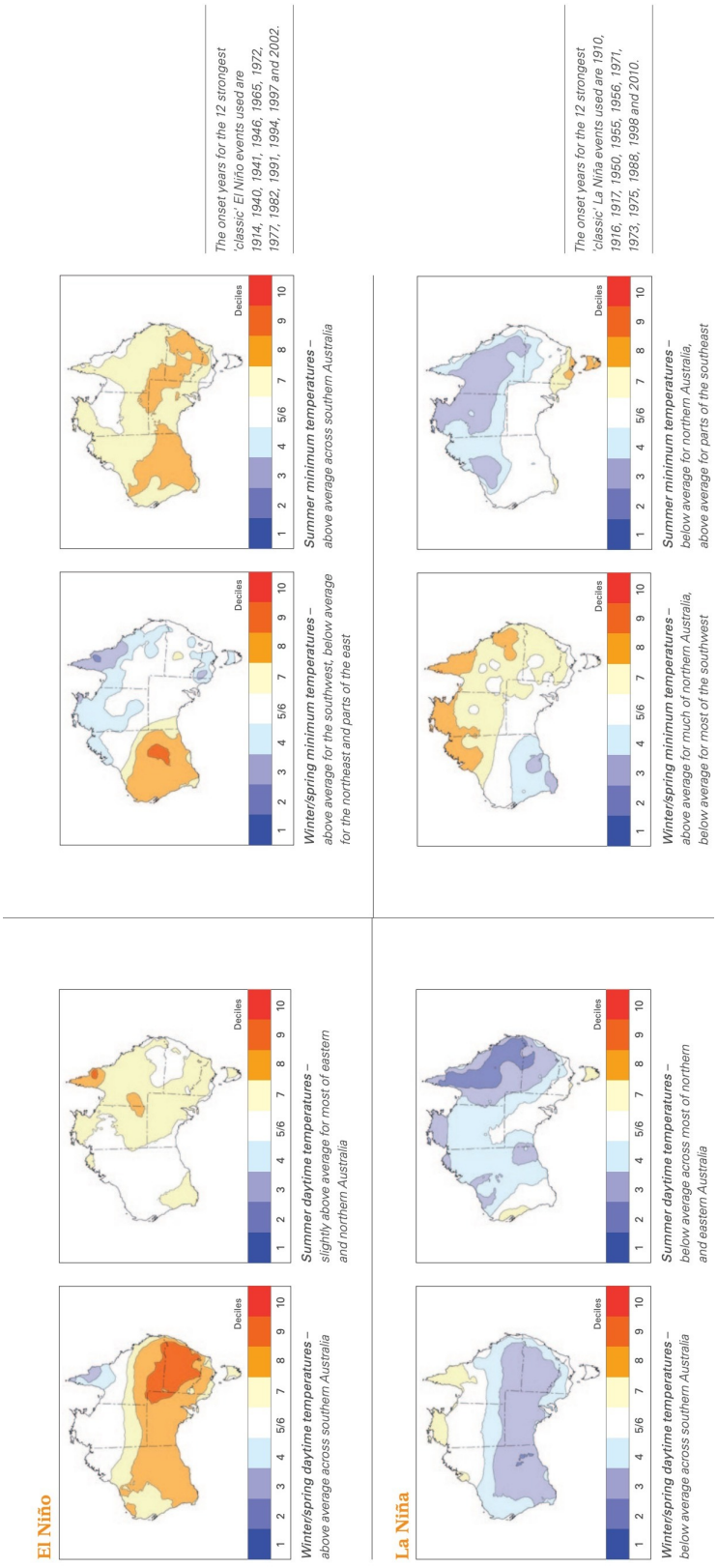


Figure 3.3: Temperature patterns during ENSO events in Australia. (Images from [Autralian Government - Bureau of Meteorology 2012])

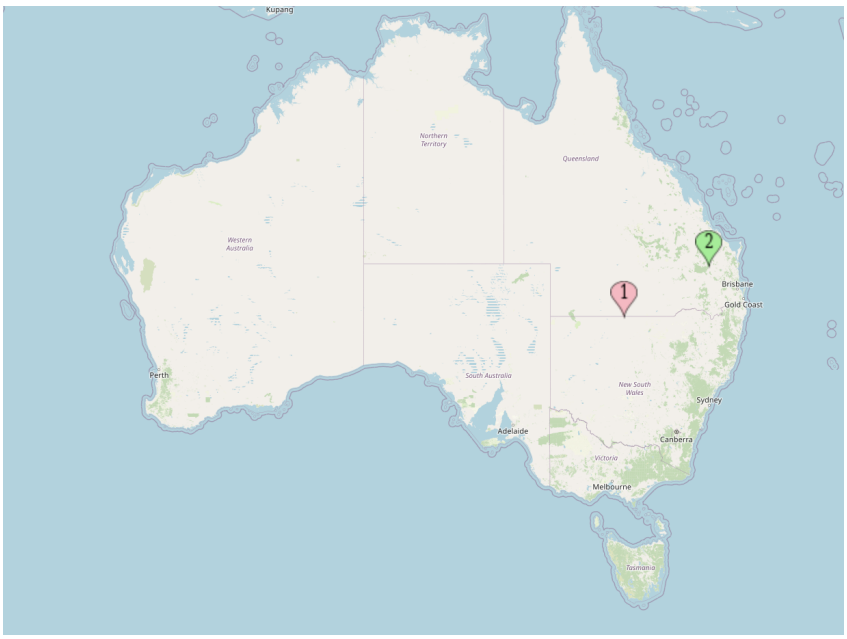


Figure 3.4: Positions of the Australian stations. The first and pink indicator is station 44026 and the second and green indicator is station 39000.

3.2.3 Causal analysis

For each station the following information have been gathered[1] for each month and each station:

- The lowest temperature observed during the considered month, in what follows all temperature are in Celsius degrees,
- the highest temperature observed during the considered month,
- the mean of the lowest temperature observed each day during the considered month,
- the mean of the highest temperature observed each day during the considered month,
- the monthly mean daily global solar exposure, this is the total amount of solar energy falling on a horizontal surface of unit area over a day, it is expressed in MJ/m*m (megajoules per square metre),
- the daily rainfalls in millimeters, in what follow we will consider alternatively the mean of these precipitations over the given month, or the maximum observed during a day of the month.

[1]All these data come from the Australian Bureau of Meteorology: http://www.bom.gov.au/climate/data/

For station 44026 the data are available from 1990 to 2019, for station 39000 (and 39089) the data are available from 1992 to 2019 with 2016 and 2017 missing.

In order to apply HTERF to study the causal effect of ENSO events, we considered the rainfall as the outcome Y, with alternatively the monthly mean and the monthly maximum as described above. We also considered different values for the treatment variable:

- Setting 1: Let $T = 0$ when La Niña occurs and let $T = 1$ when El Niño occurs.
- Setting 2: Let $T = 0$ when neutral conditions occur and let $T = 1$ when El Niño occurs.
- Setting 3: Let $T = 0$ when La Niña or neutral conditions occur and let $T = 1$ when El Niño occurs.

To define when an ENSO event occurs, the ONI previously designed has been used. Thus between 1990 and 2022, the the years have classified as follow using Golden Gate Weather Services[2].

El Niño	La Niña	Neutral
2004, 2006, 2014, 2018	2000, 2005, 2008, 2016	1990, 1992, 1993, 1996
1994, 2002, 2009, 1991, 2015	2017, 2022, 1995, 2011	1997, 1998, 2001, 2003
2020, 2021, 1999, 2007, 2010		2012, 2013, 2019

Table 3.3: Classifications of years, regarding the ENSO event occurring during summer.

HTERF algorithm has been applied to the data from station 44026 during spring and summer. The importances of the exogenous variables have been retrieved from the causal forest and an ATE has been estimated by averaging the CATE obtained with the causal forest, see Table 3.4.

The causal effect of El Niño on rainfall has the starkest contrast (for both max and mean configurations), when the baseline case is La Niña. Which is consistent with the expected effects of ENSO events. Across all settings considered, the most important variables seem to be be the solar exposure as well as the lowest temperature observed during the month.

The same study has been applied to the station 39000, the results are presented in Table 3.5.

This time the results are not consistent between the choices of Y, the impact of El Niño is never in the same direction depending on which outcome Y is chosen. Once again the solar exposure seems to be the most important variable. These inconsistent results might be explained by the lowest quality of data, since the temperature data comes from a proxy station.

[2]Data available here: https://ggweather.com/enso/oni.htm.

Setting	ATE	Imp. lowT	Imp. hiT	Imp. mlowT	Imp. mhiT	Imp. solar
S1 max	-5.3	0.17	0.19	0.24	0.14	0.25
S1 mean	-0.29	0.20	0.23	0.22	0.14	0.20
S2 max	-4.9	0.27	0.18	0.16	0.15	0.24
S2 mean	-0.17	0.21	0.24	0.15	0.15	0.25
S3 max	-3.6	0.20	0.24	0.19	0.16	0.21
S3 mean	-0.14	0.24	0.19	0.18	0.15	0.24

Table 3.4: Results of HTERF on data from station 44026. The tree settings described above are considered while alternatively considering the daily maximum of precipitation during the month (max) or the daily mean of precipitation (mean). The importance of each variable is also displayed: lowest temperature of the month (lowT), highest temperature of the month (hiT), mean of lowest temperature of each day(mlowT), mean of highest temperature of each day(mhiT) and the monthly mean daily global solar exposure (solar). Causal forests of 500 trees have been used with regression random forests of size 500 for the prior centering.

Setting	ATE	Imp. lowT	Imp. hiT	Imp. mlowT	Imp. mhiT	Imp. solar
S1 max	0.87	0.15	0.22	0.14	0.16	0.33
S1 mean	-0.081	0.22	0.17	0.23	0.15	0.23
S2 max	-2.11	0.17	0.30	0.17	0.10	0.27
S2 mean	0.087	0.26	0.18	0.20	0.09	0.27
S3 max	-0.65	0.19	0.25	0.23	0.09	0.23
S3 mean	0.16	0.27	0.18	0.15	0.11	0.29

Table 3.5: Results of HTERF on data from station 39000. The tree settings described above are considered while alternatively considering the daily maximum of precipitation during the month (max) or the daily mean of precipitation (mean). The importance of each variable is also diplayed: lowest temperature of the month (lowT), highest temperature of the month (hiT), mean of lowest temperature of each day(mlowT), mean of highest temperature of each day(mhiT) and the monthly mean daily global solar exposure (solar). Causal forests of 500 trees have been used with regression random forests of size 500 for the prior centering.

3.2.4 Discussion

The inland station (station 44026) showed results which are consistent with the expected effect of El Niño on Australian rainfalls. This study could be extended to more meteorological stations to obtain a cartography of the causal effect of El Niño on rainfalls depending on the geographical area. We also notice that there is not a very informative variables that is being highlighted; furthermore the temperature related variable are quiet similar due to their construction. The addition of new complementary exogenous variables could greatly increase the quality of these results, and allow to gain a greater understanding on how these explanatory variables have an impact on rainfalls. Finally, several intensity of ENSO events can be defined with the ONI indicator, with more data available we could imagine studying other contrasts like "Strong El Niño" vs "Strong La Niña".

Chapter 4

Transfer learning for causal forest

This chapter consists in an article in preparation.

Abstract
Transfer learning addresses the challenge of transferring knowledge from one domain to another. Traditional transfer learning focuses on adapting models trained on a source domain (with a lot of observations) to improve performance on a target domain (with few observations). In this work we consider the case of a model shift and we focus on the transfer learning applied to a causal forest namely HTERF. This causal forest aims to estimate the Conditional Average Treatment Effect (CATE). The approach considered is the offset method presented by [Wang 2016] adapted to a causal context. This method relies on the use of intermediate models in order to estimate the offset between source and target distributions. Our main result is a bound on the CATE error of HTERF on target depending on the error of the intermediates models. Simulation studies show the good performances of this approach in different settings on simulations and on a real-world dataset.

Contents

4.1 Introduction

For completeness, let us recall the causal framework that we are considering. We also give a presentation of transfer learning, with a focus on domain adaptation.

4.1.1 The causal framework

Following the framework outlined in [Imbens & Rubin 2015], we define potential outcomes denoted as $Y(1)$ and $Y(0)$ corresponding to the outcome that would have been observed if treatment or control had been assigned to the quantity of interest Y, respectively. Let $Y = Y(W)$ be the observed outcome, where W represents a binary treatment. Additionally, we incorporate a set of covariates $\mathbf{X} \in \mathbb{R}^d$. The conditional average treatment effect (CATE) at $\mathbf{x}$ is characterized as follows:

$$\tau(\mathbf{x}) = \mathbb{E}\left[Y(1) - Y(0) | \mathbf{X} = \mathbf{x}\right]. \tag{4.1.1}$$

The average treatment effect (ATE) can also be defined:

$$\tau = \mathbb{E}\left[Y(1) - Y(0)\right]. \tag{4.1.2}$$

A standard assumption for identifiability of CATE is unconfoundedness ([Rosenbaum & Rubin 1983]), meaning that conditionally on $\mathbf{X}$ the treatment assignment W is independent of the potential outcomes for Y :

$$\{Y(1), Y(0)\} \perp\!\!\!\perp W|\mathbf{X}. \tag{4.1.3}$$

Many algorithms in the literature allow to evaluate CATE: causal forests, metalearners, causal neural networks as examples. They have been presented earlier in this manuscript, in what follows we focus on HTERF, which is the object of the two previous chapters.

4.1.2 Transfer learning

Transfer learning is a machine learning technique that leverages knowledge gained from solving one problem and applies it to a different but related problem. In traditional machine learning approaches, models are trained from scratch for each task, requiring substantial amounts of labeled data and computational resources. However, in real-world scenarios, labeled data might be scarce or expensive to acquire, hindering the effectiveness of such methods. Transfer learning addresses these limitations by transferring knowledge from a source domain where labeled data is abundant to a target domain where labeled data is scarce. This approach allows models to generalize better and achieve improved performance, particularly in situations where limited labeled data is available for training.

Domain adaptation is a special case of transfer learning. In domain adaptation, the source and target domains all have the same feature space (but different distributions), while transfer learning includes cases where the target domain's feature space is different from the source feature space or spaces. In what follows, the problem of supervised domain adaptation is considered, where both source and target dataset are labeled.

According to [Huyen 2022], in a supervised machine learning problem, the training dataset can be viewed as a set of samples from a joint distribution of $P(X, Y)$, where X is the input and Y is the output. We are interested in modelling $P(Y|X)$. $P(X, Y)$ can be decomposed as $P(X|Y) \times P(Y)$ or $P(Y|X) \times P(X)$. Different problems are treated in transfer learning. The most common is covariate shift where the marginal distribution $P(X)$ differs between source and target domains but the conditional distribution $P(Y|X)$ stays the same across the domains. Similarly label shift can be defined as the case where $P(Y)$ differs between source and target domains but the conditional distribution $P(X|Y)$ stays the same across the domains. Finally the model shift or concept drift concerns the cases where $P(Y|X)$ changes but $P(X)$ remains the same.

Different strategies are presented in [Huyen 2022] to address these data distribution shifts. The first strategy and the simplest is to train models and large and rich datasets hoping that points following both source and target distribution will be present in this large dataset. This method requires to have access to large external

datasets susceptible to contain both source and target distributions. Furthermore it can be costly to train models on very large datasets. A second approach is to use algorithms dedicated to take into account a certain type of shift, for example the kernel mean matching (KMM) method ([Huang *et al.* 2006], [Gretton *et al.* 2006]) allows to deal with covariate shift. [Zhang *et al.* 2013] proposes an approach to correct both covariate shift and label shift without using labels from target distribution (unsupervised domain adaptation problem), in a similar fashion [Zhao *et al.* 2019] proposed domain-invariant representation learning. [Wang *et al.* 2014] introduces two methods to deal with covariate shift in real regression cases, they use labeled source data. Finally a third kind of approach to deal with data distribution shift is to retrain the model with labeled target data, either the model is retrained from scratch with both source and target data or the existing model trained on source resumes its training on target data. This second option named fine tuning is easily applicable on neural networks by using procedures such as freezing layers or warm starting.

This can be extended to the causal context. In this case we consider the source domain (X^s, Y^s, T^s) and the target domain (X^t, Y^t, T^t). We study the model shift case where $P(Y^t(1)|X^t) \neq P(Y^s(1)|X^s)$ and $P(Y^t(0)|X^t) \neq P(Y^s(0)|X^s)$. We assume same distribution for X^s and X^t (respectively T^s and T^t) in what follows. If necessary the distributions of X^s and X^t could be matched by various methods dealing with covariate shift (e.g. KMM) without the use of Y. The goal is then to evaluate the CATE on the target population. A recent work has been done to estimate ATE in a supervised domain adaptation setup in [Wei *et al.* 2024], the nuisance parameters (such as the propensity score) are estimated using $\updownarrow^1$ regularised transfer learning, and then plugged in an ATE estimator. We can also mention [Künzel *et al.* 2018], who proposed to transfer knowledge by using several strategies such as: using neural network (NN) weights estimated from the source domain as the warm start of the subsequent target domain NN training, using NN weights estimated from the source domain and freezing some of its layers before backpropagating through the unfrozen ones when training on target dataset. Neural networks with an architecture dedicated to causal transfer learning have also been proposed by [Bica & van der Schaar 2022]. The method we propose is innovative, since it allows transfer learning on CATE estimation without using a neural network.

4.2 The offset approach

In what follows, we will be concerned with the offset approach that we describe now.

4.2.1 Presentation

Let $\mathcal{X} \in \mathbb{R}^d$ and $\mathcal{Y} \in \mathbb{R}$ the input and output spaces for the regression task for both source and target domains. Let $(\mathbf{Z}_i^s)_{i\in\{1,\dots,n\}} = ((\mathbf{X}_i^s, Y_i^s))_{i\in\{1,\dots,n\}}$ the source data set of size n, we also consider $(\mathbf{Z}_i^{tL})_{i\in\{1,\dots,n_l\}} = ((\mathbf{X}_i^{tL}, y_i^{tL}))_{i\in\{1,\dots,n_l\}}$ the labeled

target data set of size n_l. There is also an unlabeled target dataset on which we want to test the performance of transfer learning we denote it $(\mathbf{X}_i^{tU})_{i\in\{1,\dots,n_u\}}$ of size n_u.

The offset algorithm (Algorithm 6) introduced by [Wang *et al.* 2014] can be used with any regression machine learning algorithm for each estimator (namely $\hat{f}^s, \hat{f}^o, M$).

Algorithm 6 Offset algorithm

Input: A source data set $\{X_i^s, Y_i^s\}_{i\in\{1,\dots,n\}}$, a labeled target data set $\{X_i^{tL}, Y_i^{tL}\}_{i\in\{1,\dots,n_l\}}$ and an unlabeled target data set $\{X_i^{tU}\}_{i\in\{1,\dots,n_u\}}$.
Estimate a model $\hat{f}^s$ that regresses $\{Y_i^s\}$ against $\{X_i^s\}$.
Estimate a model $\hat{f}^o$ that regresses $\{\hat{Y}_i^o\} = \{Y_i^{tL} - \hat{f}^s(X_i^{tL})\}$ against $\{X_i^t\}$.
$\{Y_i^{new}\} \leftarrow \{Y_i^s + \hat{f}^o(X_i^s)\}$
Train a model M on $\{X_i^s, Y_i^{new}\} \cup \{X_i^{tL}, Y_i^{tL}\}$.
Output: $\{\hat{Y_i^{tU}}\} \leftarrow \{M(X_i^{tU})\}$

4.2.2 Using Kernel Ridge Regression

A generalisation bound is proposed in [Wang & Schneider 2015] when Kernel Ridge Regression (KRR) is used in the offset algorithm.

KRR and its associated notations are defined the following way.

Definition 4.1 ([Bousquet & Elisseeff 2002]). *Let $\mathcal{S}_T = \{\mathbf{Z}_1 = (\mathbf{X}_1, Y_1), \dots, \mathbf{Z}_n = (\mathbf{X}_n, Y_n)\}$ a training sample sample for a regression task in a reproducing kernel Hilbert space (see [Wahba 2003]) $\mathcal{H}$ with kernel K, scalar product k and associated norm $||.||_k$. Let ℓ be the l^2 loss function, then the KRR estimator is:*

$$\arg\min_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} \ell(h, Z_i) + \lambda||h||_k^2. \tag{4.2.1}$$

Two errors are defined:

- $R = \mathbb{E}[\ell(\mathcal{S}_T, Z)]$, *the generalisation error,*
- $R_{emp} = \frac{1}{n}\sum_{i=1}^{n} \ell(\mathcal{S}_T, Z_i)$ *the empirical error.*

Theorem ([Wang & Schneider 2015]). *If KRR is used to estimate the three functions in the offset method, let R^t be the generalisation error on the target dataset of the final model M, R^s_{emps} the empirical error on the source model and $\bar{R}^o_{emp}$ is the empirical error of the estimator $\hat{f}^o$ against $\{X^{tL}, \hat{Y}^o\}$, then*

$$R^t - 2\left(R^s_{emps} - \bar{R}^o_{emp}\right) = O\left(\frac{1}{\sqrt{\lambda_o n_l}}\right). \tag{4.2.2}$$

This result relies on Theorem 12 in [Bousquet & Elisseeff 2002] which gives a property of uniform stability for the KRR algorithm. However this property is not

known for many other algorithms than KRR (or only a weaker version of stability is obtained) which makes difficult extensions of this result to the causal case presented in the following section.

4.3 Causal adaptation

We propose an offset method adapted to the causal framework.

4.3.1 Overview

Two causal adaptation to the offset method are proposed, in Algorithm 7 the treated and the control populations are processed separately in order to estimate source and offset functions (one estimator for each group). In Algorithm 8 the treatment variable is considered as an additional covariate for the source and offset functions.

Algorithm 7 Offset causal algorithm separate models

Input: A source data set $\{W_i^s, X_i^s, Y_i^s\}_{i\in\{1,\dots,n\}}$, a labeled target data set $\{W_i^{tL}, X_i^{tL}, Y_i^{tL}\}_{i\in\{1,\dots,n_l\}}$ and an unlabeled target data set $\{X_i^{tU}\}_{i\in\{1,\dots,n_u\}}$.
Estimate a model $\hat{f}_0^s$ that regresses $\{Y_i^s\}_{W_i^s=0}$ against $\{X_i^s\}_{W_i^s=0}$ and a model $\hat{f}_1^s$ that regresses $\{Y_i^s\}_{W_i^s=1}$ against $\{X_i^s\}_{W_i^s=1}$.
Estimate a model $\hat{f}_0^o$ that regresses $\{Y_i^{tL} - \hat{f}_0^s(X_i^{tL})\}_{W_i^{tL}=0}$ against $\{X_i^{tL}\}_{W_i^{tL}=0}$ and a model $\hat{f}_1^o$ that regresses $\{Y_i^{tL} - \hat{f}_1^s(X_i^{tL})\}_{W_i^{tL}=1}$ against $\{X_i^{tL}\}_{W_i^{tL}=1}$.
$\{Y_i^{new}\} \leftarrow \{Y_i^s + \hat{f}_{W_i^s}^o(X_i^s)\}$
Train an HTERF model M on $\{W_i^s, X_i^s, Y_i^{new}\} \cup \{W_i^{tL}, X_i^{tL}, Y_i^{tL}\}$.
Output: $\{\hat{\tau}^t(X_i^{tU})\} \leftarrow \{M(X_i^{tU})\}$

Algorithm 8 Offset causal algorithm unique models

Input: A source data set $\{W_i^s, X_i^s, Y_i^s\}_{i\in\{1,\dots,n\}}$, a labeled target data set $\{W_i^{tL}, X_i^{tL}, Y_i^{tL}\}_{i\in\{1,\dots,n_l\}}$ and an unlabeled target data set $\{X_i^{tU}\}_{i\in\{1,\dots,n_u\}}$.
Estimate a model $\hat{f}^s$ that regresses $\{Y_i^s\}$ against $\{X_i^s, W_i^s\}$.
Estimate a model $\hat{f}^o$ that regresses $\{Y_i^{tL} - \hat{f}^s(X_i^{tL}, W_i^{tL})\}$ against $\{X_i^{tL}\}$.
$\{Y_i^{new}\} \leftarrow \{Y_i^s + \hat{f}^o(X_i^s, W_i^s)\}$
Train an HTERF model M on $\{W_i^s, X_i^s, Y_i^{new}\} \cup \{W_i^{tL}, X_i^{tL}, Y_i^{tL}\}$.
Output: $\{\hat{\tau}^t(X_i^{tU})\} \leftarrow \{M(X_i^{tU})\}$

Any algorithm could be used to estimate the source and offset functions, in practice we obtained good results when using regression random forests.

4.3.2 Convergence result

Since the size of the source dataset is assumed to be large compared to the target data set, we write a convergence theorem and a generalisation bound on the causal offset

algorithm if the HTERF model in the last step is only fit on the set $\{W^s, X^s, Y^{new}\}$ of size $n = n_s$.

Let $\tau_1^t(\mathbf{x}^t) = \mathbb{E}\left[Y^t(1)|\mathbf{X}^t = \mathbf{x}\right]$ and $\tau_0^t(\mathbf{x}) = \mathbb{E}\left[Y^t(0)|\mathbf{X}^t = \mathbf{x}\right]$ and similarly $\hat{\tau}_1^{new}(\mathbf{x}) = \sum_{i:W_i=1} \alpha_i(\mathbf{x}) Y_i^{new}$ and $\hat{\tau}_0^{new}(\mathbf{x}) = \sum_{i:W_i=0} \alpha_i'(\mathbf{x}) Y_i^{new}$. We shall make the following assumptions.

Assumption 4.1.

- $Y^t = \tau^t(\mathbf{X}^t)g(\mathbf{W}^t) + \gamma^t(\mathbf{X}^t) + \varepsilon^t$ and $Y^s = \tau^s(\mathbf{X}^s)g(\mathbf{W}^s) + \gamma^s(\mathbf{X}^s) + \varepsilon^s$.
- $\forall i$ *such that* $W_i = 1, Y_i^s = f_1^s(X_i^s) + \varepsilon_{1,i}^s, \varepsilon_1^s \perp\!\!\!\perp X^s$ *and* $Y_i^t - f_1^s(X_i^t) = f_1^o(X_i^t) + \varepsilon_{1,i}^t, \varepsilon_1^t \perp\!\!\!\perp X^s, X^t$. ε_1^s *and* ε_1^t *are continuous centered random variables.*
- $\forall i$ *such that* $W_i = 0, Y_i^s = f_0^s(X_i^s) + \varepsilon_{0,i}^s, \varepsilon_0^s \perp\!\!\!\perp X^s$ *and* $Y_i^t - f_0^s(X_i^t) = f_0^o(X_i^t) + \varepsilon_{0,i}^t, \varepsilon_0^t \perp\!\!\!\perp X^s, X^t$. ε_0^s *and* ε_0^t *are continuous centered random variables.*
- $\mathbf{X}^s$ *and* $\mathbf{X}^t$ *are distributed as* $\mathbf{X} = (X_1, \ldots, X_d)$*, which is a continuous random vector with independent coordinates. The density of* $\mathbf{X}$ *is positive and bounded from above and below by positive constants.*
- W^s *and* W^t *conditionally on* $\mathbf{X}$ *have the same distribution.*
- $\mathbf{X}$ *takes its values in* $\mathcal{X}$ *which is assumed to be a compact hyper-rectangle of* $\mathbb{R}^d$*:* $\mathcal{X} = \prod_{i=1}^{d} [u_i, v_i]$, $-\infty < u_i \leq v_i < \infty$.
- $\mathbf{x} \mapsto \gamma^t(\mathbf{x})$, $\mathbf{x} \mapsto \tau_1^s(\mathbf{x})$, $\mathbf{x} \mapsto \tau_0^s(\mathbf{x})$, $\mathbf{x} \mapsto \tau_1^t(\mathbf{x})$ *and* $\mathbf{x} \mapsto \tau_0^t(\mathbf{x})$ *are continuous. So in particular* $\mathbf{x} \mapsto \tau^t(\mathbf{x})$ *and* $\mathbf{x} \mapsto \tau^s(\mathbf{x})$ *are continuous.*

Assumption 4.2. *The following assumptions are made on* B *(number of trees in HTERF),* $N_{n,1}(\mathbf{x}; \Theta, \mathcal{D}_n)$ *resp.* $N_{n,0}(\mathbf{x}; \Theta, \mathcal{D}_n)$ *(number of observations in a leaf node such that* $W = 1$*, resp.* $W = 0$*) and on the construction of the trees:*

1. $B = \mathcal{O}(\sqrt{n})$ and $B = \Theta\left(\frac{\sqrt{n}}{(\ln(n))^\beta}\right)$, with $\beta > 1$.
2. $\forall \mathbf{x} \in \mathcal{X}, \quad \mathbb{E}\left[N_{n,1}(\mathbf{x}; \Theta, \mathcal{D}_n)\right] = \Omega\left(\sqrt{n}(\ln(n))^\beta\right)$.
3. $\forall \mathbf{x} \in \mathcal{X}, \quad \mathbb{E}\left[N_{n,0}(\mathbf{x}; \Theta, \mathcal{D}_n)\right] = \Omega\left(\sqrt{n}(\ln(n))^\beta\right)$.
4. $\max_{x,\Theta} N_{n,1}(\mathbf{x}; \Theta, \mathcal{D}_n) = o(n)$.
5. $\max_{x,\Theta} N_{n,0}(\mathbf{x}; \Theta, \mathcal{D}_n) = o(n)$.
6. *At every step of the tree building procedure, the probability that the next split is done along the* $j-th$ *feature is bounded below by* π/d *for some* $0 < \pi \leq 1$ *for all* $j = 1, \ldots, d$.

7. *$\mathcal{I}_2$ verifies that each split leaves at least a fraction α of the available training sample such that $W = 1$ (resp. $W = 0$) on each side of the split, for some $0 < \alpha \leq 0.5$.*

The function $\hat{f}_1^s$ is an estimator of f_1^s in the first step:

$$\forall i \text{ such that } W_i = 1, Y_i^s = \hat{f}_1^s(X_i^s) + \varepsilon_{1,i}^s + E_{1,i}^s(\mathcal{D}_s, \mathbf{X}_i^s). \tag{4.3.1}$$

The function $\hat{f}_1^o$ is an estimator of f_1^o in the second step:

$$\forall i \text{ such that } W_i = 1, Y_i^t - \hat{f}_1^s(\mathbf{X}_i^t) - E_1^s(\mathcal{D}_s, \mathbf{X}_i^t) = f_1^o(\mathbf{X}_i^t) + \varepsilon_{1,i}^t. \tag{4.3.2}$$

So:

$$\forall i \text{ such that } W_i = 1, Y_i^t - \hat{f}_1^s(\mathbf{X}_i^t) - E_1^s(\mathcal{D}_s, \mathbf{X}_i^t) = \hat{f}_1^o(\mathbf{X}_i^t) + E_1^o(\mathcal{D}, \mathbf{X}_i^t) + \varepsilon_{1,i}^t. \tag{4.3.3}$$

Finally the third step leads to:

$$\begin{aligned}
\forall i \text{ such that } W_i = 1, Y_i^{new} &= Y_i^s + \hat{f}_1^o(\mathbf{X}_i^s) \\
&= Y_i^s + Y_i^t - \hat{f}_1^s(\mathbf{X}_i^s) - E_1^s(\mathcal{D}_s, \mathbf{X}_i^s) - E_1^o(\mathcal{D}, \mathbf{X}_i^t) - \varepsilon_{1,i}^t \\
&= Y_i^t + (Y_i^s - \hat{f}_1^s(\mathbf{X}_i^s)) - E_1^s(\mathcal{D}_s, \mathbf{X}_i^s) - E_1^o(\mathcal{D}, \mathbf{X}_i^t) - \varepsilon_{1,i}^t \\
&= Y_i^t + \varepsilon_{1,i}^s + E_1^s(\mathcal{D}_s, \mathbf{X}_i^s) - E_1^s(\mathcal{D}_s, \mathbf{X}_i^s) - E_1^o(\mathcal{D}, \mathbf{X}_i^t) - \varepsilon_{1,i}^t \\
Y_i^{new} &= Y_i^t + \varepsilon_{1,i}^s - \varepsilon_{1,i}^t - E_1^o(\mathcal{D}, \mathbf{X}_i^t).
\end{aligned}$$

In a similar fashion we have:

$$\forall i \text{ such that } W_i = 0, Y_i^{new} = Y_i^t + \varepsilon_{0,i}^s - \varepsilon_{0,i}^t - E_0^o(\mathcal{D}, \mathbf{X}_i^t) \tag{4.3.4}$$

Then HTERF is trained on $\{W_i^s, \mathbf{X}_i^s, Y_i^{new}\}$, which gives the following estimator $\hat{\tau}_{B,n}^{new}(\mathbf{X}) = \hat{\tau}_1^{new}(\mathbf{x}) - \hat{\tau}_0^{new}(\mathbf{x})$, where $\hat{\tau}_1^{new}(\mathbf{x}) = \sum_{i:W_i=1} \alpha_i(\mathbf{x}) Y_i^{new}$ and $\hat{\tau}_0^{new}(\mathbf{x}) = \sum_{i:W_i=0} \alpha_i'(\mathbf{x}) Y_i^{new}$.

Theorem 4.1. *Let Assumptions 4.1 and 4.2 be verified, assume that for a fixed $\beta > \frac{5}{2}$, $C > 0$, each HTERF tree is the highest such that $C\sqrt{n}(\ln n)^\beta \leq N_{n,0}(\mathbf{x}; \Theta_\ell, \mathcal{D}_n), N_{n,1}(\mathbf{x}; \Theta_\ell, \mathcal{D}_n)$. Assume that $\mathbb{E}\left[\max\left(\varepsilon_{1,i}^s\right)^2\right], \mathbb{E}\left[\max\left(\varepsilon_{1,i}^t\right)^2\right], \mathbb{E}\left[\max\left(\varepsilon_{0,i}^s\right)^2\right], \mathbb{E}\left[\max\left(\varepsilon_{0,i}^t\right)^2\right] \leq K(\ln n)^u$ with $\beta - u > \frac{1}{2}$ and K is a positive constant. Also assume that Y and E^o error term are bounded and that E^o converges to 0 in L^2 as n_l tends to $+\infty$. Then*

$$\mathbb{E}\left[\left|\hat{\tau}_{B,n}^{new}(\mathbf{X}) - \tau^t(\mathbf{X})\right|\right] \underset{n, n_l \to \infty}{\longrightarrow} 0.$$

The proof is described in Appendix A.

Remark 4.1. *With an estimator $\hat{f}^s$ of the form $\sum \omega_i Y_i^s$, since Y is bounded, so are $\hat{f}^s$ and E^s. With $\hat{f}^o$ of the form $\sum \omega_i (Y_i^t - \hat{f}^s(X_i^t))$, the error term E^o is also bounded. Most of the classical regression algorithm provide estimators of this form: random forest, linear regression, neural network...*

Remark 4.2. *Following what is done in the proof of Theorem 4.1, the following quantity can be bounded this way (same rationale applies for τ_0), let $\mathbf{x} \in \mathcal{X}$:*

$$\left|\hat{\tau}_1^{new}(\mathbf{x}) - \tau_1^t(\mathbf{x})\right| \leq Bound_{offset} + Bound_{HTERF}. \tag{4.3.5}$$

Overall this bound tends to 0, the first term is the bound of HTERF on a sample of size n and the second term is introduced by the offset method, the rate of convergence of this quantity only depends on the rate of convergence of $\hat{f}^o$. More details of this bound are présented in Appendix B.

4.4 Simulation results

In the following examples causal offset with unique and distinct models is compared to the baseline case where HTERF is simply trained on available target data. Two choices of algorithms are considered to estimate functions f^s and f^o, namely Kernel Ridge Regression and Regression Random Forest.

4.4.1 One dimensional example

Firstly the source domain is defined, let $\mathbf{X}_i^s \sim U([0,1])$, $W_i^s \sim Bern(0.5)$ and $Y_i^s = \tau^s(\mathbf{X}_i^s)W_i^s + \gamma^s(\mathbf{X}_i^s)$. Where $\tau^s(\mathbf{x}) = \sin(\mathbf{x})$ and $\gamma^s(\mathbf{x}) = \cos(\mathbf{x})$. A source sample of 10000 units in considered. The target domain is defined as $\mathbf{X}_i^t \sim U([0,1])$, $W_i^t \sim Bern(0.5)$ and $Y_i^t = \tau^t(\mathbf{X}_i)W_i + \gamma^t(\mathbf{X}_i)$. Where $\tau^t(\mathbf{x}) = \cos(\mathbf{x})$ and $\gamma^t(\mathbf{x}) = \cos(\mathbf{x})$. The unlabeled target dataset and the labeled target dataset are both of size 500.
In Table 4.1, the performance of HTERF on source model is presented in the first line. Then we present the results of offset method with KRR used to estimate the functions f^s and f^o. Two cases have been considered in the first one, separate model are trained respectively for treated and control groups, in the second case the treatment variable T is considered as a covariate for f^s and f^o. Finally a no transfer strategy is considered where HTERF is trained only on the target data set. Figure 4.1 offers a graphical illustration of this example.

Both causal offset methods have better performances than the baseline method. In this example using a single model for treated and untreated individuals is the most efficient.

4.4.2 Multi-dimensional example

A multi-dimensional example inspired by the previous one is proposed, for the source domain let $\mathbf{X}_i^s \sim U([0,1]^{10})$, $W_i^s \sim Bern(0.5)$ and $Y_i^s = \tau^s(\mathbf{X}_i^s)W_i^s + \gamma^s(\mathbf{X}_i^s)$. Where $\tau^s(\mathbf{x}) = \sin(\mathbf{x}^{(1)})$ and $\gamma^s(\mathbf{x}) = \cos(\mathbf{x}^{(1)})$. A source sample of 10000 units in considered. The target domain is defined as $\mathbf{X}_i^t \sim U([0,1]^{10})$, $W_i^t \sim Bern(0.5)$ and $Y_i^t = \tau^t(\mathbf{X}_i)W_i + \gamma^t(\mathbf{X}_i)$. Where $\tau^t(\mathbf{x}) = \cos(\mathbf{x}^{(1)})$ and $\gamma^t(\mathbf{x}) = \cos(\mathbf{x}^{(2)})$. The

Method	RMSE
Source	0.003
Offset separate models	0.015
Offset unique model	0.009
No transfer	0.205

Table 4.1: Root mean squared errors of CATE on source and on target with three different methods namely, offset causal with separate KRR models, offset causal with unique KRR model and HTERF only trained on target data (baseline method). HTERF causal forests have 500 trees, the forest of the first step in HTERF have 500 trees. The results are aggregated over 50 simulation replications with 500 test points each (the source dataset stay unchanged only the target training and test dataset are modified).

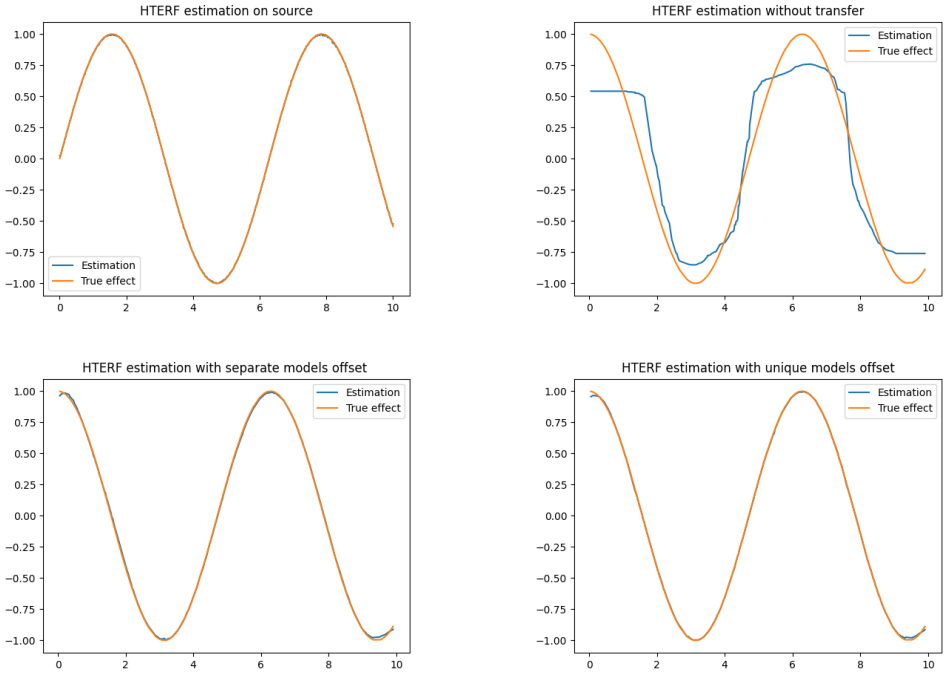


Figure 4.1: Graphical illustration for one dimensional example
Top left: HTERF CATE estimation on source
Top right: HTERF CATE estimation on target using only target data
Bottom left: HTERF CATE estimation on target using offset causal with separate KRR models
Bottom left: HTERF CATE estimation on target using offset causal with unique KRR model

Method	RMSE
Source	0.004
Offset separate KRR models	0.957
Offset unique KRR model	0.960
Offset separate RF models	0.120
Offset unique RF model	0.135
No transfer	0.348

Table 4.2: Root mean squared errors of CATE on source and on target with five different methods namely, offset causal with separate KRR models, offset causal with unique KRR model, offset causal with separate random forest (RF) models, offset causal with unique RF model and HTERF only trained on target data (baseline method). HTERF causal forests have 500 trees, the forest of the first step in HTERF have 500 trees. The results are aggregated over 50 simulation replications with 500 test points each (the source dataset stay unchanged only the target training and test dataset are modified).

unlabeled target dataset and the labeled target dataset are both of size 500.

Figure 4.4.2 illustrate the poor performance of KKR as the algorithm for f^s and f^o. In the top right image, the KRR estimator of the function f^s fails to capture the variations of Y_0^s against $\mathbf{X}^{s,(2)}$. however using random forest to estimate f^s and f^o, causal offset algorithms are more efficient than the baseline method. This time using separate models for treated and control units is the best strategy.

4.4.3 Semi-synthetic dataset

A last example is presented, using the IHDP dataset already introduced in Chapter 2. This dataset has been studied in [Wei *et al.* 2024] to compare accuracy of various ATE estimators in a transfer learning context. In addition to RMSE two additional indicators of performances for ATE estimation are added, namely:

- ATE1: $\frac{1}{n_u}\left|\sum_{i=1}^{n_u}\hat{\tau}_{B,n}^{new}(\mathbf{X}_i^{tU}) - \tau^t(\mathbf{X}_i^{tU})\right|$,
- ATE2: $\frac{1}{n_u}\sum_{i=1}^{n_u}\left|\hat{\tau}_{B,n}^{new}(\mathbf{X}_i^{tU}) - \tau^t(\mathbf{X}_i^{tU})\right|$.

The dataset has already been presented in the Chapter 2. To create the source and target domains, the binary variable "The mother drank alcohol during pregnancy" has been used: the source domain consists of all children whom mothers did not drink and the target domain consists of the children whom mothers drank alcohol.

In this example, causal offset with a unique random forest model is the most efficient for CATE and ATE estimation, followed by causal offset with two separate models. Both algorithms outperform the baseline method without transfer.

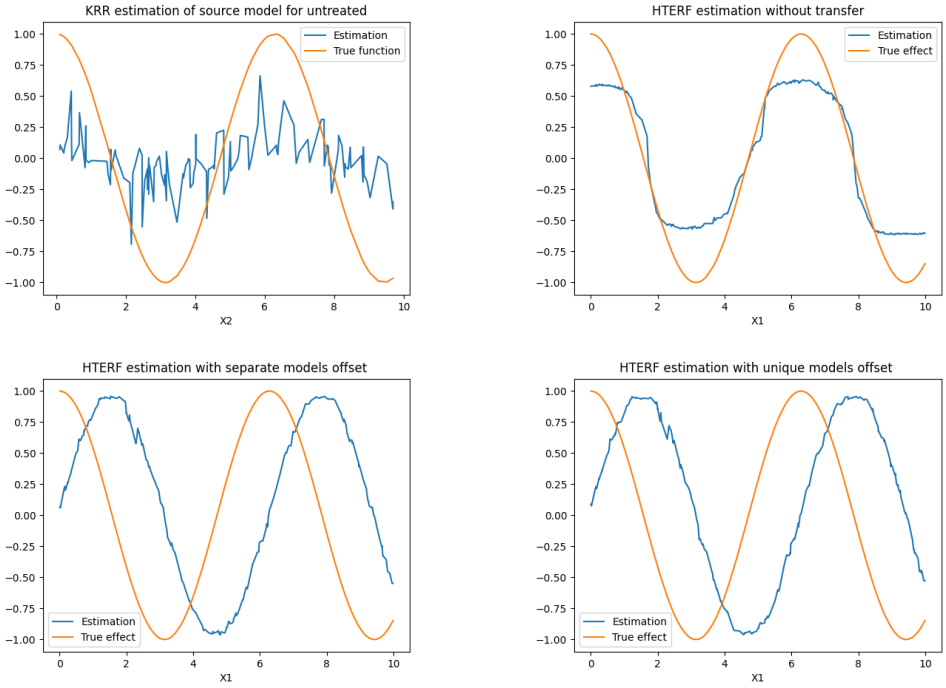


Figure 4.2: Graphical illustration for multi-dimensional example
Top left: KRR estimation on source of the function f_0^s
Top right: HTERF CATE estimation on target using only target data
Bottom left: HTERF CATE estimation on target using offset causal with separate KRR models
Bottom left: HTERF CATE estimation on target using offset causal with unique KRR model

Method	CATE	ATE1	ATE2
Offset separate models	0.808	0.292	0.561
Offset unique model	0.734	0.180	0.491
No transfer	1.016	0.351	0.732

Table 4.3: RMSE on CATE and two different errors on ATE with three different methods namely, offset causal with separate RF models, offset causal with unique RF model and HTERF only trained on target data (baseline method).HTERF causal forests have 500 trees, the forest of the first step in HTERF have 500 trees. The results are aggregated over 50 simulation replications, the source dataset stays unchanged but for each replication the labeled target and unlabeled target are modified.

4.5 Discussion

We have presented in this work an algorithm to perform transfer learning on the causal inference problem. This approach combines the offset algorithm already used on regression problems and the HTERF causal forest. The combination of these two methods allows to have a consistency result on the CATE estimation in the target domain. A generalisation bound is also shown, these results rely on stronger assumptions than the classical HTERF consistency, especially regarding the number of trees in the forest which needs to be large ($> C\frac{\sqrt{n}}{(\ln n)^\beta}$).
Additional work could be done on the proof on consistency to lighten the assumptions. An almost sure convergence might also be obtained instead of a L^1 convergence.

A Proof of results

Proof of Theorem 4.1.

Let define a diamond dataset $\mathcal{D}^\diamond = (Y_i^\diamond, \mathbf{X}_i^\diamond, W_i^\diamond)_{i=1,\dots,n}$ that is a sample of $(Y^t, \mathbf{X}, W)$, being independent of $\mathcal{D}^t$ and $\mathcal{D}^s$. This new sample is used to build $(Y_i^{new,\diamond})_{i=1,\dots,n}$ using the estimators $\hat{f}^s$ and $\hat{f}^o$ previously build:

$$\forall i \text{ such that } W_i^\diamond = 1, Y_i^{s,\diamond} = \hat{f}_1^s(\mathbf{X}_i^\diamond) + \varepsilon_{1,i}^{s,\diamond} + E_1^{s,\diamond}(\mathcal{D}_s, \mathbf{X}_i^\diamond) \tag{A.1}$$

$$Y_i^\diamond - \hat{f}_1^s(\mathbf{X}_i^\diamond) - E_1^{s,\diamond}(\mathcal{D}_s, \mathbf{X}_i^\diamond) = \hat{f}_1^o(\mathbf{X}_i^\diamond) + E_1^{o,\diamond}(\mathcal{D}, \mathbf{X}_i^\diamond) + \varepsilon_{1,i}^{t,\diamond} \tag{A.2}$$

$$Y_i^{new,\diamond} = Y_i^\diamond + \varepsilon_{1,i}^{s,\diamond} - \varepsilon_1^{t,\diamond} - E_1^{o,\diamond}(\mathcal{D}, \mathbf{X}_i^\diamond) \tag{A.3}$$

As in HTERF consistency proof, the trees are grown using $\mathcal{D}_n = \mathcal{D}^s \cup \mathcal{D}^t$, but the sample $\mathcal{D}_n^\diamond$ (independent of $\mathcal{D}_n$ and Θ) is used to define a dummy estimator

$$\begin{aligned}
&\tau_{B,n}^{new,\diamond}(\mathbf{x}; \Theta_1, \dots, \Theta_k, \mathcal{D}_n^\diamond, \mathcal{D}_n) \\
=&\sum_{j=1}^n \alpha_{n,j}^\diamond(\mathbf{x}; \Theta_1, \dots, \Theta_k, \mathbf{X}_1^\diamond, \dots, \mathbf{X}_n^\diamond, W_1^\diamond, \dots, W_n^\diamond, \mathcal{D}_n)\, Y_j^{\diamond,new} \\
&- \sum_{j=1}^n \alpha_{n,j}^{'\diamond}(\mathbf{x}; \Theta_1, \dots, \Theta_k, \mathbf{X}_1^\diamond, \dots, \mathbf{X}_n^\diamond, W_1^\diamond, \dots, W_n^\diamond, \mathcal{D}_n)\, Y_j^{\diamond,new},
\end{aligned}$$

where the weights are

$$\begin{aligned}
&\alpha_{n,j}^\diamond(\mathbf{x}; \Theta_1, \dots, \Theta_k, \mathbf{X}_1^\diamond, \dots, \mathbf{X}_n^\diamond, W_1^\diamond, \dots, W_n^\diamond, \mathcal{D}_n) \\
=&\frac{1}{B}\sum_{\ell=1}^B \frac{\mathbb{1}_{\{\mathbf{X}_j^\diamond \in A_n(\mathbf{x}; \Theta_\ell, \mathcal{D}_n)\} \cap W_j^\diamond = 1}}{N_{n,1}^\diamond(\mathbf{x}; \Theta_\ell, \mathbf{X}_1^\diamond, \dots, \mathbf{X}_n^\diamond, W_1^\diamond, \dots, W_n^\diamond, \mathcal{D}_n)}, \ j = 1, \dots, n.
\end{aligned}$$

with $N_{n,1}^\diamond(\mathbf{x}; \Theta_\ell, \mathbf{X}_1^\diamond, \dots, \mathbf{X}_n^\diamond, W_1^\diamond, \dots, W_n^\diamond, \mathcal{D}_n)$, the number of elements of $\mathcal{D}_n^\diamond$ that fall into $A_n(\mathbf{x}; \Theta_\ell, \mathcal{D}_n)$ such that $W^\diamond = 1$. Throughout this section, we shall use the convention $\frac{0}{0} = 0$ in case $N_{n,1}^\diamond(\mathbf{x}; \Theta_\ell, \mathbf{X}_1^\diamond, \dots, \mathbf{X}_n^\diamond, W_1^\diamond, \dots, W_n^\diamond, \mathcal{D}_n) = 0$ and thus $\mathbb{1}_{\{\mathbf{X}_j^\diamond \in A_n(\mathbf{x}; \Theta_\ell, \mathcal{D}_n)\} \cap W_j^\diamond = 1} = 0$ for $j = 1, \dots n$.
Similarly we have:

$$\begin{aligned}
&\alpha_{n,j}^{'\diamond}(\mathbf{x}; \Theta_1, \dots, \Theta_k, \mathbf{X}_1^\diamond, \dots, \mathbf{X}_n^\diamond, W_1^\diamond, \dots, W_n^\diamond, \mathcal{D}_n) \\
=&\frac{1}{B}\sum_{\ell=1}^B \frac{\mathbb{1}_{\{\mathbf{X}_j^\diamond \in A_n(\mathbf{x}; \Theta_\ell, \mathcal{D}_n)\} \cap W_j^\diamond = 0}}{N_{n,0}^\diamond(\mathbf{x}; \Theta_\ell, \mathbf{X}_1^\diamond, \dots, \mathbf{X}_n^\diamond, W_1^\diamond, \dots, W_n^\diamond, \mathcal{D}_n)}, \ j = 1, \dots, n.
\end{aligned}$$

To lighten the notation in the sequel, we will simply write $\tau_{B,n}^{new,\diamond}(\mathbf{x}) = \sum_{j=1}^n \alpha_j^\diamond(\mathbf{x})\, Y_j^{\diamond,new} - \sum_{j=1}^n \alpha_j^{'\diamond}(\mathbf{x})\, Y_j^{\diamond,new} = \tau_1^{new,\diamond}(\mathbf{x}) - \tau_0^{new,\diamond}(\mathbf{x})$.
Let $\mathbf{x} \in \mathcal{X}$, we have:

$$\begin{aligned}
\left|\hat{\tau}^{new}(\mathbf{x}) - \tau^t(\mathbf{x})\right| \leq &\left|\hat{\tau}^{new}(\mathbf{x}) - \tau^{new,\diamond}(\mathbf{x})\right| \\
&+ \left|\tau^{new,\diamond}(\mathbf{x}) - \tau^t(\mathbf{x})\right|.
\end{aligned}$$

Let $\mathbf{x}$ in $\mathcal{X}$: $\left|\tau^{new,\diamond}(\mathbf{x}) - \tau^{t}(\mathbf{x})\right| \leq \left|\tau_1^{new,\diamond}(\mathbf{x}) - \tau_1^{t}(\mathbf{x})\right| + \left|\tau_0^{new,\diamond}(\mathbf{x}) - \tau_0^{t}(\mathbf{x})\right|$ Each of the two terms will be treated the same way.

$$\begin{aligned}\left|\tau_1^{new,\diamond}(\mathbf{x}) - \tau_1^{t}(\mathbf{x})\right| \leq & \left| \sum_{\substack{i=1 \\ W_i^{\diamond}=1}}^{n} \alpha_i^{\diamond}(\mathbf{x}) \left[\left(Y_i^{new,\diamond}\right) - \mathbb{E}\left[Y^t(1)|\mathbf{X}_i^{\diamond}\right]\right] \right| \\ & + \left| \sum_{\substack{i=1 \\ W_i^{\diamond}=1}}^{n} \alpha_i^{\diamond}(\mathbf{x}) \left[\mathbb{E}\left[Y^t(1)|\mathbf{X}_i^{\diamond}\right] - \mathbb{E}\left[Y^t(1)|\mathbf{X}=\mathbf{x}\right]\right] \right| \\ \leq & U_n + V_n.\end{aligned}$$

The U_n term gives:

$$\begin{aligned}U_n = & \left| \sum_{\substack{i=1 \\ W_i^{\diamond}=1}}^{n} \alpha_i^{\diamond} \left(\varepsilon_{1,i}^{\diamond,s} - \varepsilon_{1,i}^{\diamond,t} - E_1^{o,\diamond}(\mathcal{D}, \mathbf{X}_i^{\diamond}) \right) \right| \\ \leq & \left| \sum_{\substack{i=1 \\ W_i^{\diamond}=1}}^{n} \alpha_i^{\diamond} \varepsilon_{1,i}^{\diamond,s} \right| + \left| \sum_{\substack{i=1 \\ W_i^{\diamond}=1}}^{n} \alpha_i^{\diamond} \varepsilon_{1,i}^{\diamond,t} \right| + \left| \sum_{\substack{i=1 \\ W_i^{\diamond}=1}}^{n} \alpha_i^{\diamond} E_1^{o,\diamond}(\mathcal{D}, \mathbf{X}_i^{\diamond}) \right|.\end{aligned}$$

Since Y is supposed to bounded, in addition to the fact that τ_1 and γ are continuous and $\mathbf{X}$ lives in a compact space, necessarily ε_1^s and ε_1^t are bounded. Following the HTERF consistency proof we have:

$$\begin{aligned}\mathbb{E}\left[\left| \sum_{\substack{i=1 \\ W_i^{\diamond}=1}}^{n} \alpha_i^{\diamond} \varepsilon_{1,i}^{\diamond,s} \right|^2 \right] \leq & \mathbb{E}\left[\max \varepsilon_j^{\diamond 2}\right] \left(\frac{4}{K\sqrt{n}(\ln n)^{\beta}} + 16C\sqrt{n}(n+1)^{2d} e^{-K^2(\ln n)^{2\beta}/2048} \right) \\ \leq & \frac{4}{K\sqrt{n}(\ln n)^{\beta-u}} + 16C\sqrt{n}(n+1)^{2d}(\ln n)^{u} e^{-K^2(\ln n)^{2\beta}/2048} \\ \to & 0.\end{aligned}$$

For the last term since the assumption is made that $E_1^o(\mathcal{D}, \mathbf{X}^s)$ is L^1 consistent, it can be bounded the following way

$$\mathbb{E}\left[\left| \sum_{\substack{i=1 \\ W_i^{\diamond}=1}}^{n} \alpha_i^{\diamond} E_1^{o,\diamond}(\mathcal{D}, \mathbf{X}_i^{\diamond}) \right| \right] \leq \mathbb{E}\left[\alpha_1^{\diamond} \left| E_1^{o,\diamond}(\mathcal{D}, \mathbf{X}_1^{\diamond}) \right| \right]. \tag{A.4}$$

Notice the decomposition:

$$\mathbb{E}\left[\alpha_1^{\diamond} \left| E_1^{o,\diamond}(\mathcal{D}, \mathbf{X}_1^{\diamond}) \right| \right] \leq \mathbb{E}\left[\left| \alpha_i^{\diamond} - \frac{1}{n} \right| \left| E_1^{o,\diamond}(\mathcal{D}, \mathbf{X}_i^{\diamond}) \right| \right] + \frac{1}{n} \mathbb{E}\left[\left| E_1^{o,\diamond}(\mathcal{D}, \mathbf{X}_i^{\diamond}) \right| \right]. \tag{A.5}$$

Since $\sum_{j=1}^{n} \alpha_j^{\diamond} = 1$ and $\alpha^{\diamond}$ are identically distributed, we have $\mathbb{E}[\alpha_i^{\diamond}] = \frac{1}{n}$ and $\mathbb{E}[\alpha_i^{\diamond}|\mathcal{D}] = \frac{1}{n}$. With Hölder inequality:

$$\mathbb{E}\left[\left|\alpha_i^{\diamond} - \frac{1}{n}\right| \left|E_1^{o,\diamond}(\mathcal{D}, \mathbf{X}_i^{\diamond})\right|\right] \leq \sqrt{Var(\alpha_i^{\diamond})}\sqrt{\mathbb{E}\left[\left|E_1^{o,\diamond}(\mathcal{D}, \mathbf{X}_i^{\diamond})\right|^2\right]}. \tag{A.6}$$

Using the total variance formula:

$$Var(\alpha_i^{\diamond}) = \mathbb{E}\left[Var(\alpha_i^{\diamond}|\mathcal{D})\right]. \tag{A.7}$$

We can rewrite:

$$\alpha_i^{\diamond} = \frac{1}{B}\sum_{l=1}^{B} \frac{\mathbb{1}_{\mathbf{X}_i^{\diamond} \in A_n(l)}}{N_{n,1}^{\diamond}(\mathbf{x}; \Theta_l, \mathcal{D}_n)} = \frac{1}{B}\sum_{l=1}^{B} Z_l, \tag{A.8}$$

where conditionally on $\mathcal{D}$ the $(Z_l)_{l\in 1,\dots,n}$ are independent and identically distributed, this leads to

$$\begin{aligned}
Var(\alpha_i^{\diamond}|\mathcal{D}) &= \frac{1}{B}Var(Z_1) \\
&\leq \frac{1}{B}\mathbb{E}[Z_1^2|\mathcal{D}] \\
&\leq \frac{1}{B}\mathbb{E}\left[\frac{\mathbb{1}_{\mathbf{X}_i^{\diamond} \in A_n(1)}}{(N_{n,1}^{\diamond}(\mathbf{x}; \Theta_1, \mathcal{D}_n))^2}\Big|\mathcal{D}\right] \\
&\leq \frac{1}{B}\mathbb{E}\left[\frac{\mathbb{1}_{\mathbf{X}_i^{\diamond} \in A_n(1)}\mathbb{1}_{\{N_{n,1}^{\diamond}(\mathbf{x};\Theta_1,\mathcal{D}_n) \geq \lambda\}}}{(N_{n,1}^{\diamond}(\mathbf{x}; \Theta_1, \mathcal{D}_n))^2}\Big|\mathcal{D}\right] + \frac{1}{B}\mathbb{E}\left[\frac{\mathbb{1}_{\mathbf{X}_i^{\diamond} \in A_n(1)}\mathbb{1}_{\{N_{n,1}^{\diamond}(\mathbf{x};\Theta_1,\mathcal{D}_n) < \lambda\}}}{(N_{n,1}^{\diamond}(\mathbf{x}; \Theta_1, \mathcal{D}_n))^2}\Big|\mathcal{D}\right].
\end{aligned}$$

Let $\lambda = \frac{\sqrt{n}(\ln n)^{\beta}}{2}$,

$$\begin{aligned}
Var(\alpha_1^{\diamond}|\mathcal{D}) &\leq \frac{1}{B\lambda}\mathbb{E}\left[\frac{\mathbb{1}_{\mathbf{X}_i^{\diamond} \in A_n(1)}}{N_{n,1}^{\diamond}(\mathbf{x}; \Theta_1, \mathcal{D}_n)}\Big|\mathcal{D}\right] + \frac{1}{B}\mathbb{P}\left(N_{n,1}^{\diamond}(\mathbf{x}; \Theta_1, \mathcal{D}_n) < \lambda|\mathcal{D}\right) \\
&\leq \frac{1}{B\lambda}\mathbb{E}\left[\alpha_i^{\diamond}|\mathcal{D}\right] + \frac{1}{B}\mathbb{P}\left(N_{n,1}^{\diamond}(\mathbf{x}; \Theta_1, \mathcal{D}_n) < \lambda|\mathcal{D}\right) \\
&\leq \frac{1}{B\lambda n} + \mathbb{P}\left(N_{n,1}^{\diamond}(\mathbf{x}; \Theta_1, \mathcal{D}_n) < \lambda|\mathcal{D}\right).
\end{aligned}$$

Remark that

$$\left\{N_{n,1}^{\diamond}(\mathbf{x}; \Theta_1, \mathcal{D}_n) < \frac{\sqrt{n}(\ln n)^{\beta}}{2}\right\} \subset \left\{\left|N_{n,1}(\mathbf{x}; \Theta_1, \mathcal{D}_n) - N_{n,1}^{\diamond}(\mathbf{x}; \Theta_1, \mathcal{D}_n)\right| > \frac{\sqrt{n}(\ln n)^{\beta}}{2}\right\},$$

thus we have

$$\mathbb{P}\left(N_{n,1}^{\diamond}(\mathbf{x}; \Theta_1, \mathcal{D}_n) < \lambda|\mathcal{D}\right) \leq \mathbb{P}\left(\left|N_{n,1}(\mathbf{x}; \Theta_1, \mathcal{D}_n) - N_{n,1}^{\diamond}(\mathbf{x}; \Theta_1, \mathcal{D}_n)\right| > \lambda|\mathcal{D}\right).$$

Lemma A.1. *Let $u \in \{0, 1\}$, as before, $N_{n,u}(A_n(\Theta)) = N_{n,u}(\mathbf{x}; \Theta, \mathcal{D}_n)$ is the number of observations of $\mathcal{D}_n$ such that $W = u$ that fall into in $A_n(\Theta) = A_n(\mathbf{x}; \Theta, \mathcal{D}_n)$ and $N_{n,u}^{\diamond}(A_n(\Theta)) = N_{n,u}^{\diamond}\left(\mathbf{x}; \Theta, \mathbf{X}^{\diamond 1}, \dots, \mathbf{X}^{\diamond n}, \mathcal{D}_n\right)$, the number of observations of $\mathcal{D}_n^{\diamond}$ such that $W = u$ that fall into $A_n(\Theta)$. Then,*

$$\forall \varepsilon > 0, \quad \mathbb{P}\left(\left|N_{n,u}(A_n(\Theta)) - N_{n,u}^{\diamond}(A_n(\Theta))\right| > \varepsilon\right) \leqslant 16(n+1)^{2d}e^{-\varepsilon^2/128n}.$$

Using Assumption 4.2 and Lemma A.1, there exists C and M positive constants such that:

$$\begin{aligned} Var(\alpha^{\diamond}) \leq & \frac{1}{B\lambda n} + \mathbb{E}\left[\mathbb{P}\left(\left|N_{n,1}\left(\mathbf{x};\Theta_1,\mathcal{D}_n\right) - N^{\diamond}_{n,1}\left(\mathbf{x};\Theta_1,\mathcal{D}_n\right)\right| > \frac{\sqrt{n}\left(\ln n\right)^{\beta}}{2}\middle|\mathcal{D}\right)\right] \\ \leq & \frac{2}{Bn^{3/2}\left(\ln n\right)^{\beta}} + \mathbb{P}\left(\left|N_{n,1}\left(\mathbf{x};\Theta_1,\mathcal{D}_n\right) - N^{\diamond}_{n,1}\left(\mathbf{x};\Theta_1,\mathcal{D}_n\right)\right| > \frac{\sqrt{n}\left(\ln n\right)^{\beta}}{2}\right) \\ \leq & \frac{2}{Mn^2} + 4C(n+1)^{2d}e^{-(\ln n)^{2\beta}/512} = \mathcal{O}\left(\frac{1}{n^2}\right). \end{aligned}$$

Thus since E_1^o converges to 0 in L^2 which implies that it also converges in L^1:

$$\begin{aligned} \mathbb{E}\left[\left|\sum_{\substack{i=1\\W_i^{\diamond}=1}}^{n}\alpha_i^{\diamond}E_1^{o,\diamond}(\mathcal{D},\mathbf{X}_i^{\diamond})\right|\right] \leq & \sum_{\substack{i=1\\W_i^{\diamond}=1}}^{n}\sqrt{Var(\alpha_i^{\diamond})}\sqrt{\mathbb{E}\left[\left|E_1^{o,\diamond}(\mathcal{D},\mathbf{X}_i^{\diamond})\right|^2\right]} + \mathbb{E}\left[\left|E_1^{o,\diamond}(\mathcal{D},\mathbf{X}_i^{\diamond})\right|\right] \\ \leq & n\sqrt{Var(\alpha_i^{\diamond})}\sqrt{\mathbb{E}\left[\left|E_1^{o,\diamond}(\mathcal{D},\mathbf{X}_i^{\diamond})\right|^2\right]} + \mathbb{E}\left[\left|E_1^{o,\diamond}(\mathcal{D},\mathbf{X}_i^{\diamond})\right|\right] \\ \longrightarrow & 0. \end{aligned}$$

The term V_n can now be treated:

$$\begin{aligned} V_n = & \left|\sum_{\substack{i=1\\W_i^{\diamond}=1}}^{n}\alpha_i^{\diamond}(\mathbf{x})\left[\mathbb{E}\left[Y^t(1)|\mathbf{X}_i^{\diamond}\right] - \mathbb{E}\left[Y^t(1)|\mathbf{X}=\mathbf{x}\right]\right]\right| \\ \leq & \left|\sum_{\substack{i=1\\W_i^{\diamond}=1}}^{n}\alpha_i^{\diamond}(\mathbf{x})\left[\mathbb{E}\left[Y^{new}(1)|\mathbf{X}_i^{\diamond}\right] - \mathbb{E}\left[Y^{new}(1)|\mathbf{X}=\mathbf{x}\right]\right]\right| \\ & + \left|\sum_{\substack{i=1\\W_i^{\diamond}=1}}^{n}\alpha_i^{\diamond}(\mathbf{x})\left[E_1^o(\mathcal{D},\mathbf{X}_i^{\diamond}) - E_1^o(\mathcal{D},\mathbf{x})\right]\right|. \end{aligned}$$

We can state the following lemma from Lemma 2 in [Meinshausen & Ridgeway 2006] and similar to Lemma 5 in [Bénard *et al.* 2022].

Lemma A.2. *Let Assumptions 4.1 and 4.2 be verified, let* $\mathbf{x} \in \mathcal{X}$ *and* $\ell \in [1, B]$. *Denote* $A_n(\mathbf{x}, \Theta_\ell, \mathcal{D}_n) = \bigotimes_{j=1}^{d} I(\mathbf{x}, \Theta_\ell, \mathcal{D}_n)$, *where* $I(\mathbf{x}, \Theta_\ell, \mathcal{D}_n)$ *are intervals, then*

$$\max_{j=1,\ldots,d} |I(\mathbf{x}, \mathbf{\Theta_e ll}, \mathcal{D_n})| = o(1).$$

Combining the Lemma A.2 with the continuity of τ_1, we get

$$\forall \ell \in [1, B], \forall \mathbf{x} \in \mathcal{X}, \sup_{\mathbf{z} \in A_n(\mathbf{x}, \Theta_e ll, \mathcal{D}_n)} |\tau_1(\mathbf{z}) - \tau_1(\mathbf{x})| \xrightarrow[n\to\infty]{a.s.} 0. \tag{A.9}$$

Using this result we get that the first term tends to 0 almost surely:

$$
\begin{aligned}
&\left| \sum_{\substack{i=1 \\ W_i^\diamond=1}}^{n} \alpha_i^\diamond(\mathbf{x}) \left[\mathbb{E}\left[Y^{new}(1)|\mathbf{X}_i^\diamond\right] - \mathbb{E}\left[Y^{new}(1)|\mathbf{X}=\mathbf{x}\right] \right] \right| \\
= &\left| \sum_{\substack{i=1 \\ W_i^\diamond=1 \\ \exists l | \mathbf{X}_i^\diamond \in A_n(\mathbf{x},\Theta_l)}}^{n} \alpha_i^\diamond(\mathbf{x}) \left[\mathbb{E}\left[Y^{new}(1)|\mathbf{X}_i^\diamond\right] - \mathbb{E}\left[Y^{new}(1)|\mathbf{X}=\mathbf{x}\right] \right] \right| \\
\leq &\sum_{\substack{i=1 \\ W_i^\diamond=1 \\ \exists l | \mathbf{X}_i^\diamond \in A_n(\mathbf{x},\Theta_l)}}^{n} \left| \alpha_i^\diamond(\mathbf{x}) \left[\mathbb{E}\left[Y^{new}(1)|\mathbf{X}_i^\diamond\right] - \mathbb{E}\left[Y^{new}(1)|\mathbf{X}=\mathbf{x}\right] \right] \right| \\
\leq &\sup_{\mathbf{z} \in A_n(\mathbf{x})} \left| \tau_1(\mathbf{z}) - \tau_1(\mathbf{x}) \right| \xrightarrow[n \to +\infty]{} 0.
\end{aligned}
$$

Since each τ term is bounded, by dominated convergence theorem we have the L^1 convergence of this quantity.

The second term can be shown to be L^1 convergent to 0 using the same rationale than for U_n.

The quantity $|\hat{\tau}^{new}(\mathbf{x}) - \tau^{new,\diamond}(\mathbf{x})|$ is now treated. We use the same decomposition and consider separately but in similar fashion $\left|\hat{\tau}_1^{new}(\mathbf{x}) - \tau_1^{new,\diamond}(\mathbf{x})\right|$ and $\left|\hat{\tau}_0^{new}(\mathbf{x}) - \tau_0^{new,\diamond}(\mathbf{x})\right|$:

$$
\left|\hat{\tau}_1^{new}(\mathbf{x}) - \tau_1^{new,\diamond}(\mathbf{x})\right| = \left| \frac{1}{B} \sum_{l=1}^{B} \sum_{j=1}^{n} \frac{\mathbb{1}_{\mathbf{X}_j \in A_n(l)} \mathbb{1}_{W_j=1}}{N_{n,1}(\mathbf{x};\Theta_l,\mathcal{D}_n)} Y_j^{new} - \frac{\mathbb{1}_{\mathbf{X}_j^\diamond \in A_n(l)} \mathbb{1}_{W_j^\diamond=1}}{N_{n,1}^\diamond(\mathbf{x};\Theta_l,\mathcal{D}_n)} Y_j^{new,\diamond} \right|.
$$

Following the HTERF consistency proof, this term converges to 0 almost surely. Since all the Y terms are bounded, with dominated convergence theorem this term tends to 0 in L^1.

□

B Generalisation bound

Using the proof in Section A, we get the following decomposition:

$$
\begin{aligned}
\left|\hat{\tau}_1^{new}(\mathbf{x}) - \tau_1^t(\mathbf{x})\right| \leq & \left|\hat{\tau}_1^{new}(\mathbf{x}) - \tau_1^{new,\diamond}(\mathbf{x})\right| + \left|\tau_1^{new,\diamond}(\mathbf{x}) - \tau_1^t(\mathbf{x})\right| \\
\leq & \left| \sum_{\substack{i=1 \\ W_i^\diamond=1}}^{n} \alpha_i^\diamond E_1^{o,\diamond}(\mathcal{D}, X_i^\diamond) \right| \\
& + \left| \sum_{\substack{i=1 \\ W_i^\diamond=1}}^{n} \alpha_i^\diamond(\mathbf{x}) \left[E_1^o(\mathcal{D}, \mathbf{X}_i^\diamond) - E_1^o(\mathcal{D}, \mathbf{x})\right] \right| \\
& + \left| \sum_{\substack{i=1 \\ W_i^\diamond=1}}^{n} \alpha_i^\diamond(\mathbf{x}) \left[\mathbb{E}\left[Y^{new}(1)|\mathbf{X}_i^\diamond\right] - \mathbb{E}\left[Y^{new}(1)|\mathbf{X}=\mathbf{x}\right]\right] \right| \\
& + \left| \sum_{\substack{i=1 \\ W_i^\diamond=1}}^{n} \alpha_i^\diamond \varepsilon_{1,i}^{\diamond,s} \right| + \left| \sum_{\substack{i=1 \\ W_i^\diamond=1}}^{n} \alpha_i^\diamond \varepsilon_{1,i}^{\diamond,t} \right| \\
& + \left| \frac{1}{B} \sum_{l=1}^{B} \sum_{j=1}^{n} \frac{\mathbb{1}_{\mathbf{X}_j \in A_n(l)} \mathbb{1}_{W_j=1}}{N_{n,1}(\mathbf{x}; \Theta_l, \mathcal{D}_n)} Y_j^{new} - \frac{\mathbb{1}_{\mathbf{X}_j^\diamond \in A_n(l)} \mathbb{1}_{W_j^\diamond=1}}{N_{n,1}^\diamond(\mathbf{x}; \Theta_l, \mathcal{D}_n)} Y_j^{new,\diamond} \right| \\
\leq & Bound_{offset} + Bound_{HTERF},
\end{aligned}
$$

where

$$
\begin{aligned}
Bound_{offset} = & \left| \sum_{\substack{i=1 \\ W_i^\diamond=1}}^{n} \alpha_i^\diamond E_1^{o,\diamond}(\mathcal{D}, X_i^\diamond) \right| \\
& + \left| \sum_{\substack{i=1 \\ W_i^\diamond=1}}^{n} \alpha_i^\diamond(\mathbf{x}) \left[E_1^o(\mathcal{D}, \mathbf{X}_i^\diamond) - E_1^o(\mathcal{D}, \mathbf{x})\right] \right|
\end{aligned}
$$

and

$$
\begin{aligned}
Bound_{HTERF} = &+ \left| \sum_{\substack{i=1 \\ W_i^\diamond = 1}}^{n} \alpha_i^\diamond(\mathbf{x}) \left[\mathbb{E}\left[Y^{new}(1) | \mathbf{X}_i^\diamond \right] - \mathbb{E}\left[Y^{new}(1) | \mathbf{X} = \mathbf{x} \right] \right] \right| \\
&+ \left| \sum_{\substack{i=1 \\ W_i^\diamond = 1}}^{n} \alpha_i^\diamond \varepsilon_{1,i}^{\diamond,s} \right| + \left| \sum_{\substack{i=1 \\ W_i^\diamond = 1}}^{n} \alpha_i^\diamond \varepsilon_{1,i}^{\diamond,t} \right| \\
&+ \left| \frac{1}{B} \sum_{l=1}^{B} \sum_{j=1}^{n} \frac{\mathbb{1}_{\mathbf{X}_j \in A_n(l)} \mathbb{1}_{W_j = 1}}{N_{n,1}(\mathbf{x}; \Theta_l, \mathcal{D}_n)} Y_j^{new} - \frac{\mathbb{1}_{\mathbf{X}_j^\diamond \in A_n(l)} \mathbb{1}_{W_j^\diamond = 1}}{N_{n,1}^\diamond(\mathbf{x}; \Theta_l, \mathcal{D}_n)} Y_j^{new,\diamond} \right|.
\end{aligned}
$$

Conclusion et perspectives

Dans ce travail de thèse, nous avons apporté des contributions aux forêts causales et à l'apprentissage par transfert dans un contexte causal. Une attention particulière a été portée à la notion d'interprétabilité lors du développement du modèle de forêt.

Conclusions

Dans le Chapitre 2, nous avons développé un nouvel algorithme de forêt causale nommé HTERF, qui permet d'estimer la quantité CATE sous un traitement binaire. Cette approche obtient de meilleurs résultats en termes de qualité de l'estimation de CATE ainsi qu'en termes d'interprétabilité, en comparaison avec GRF sur des jeux de données synthétiques et semi-synthétiques. Nous avons obtenu un résultat de convergence presque sûre pour ce nouvel estimateur. Ce résultat est obtenu sous des hypothèses plus faibles et en alignement avec ce qui se fait en pratique, en comparaison avec le résultat de convergence en loi de GRF. Nous avons également établi un résultat d'interprétabilité, caractérisant la surreprésentation des variables informatives dans les divisions des premiers étages des arbres causaux. Dans le cadre de ces travaux, un package en `Julia` a été développé et est présenté en appendice.

Dans le Chapitre 3, deux applications de HTERF sont proposées sur des jeux de données réels. La première application concerne une quantité relative au risque de crédit mesurée chez Natixis, le coût du risque de crédit. Nous proposons d'évaluer l'impact des crises économiques et financières sur cette variable, conditionnellement à un ensemble de variables macro-économiques. Le backtesting des performances du modèle est de qualité hétérogène selon les années. Cependant, les résultats sur l'identification des variables informatives permettent d'expliquer les variations du coût du risque selon les différents types de crises. La seconde application concerne le phénomène climatique El Niño, et plus particulièrement son impact sur les précipitations dans l'est australien. Pour une des deux stations météorologiques considérées, nous avons obtenu des résultats cohérents avec le comportement attendu d'El Niño, que nous avons donc pu quantifier.

Dans le Chapitre 4, nous proposons l'algorithme offset causal qui est une adaptation de l'algorithme offset de [Wang 2016], permettant de faire de l'adaptation de domaine par changement de modèle. Nous obtenons un résultat de convergence en norme L^1 de cet algorithme sous des conditions plus strictes que HTERF. De bons résultats ont également été obtenus sur des jeux de données synthétiques et semi-synthétiques. Une borne de généralisation peut également être exprimée, permettant de décomposer l'erreur en une partie liée à la méthode offset et une autre partie liée à l'utilisation de HTERF.

Perspectives

Des perspectives à la fois théoriques et pratiques peuvent être envisagées concernant ces travaux.

Dans le Chapitre 2, nous avons mis en évidence l'impact de l'étape préliminaire de centrage sur les performances de HTERF. L'utilisation d'un algorithme de régression adapté dans cette étape permet d'améliorer drastiquement les performances de HTERF. Concernant le package `Julia` implémentant cet algorithme, bien qu'il mette déjà en place de la parallélisation, il pourrait être davantage optimisé pour obtenir de meilleurs temps d'exécution, que ce soit lors de l'étape de construction des arbres ou lors de l'estimation de CATE sur des points de test.

Dans le Chapitre 3, l'application sur le risque de crédit pourrait être affinée en proposant des variables explicatives plus fines et spécifiques à chaque contrepartie considérée, en utilisant par exemple des variables sectorielles et géographiques. Concernant l'application climatique, l'introduction de nouvelles variables explicatives permettrait également d'obtenir de meilleurs résultats. L'étude pourrait également être étendue à de nouvelles stations météorologiques afin d'obtenir des données de meilleure qualité d'une part, et d'étudier l'impact d'El Niño sur des régions plus variées d'Australie.

Concernant le Chapitre 4, un travail additionnel pourrait être effectué pour alléger les hypothèses nécessaires à la convergence. Un nouveau jeu d'hypothèses minimales pourrait également être identifié afin d'obtenir la convergence presque sûre de la méthode offset causale.

Appendices

Appendix A

Example of code used for causal transfer learning

In this appendix we present a code sample used in Section 4.4.2, to illustrate how to implement the offset causal method in `Python` with a `Julia` code.

Causal transfer learning, code example

April 4, 2024

[12]:
```
import pandas as pd
import numpy as np
from sklearn import svm
from sklearn.ensemble import RandomForestRegressor
from sklearn import kernel_ridge
import copy
```

[13]:
```
import sys
print(sys.executable)
```

```
/Users/jocteur/PycharmProjects/transfer final/venvtrans/bin/python
```

[14]:
```
import julia as ju
ju.install(julia="/Applications/Julia-1.7.app/Contents/Resources/julia/bin/
 ↪julia")
```

```
[ Info: Julia version info

Julia Version 1.7.2
Commit bf53498635 (2022-02-06 15:21 UTC)
Platform Info:
  OS: macOS (arm64-apple-darwin21.2.0)
  uname: Darwin 23.3.0 Darwin Kernel Version 23.3.0: Wed Dec 20 21:30:27 PST
2023; root:xnu-10002.81.5~7/RELEASE_ARM64_T8103 arm64 arm
  CPU: Apple M1:
              speed         user         nice          sys         idle
irq
       #1    24 MHz     543827 s          0 s     262751 s    3133942 s
0 s
       #2    24 MHz     515682 s          0 s     254138 s    3187224 s
0 s
       #3    24 MHz     465717 s          0 s     226089 s    3275080 s
0 s
       #4    24 MHz     425306 s          0 s     200166 s    3349763 s
0 s
       #5    24 MHz     311702 s          0 s      57000 s    3633457 s
0 s
       #6    24 MHz     180331 s          0 s      40341 s    3785014 s
0 s
```

```
      #7    24 MHz     135248 s          0 s      26341 s    3845727 s
0 s
      #8    24 MHz     108617 s          0 s      20345 s    3879090 s
0 s

  Memory: 16.0 GB (45.75 MB free)
  Uptime: 1.147224e6 sec
  Load Avg:  2.20947265625  2.41650390625  2.14892578125
  WORD_SIZE: 64
  LIBM: libopenlibm
  LLVM: libLLVM-12.0.1 (ORCJIT, cyclone)
Environment:
  MANPATH = /opt/homebrew/share/man::
  TERM = xterm-color
  HOMEBREW_REPOSITORY = /opt/homebrew
  PATH = /Users/jocteur/PycharmProjects/transfer final/venvtrans/bin:/Library/Fr
ameworks/Python.framework/Versions/2.7/bin:/opt/homebrew/bin:/opt/homebrew/sbin:
/usr/local/bin:/System/Cryptexes/App/usr/bin:/usr/bin:/bin:/usr/sbin:/sbin:/var/
run/com.apple.security.cryptexd/codex.system/bootstrap/usr/local/bin:/var/run/co
m.apple.security.cryptexd/codex.system/bootstrap/usr/bin:/var/run/com.apple.secu
rity.cryptexd/codex.system/bootstrap/usr/appleinternal/bin:/Library/TeX/texbin
  XPC_FLAGS = 0x0
  HOME = /Users/jocteur
  HOMEBREW_PREFIX = /opt/homebrew
  HOMEBREW_CELLAR = /opt/homebrew/Cellar
  INFOPATH = /opt/homebrew/share/info:

[ Info: Julia executable:
/Applications/Julia-1.7.app/Contents/Resources/julia/bin/julia
[ Info: Trying to import PyCall…
 Info: PyCall is already installed and compatible with Python executable.

 PyCall:
     python: /Users/jocteur/PycharmProjects/transfer final/venvtrans/bin/python
     libpython: /Library/Developer/CommandLineTools/Library/Frameworks/Python3.
framework/Versions/3.9/Python3
 Python:
     python: /Users/jocteur/PycharmProjects/transfer final/venvtrans/bin/python
     libpython: /Library/Developer/CommandLineTools/Library/Frameworks/Python3.
framework/Versions/3.9/Python3
```

[15]:
```
from julia import CausalForest
from julia import StatsBase
```

[16]:
```
rand = np.random.RandomState(10)

xtrain = rand.uniform(0,10,100000).reshape((10000,10))
ttrain = rand.binomial(1,0.5,10000)
```

```
ytrain = np.sin(xtrain[:,0])*ttrain  + np.cos(xtrain[:,1])
```

source Y = sin(X1) * T + cos(X2) 10000 units

[17]:
```
xnew = rand.uniform(0,10,5000).reshape((500,10))
tnew = rand.binomial(1,0.5,500)
ynew = np.cos(xnew[:,0])*tnew  + np.cos(xnew[:,1])
```

target Y = cos(X1) * T + cos(X2) 500 units

[18]:
```
def rmse(predictions, targets):
    return np.sqrt(((predictions - targets) ** 2).mean())
```

[19]:
```
from julia import Main

Main.xtrainj = xtrain
Main.ttrainj = ttrain
Main.ytrainj = ytrain

Main.eval('cf = CausalForest.build_forest(false, true, ytrainj, ttrainj,␣
 ↪xtrainj, true, 10, 500, 500)')
Main.eval('tauhat = CausalForest.apply_forest(cf, xtrainj)')
pred = Main.tauhat
print(rmse(pred, np.sin(xtrain[:,0])))
```

```
0.004171521013063611
```

Above is HTERF error on source data

[20]:
```
errCATE_offKRR = []
errCATE_offsingleKRR = []
errCATE_naive = []
errCATE_offRF = []
errCATE_offsingleRF = []

xtrain0 = xtrain[ttrain==0,:]
xtrain1 = xtrain[ttrain==1,:]
ytrain0 = ytrain[ttrain==0]
ytrain1 = ytrain[ttrain==1]
ttrain0 = ttrain[ttrain==0]
ttrain1 = ttrain[ttrain==1]

np.random.seed(78)

fs_model0rf = RandomForestRegressor(n_estimators=500)
fs_model0rf.fit(xtrain0, ytrain0)

fs_model1rf = RandomForestRegressor(n_estimators=500)
fs_model1rf.fit(xtrain1, ytrain1)
```

```
fs_modelrf = RandomForestRegressor(n_estimators=500)
fs_modelrf.fit(np.c_[xtrain, ttrain ], ytrain)

fs_model0krr = kernel_ridge.KernelRidge(alpha=0.1, kernel="rbf")
fs_model0krr.fit(xtrain0, ytrain0)

fs_model1krr = kernel_ridge.KernelRidge(alpha=0.1, kernel="rbf")
fs_model1krr.fit(xtrain1, ytrain1)

fs_modelkrr = kernel_ridge.KernelRidge(alpha=0.1, kernel="rbf")
fs_modelkrr.fit(np.c_[xtrain, ttrain ], ytrain)


for i in range(50):

    xnew = rand.uniform(0,10,5000).reshape((500,10))
    tnew = rand.binomial(1,0.5,500)
    ynew = np.cos(xnew[:,0])*tnew  + np.cos(xnew[:,1])

    indices = np.random.permutation(xnew.shape[0])
    sz_h = int(xnew.shape[0] / 2)
    xtestL = xnew[indices[:sz_h],:]
    xtestU = xnew[indices[sz_h:],:]
    ytestL = ynew[indices[:sz_h]]
    ytestU = ynew[indices[sz_h:]]
    ttestL = tnew[indices[:sz_h]]
    ttestU = tnew[indices[sz_h:]]


    xtestL0 = xtestL[ttestL==0,:]
    xtestU0 = xtestU[ttestU==0,:]
    ytestL0 = ytestL[ttestL==0]
    ytestU0 = ytestU[ttestU==0]
    xtestL1 = xtestL[ttestL==1,:]
    xtestU1 = xtestU[ttestU==1,:]
    ytestL1 = ytestL[ttestL==1]
    ytestU1 = ytestU[ttestU==1]


    predtestL0 = fs_model0rf.predict(xtestL0)
    offsetobj0 = ytestL0 - predtestL0
    fo_model0 = RandomForestRegressor(n_estimators=500)
    fo_model0.fit(xtestL0, offsetobj0)
    predtrain0 = fo_model0.predict(xtrain0)
    ynew0 = ytrain0 + predtrain0
```

```
    predtestL1 = fs_model1rf.predict(xtestL1)
    offsetobj1 = ytestL1 - predtestL1
    fo_model1 = RandomForestRegressor(n_estimators=500)
    fo_model1.fit(xtestL1, offsetobj1)
    predtrain1 = fo_model1.predict(xtrain1)
    ynew1 = ytrain1 + predtrain1

    X = np.vstack([xtrain0, xtrain1, xtestL])
    Y = np.hstack([ynew0, ynew1, ytestL])
    T = np.hstack([ttrain0, ttrain1, ttestL])


    Main.Xj = X
    Main.Tj = T
    Main.Yj = Y

    Main.testUj = xtestU

    Main.eval('cf = CausalForest.build_forest(false, true, Yj, Tj, Xj, true,␣
↪10, 500, 500)')
    Main.eval('tauhat = CausalForest.apply_forest(cf, testUj)')
    pred = Main.tauhat
    errCATE_offRF.append(rmse(pred, np.cos(xtestU[:,0])))



    predtestL = fs_modelrf.predict(np.c_[xtestL, ttestL])
    offsetobj = ytestL - predtestL
    fo_model = RandomForestRegressor(n_estimators=500)
    fo_model.fit(np.c_[xtestL, ttestL ], offsetobj)
    predtrain = fo_model.predict(np.c_[xtrain, ttrain])
    ynew = ytrain + predtrain


    X = np.vstack([xtrain, xtestL])
    Y = np.hstack([ynew, ytestL])
    T = np.hstack([ttrain, ttestL])


    Main.Xj = X
    Main.Tj = T
    Main.Yj = Y

    Main.testUj = xtestU
```

```
    Main.eval('cf = CausalForest.build_forest(false, true, Yj, Tj, Xj, true,␣
↪10, 500, 500)')
    Main.eval('tauhat = CausalForest.apply_forest(cf, testUj)')
    pred = Main.tauhat
    errCATE_offsingleRF.append(rmse(pred, np.cos(xtestU[:,0])))


    Main.xtestLj = xtestL
    Main.ttestLj = ttestL
    Main.ytestLj = ytestL

    Main.testUj = xtestU

    Main.eval('cf = CausalForest.build_forest(false, true, ytestLj, ttestLj,␣
↪xtestLj, true, 10, 500, 500)')
    Main.eval('tauhat = CausalForest.apply_forest(cf, testUj)')
    pred = Main.tauhat
    errCATE_naive.append(rmse(pred, np.cos(xtestU[:,0])))




    predtestL0 = fs_model0krr.predict(xtestL0)
    offsetobj0 = ytestL0 - predtestL0
    fo_model0 = kernel_ridge.KernelRidge(alpha=0.1, kernel="rbf")
    fo_model0.fit(xtestL0, offsetobj0)
    predtrain0 = fo_model0.predict(xtrain0)
    ynew0 = ytrain0 + predtrain0


    predtestL1 = fs_model1krr.predict(xtestL1)
    offsetobj1 = ytestL1 - predtestL1
    fo_model1 = kernel_ridge.KernelRidge(alpha=0.1, kernel="rbf")
    fo_model1.fit(xtestL1, offsetobj1)
    predtrain1 = fo_model1.predict(xtrain1)
    ynew1 = ytrain1 + predtrain1

    X = np.vstack([xtrain0, xtrain1, xtestL])
    Y = np.hstack([ynew0, ynew1, ytestL])
    T = np.hstack([ttrain0, ttrain1, ttestL])


    Main.Xj = X
    Main.Tj = T
    Main.Yj = Y

    Main.testUj = xtestU
```

```
    Main.eval('cf = CausalForest.build_forest(false, true, Yj, Tj, Xj, true,␣
↪10, 500, 500)')
    Main.eval('tauhat = CausalForest.apply_forest(cf, testUj)')
    pred = Main.tauhat
    errCATE_offKRR.append(rmse(pred, np.cos(xtestU[:,0])))



    predtestL = fs_modelkrr.predict(np.c_[xtestL, ttestL])
    offsetobj = ytestL - predtestL
    fo_model = kernel_ridge.KernelRidge(alpha=0.1, kernel="rbf")
    fo_model.fit(np.c_[xtestL, ttestL ], offsetobj)
    predtrain = fo_model.predict(np.c_[xtrain, ttrain])
    ynew = ytrain + predtrain


    X = np.vstack([xtrain, xtestL])
    Y = np.hstack([ynew, ytestL])
    T = np.hstack([ttrain, ttestL])


    Main.Xj = X
    Main.Tj = T
    Main.Yj = Y

    Main.testUj = xtestU

    Main.eval('cf = CausalForest.build_forest(false, true, Yj, Tj, Xj, true,␣
↪10, 500, 500)')
    Main.eval('tauhat = CausalForest.apply_forest(cf, testUj)')
    pred = Main.tauhat
    errCATE_offsingleKRR.append(rmse(pred, np.cos(xtestU[:,0])))
```

[22]:
```
print(np.mean(errCATE_offKRR))
print(np.mean(errCATE_offsingleKRR))
print(np.mean(errCATE_naive))
print(np.mean(errCATE_offRF))
print(np.mean(errCATE_offsingleRF))
```

```
0.9574646152746428
0.9602047043555467
0.34800205906959897
0.1203190149657811
0.1354537317550674
```

Above : errors on simulations using KRR and regression random forests

Bibliography

[AFP 2023] *Dossier AFP réchauffement climatique*. https://factuel.afp.com/doc.afp.com.33B93NE, 2023. Accessed: 2023-04-25. (Cited on page 83.)

[Athey & Imbens 2016] Susan Athey and Guido Imbens. *Recursive partitioning for heterogeneous causal effects*. Proceedings of the National Academy of Sciences, vol. 113, no. 27, pages 7353–7360, 2016. (Cited on pages 23, 36 and 39.)

[Athey *et al.* 2019] Susan Athey, Julie Tibshirani and Stefan Wager. *Generalized random forests*. The Annals of Statistics, vol. 47, no. 2, pages 1148–1178, 2019. (Cited on pages v, vii, 19, 23, 25, 33, 35, 39, 40, 42, 43, 44, 46, 47, 48 and 53.)

[Autralian Government - Bureau of Meteorology 2012] Autralian Government - Bureau of Meteorology. *Record-breaking La Nina events*, 2012. (Cited on pages 82, 85 and 86.)

[Bénard *et al.* 2022] Clément Bénard, Sébastien Da Veiga and Erwan Scornet. *Mean decrease accuracy for random forests: inconsistency, and a practical solution via the Sobol-MDA*. Biometrika, vol. 109, no. 4, pages 881–900, 2022. (Cited on page 107.)

[Bica & van der Schaar 2022] Ioana Bica and Mihaela van der Schaar. *Transfer learning on heterogeneous feature spaces for treatment effects estimation*. Advances in Neural Information Processing Systems, vol. 35, pages 37184–37198, 2022. (Cited on page 94.)

[Boucheron *et al.* 2013] Stéphane Boucheron, Gábor Lugosi and Pascal Massart. Concentration inequalities: A nonasymptotic theory of independence. Oxford university press, 2013. (Cited on page 43.)

[Bousquet & Elisseeff 2002] Olivier Bousquet and André Elisseeff. *Stability and generalization*. The Journal of Machine Learning Research, vol. 2, pages 499–526, 2002. (Cited on page 95.)

[BPCE 2024] Groupe BPCE. *Groupe BPCE Rapport Pilier III 2023*, 2024. Accessed: 2024-04-01. (Cited on pages 3 and 4.)

[Breiman 2001] Leo Breiman. *Random forests*. Machine learning, vol. 45, no. 1, pages 5–32, 2001. (Cited on pages 19, 23 and 35.)

[Cartwright 2012] Nancy Cartwright. *Causal laws and effective strategies*. In Arguing About Science, pages 466–479. Routledge, 2012. (Cited on page 11.)

[Charig *et al.* 1986] Clive Charig, David R. Webb, Stephen R. Payne and J. E. A. Wickham. *Comparison of treatment of renal calculi by open surgery, percutaneous nephrolithotomy, and extracorporeal shockwave lithotripsy.* British Medical Journal (Clinical research ed.), vol. 292, pages 879 – 882, 1986. (Cited on page 11.)

[Chipman *et al.* 2010a] Hugh A. Chipman, Edward I. George and Robert E. McCulloch. *BART: Bayesian additive regression trees.* The Annals of Applied Statistics, vol. 4, no. 1, pages 266 – 298, 2010. (Cited on page 20.)

[Chipman *et al.* 2010b] Hugh A Chipman, Edward I George and Robert E McCulloch. *BART: Bayesian additive regression trees.* The Annals of Applied Statistics, vol. 4, no. 1, pages 266–298, 2010. (Cited on page 35.)

[Curth & van der Schaar 2021] Alicia Curth and Mihaela van der Schaar. *Nonparametric estimation of heterogeneous treatment effects: From theory to learning algorithms.* In International Conference on Artificial Intelligence and Statistics, pages 1810–1818. PMLR, 2021. (Cited on pages 19 and 20.)

[Drouet 2012] Isabelle Drouet. Causes, probabilités, inférences. Vuibert, 2012. (Cited on page 10.)

[Elie-Dit-Cosaque & Maume-Deschamps 2022] Kévin Elie-Dit-Cosaque and Véronique Maume-Deschamps. *Random forest estimation of conditional distribution functions and conditional quantiles.* Electronic Journal of Statistics, vol. 16, no. 2, pages 6553–6583, 2022. (Cited on pages 42, 43, 52, 53, 59, 60, 61 and 63.)

[Green & Kern 2012] Donald P Green and Holger L Kern. *Modeling heterogeneous treatment effects in survey experiments with Bayesian additive regression trees.* Public opinion quarterly, vol. 76, no. 3, pages 491–511, 2012. (Cited on page 35.)

[Gretton *et al.* 2006] Arthur Gretton, Karsten Borgwardt, Malte Rasch, Bernhard Schölkopf and Alex Smola. *A kernel method for the two-sample-problem.* Advances in neural information processing systems, vol. 19, 2006. (Cited on page 94.)

[Györfi *et al.* 2002] László Györfi, Michael Kohler, Adam Krzyzak, Harro Walk*et al.* A distribution-free theory of nonparametric regression, volume 1. Springer, 2002. (Cited on page 58.)

[Hahn *et al.* 2020] P Richard Hahn, Jared S Murray and Carlos M Carvalho. *Bayesian regression tree models for causal inference: Regularization, confounding, and heterogeneous effects (with discussion).* Bayesian Analysis, vol. 15, no. 3, pages 965–1056, 2020. (Cited on pages 16 and 21.)

[Hill & Su 2013] Jennifer Hill and Yu-Sung Su. *Assessing lack of common support in causal inference using Bayesian nonparametrics: Implications for evaluating the effect of breastfeeding on children's cognitive outcomes.* The Annals of Applied Statistics, pages 1386–1420, 2013. (Cited on page 35.)

[Hill 2011] Jennifer L Hill. *Bayesian nonparametric modeling for causal inference.* Journal of Computational and Graphical Statistics, vol. 20, no. 1, pages 217–240, 2011. (Cited on pages 21, 35, 43, 47 and 51.)

[Huang *et al.* 2006] Jiayuan Huang, Arthur Gretton, Karsten Borgwardt, Bernhard Schölkopf and Alex Smola. *Correcting sample selection bias by unlabeled data.* Advances in neural information processing systems, vol. 19, 2006. (Cited on page 94.)

[Hume 1995] David Hume. Traité de la nature humaine, trad. Flammarion, 1995. translated by P. Baranger, P. Saltel. (Cited on page 10.)

[Huyen 2022] Chip Huyen. Designing machine learning systems. " O'Reilly Media, Inc.", 2022. (Cited on page 93.)

[Im *et al.* 2021] Daniel Jiwoong Im, Kyunghyun Cho and Narges Razavian. *Causal effect variational autoencoder with uniform treatment.* arXiv preprint arXiv:2111.08656, 2021. (Cited on page 51.)

[Imbens & Rubin 2015] Guido W Imbens and Donald B Rubin. Causal inference in statistics, social, and biomedical sciences. Cambridge University Press, 2015. (Cited on pages 14, 33 and 92.)

[Ishigami & Homma 1990] Tsutomu Ishigami and Toshimitsu Homma. *An importance quantification technique in uncertainty analysis for computer models.* In [1990] Proceedings. First International Symposium on Uncertainty Modeling and Analysis, pages 398–403. IEEE, 1990. (Cited on page 51.)

[Jacob 2021] Daniel Jacob. *CATE meets ML: Conditional average treatment effect and machine learning.* Digital Finance, vol. 3, no. 2, pages 99–148, 2021. (Cited on page 15.)

[Jocteur *et al.* 2024] Bérénice-Alexia Jocteur, Véronique Maume-Deschamps and Pierre Ribereau. *Heterogeneous Treatment Effect-based Random Forest: HTERF.* Computational Statistics & Data Analysis, page 107970, 2024. (Cited on page 31.)

[Johansson *et al.* 2016] Fredrik Johansson, Uri Shalit and David Sontag. *Learning representations for counterfactual inference.* In International conference on machine learning, pages 3020–3029. PMLR, 2016. (Cited on page 51.)

[Kennedy 2020] Edward H Kennedy. *Towards optimal doubly robust estimation of heterogeneous causal effects.* arXiv preprint arXiv:2004.14497, 2020. (Cited on page 16.)

[Kennedy 2022] Edward H Kennedy. *Semiparametric doubly robust targeted double machine learning: a review.* arXiv preprint arXiv:2203.06469, 2022. (Cited on page 16.)

[Klusowski 2019] Jason M Klusowski. *Analyzing cart.* arXiv preprint arXiv:1906.10086, 2019. (Cited on page 42.)

[Künzel *et al.* 2018] Sören R Künzel, Bradly C Stadie, Nikita Vemuri, Varsha Ramakrishnan, Jasjeet S Sekhon and Pieter Abbeel. *Transfer learning for estimating causal effects using neural networks.* arXiv preprint arXiv:1808.07804, 2018. (Cited on page 94.)

[Künzel *et al.* 2019] Sören R Künzel, Jasjeet S Sekhon, Peter J Bickel and Bin Yu. *Metalearners for estimating heterogeneous treatment effects using machine learning.* Proceedings of the national academy of sciences, vol. 116, no. 10, pages 4156–4165, 2019. (Cited on pages 15, 18 and 34.)

[Lewis 1973] David Lewis. *Causation.* The journal of philosophy, vol. 70, no. 17, pages 556–567, 1973. (Cited on page 12.)

[Louizos *et al.* 2017] Christos Louizos, Uri Shalit, Joris M Mooij, David Sontag, Richard Zemel and Max Welling. *Causal effect inference with deep latent-variable models.* Advances in neural information processing systems, vol. 30, 2017. (Cited on page 51.)

[Meinshausen & Ridgeway 2006] Nicolai Meinshausen and Greg Ridgeway. *Quantile regression forests.* Journal of machine learning research, vol. 7, no. 6, 2006. (Cited on page 107.)

[Nie & Wager 2021] Xinkun Nie and Stefan Wager. *Quasi-oracle estimation of heterogeneous treatment effects.* Biometrika, vol. 108, no. 2, pages 299–319, 2021. (Cited on pages 15, 17 and 46.)

[Park & Sandberg 1991] Jooyoung Park and Irwin W Sandberg. *Universal approximation using radial-basis-function networks.* Neural computation, vol. 3, no. 2, pages 246–257, 1991. (Cited on page 42.)

[Pearl 2009] Judea Pearl. Causality. Cambridge University Press, Cambridge, UK, 2 édition, 2009. (Cited on page 14.)

[Powers *et al.* 2018] Scott Powers, Junyang Qian, Kenneth Jung, Alejandro Schuler, Nigam H Shah, Trevor Hastie and Robert Tibshirani. *Some methods for heterogeneous treatment effect estimation in high dimensions.* Statistics in medicine, vol. 37, no. 11, pages 1767–1787, 2018. (Cited on pages 21 and 22.)

[Robinson 1988] Peter M Robinson. *Root-N-consistent semiparametric regression.* Econometrica: Journal of the Econometric Society, pages 931–954, 1988. (Cited on page 17.)

[Rosenbaum & Rubin 1983] Paul R Rosenbaum and Donald B Rubin. *The central role of the propensity score in observational studies for causal effects.* Biometrika, vol. 70, no. 1, pages 41–55, 1983. (Cited on pages 34 and 93.)

[Rubin 1974] Donald B Rubin. *Estimating causal effects of treatments in randomized and nonrandomized studies.* Journal of educational Psychology, vol. 66, no. 5, page 688, 1974. (Cited on page 12.)

[Scornet *et al.* 2015] Erwan Scornet, Gérard Biau and Jean-Philippe Vert. *Consistency of random forests.* The Annals of Statistics, vol. 43, no. 4, pages 1716–1741, 2015. (Cited on pages 25, 33, 37, 40 and 44.)

[Shalit *et al.* 2017] Uri Shalit, Fredrik D Johansson and David Sontag. *Estimating individual treatment effect: generalization bounds and algorithms.* In International conference on machine learning, pages 3076–3085. PMLR, 2017. (Cited on pages 20 and 51.)

[Shi *et al.* 2019] Claudia Shi, David Blei and Victor Veitch. *Adapting neural networks for the estimation of treatment effects.* Advances in neural information processing systems, vol. 32, 2019. (Cited on page 20.)

[Skyrms 1980] Brian Skyrms. Causal necessity: A pragmatic investigation of the necessity of laws. Yale University Press, New Haven, 1980. (Cited on page 11.)

[Su *et al.* 2009] Xiaogang Su, Chih-Ling Tsai, Hansheng Wang, David M Nickerson and Bogong Li. *Subgroup analysis via recursive partitioning.* Journal of Machine Learning Research, vol. 10, no. 2, 2009. (Cited on page 35.)

[Suppes 1970] P. Suppes. A probabilistic theory of causality:. Number n° 24 de A Probabilistic Theory of Causality. North-Holland Publishing Company, 1970. (Cited on page 11.)

[Wager & Athey 2018] Stefan Wager and Susan Athey. *Estimation and inference of heterogeneous treatment effects using random forests.* Journal of the American Statistical Association, vol. 113, no. 523, pages 1228–1242, 2018. (Cited on pages 23, 35, 41 and 42.)

[Wahba 2003] Grace Wahba. *An introduction to reproducing kernel hilbert spaces and why they are so useful.* In Proceedings of the 13th IFAC Symposium on System Identification (SYSID 2003), pages 525–528, 2003. (Cited on page 95.)

[Wang & Schneider 2015] Xuezhi Wang and Jeff G Schneider. *Generalization Bounds for Transfer Learning under Model Shift.* In UAI, pages 922–931, 2015. (Cited on pages 28 and 95.)

[Wang *et al.* 2014] Xuezhi Wang, Tzu-Kuo Huang and Jeff Schneider. *Active transfer learning under model shift.* In International Conference on Machine Learning, pages 1305–1313. PMLR, 2014. (Cited on pages 94 and 95.)

[Wang 2016] Xuezhi Wang. *Active Transfer Learning.* PhD thesis, Ph. D. Dissertation. BAE Systems, 2016. (Cited on pages vi, vii, 28, 91 and 111.)

[Wei *et al.* 2024] Song Wei, Hanyu Zhang, Ronald Moore, Rishikesan Kamaleswaran and Yao Xie. *Transfer Learning for Causal Effect Estimation*, 2024. (Cited on pages 94 and 101.)

[Woodward 2003] James F. Woodward. Making things happen: A theory of causal explanation. Oxford University Press, New York, 2003. (Cited on page 13.)

[Yang *et al.* 2020] Qiang Yang, Yu Zhang, Wenyuan Dai and Sinno Jialin Pan. Transfer learning. Cambridge University Press, 2020. (Cited on page 22.)

[Zhang *et al.* 2013] Kun Zhang, Bernhard Schölkopf, Krikamol Muandet and Zhikun Wang. *Domain adaptation under target and conditional shift.* In International conference on machine learning, pages 819–827. Pmlr, 2013. (Cited on page 94.)

[Zhao *et al.* 2019] Han Zhao, Remi Tachet Des Combes, Kun Zhang and Geoffrey Gordon. *On learning invariant representations for domain adaptation.* In International conference on machine learning, pages 7523–7532. PMLR, 2019. (Cited on page 94.)

Forêts causales et transfert par changement de modèle pour l'estimation d'effets de traitement hétérogènes

Résumé. Cette thèse a été réalisée dans le cadre d'un partenariat CIFRE entre l'Université Lyon 1 et Natixis. Elle a pour objectif de développer des méthodes d'apprentissage statistique permettant l'estimation d'effets causaux. Pour ce faire un modèle spécifique de forêt aléatoire a été développé et ses propriétés asymptotiques ont été étudiées. Puis des applications sur des données réelles ont été proposées, notamment sur une quantité d'intérêt pour la direction des risques de Natixis, mais aussi sur une problématique climatique. Enfin une méthode d'apprentissage par transfert sur la forêt précédemment introduite est proposée et des propriétés de convergence ainsi qu'une borne de généralisation sont établies.

Mots-clés : forêt causale, inférence causale, effet de traitement hétérogène, résultats contrefactuels, apprentissage par transfert, décalage de modèle, risque de crédit, El Niño

Abstract. This thesis was carried out within the framework of a CIFRE partnership between University Lyon 1 and Natixis. Its aim is to develop statistical learning methods for the estimation of causal effects. To this end, a specific model of random forest was developed, and its asymptotic properties were studied. Subsequent applications on real data were proposed, mainly on data of interest for the risk management department of Natixis, but also on a climate-related issue. Finally, a transfer learning method on the previously introduced forest is proposed, and convergence properties as well as a generalization bound are established.

Keywords: causal forest, causal inference, heterogeneous treatment effect, potential outcomes, transfer learning, model shift, credit risk, El Niño

Image de couverture: Réalisée par DALL·E 3.

UNIVERSITÉ DE LYON