# Statistical learning and causal inference for energy production

Naoufal Acharki

INSTITUT POLYTECHNIQUE DE PARIS

NNT : 2022IPPAX101

Thèse de doctorat

# Statistical learning and causal inference for energy production

Thèse de doctorat de l'Institut Polytechnique de Paris
préparée à l'École polytechnique

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

Thèse présentée et soutenue à Palaiseau, le 22 Novembre 2022, par

**NAOUFAL ACHARKI**

Composition du Jury :

| | |
|---|---|
| Rémi Flamary<br>Professeur, École polytechnique (CMAP) | Président |
| Marianne Clausel<br>Professeure, Université de Lorraine (IECL) | Rapporteure |
| Tim Sullivan<br>Associate Professor, University of Warwick | Rapporteur |
| Olivier Roustant<br>Professeur, INSA Toulouse (IMT) | Examinateur |
| Michèle Sebag<br>Directrice de recherche, CNRS (LISN) | Examinatrice |
| Josselin Garnier<br>Professeur, École polytechnique (CMAP) | Directeur de thèse |
| Antoine Bertoncello<br>Ingénieur de recherche, TotalEnergies OneTech (R&D Power) | Co-encadrant de thèse |

# Acknowledgements

Ce travail de recherche n'aurait pas pu être effectué sans l'aide, la patience et le soutien de nombreuses personnes. Bien que la moitié de cette thèse se soit déroulée sous format inhabituel à cause du COVID et des confinements, je suis très satisfait de cette expérience formidable et enrichissante.


Je remercie d'abord et avant tout mon directeur de thèse Josselin Garnier et je voudrais exprimer ma gratitude (que je résume dans la photo ci-dessus) pour m'avoir encadré tout au long de ces trois années de thèse. Il est était toujours disponible pour suivre mes travaux de près et veiller à leur rigueur en me faisant des retours et des corrections en permanence, au matin comme au soir, semaine après semaine. Il était toujours là pour m'aider à me former et m'améliorer avec ses connaissances et son expertise. J'ai appris à ne pas "faire de la cuisine" et être plus rigoureux dans mon esprit, mes démarches et preuves. Je fais partie maintenant de la nouvelle génération qui sait manipuler des réseaux de neurones et je sais en plus appliquer le théorème de convergence dominée proprement. Je n'oublierai pas nos discussions et nos débats sur les démonstrations et les maths et ceci, même si j'avais un taux de réussite de 10% à 15% (et ce qui est déjà pas mal). Je lui suis également reconnaissant de m'avoir aidé à traverser des

moments extrêmement difficiles au cours des confinements, de l'analyse et de la rédaction de la thèse. Je le remercie sincèrement pour sa confiance en moi.

Je remercie mon co-encadrant de thèse, Antoine Bertoncello, qui est à l'initiative de cette thèse (avec Eric Chaput) et m'a initié à la recherche appliquée à l'issue de mon stage avec lui en été 2018 et après les deux conférences MATHIAS 2018 et MATHIAS 2019 et la discussion avec Georges Oppenheim et Mark Asch. Son parcours m'a été également une source de motivation et d'inspiration. Il a manifesté un intérêt pour valoriser mon travail et communiquer à son sujet auprès des ingénieurs et des chercheurs en interne au sein des autres équipes. Il m'a permis aussi de découvrir une grande variété de problèmes statistiques issus de la R&D au sein de TotalEnergies.

Je remercie mes rapporteurs de thèse, Marianne Clausel et Tim Sullivan, d'avoir pris le temps de lire attentivement mon manuscrit et pour leur rapports très détaillés. Je remercie Marianne d'avoir apprécié la qualité de mon travail et d'avoir échangé avec elle pour l'améliorer.*I also thank Tim Sullivan for his detailed report and his suggestion for improvment of the manuscript. Working on his comments allowed me to make this work more clearer.* Je remercie aussi Rémy Flamary, et Michèle Sebag et Olivier Roustant qui m'ont fait l'honneur d'être membres de mon jury de thèse pour conclure ces trois années. Je remercie particulièrement Olivier Roustant, que j'ai eu le plaisir d'avoir en cours pendant mes deux premières années aux Mines de Saint-Etienne et lors du Challenge Data Science (avec Michel Lutz), de m'avoir transmis la passion pour les statistiques et le Machine Learning, ainsi pour ses conseils et ses recommandations pour aller plus loin dans ce domaine.

Je remercie les membres du GdR MASCOT-NUM et du laboratoire commun SINCLAIR avec qui j'ai pu échanger et qui m'ont fait profiter de leur expérience, sur les plans scientifique ou professionnel. Je pense ici en particulier à Bertrand Iooss, Vincent Chabridon, Francois Bachoc, David Ginsbourger et Rodolphe Leriche. Je remercie également les membres de CAUSAL TAU: Alessandro Leite, Georges Oppenheim et Audrey Poinsot pour leur remarques et suggestions d'amélioration de mon travail.

Je remercie tous mes collègues et ex-collègues de la plateforme numérique de TotalEnergies OneTech R&D, pour l'ambiance de travail (All-Hands, MATHIAS, Discovery sessions etc.) malgré les confinements et les restrictions liéees au Covid-19, en passant des bureaux de Nano-Innov aux bureaux du PlayGround. Je tiens à remercier particulièrement Cécile, Sébastien G, Marko, Sanjay, Romain, Antoine M, Philippe C et Sébastien G qui m'ont souvent donné de judicieux conseils en me faisant profiter de leur expérience professionnelle et personnelle. Je pense également aux doctorants de l'équipe: Amin, Benjamin, Cheikhna, Baptiste, Ali, Yagnik, Wassil, Elie ainsi que les stagiaires et postdocs comme Mohammed et Houssein avec qui j'ai passé des moments d'échanges et de discussions intéressants et amusants, et à qui je souhaite la meilleure réussite pour la suite de leurs thèses.

Je remercie tous les thésards du CMAP avec qui j'ai pu échanger souvent lors des déjeuners, des séminaires et Simpas Group Meeting: Les nombreux doctorants de Josselin (on devrait former l'équipe FC Josselin), en particulier Baptiste K, Corentin et Guillaume. Je remercie également Arthur, Louis, Leila, Achille, Benjamin et enfin Constantin, Clément M, Pablo, Baptiste G et Vincent pour ces moments aux workshops SYMPA au CIRM ou à Font Romeu.

Je remercie tous amis pour leur soutien pendant ma thèse, en commençant par mes amis d'enfance: Omar Ben, Tarik, Mohammed Diae, Salah, Simou, Nawfal et Nizar dont la majorité n'ont pas eu l'occasion d'assister à mes travaux et ma soutenance mais que je n'oublierai pas les souvenirs qu'on a vécu ensemble. Je remercie aussi mes meilleurs amis: Amine Barnicha, Omar Bouidra, Amine El Rhatrif, Oussama M, Hamza O, Taha Yassine, Amine Zaza, Amine Ben Yusef, Taha El Ouahabi, Anass E, Ahmed et Ayoub Saadi, Abdessamad L, Youssef H, Omar El Idrissi, Mehdi T, Ayoub B et Amine Boudlal pour leur fidélité, pour les rires, les séances de sport, les discussions et débats, les memes/reels qu'on partageait et le divertissement pour me faire oublier les maths. Je n'oublierai pas le soutien constant de plusieurs personnes pendant les moments de la rédaction de la thèse dont Houda et Aymen. Je suis aussi reconnaissant à mon parrain Reda A et mon grand-parrain de promo Alae H pour leur conseils. Enfin, je remercie mes amis doctorants et ceux qui font de la recherche: Aymen, Ali, Youssef, Rachida, Abdelhakim et Outmane avec qui je discutais souvent des mathématiques et de la science et on s'amusait à faire des blagues sur la vie des doctorants (High impact PhD memes).

وفي النهاية بغيت نشكر خوتي و بالاخص واليديا على الصبر و الدعم و الدعاوي ديالهم هاد ثلث سنين كاملة، طبعا كلمة الشكر قليلة و كنبقا مقصر في حقكم في هاد ثلث السنين خصوصا مع كورونا و لكن كنواعدكم، الا طول لكم ربي في العمر و زادكم في طحكتم، انبقا ديما مكبر بكم و نفرحكم كيما كنتوا كتفرحوني و انا صغير و كنتمنى من هاد الخدمة و هاد ٢٥٠ صفحة تكون مصدر الهام و فخر لكم.

# Résumé

Grace à la croissance importante des données générées par le secteur, les entreprises s'appuient davantage sur l'intelligence artificielle pour développer leur activité. En effet, l'application des modèles d'apprentissage automatique à ces données leur permet de gérer la demande d'énergie, la consommation et anticiper les défaillances de manière efficace en termes de temps et du coût. L'apprentissage automatique présente un outil puissant pour découvrir de nouvelles sources d'énergies durables et optimiser l'utilisation des énergies traditionnelles.

Ces dernières années, l'apprentissage automatique a conduit à de nombreuses applications et avancées réussies dans le domaine de l'énergie. Cependant, et malgré leur précision, plusieurs difficultés apparaissent avec les modèles utilisés: leur prédictions sont parfois insatisfaisantes et manquent d'interprétabilité. En effet, la plupart des modèles d'apprentissage automatiques sont considérés comme des boîtes noires. Nous n'avons pas d'idée de (i) l'incertitude de la prédiction ni (ii) de l'impact réel des changements de variables et d'interventions à travers ces boîtes noires. Il en résulte la sur/sous-estimation de l'incertitude du modèle, ou des prédictions trompeuses qui contredisent les connaissances des ingénieurs et des experts. Ce problème est assez critique dans les systèmes énergétiques où la gestion des risques et l'interprétabilité des prédictions sont primordiales pour des raisons économiques, environnementales et opérationnelles.

Dans la première partie de cette thèse, nous considérons le problème de la quantification des incertitudes. Le modèle de processus gaussiens est connu comme l'une des méthodes d'apprentissage automatique bayésien les plus performantes pour quantifier les incertitude. Les méthodes d'estimation par maximum de vraisemblance ou de validation croisée sont fréquemment utilisées pour identifier ses paramètres. Néanmoins, elles peuvent échouer et ne pas estimer correctement les intervalles de prédiction si certaines hypothèses sur le modèle ne sont pas vérifiées, typiquement la bonne spécification du modèle.

Concernant le problème des modèles de processus gaussiens mal-spécifiés, une approche robuste en deux étapes est développée pour ajuster et calibrer les intervalles de prédiction du modèle. La méthode permet d'obtenir des intervalles de prédiction de petites largeurs avec des probabilités de couverture appropriées. Elle se base sur la validation croisée comme métrique pour ajuster les hyperparamètres de la covariance et assurer que la probabilité de couverture du modèle final atteigne le niveau nominal.

Dans la deuxième partie, nous considérons le problème de l'inférence causale et l'estimation des effets d'interventions. Le modèle causal de Neyman-Rubin est largement utilisé par les statisticiens pour faire estimer les effets d'un traitement. Cependant, la plupart des considérations de ce modèle se limitent à un traitement binaire. Or, dans de nombreuses

applications, la variable d'intérêt peut être discrète ou même continue. En outre, les effets du traitement varient selon les caractéristiques des unités. L'hétérogénéité du traitement doit être explorée pour personnaliser mieux la politique d'intervention et optimiser les résultats.

Pour résoudre le problème de l'estimation des effets hétérogènes du traitement, un cadre bien connu d'estimateurs statistiques, appelé méta-apprenants, est étendu aux traitements multiples et continus. La discussion sur la consistance des méta-apprenants et l'analyse de leur biais et variance donne un aperçu des avantages et des inconvénients de chaque méta-apprenant. Enfin, quelques recommandations et limites ont été mises en évidence quant à l'utilisation des méta-apprenants pour les traitements continus.

Le travail effectué dans cette thèse est générique. Les applications réelles comprennent, sans s'y limiter, les puits de gaz conventionnels, les batteries et les systèmes géothermiques améliorés.

# Abstract

With the significant growth of the data generated by the sector, energy companies are relying more on Artificial Intelligence for their business and development. Indeed, applying Machine Learning algorithms to this data can help them to predict energy demand and consumption and anticipate its failures efficiently, with less time and at low cost. Machine Learning presents a powerful tool to search for new sustainable energy sources and optimize the use of current traditional sources.

In recent years, Machine Learning has seen many successful applications and advances in the energy field. However, several difficulties arise despite its accuracy: Machine Learning models' predictions are sometimes unreliable and lack interpretability. Indeed, most Machine Learning models are black boxes. We have no idea of (i) the uncertainty of the prediction nor (ii) the real impact of changes in variables and interventions through these black boxes. This may produce an over/underestimation of the model uncertainty or misleading predictions that contradict engineers' and experts' knowledge. This problem is quite critical in energy systems where risk management and interpretability of predictions are vital for economic, environmental and operational reasons.

In the first part of the thesis, we consider the problem of Uncertainty Quantification. The Gaussian Process model is known to be one of the most powerful Bayesian Machine Learning methods for quantifying the uncertainty of predictions. Maximum Likelihood estimation or Cross-Validation methods are widely used to fit parameters. Nevertheless, they may fail to fit the optimal model that estimates Prediction Intervals correctly if some assumptions do not hold, typically the well-specification of the Gaussian Process model.

Concerning the problem of Gaussian process misspecified models, a robust two-step approach is developed to adjust and calibrate Prediction Intervals for Gaussian Processes Regression. The method gives prediction Intervals with appropriate coverage probabilities and small widths. It uses the Cross-Validation and the Leave-One-Out Coverage Probability as a metric to fit covariance hyperparameters and assess the Coverage Probability to a nominal level.

In the second part, we consider the problem of Causal Inference of interventions. The Neyman-Rubin Causal model is widely used by statisticians to make Causal Inference and estimate the effects of a treatment on the outcome. Unfortunately, most considerations of this model are limited to the setting of a binary treatment. In many real-world applications, the variable of interest can be multi-valued or even continuous. Furthermore, treatment effects vary across units with different characteristics. The heterogeneity should be explored to personalize the intervention policy and optimize the outcome.

A well-known framework of statistical estimators, called meta-learners, is extended to multiple and continuous treatments to solve the problem of heterogeneous treatment effects. The discussion about the consistency of meta-learners and the analysis of their bias and variance gives an overview of the advantages and disadvantages of each meta-learner. Finally, some recommendations and limits are highlighted about the use of meta-learners for continuous treatments.

The proposed methods and contributions of the thesis are generic and can be applied to any industrial problem. The actual applications include, but are not limited to, unconventional gas wells, batteries and enhanced geothermal systems.

# Contents

# List of Figures

# List of Tables

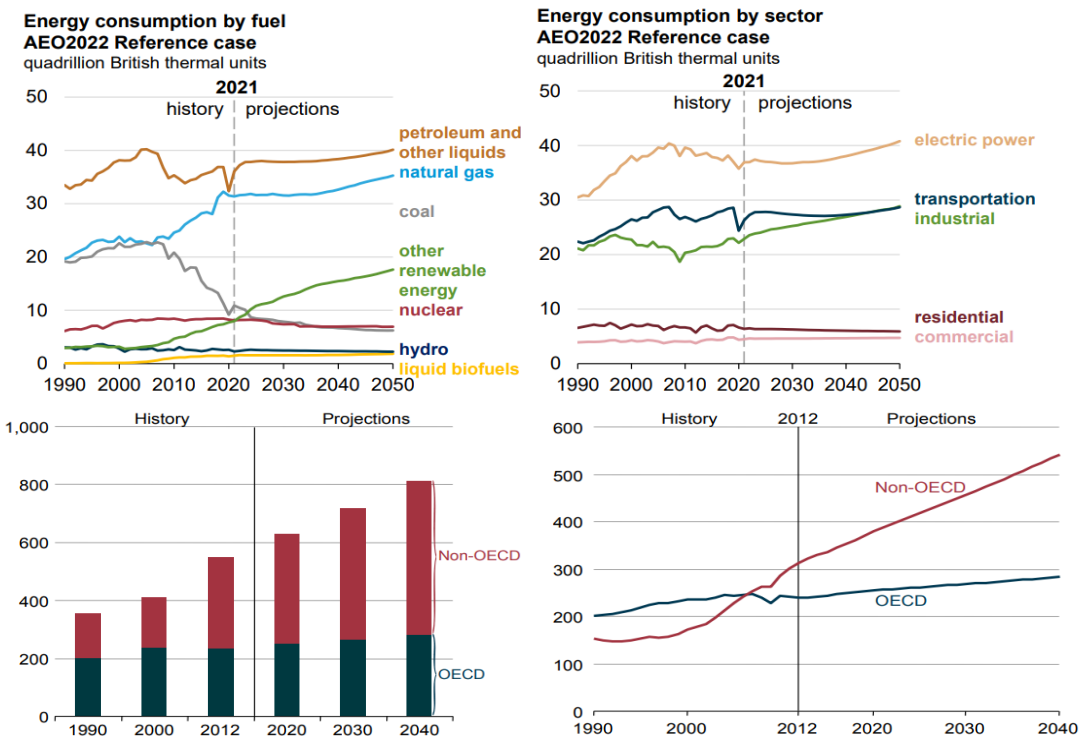
# CHAPTER 1

# Introduction

## 1.1 Context

### 1. Energy challenges and sustainability

In recent decades, the development of human society has shown a vital need for energy. Indeed, energy is considered as the lifeblood of any economic development and one of the main pillars for increasing the wealth and growth of any nation (Arto et al., 2016; Cottrell, 2009). Moreover, its importance has become evident in the development process due to its close connection with various fields, particularly in the industrial, transport and residential sectors (Ministère de la transition écologique., 2021).




(a) World energy consumption in quadrillion Btu (b) World energy consumption by region

Figure 1.1: From World Energy Outlook 2021 International Energy Agency (2016, 2021)

However, the energy sector faces an increasingly critical set of economic, geopolitical, technological and environmental challenges. The world's population is still growing, and, thus, the energy needs of billions of people in rural and urban areas, particularly in emerging and non-OECD countries, must be met (International Energy Agency, 2016, 2017, 2021).

Among the essential and main issues related to energy and power plants are their ability to respond to supply and demand, having optimal performance, and minimal environmental impact (Bruckner et al., 2014). Nevertheless, up to nowadays, the global energy mix is still provided by fossil energy sources and hydrocarbons, including oil and natural gas (British Petroleum, 2020). These resources, by nature, have high costs and high potential environmental risks, are expected to decline in the not-too-distant future and, therefore, are unable to cope with the continuous rise in demand for energy (International Energy Agency, 2022). Furthermore, the sustained and excessive consumption of fossil resources has threatened global energy security and caused severe environmental issues and negative impacts on ecosystems and society, such as greenhouse CO2 gas emissions and global warming problems (Intergovernmental Panel on Climate Change, 2015; Jarvis et al., 2012).

Consequently, facing these challenges became no longer an option but an emergency. The current situation calls for the importance of Energy Transition and the necessity of renewing the existing energy production and consumption patterns (OCDE, 1999). The benefits seem to be valuable (UN General Assembly, 2015): environmental balance, sustainable growth and maintaining a strategic reserve of natural resources for the coming generations.

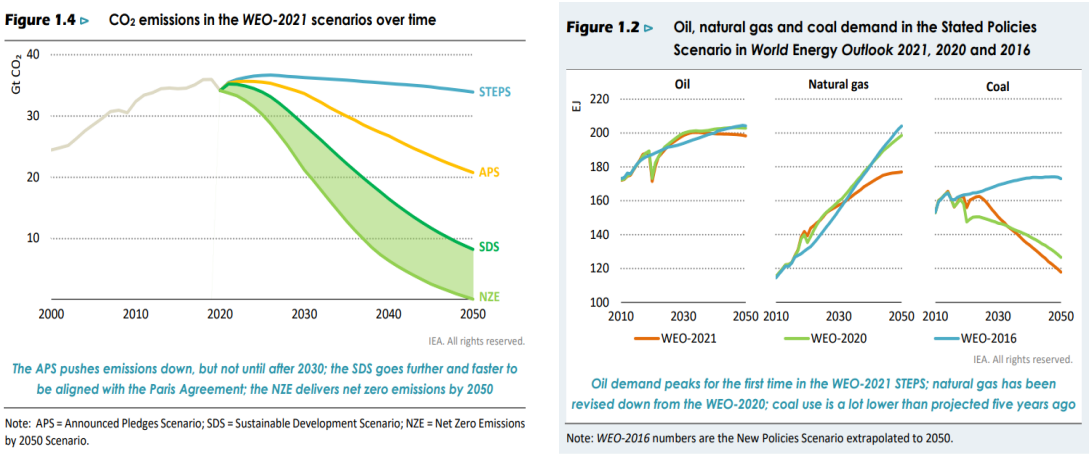



Figure 1.2: From World Energy Outlook 2021 International Energy Agency (2021)

In addition, a series of local and international conferences (COP'21, COP'22, COP'26 etc.) and agreements (Paris Agreement 2015) have taken place to set effective steps to face climate change and search for alternative solutions to ease the pressure on the environment. There was a global consensus and ambition of the international community to ensure access to affordable, reliable, sustainable and modern energy (United Nations., 2021). Thereby, achieving environmental and social sustainability requires supporting renewable energy, reducing energy demand and reducing the dependence on fossil energy sources.

Based on these recommendations, governments and private organizations, including energy companies, are expected in their development and action strategies to make a clear and responsible commitment to preserving the climate for future generations. The action plans

should consider: fewer CO2 emissions, providing sustainable energy and taking care of the environment.

In the same context, TotalEnergies revealed its strategy of becoming a multi-energy company that provides reliable, sustainable and affordable energy (TotalEnergies, 2020). The company's ambition is to place sustainable development at the heart of its strategies, projects and operations and, by doing so, become a major player in Energy Transition and carbon neutrality by 2050 (TotalEnergies, 2022). To this end, TotalEnergies sets an ambitious target in 2050: Produce 50% renewable electricity (solar, wind), 25% new low-carbon molecules from biomass (biofuels, biogas) or renewable electricity (hydrogen, e-fuels) and 25% hydrocarbons (oil and gas).

Currently, the company is positioning itself for future energy supplies and diversifying its energy mix by reducing the share of oil products, increasing natural gas and renewable electricity, and promoting transitional energy. For example, one short-term goal is to maintain and adapt existing hydrocarbon capabilities and invest in new low-cost and low-emission fields (TotalEnergies, 2021).

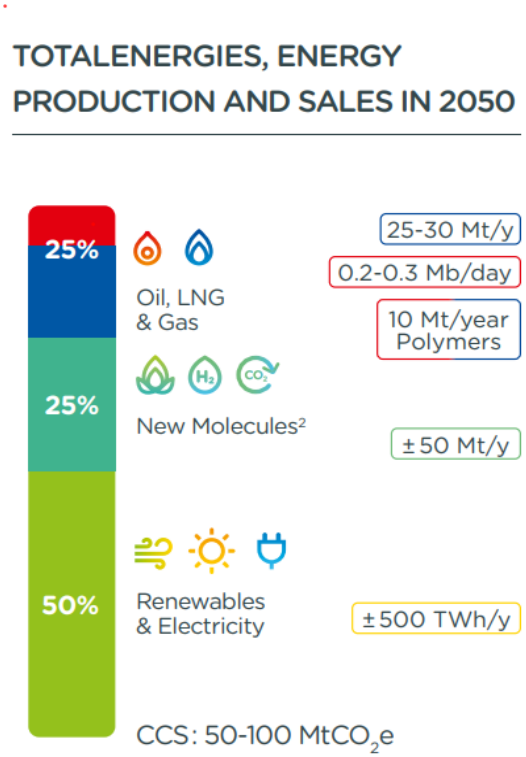



Figure 1.3: From TotalEnergies Sustainability & Climate 2022 Progress Report (TotalEnergies, 2022).

### 2. Artificial Intelligence (AI) in Energy

In his essay "Dear class of 2017", Gates (2017) wrote about *"things he wishes he'd known when he left college"*. He called graduate students, seeking advice on which path to take to maximize their impact in the world, that he *"would consider three fields. One is Artificial Intelligence, [...] it will make people's lives more productive and creative. The second is energy because making it clean, affordable, and reliable will be essential for fighting poverty and climate*

*change"*. Bill Gates may not be the first to prefer this path of Artificial Intelligence (AI), and this is not random. Many scientists and experts are conscious and agree that AI will be a revolutionary tool to kick in the energy sector, address its challenges and even go beyond climate and environmental issues.

Indeed, Artificial Intelligence and Machine Learning can support the energy industry by providing clean, cheap and sustainable energy needed for humanity's development (United Nations. (2021) 7th goal). The extraction, analysis, and evaluation of large volumes of data with statistics and computer science tools have made it much easier to get meaningful information that contributes to developing new solutions and assists in decision-making. Specifically, Artificial Intelligence and Machine Learning can leverage massive data and build models that can significantly impact energy's production and consumption by enhancing its performance, cutting energy waste, reducing operating costs and maximizing profitability. It can also help improve safety measures, maintain resources' sustainability and achieve better demand-response management (Makala & Bakovic, 2020).

Bughin et al. (2017) of McKinsey Global Institute, for instance, examine investment in AI and its return on investment. According to them, in electric utilities, AI and digitization increase energy productivity by up to 20%, reduce energy waste and CO2 emissions, and improve Earnings Before Interest, Taxes, Depreciation, and Amortization (EBITDA) by 10% to 20%. Another successful example is DeepMind (Gao, 2014; Gao & Evans, 2014). With AI and Machine Learning, they manage to enhance the Google Data center's efficiency and reduce energy consumption by 15%. It is, therefore, not a surprise that The World Economic Forum (2021) calls governments and companies to invest more in AI as it finds a *"tremendous potential"* in AI *"to accelerate a global reliable and lowest-cost energy transition"*. So also did Villani et al. (2018) in his book when he was in charge of the implementation of a French and European strategy in AI.

Besides, the outstanding performances of Machine Learning models in analyzing and predicting outcomes from complex and multi-dimensional data have made them very popular in many areas. This popularity has led to many studies and applications with significant and valuable impacts in the energy field. The applications include but are not limited to Oil and Gas industry (Alvarado et al., 2002; Cao et al., 2016; Mohaghegh et al., 2011), geothermal energy (Arslan & Yetik, 2011; Assouline et al., 2019), well performance analysis (Fulford et al., 2015; Nwachukwu et al., 2018), nuclear energy and power plants (Iooss & Le Gratiet, 2019; Santosh et al., 2007), solar power forecasting (Gensler et al., 2016; Li et al., 2016; Voyant et al., 2017), wind power forecasting (Foley et al., 2012; Heinermann & Kramer, 2016; Jursa & Rohrig, 2008), batteries lifetime capacity (Li et al., 2019; Ng et al., 2020; Severson et al., 2019), fault detection and prediction in Energy systems (Dhaou et al., 2021; Gupta et al., 2015; Zhao et al., 2019), energy load forecasting and demand (Ahmad & Chen, 2018; Bouktif et al., 2018; Raza & Khosravi, 2015), buildings thermal load (Idowu et al., 2016; Jovanović et al., 2015; Wang et al., 2020) and comfort prediction (Han et al., 2020; Yuce et al., 2014), and enhancing building's efficiency and control (Drgoňa et al., 2018; Yang et al., 2020).

## 1.2 Problem statement

While most people agree that Machine Learning has become a valuable tool for solving business problems, it is essential to mention that most of these methods focus solely on answering

predictive questions using regression or classification methods.

However, making predictions without quantifying their uncertainty is generally not trustworthy, especially for decision-making. For example, energy companies' investment strategy and production capacities cannot be planned solely based on the mean predictions or the average scenario. Therefore, considering uncertainties (weather, terrain, etc.) and risks (economic, environmental, etc.) is required and reliable forecasts are highly desirable.

In addition, many problems in the energy industry are not always about predicting outcomes based on the correlation between variables. One must consider causal effects and answer questions about what would be the effect on the production if just one variable involved in the process is changed. In other words, it is important to obtain valuable information from the data and move beyond prediction to causal inference to interpret and understand the results before using them in decision-making.

For the previous considerations, we raise the following challenges in the R&D:

- It is difficult in multivariate and small data contexts to make reliable predictions for decision-making.
- Standard statistical inference and Machine Learning models cannot distinguish between correlation and causation.

## 1.3 Objective of the thesis

This thesis does not aim to solve a specific industrial problem. It is more concerned with developing new *data-driven* approaches to answer generic problems on statistical learning and causal inference in potentially uncertain environment settings. The approaches will be used principally for optimization and decision-making purposes, particularly for energy production analysis and forecasting. The thesis is divided into two major parts:

- The predictive part: predicting outcome with the associated uncertainties.
- The causal part: conducting a causal study and inferring the effects of interventions.

## 1.4 Outline

The main focus of this thesis lies in the research field of statistics and Machine Learning, more specifically: statistical learning and causal inference for optimization and decision-making. The dissertation is organized into two parts with an introduction and a conclusion. Both Chapters 2 and 3 form the first part of the thesis, dedicated to statistical learning and Uncertainty Quantification. The second part gathers Chapters 4 to 6 and tackles Causal Inference and the estimation of intervention effects. To be precise, the thesis is structured as follows:

**Chapter 1** : In this chapter, we introduce the main context and the industrial motivation of the thesis.

**Chapter 2** : In this chapter, we review some existing approaches for Uncertainty Quantification. We present the Gaussian process model, its properties for prediction with uncertainty and common methods to learn the Gaussian Process model. Unfortunately, these methods do not always make correct predictions. This situation typically happens for a misspecified Gaussian Process model.

**Chapter 3** : In this chapter, we propose a method to overcome the problem of quantifying the uncertainty with a misspecified Gaussian Process model. Our approach uses Cross-Validation and the Gaussian Process model to calibrate Prediction Intervals. By adjusting the upper and lower bounds, the method gives the appropriate uncertainty, that is, Prediction Intervals respecting the targeted confidence level and having small widths.

**Chapter 4** : In this chapter, we present the state-of-the-art of Causality and Causal Inference. We review, in particular, the potential outcomes theory and the Rubin Causal model as one of the most popular models for evaluating the impact of interventions (usually called treatment effects) on a given outcome.

**Chapter 5** : In this chapter, we study the problem of estimating heterogeneous treatment effects: the effect of interventions across sub-groups of units. We develop statistical frameworks, called meta-learners, for evaluating heterogeneous effects under multi-valued treatments. We provide some meta-learners' error bounds and highlight their performances. We also describe a semi-synthetic dataset that serves to validate Causal Inference methods and present our results on it.

**Chapter 6** : In this chapter, we extend the estimation of heterogeneous effects to a continuous treatment (intervention variable). Based on a detailed theoretical analysis, we discuss the generalization of the so-called meta-learners. We underline the limits they may have and make some recommendations on their use.

**Chapter 7** : In this chapter, we present our conclusion of this thesis and its perspectives.

## 1.5 Contributions

The contributions mentioned above are included in the following published or to be submitted peer-reviewed papers:

- Acharki, N., Bertoncello A., and Garnier J. Robust prediction interval estimation for Gaussian processes by cross-validation method. *Computational Statistics & Data Analysis*, 178:107597, 2023. ISSN 0167-9473. DOI: 10.1016/j.csda.2022.107597.
- Acharki, N., Garnier J., Bertoncello, A., and Lugo, R. Heterogeneous treatment effects estimation: When machine learning meets multiple treatment regime. *arXiv preprint arXiv:2205.14714*, 2022. *Submitted.*
- Acharki, N., Bertoncello, A., and Garnier, J. Pseudo-outcome representations for heterogeneous effects inference: challenges and limits under continuous treatment. *Ongoing work, to be submitted.*

The author of this thesis is responsible for reviewing the state-of-the-art, defining the basic concept, the mathematical research, writing the manuscripts and carrying out the numerical experiments. Both supervisors of this thesis are responsible for supervising, providing critical feedback, and verifying and validating proofs and results. Ramiro Lugo was responsible for simulating and generating the semi-synthetic dataset described in Chapter 5.

# PART I

# Statistical Learning and Uncertainty Quantification

*Uncertainty is NOT "I don't know". It is "I can't know". "I am uncertain" does not mean "I could be certain".*

— Werner Heisenberg

# CHAPTER 2

# Tools and State of the art

In this chapter, we begin by presenting a quick introduction to Uncertainty Quantification, and we review some existing methods used for regression including Bayesian approaches. In Section 2.2, we introduce the Gaussian process model, and we show how its properties are useful for prediction with uncertainty. In Section 2.3, we present a set of methods used to estimate the Gaussian Process model's parameters. We conclude by presenting the ongoing research topics related to Gaussian Process in Section 2.4.

## 2.1 Introduction to Uncertainty Quantification

### The regression problem

We consider $n$ observations of the output of a model (e.g. an empirical model, computer code etc.). Each observation of the output corresponds to a $d$-dimensional input vector $\boldsymbol{x} = (x_1, \dots, x_d)$ in a domain $\mathcal{D} \subseteq \mathbb{R}^d$. The $n$ points corresponding to measurement points (i.e. the model/code runs) are called an experimental design $\mathbf{X} = (\boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(n)})$ where $\boldsymbol{x}^{(i)} = (x_1^{(i)}, \dots, x_d^{(i)})^\top \in \mathcal{D}$. The outputs are denoted by $\boldsymbol{y} = (y_1, \dots, y_n) \in \mathbb{R}^n$. It is common in regression setting to assume that some data generating function $f$ and an additive noise $\epsilon$ exist. The combination of these two quantities produce the observed outcome $\boldsymbol{y}$ for the input design $\mathbf{X}$. For $i = 1, \dots n$, we write:

$$y_i = f(\boldsymbol{x}^{(i)}) + \epsilon_i. \tag{2.1}$$

Following from above, $y$ might represent a solar panel delivered energy, $\boldsymbol{x}$ is information about its location, design characteristics, surface pressure and other factors, and $f$ could be, for example, the physical model behind that generates energy given these parameters. Generally, it is not possible to observe the *exact* value $f(\boldsymbol{x})$. This is mainly due to the presence of the noise $\epsilon$. In many cases, this noise $\epsilon$ may be explained by the exclusion of some explanatory variables (e.g. unmeasured weather conditions) or by the presence of inherently stochastic effects (e.g. measurement errors).

Given a set of data points $\mathbf{D} = (\mathbf{X}, \boldsymbol{y}) = (\boldsymbol{x}^{(1)}, y_1), \dots, (\boldsymbol{x}^{(n)}, y_n)$ , one would like to know what would be the associated outcome $f(\boldsymbol{x}_{\text{new}})$ for a new point $\boldsymbol{x}_{\text{new}}$?

**Remark 2.1.1.** *In Machine Learning framework, it is usual to assume that* $(\boldsymbol{x}^{(i)}, y_i)_{i=1}^n$ *are independent and identically distributed (i.i.d.) for estimation's consistency and model assessment.*

This is a classical task that many statisticians, engineers and specialists realize, known as *regression problem*: Estimate the unknown function $\boldsymbol{x} \in \mathcal{D} \mapsto f(\boldsymbol{x})$ in (2.1) given a data $\mathbf{D}$ and make accurate predictions with the associated uncertainty.

Regression problems are at the core of Machine Learning to build a model allowing the prediction of the output. More formally, we assume that both the inputs and the output are random variables. We denote $\boldsymbol{X} \in \mathcal{D}$ and $Y \in \mathbb{R}$ to indicate the stochastic character. The noise $\epsilon$ associated with the output $Y$ is also random and has the same distribution as $\epsilon_1$.

In this setting, we write

$$Y = f(\boldsymbol{X}) + \epsilon, \tag{2.2}$$

with

$$\mathbb{E}(Y \mid \boldsymbol{X} = \boldsymbol{x}_{\text{new}}) = f(\boldsymbol{x}_{\text{new}}). \tag{2.3}$$

The goal of regression is to produce a point estimate of $f(\boldsymbol{x}_{\text{new}})$, corresponding to the mean prediction value of $Y$ given $\boldsymbol{X} = \boldsymbol{x}_{\text{new}}$. The point estimates are adequate to evaluate the accuracy of predictions but, unfortunately, give no guidance about their reliability or the range of uncertainty.

### From regression to uncertainty quantification

Decision-makers are increasingly relying on Machine Learning models as a result of the successful applications to real-world prediction problems (van Asselt & Rotmans, 1996). However, the growing importance of Machine Learning necessitates the ability to reduce and quantify the uncertainty in model predictions. The point estimates are not sufficient and do not provide necessary information for risk management. Hence, combining the predictive performance of such complex models with practical guarantees of the reliability of their results becomes critical.

In recent years, the concept of uncertainty has received increased attention in Machine Learning research (Sullivan, 2015). Any Machine Learning method should consider a trustworthy representation of uncertainty as a key feature. Indeed, we need to know how certain we are about this prediction. This is especially important in high-stakes applications where machine learning outputs will be used to inform critical decision-making, such as medicine (Begoli et al., 2019; Wiens et al., 2019), safety (Varshney, 2016) and civil and nuclear engineering (Briggs & Division, 2009; Podofillini et al., 2015).

Uncertainty Quantification (UQ) is the end-to-end study of the reliability of scientific inferences (Washington et al., 2008). In the modelling context, Uncertainty Quantification covers the different dimensions of uncertainty. It is concerned with estimating the impact of uncertain input data on the model parameter and prediction.

Uncertainty Quantification problems are typically comprised of a mathematical model representing the system under consideration, which is subject to uncertainty due to uncertain input values and model parameters. Uncertainty Quantification also entails determining how these uncertainties propagate throughout the model. The propagation of uncertainty across the model can be addressed through forwarding or backward modelling. These uncertainties are then quantified using a probabilistic framework (Ghahramani, 2015).

To understand the concept of uncertainty in observed outputs or phenomena, one should identify the various sources of uncertainty. Generally, there exist two major sources of uncertainty associated to an observed outcome (Kendall & Gal, 2017; Morgan & Henrion, 1990).

**Definition 2.1.2** (Epistemic uncertainty (Hüllermeier & Waegeman, 2021))**.** *The Epistemic uncertainty refers to uncertainty caused by a lack of knowledge. This uncertainty can in principle be reduced on the basis of additional information, e.g. more observations and insights about the physical phenomenon.*

**Definition 2.1.3** (Aleatoric uncertainty (Hüllermeier & Waegeman, 2021))**.** *The Aleatoric (stochastic) uncertainty refers to the notion of randomness, that is, the natural variability in the outcome of an experiment (done under the same conditions) which is due to inherently random effects.*

Our aim is to capture the uncertainty of the response $Y$ in (2.1), which can be quantified by its variance for example.

In a particular case, the epistemic uncertainty is captured in the $f(\boldsymbol{X})$ component, while the aleatoric uncertainty is considered in the $\epsilon$ term. Indeed, given that both terms of (2.1) have associated sources of uncertainty, and assuming they are independent, the uncertainty of the observations $\sigma_y^2$ can be decomposed into aleatoric $\sigma_{noise}^2$ and epistemic $\sigma_{model}^2$ uncertainties as:

$$\text{Total Uncertainty} = \text{Epistemic} + \text{Aleatoric} \tag{2.4}$$

$$\sigma_y^2 = \sigma_{model}^2 + \sigma_{noise}^2, \tag{2.5}$$

where $\sigma_y^2 = \text{Var}(Y)$, $\sigma_{model}^2 = \text{Var}\left[f(\boldsymbol{X})\right]$ and $\sigma_{noise}^2 = \text{Var}(\epsilon)$.

**Remark 2.1.4.** *While most of the work on Uncertainty quantification focuses on the epistemic uncertainty of the model, the aleatoric uncertainty can also be estimated as part of the model by learning the errors $\epsilon$.*

## Description of prediction intervals

The goal of Uncertainty Quantification is to enhance the model's reliability by producing the output in a probabilistic framework. One common way to quantify the uncertainty is to use the notion of Prediction Intervals (PI).

Let $\boldsymbol{X} \in \mathcal{D}$ be a $d$-dimensional random vector and $Y \in \mathbb{R}$ be a random variable whose distributions are denoted $\pi_{\boldsymbol{X}}$ and $\pi_Y$. Let $\mathbf{D}$ be the random data set with distribution $\pi_{\mathbf{D}}\big(\{(\boldsymbol{x}^{(i)}, y_i)\}_{i=1}^n\big) = \pi_{\mathbf{D}_{\boldsymbol{X}}}(\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(n)}) \prod_i \pi_{Y|\boldsymbol{X}}(y_i \mid \boldsymbol{x}^{(i)})$, where $\pi_{\mathbf{D}_{\boldsymbol{X}}}$ is the joint distribution of $\boldsymbol{X}$ on $\mathbf{D}$ and $\pi_{Y|\boldsymbol{X}}$ is the conditional distribution of $Y$ given $\boldsymbol{X}$. Let $(\boldsymbol{x}_{\text{new}}, Y_{\text{new}})$ be the random vector of interest independent of $\mathbf{D}$ with distribution $\pi_{\boldsymbol{X},Y}(\boldsymbol{x}, y) = \pi_{\boldsymbol{X}}(\boldsymbol{x})\pi_{Y|\boldsymbol{X}}(y \mid \boldsymbol{x})$. It is possible to assume that $\pi_{\mathbf{D}_{\boldsymbol{X}}}(\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(n)}) = \prod_i \pi_{\boldsymbol{X}}(\boldsymbol{x}^{(i)})$, which means that the observations $(\boldsymbol{x}^{(i)}, y_i)$ are *i.i.d.* drawn from the distribution $\pi_{\boldsymbol{X},Y}$, but this is not necessary in this setting. Finally, let $(1-\alpha)$ with $0 < \alpha < 1$ define a level of confidence. The following definition is inspired from (Beran, 1992) and Chen et al. (2021a).

**Definition 2.1.5** (Prediction Interval)**.** *A Prediction Interval $\mathcal{PI}_{1-\alpha} \subseteq \mathbb{R}$ is an interval-valued function $\mathcal{PI}_{1-\alpha}(\boldsymbol{x}_{\text{new}}; \mathbf{D}) = \mathcal{PI}_{1-\alpha}(\boldsymbol{x}_{\text{new}})$ depending on $\mathbf{D}$ where you expect, with a confidence of $(1-\alpha)\times 100\%$, a new observation of the outcome $Y_{\text{new}}$ to fall. In other terms,*

$$\mathbb{P}\left(Y \in \mathcal{PI}_{1-\alpha}(\boldsymbol{X})\right) = 1 - \alpha, \tag{2.6}$$

*where $\mathbb{P}$ can be taken with respect to the distributions $\mathbf{D}$ and $(\boldsymbol{X}, Y)$ but conditional versions can also be studied (see below). The Prediction Interval is given by*

$$\mathcal{PI}_{1-\alpha}(\boldsymbol{x}_{\text{new}}) = \left[l_{1-\alpha}(\boldsymbol{x}_{\text{new}}), u_{1-\alpha}(\boldsymbol{x}_{\text{new}})\right], \tag{2.7}$$

*where* $l_{1-\alpha}, u_{1-\alpha} : \mathcal{D} \mapsto \mathbb{R}$ *such that* $l_{1-\alpha} \leq u_{1-\alpha}$ *are two mappings trained on the dataset* $\mathbf{D}$ *and define the upper and lower bounds of the Prediction Interval* $\mathcal{PI}_{1-\alpha}$.

Formally, a prediction interval is a range of values that is likely to contain the value of the new observation $Y_{\text{new}}$, given the training set $\mathbf{D}$ with inputs $\mathbf{X}$, the output $\boldsymbol{y}$ and a given degree of confidence. The bounds $u_{1-\alpha}(\boldsymbol{x}_{\text{new}})$ and $l_{1-\alpha}(\boldsymbol{x}_{\text{new}})$ represent the range of uncertainty, and therefore, the reliability of the estimation of the outcome $Y_{\text{new}}$.

The notion of Prediction Intervals is not recent, it dates back many decades. Based on Fisher (1925) methods for statistical inference, Baker (1935) considers predicting a future sample mean and how it is being expected to differ from the set of observations available. However, the term *prediction interval* seems to have been introduced a little later. Using a frequentist approach, Proschan (1953) derives the same interval as Fisher and writes: *such an interval might more properly be called a prediction interval, since the term 'confidence interval' usually refers to population parameters, which are not random.* Thatcher (1964) studies the prediction of the binomial distribution but refers to the prediction interval as the *"confidence limit for the prediction"*. Nelson (1968) provides an overview of general theory and methods for computing prediction intervals. In a detailed review of literature, Patel (1989) states that, in the late 1960s, many articles in engineering and applied statistics journals presented methods for some specific prediction problems and used the term *prediction interval*.

Unfortunately, as described before, there is often some confusion about the difference between a confidence and a prediction interval, leading to a misinterpretation of predictions. A confidence interval is an interval that does contain, with a given degree of confidence, a deterministic parameter of interest. For example, if the parameter of interest is the mean of a population $\mu = \mathbb{E}(Y)$, the confidence interval tells you where the population mean $\mu$ is, with a given confidence level.

**Remark 2.1.6.** *In Prediction Intervals, the quantity of interest we are looking for is a random variable, the outcome* $Y_{\text{new}}$ *for instance. In confidence intervals, it is a deterministic parameter.*

Prediction Intervals are wider than confidence intervals, since the prediction interval must also include total uncertainty in the output, while the confidence interval does only include the epistemic uncertainty and excludes the noise. This is why using Prediction Intervals is a meaningful way of providing information about the uncertainty of predictions. They capture the contributions from both types of uncertainty on the response.

### Uncertainty Quantification with Prediction Intervals

Assessing the quality of Prediction Intervals is not very common in regression, unlike point-wise prediction metrics. The state-of-the-art offers limited options that are not examined further in the theoretical aspects (Pearce et al., 2018) as point-wise metrics. In particular, Prediction Interval's performances can be measured by two main quantities: their width and coverage probability (Khosravi et al., 2010; Pearce et al., 2018; Shrestha & Solomatine, 2006).

Zhang et al. (2020) introduce a taxonomy with four coverage types of Prediction Intervals. The coverage can be marginal (Type I, as defined in 2.1.5) or conditional (on the training set Type II, on the new point Type III, or both Type IV). While most studies in the literature on the construction of Prediction Intervals fall in the type I coverage, only a few authors dealt with conditional coverage (Vovk (2012) for type II and Foygel Barber et al. (2020) for type

IV). In fact, the conditional coverage (especially type IV) is a much stricter definition than the marginal coverage and is much more challenging to satisfy, especially in distribution-free settings (Xu & Xie, 2021). We refer to Barber et al. (2021) for more details between marginal and conditional coverage.

In the following, we consider the type II conditional coverage of Prediction Intervals.

**Definition 2.1.7** (The type II Coverage Probability (CP))**.** *The conditional Coverage Probability given the training set* $\mathbf{D}$*, also known as type II Coverage Probability, is the probability*

$$\mathbb{P}_{\pi}\left(Y \in \mathcal{PI}_{1-\alpha}(\boldsymbol{X}) \mid \mathbf{D}\right), \tag{2.8}$$

*where* $\mathbb{P}_{\pi}$ *denotes the probability with respect to the joint distribution* $\pi$ *of* $(\boldsymbol{X}, Y)$*.*

The goal is to construct $\mathcal{PI}_{1-\alpha}$ so that the conditional coverage probability becomes as close as possible to $1-\alpha$ (and converges to $1-\alpha$ in probability when $n$ increases). Earlier, Cox (1975) studied this estimator and developed an algebraic adjustment to reduce its bias, Guttman (1970) used the coverage probability notion to identify the tolerance region in regression. Now, the coverage Probability is gaining more popularity in regression problems and Machine Learning whenever the uncertainty of prediction is raised.

**Remark 2.1.8.** *If the conditional coverage is achieved by some method, then the marginal (type I) coverage probability is also achieved. Indeed.*

$$\begin{aligned}
\mathbb{P}_{\mathbf{D}\times(\boldsymbol{X},Y)}\left(Y \in \mathcal{PI}_{1-\alpha}(\boldsymbol{X})\right) &= \mathbb{E}_{\mathbf{D}\times(\boldsymbol{X},Y)}\left[\mathbf{1}\{Y \in \mathcal{PI}_{1-\alpha}(\boldsymbol{X})\}\right] \\
&= \mathbb{E}_{\mathbf{D}\times(\boldsymbol{X},Y)}\left[\mathbb{E}_{\pi}[\mathbf{1}\{Y \in \mathcal{PI}_{1-\alpha}(\boldsymbol{X})\} \mid \mathbf{D}]\right] \\
&= \mathbb{E}_{\mathbf{D}\times(\boldsymbol{X},Y)}\left[\mathbb{P}_{\pi}\left(Y \in \mathcal{PI}_{1-\alpha}(\boldsymbol{X}) \mid \mathbf{D}\right)\right] \\
&= 1-\alpha,
\end{aligned} \tag{2.9}$$

*where the probability* $\mathbb{P}_{\mathbf{D}\times(\boldsymbol{X},Y)}$ *(respectively,* $\mathbb{P}_{\pi}$*) and the expectation* $\mathbb{E}_{\mathbf{D}\times(\boldsymbol{X},Y)}$ *(respectively,* $\mathbb{E}_{\pi}$*) are taken with respect to the distribution of* $\mathbf{D}$ *and* $(\boldsymbol{X}, Y)$ *(respectively, to the distribution* $\pi$ *of* $(\boldsymbol{X}, Y)$*).*

*However, note that the backward implication is not true.*

In other words, for a given confidence level $1-\alpha$, Prediction intervals with respect to type II Coverage Probability are also Prediction Intervals with respect to type I Coverage Probability.

**Remark 2.1.9.** *Not all Prediction Interval methods are exact and some of them are sensitive to training dataset* $\mathbf{D}$*. If* $\mathbb{P}_{\pi}\left(Y \in \mathcal{PI}_{1-\alpha}(\boldsymbol{X}) \mid \mathbf{D}\right) \geq 1-\alpha$ *the procedure is said to be conservative. If* $\mathbb{P}_{\pi}\left(Y \in \mathcal{PI}_{1-\alpha}(\boldsymbol{X}) \mid \mathbf{D}\right)$ *goes to* $1-\alpha$ *as* $n \to \infty$*, we say the method is asymptotically correct.*

**Definition 2.1.10** (Empirical Coverage Probability)**.** *For a given training dataset* $\mathbf{D} = (\mathbf{X}, \boldsymbol{y})$ *of observed inputs and output, and for a given confidence level* $1-\alpha$*, the Coverage Probability on the (testing) dataset* $\mathbf{D}' = \{(\boldsymbol{x}'^{(i)}, y'_i)\}_{i=1}^{n'}$ *of sample size* $n'$*, drawn from* $\pi$*, is the percentage of* $\boldsymbol{y}'$ *that fall inside Prediction Intervals* $\mathcal{PI}_{1-\alpha}(\boldsymbol{x}'^{(i)}; \mathbf{D})$

$$\mathrm{CP}_{1-\alpha} = \frac{1}{n'}\sum_{i=1}^{n'} \mathbf{1}\{y'_i \in \mathcal{PI}_{1-\alpha}(\boldsymbol{x}'^{(i)}; \mathbf{D})\}, \tag{2.10}$$

*where* $\mathbf{1}\{A\}$ *is the indicator function of the event* $A$*. Here we denote* $\mathcal{PI}_{1-\alpha}(\boldsymbol{x}'^{(i)}; \mathbf{D})\}$ *to indicate that Prediction Intervals are built using the training set* $\mathbf{D}$*.*

The *empirical* Coverage Probability in 2.1.10 corresponds to $\mathbb{P}_{\hat{\pi}}\left(Y \in \mathcal{PI}_{1-\alpha}(\boldsymbol{X}) \mid \mathbf{D}\right)$, where $\hat{\pi}$ is probability with respect to the empirical distribution constructed from the data $\mathbf{D}'$. It is the naive Monte-Carlo estimator of the probability $\mathbb{P}_{\pi}\left(Y \in \mathcal{PI}_{1-\alpha}(\boldsymbol{X}) \mid \mathbf{D}\right)$, with, obviously,

$$\mathrm{CP}_{1-\alpha} = \frac{1}{n'} \sum_{i=1}^{n'} \mathbf{1}\{y'_i \in \mathcal{PI}_{1-\alpha}(\boldsymbol{x}'^{(i)}; \mathbf{D})\} \overset{n' \to +\infty}{\longrightarrow} \mathbb{P}_{\pi}\left(Y \in \mathcal{PI}_{1-\alpha}(\boldsymbol{X}) \mid \mathbf{D}\right). \tag{2.11}$$

In practice, the *empirical* Coverage Probability $\mathrm{CP}_{1-\alpha}$ measures the reliability of the predictions made using some model or method. If the model is uncertain at some points, then we expect Prediction Intervals to be larger to cover the observed value.

**Remark 2.1.11.** *Computing the empirical Coverage Probability may be sensitive to the sample distribution and sample size. Other issues of under-fitting and over-fitting may also arise.*

**Definition 2.1.12** (Mean of Prediction Intervals Width (MPIW))**.** *The Mean of the Prediction Interval Width (MPIW) is the average width of the prediction intervals, defined as:*

$$\mathrm{MPIW}_{1-\alpha} = \frac{1}{n'} \sum_{i=1}^{n'} \left| \mathcal{PI}_{1-\alpha}(\boldsymbol{x}'^{(i)}; \mathbf{D}) \right|, \tag{2.12}$$

*where* $\left| \mathcal{PI}_{1-\alpha}(\boldsymbol{x}'^{(i)}; \mathbf{D}) \right| = \left| u(\boldsymbol{x}'^{(i)}) - l(\boldsymbol{x}'^{(i)}) \right|$ *is the length of the interval.*

**Definition 2.1.13** (Standard-deviation of Prediction Intervals Width (SdPIW))**.** *The Standard-deviation of the Prediction Interval Width (SdPIW) is the average dispersion of the prediction intervals, defined as:*

$$\mathrm{SdPIW}_{1-\alpha} = \sqrt{\frac{1}{n'} \sum_{i=1}^{n'} \left[ \left| \mathcal{PI}_{1-\alpha}(\boldsymbol{x}'^{(i)}; \mathbf{D}) \right| - \mathrm{MPIW}_{1-\alpha} \right]^2}, \tag{2.13}$$

*where* $\left| \mathcal{PI}_{1-\alpha}(\boldsymbol{x}'^{(i)}; \mathbf{D}) \right| = \left| u(\boldsymbol{x}'^{(i)}) - l(\boldsymbol{x}'^{(i)}) \right|$ *is the length of the interval.*

Other criteria for quantifying Prediction Intervals are also possible. This includes, for example, the normalized mean Prediction Interval width (NMPIW) (Khosravi et al., 2010), the Coverage Width-based Criterion (CWC) (Khosravi et al., 2011), some hybrid loss functions defined on the CWC with a Lagrangian Hu et al. (2019); Pearce et al. (2018), or a graphic indicator on Characteristic curve (ROC-PI) Pang et al. (2018).

## Existing methods

In the following decades, statisticians developed general methods to construct prediction intervals. We briefly describe the most used methods, and we refer to Dewolf & Baets (2022); Patel (1989); Tian et al. (2020) for a review of these methods.

### Frequentist methods

*Ensemble learning.*

Ensemble learning is a popular approach to enhancing predictions by training multiple models (Dietterich, 2000; Heskes, 1996).

In traditional statistical learning (e.g. Random Forests), Ensemble learning is known as *bagging* (Breiman, 2001). The goal is to aggregate the individual predictions among a large sample of models. This allows a naive construction of a Prediction Interval by treating the predictions of the individual models in the ensemble as elements of a data sample. The empirical mean and variance are computed and used as moment estimators for a normal distribution. The Prediction Intervals bounds can be determined using the $z$-score corresponding to a significance level for the standard normal distribution.

For Deep Learning algorithms, the idea behind deep ensembles is also similar Hansen & Salamon (1990); Lakshminarayanan et al. (2017): training multiple models to obtain a better and more robust prediction. The loss functions of these ensembled deep models are aggregated to predict the mean and variance of the output (Nix & Weigend, 1994).

*Bootstrap.*

The Bootstrap method is initially introduced by Efron (1979); Efron & Gong (1983) for independent variables and later extended to deal with more complex dependent variables. It is a class of nonparametric methods that allow statisticians to conduct statistical inference on a wide range of problems without imposing structural assumptions on the underlying data-generating random process.

In the regression setting, the Bootstrap method estimates model uncertainty by constructing multiple models, with different parameter initialization, on different resampled versions of the training dataset (Heskes, 1996). It is considered one of the most used methods Efron & Tibshirani (1993) for estimating empirical variances and constructing Predictions Intervals. It is claimed to generate valid prediction intervals under some asymptotic frameworks. More precisely, the bootstrap estimator is $\sqrt{n}$-asymptotically normal and consistent.

*Jackknife.*

Jackknife resampling, initially developed by Quenouille (1949) for reducing the bias of an estimator of a serial correlation coefficient by splitting the sample and refined later by Tukey (1958), is a nonparametric method used for estimating sampling distributions (variance and bias) of a large population. It involves a Leave-One-Out strategy of estimating a parameter (e.g., the variance) in a data set of $n$ observations by $n-1$ models.

The jackknife method went over continuous improvements. Kunsch (1989) proposed a variant of the jackknife for general stationary observations rather than *i.i.d.* data. Jackknife-after-Bootstrap method (Efron, 1992) was proposed to improve the variance estimate of a bootstrap estimate. The infinitesimal jackknife method was previously used for quantifying the predictive uncertainty in random forests (Wager et al., 2014). These methods are, however, bespoke to bagging predictors.

More recently, general-purpose jackknife estimators were developed in (Barber et al., 2021). Two specific leave-one-out procedures: the Jackknife+ and the Jackknife-minmax, were shown to have assumption-free worst-case coverage guarantees.

*Quantile regression.*

The quantile regression is a type of regression analysis used in statistics and econometrics. This method estimate in particular the quantile of the conditional distribution of $Y$ given $\boldsymbol{X} = \boldsymbol{x}$ (instead of estimating the conditional mean as standard regression). It was introduced by Koenker & Bassett (1978), and developed later Koenker & Hallock (2001), by extending the

regression to the estimation of conditional quantile functions,

Meinshausen (2006) provides a new method, *Quantile Random Forest*, for estimating prediction intervals for ensemble methods (Random Forest, for instance). The idea consists of replacing the mean-squared error with the *Pinball loss* or quantile loss (Koenker, 2005) that targets the $a$-quantile efficiently.

The main advantage of this approach is that it does not depend on any assumed distribution of the outcome $Y$. It is, therefore, a nonparametric tool for estimating Prediction Intervals. However, it also has some disadvantages: the quantile regression targets a specific quantile $a$ at each time, which means that the model needs to be retrained if one is interested in different values of $a$. Moreover, the quantile regression is only able to capture the aleatoric uncertainty.

*High-Quality principle methods.*

The High-Quality principle methods Pearce et al. (2018) is a class of direct interval estimators trained to output a prediction interval, given by its upper and lower bounds. The idea is to construct a loss function in such a way that the optimal Prediction Intervals achieve the optimal (marginal or conditional) Coverage Probability and minimize their average width (MPIW).

The first to propose the High-Quality principle are Khosravi et al. (2011) with the Lower Upper Bound Estimation for Neural Networks. The used loss function, Coverage Width-based Criterion, combines the coverage and the width of Prediction Intervals. Pearce et al. (2018) formalize the ideas of the High-Quality principle and developed an alternative to the Coverage Width-based Criterion, derived from a likelihood principle.

### Conformal Prediction

Conformal Prediction, introduced by (Gammerman et al., 1998; Vovk et al., 1999), became a popular statistical framework that can be used to build Prediction Intervals for arbitrary Machine Learning models for both regression and classification problems. It provides valid Prediction Intervals (i.e. achieve nominal marginal coverage, *not conditional coverage*) in a finite sample under a certain set of assumptions (e.g., exchangeable data) (Romano et al., 2019).

The original implementation had several computational issues because all calculations had to be redone for every data point. Inductive Conformal Prediction (ICP) or Split-Conformal Prediction (Lei et al., 2016; Vovk et al., 2005) was proposed as a solution. This method shed more interest in Conformal Prediction (Papadopoulos, 2008; Vovk et al., 2018). The recent development is the jackknife+ (Barber et al., 2021) which offers guarantees that are not possible for the original Jackknife, a valid coverage and a compromise between the computational and statistical costs of the two methods.

The current research on Conformal Prediction focuses more on non-exchangeable data. Tibshirani et al. (2019) introduce the concept of *weighted exchangeability* to extend conformal prediction to the non-exchangeable data setting. In a similar work, Barber et al. (2022) use weighted quantiles that do not treat data points symmetrically. On time-series, Gibbs & Candes (2021) propose a robust method for predicting distribution-shift time series. Xu & Xie (2021) consider ensembling time-series predictors that are trained over bootstrapped subsamples.

So far, Conformal prediction is used to achieve the marginal coverage, which has a weaker property than the conditional Coverage Property as defined in (2.1.7). In other words, unfortunately, Conformal Prediction methods do not provide sufficient guarantees to achieve

the nominal level with respect to the conditional coverage (Angelopoulos & Bates, 2021).

### Bayesian approaches

The Bayesian framework offers a principled framework for handling uncertainty. Indeed, unlike classical learning algorithms, Bayesian inference does not attempt to identify *best-fit* models of the data. Instead, it computes a posterior distribution over models. More specifically, one tries to model the distribution of interest (here, the distribution of $Y$ given $\boldsymbol{X} = \boldsymbol{x}_{\text{new}}$ by updating a prior (e.g. over some parameters) in light of evidence (e.g. observed data). The conditional distribution given parameters are inferred from a given parametric model or *likelihood* function using Bayes' Rule. The posterior predictive distribution is then calculated by marginalizing the parameters. The obtained posterior predictive distribution is used to make predictions at new points.

**Definition 2.1.14** (probabilistic model)**.** *In statistical Machine Learning, for a new point $\boldsymbol{x}_{\text{new}}$, a probabilistic model is a model that is able to predict a probability distribution over a set of distributions rather than only outputting a single value (corresponding to mean, median or most probable).*

Probabilistic models consider inputs and the output as random variables and assume joint probability distributions over them. Consequently, using probability theory and Bayesian inference, the model's output is also a probability distribution. This represents one of the significant benefits of probabilistic models because they show how the uncertainty is propagated in the predictions Ghahramani (2015).

One of the famous probabilistic Bayesian models are Bayesian Neural Networks (MacKay, 1992). Unfortunately, they frequently attach an over-confidence in predictions made on target data. Furthermore, the complexity of the approach (due to the number of weights and layers) led to considering Gaussian Processes prior over function. Since then, Gaussian Processes have become one of the most popular probabilistic models for regression problems (Williams & Rasmussen, 1995). The main reason for its popularity is that it is one of the few Bayesian methods where the Bayesian inference is performed exactly since the marginalization of multivariate normal distributions can be written in closed form (Dewolf & Baets, 2022).

However, probabilistic models also suffer from critical points. Firstly, they do not typically come with coverage guarantees. Secondly, the coverage of the obtained Prediction Intervals depends highly on the correctness of the model (well-specified). It can even fail in certain high-dimensional regimes where the model is well-specified (Bai et al., 2021). Finally, there is no existing unique method for calibrating Predicting Intervals for these methods. Close work was developed by Lawless & Fredette (2005) for parametric predictive distributions but not for probabilistic models.

For all these considerations, we will consider in the following the Gaussian Process regression for estimating Prediction Intervals.

## 2.2 Gaussian Process Regression

The history of Gaussian Processes began in the 1940s with the works of Wiener and Kolmogorov for predicting time series. A few years later, the Gaussian Processes regression was used in

geostatistics by Krige (1951) (to whom has credited the appellation of Kriging) to model the distribution of ore content in South African mines.

Afterwards, it became increasingly popular in geostatistics after the 1970s with Cressie (1993); Matheron (1970); Ripley (1981). It was developed for spatial interpolation problems as it considers the spatial statistical structure of the estimated variable. Sacks et al. (1989) then Oakley et al. (2004); Santner et al. (2003) have extended the kriging principles to computer experiments and surrogate modelling. It has also been used in approximation, interpolation and smoothing.

Recently, the Gaussian Process regressor, also called *Kriging model*, became popular in the machine learning community in the prediction context. Especially after the work of Williams & Rasmussen (1995) and later in Rasmussen & Williams (2005) where its basis was set up with probability theory and algebra.

The Kriging model presents several advantages, especially the interpolation and interpretability properties. Moreover, numerous authors (e.g. Currin et al. (1991); Santner et al. (2003)) show that this model can provide a statistical framework to compute an efficient predictor with associated uncertainty.

### Gaussian Process and covariance functions

This subsection defines several notions, definitions, and theorems used in Kriging with GP. Most definitions of this subsection are taken from Rasmussen & Williams (2005) and Bachoc (2013). In the following, we consider a domain of interest $\mathcal{D} \subseteq \mathbb{R}^d$,

**Definition 2.2.1** (Stochastic process)**.** *A real-valued random process (or random function) on $\mathcal{D}$ is an application $Y(\cdot)$, that associates a random variable $Y(\boldsymbol{x})$ to each $\boldsymbol{x} \in \mathcal{D}$. All the random variables $Y(\boldsymbol{x})$, for $x \in \mathcal{D}$, are defined on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$.*

In other words, a stochastic process $Y(\cdot)$ is a function on $\mathbb{R}^d$ that is unknown, or that depends on underlying random phenomena. If $\mathbb{E}\big[Y(\boldsymbol{x})^2\big] < +\infty$, we can define the mean and covariance functions of the process $Y$ as

- Mean function $m : \boldsymbol{x} \mapsto m(\boldsymbol{x}) = \mathbb{E}(Y(\boldsymbol{x}))$.
- Covariance function $k : (\boldsymbol{x}_1, \boldsymbol{x}_2) \mapsto k(\boldsymbol{x}_1, \boldsymbol{x}_2) = \mathrm{Cov}(Y(\boldsymbol{x}_1), Y(\boldsymbol{x}_2))$.

**Definition 2.2.2** (Trajectory of a random process)**.** *For each fixed $\omega \in \Omega$, the real-valued function $\mathcal{D} : \boldsymbol{x} \mapsto Y(\omega, \boldsymbol{x})$ is called a trajectory (or a realization, sample function, path) of the random process $Y(\cdot)$.*

To understand the distribution of stochastic process $Y(\cdot)$, we need to consider the finite-dimensional distribution of $Y(\cdot)$.

**Definition 2.2.3** (Finite-dimensional distribution)**.** *For any $n$ points $\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(n)}$, the multidimensional probability distribution of the random vector $Y(\boldsymbol{x}^{(1)}), \ldots, Y(\boldsymbol{x}^{(n)})$ is called the finite-dimensional distribution of the random function $Y(\cdot)$. It can be characterized, for example, by the Cumulative Distribution Function $F^Y$ such that*

$$F^Y_{\boldsymbol{x}^{(1)},\ldots,\boldsymbol{x}^{(n)}}(c_1, \ldots, c_n) = \mathbb{P}\left(Y(\boldsymbol{x}^{(1)}) \leq c_1, \ldots, Y(\boldsymbol{x}^{(n)}) \leq c_n\right). \tag{2.14}$$

The notion of finite-dimensional distribution is crucial for the predictions and conditional simulations of the process $Y(\cdot)$. Indeed, the fact that there is a probability distribution for the random vector $Y(\boldsymbol{x}^{(1)}), \ldots, Y(\boldsymbol{x}^{(n)}), Y(\boldsymbol{x})$ enables us to predict the value of $Y(\boldsymbol{x})$, after observing the values of $Y(\boldsymbol{x}^{(1)}), \ldots, Y(\boldsymbol{x}^{(n)})$ (Bachoc, 2013).

We give a short introduction to the Gaussian multidimensional distribution,

**Definition 2.2.4** (Gaussian variables)**.** *A random variable $X$ is a Gaussian variable with mean $\mu$ and variance $\sigma^2$ if its characteristic function has the form:*

$$\Phi(z) = \exp\left(iz\mu - \frac{1}{2}z^2\sigma^2\right) \quad \forall z \in \mathbb{R}. \tag{2.15}$$

When $\sigma^2 > 0$, then the probability density function of $X$ is well-defined and satisfies:

$$f(z) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(z-\mu)^2}{2\sigma^2}\right). \tag{2.16}$$

**Definition 2.2.5** (Gaussian vectors)**.** *A $n$-dimensional random vector $\boldsymbol{y} = (y_1, \ldots, y_n)$ is a Gaussian vector with mean vector $\boldsymbol{m} = \mathbb{E}(\boldsymbol{y})$ and covariance matrix $\mathbf{K} = \mathrm{Cov}(\boldsymbol{y})$ when either:*

- *Any linear combination of its components is a Gaussian random variable.*
- *The characteristic function of the random vector $\boldsymbol{y}$ has the form:*

$$\Phi(\boldsymbol{z}) = \exp\left(i\boldsymbol{z}^\top \boldsymbol{m} - \frac{1}{2}\boldsymbol{z}^\top \mathbf{K}\boldsymbol{z}\right) \quad \forall \boldsymbol{z} \in \mathbb{R}^n. \tag{2.17}$$

*We write $\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{m}, \mathbf{K})$ to specify that $\boldsymbol{y}$ is Gaussian vector.*

When $\mathbf{K}$ is non-singular, the probability density function of $\boldsymbol{y}$ can be written as

$$f(\boldsymbol{z}) = \left((2\pi)^n |\mathbf{K}|\right)^{-1/2} \exp\left(-\frac{1}{2}(\boldsymbol{z}-\boldsymbol{m})^\top \mathbf{K}^{-1}(\boldsymbol{z}-\boldsymbol{m})\right), \tag{2.18}$$

where $\boldsymbol{z} \in \mathbb{R}^n$ and $|\mathbf{K}|$ is the determinant of covariance matrix $\mathbf{K}$.

However, suppose $\mathbf{K}$ is singular. In that case, there exists a hyperplane of $\mathbb{R}^n$ which is the support of $\boldsymbol{y}$ (meaning that, almost surely, $\boldsymbol{y}$ belongs to this hyperplane) and so that, restricted on this hyperplane, $\boldsymbol{y}$ has a probability density function of the previous form (with respect to the Lebesgue measure over the hyperplane) (Bachoc, 2013).

Once the Gaussian prior is made on the observations of $\boldsymbol{y}$, the theorem below is useful to deduce the distribution of the posterior predictive distribution.

**Theorem 2.2.6** (Gaussian conditioning theorem)**.** *Let $(\boldsymbol{y}_1, \boldsymbol{y}_2)$ be a Gaussian vector such as:*

$$\begin{pmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \begin{pmatrix} \mathbf{K}_{1,1} & \mathbf{K}_{1,2} \\ \mathbf{K}_{2,1} & \mathbf{K}_{2,2} \end{pmatrix}\right) \tag{2.19}$$

*If $\mathbf{K}_{1,1}$ is invertible, then $Y_2|Y_1 = \boldsymbol{y}_1$ (i.e. $Y_2$ conditionally on $Y_1 = \boldsymbol{y}_1$) follows a Gaussian distribution*

$$Y_2|Y_1 = \boldsymbol{y}_1 \sim \mathcal{N}\left(\boldsymbol{\mu}_2 + \mathbf{K}_{2,1}\mathbf{K}_{1,1}^{-1}(\boldsymbol{y}_1 - \boldsymbol{\mu}_1), \mathbf{K}_{2,2} - \mathbf{K}_{2,1}\mathbf{K}_{1,1}^{-1}\mathbf{K}_{1,2}\right). \tag{2.20}$$

We refer to Von Mises (1964) in Section 9.3 for the proof of the theorem.

**Remark 2.2.7.** *The conditional distribution of $Y_2$ given $Y_1 = \boldsymbol{y}_1$ can be used to infer many statistical quantities of interest, such that the most probable prediction value, the threshold exceedance probability etc.*

*In particular, the conditional mean $\mathbb{E}(Y_2 \mid Y_1 = \boldsymbol{y}_1)$ is the best in the mean square sense of $Y_2$ given $Y_1$ and the conditional variance $\mathbb{V}(Y_2 \mid Y_1 = \boldsymbol{y}_1)$ quantify the degree of the approximation/prediction error.*

**Definition 2.2.8** (Gaussian Process, (Rasmussen & Williams, 2005))**.** *A stochastic process $Y$ on $\mathbb{R}^d$ is a Gaussian Process when, for all $\left(\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(n)}\right)$, the random vector $(Y(\boldsymbol{x}^{(1)}), \ldots, Y(\boldsymbol{x}^{(n)}))$ is Gaussian.*

In other words, a random process is a Gaussian process if its finite-dimensional distributions are multidimensional Gaussian distributions. Since a multidimensional Gaussian distribution is fully characterized by its mean vector $\boldsymbol{m}$ and its covariance matrix $\mathbf{K}$, a Gaussian process $Y$ is also fully characterized by its mean and covariance functions, defined in the following definitions:

**Definition 2.2.9.** *The mean function of a Gaussian process $Y$ is the map $m : \mathcal{D} \mapsto \mathbb{R}$ such that $m(\boldsymbol{x}) = \mathbb{E}(Y(\boldsymbol{x}))$.*

**Definition 2.2.10.** *The covariance function of a Gaussian process $Y$ is the application $k : \mathcal{D} \times \mathcal{D} \mapsto \mathbb{R}$ such that $k(\boldsymbol{x}_1, \boldsymbol{x}_2) = \mathrm{Cov}(Y(\boldsymbol{x}_1), Y(\boldsymbol{x}_2))$.*

The covariance function $k$ has three main properties: 1) symmetric, 2) positive semi-definite and 3) stationary. We define the two previous notions below.

**Definition 2.2.11** (Positive semi-definite)**.** *A bi-variate function $k$ is positive semi-definite if and only if, for any $\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(n)} \in \mathcal{D}$, the Gram Matrix defined by $\mathbf{K} = (k(\boldsymbol{x}^{(i)}, \boldsymbol{x}^{(j)}))_{1 \leq i,j \leq n}$ is positive semi-definite, that is, for all $\boldsymbol{a} \in \mathbb{R}^n :\ \boldsymbol{a}^\top \mathbf{K} \boldsymbol{a} \geq 0$.*

**Definition 2.2.12** (Positive definite)**.** *A bi-variate function $k$ is positive definite if and only if, for any distinct $\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(n)} \in \mathcal{D}$, its Gram Matrix $\mathbf{K}$ is positive definite, that is, for all $\boldsymbol{a} \in \mathbb{R}^n \backslash \{\mathbf{0}\} :\ \boldsymbol{a}^\top \mathbf{K} \boldsymbol{a} > 0$.*

**Definition 2.2.13** (Stationary of Gaussian Process)**.** *A random process $Y(\cdot)$ is stationary if, for all $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \in \mathcal{D}$ and for $\boldsymbol{h} \in \mathbb{R}^d$, the finite-dimensional distribution of $Y(\cdot)$ at $\boldsymbol{x}_1 + \boldsymbol{h}, \ldots, \boldsymbol{x}_n + \boldsymbol{h}$ has the same as the finite-dimensional distribution at $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \in \mathcal{D}$.*

The stationarity of a Gaussian Process can be characterized in terms of mean function and covariance kernel [Rasmussen & Williams (2005) in Chapter 4].

**Definition 2.2.14** (Stationary covariance function)**.** *A positive definite mapping $k : \mathcal{D} \times \mathcal{D} \mapsto \mathbb{R}$ is said to be stationary if there exists a mapping $r : \mathcal{D} \mapsto \mathbb{R}$ such that for all $\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{D} \times \mathcal{D}$: $k(\boldsymbol{x}, \boldsymbol{x}') = r(\boldsymbol{x} - \boldsymbol{x}')$.*

**Theorem 2.2.15.** *Let $m$ be any function from $\mathcal{D}$ to $\mathbb{R}$. Let $k$ be a function from $\mathcal{D} \times \mathcal{D}$ to $\mathbb{R}$ such that, for any $n \in \mathbb{N}$ and for any $\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(n)} \in \mathcal{D}$, the Gram matrix $\mathbf{K} = (k(\boldsymbol{x}^{(i)}, \boldsymbol{x}^{(j)}))_{1 \leq i,j \leq n}$ is symmetric and positive semi-definite. Then, there exists a Gaussian process $Y(\cdot)$ on $\mathcal{D}$ with mean function $m$ and covariance function $k$.*

We refer to Dudley (2002)[Theorem 12.1.3] for the proof, which is based on Kolmogorov's extension theorem (Billingsley, 1995).

Theorem 2.2.15 highlights a major advantage of the Gaussian Processes present: they are simple to define and simulate from their mean and covariance functions. In addition, the Gaussian distribution is reasonable for modelling a large variety of random variables. To indicate that a random function $Y(\cdot)$ follows a Gaussian process, we write:

$$Y(\cdot) \sim \mathcal{GP}(m(\cdot), k(\cdot,\cdot)), \tag{2.21}$$

where $m$ and $k$ are the mean and covariance functions of $Y(\cdot)$.

**Remark 2.2.16.** *Theorem 2.2.15 proves also the existence of one-to-one correspondence between the distribution of a Gaussian process $Y(\cdot) \sim \mathcal{GP}(m(\cdot), k(\cdot,\cdot))$ and pairs (m, k) of mean function $m$ and covariance function $k$. Therefore, most GP's properties are induced by the specification of $k$ and $m$.*

Usually, symmetric and positive definite functions are called *kernels*. We refer to the study of Schölkopf & Smola (2002) in Chapters 2 and 13, about the positive-definiteness of bi-variate mappings (kernels). We give in the following subsection a brief introduction to kernels and RKHS theory.

### RKHS Theory : Reproducing Kernel Hilbert Spaces

We begin by introducing Hilbert spaces and kernels, which form the building block of reproducing kernel Hilbert spaces as presented by Berlinet & Thomas-Agnan (2004).

In the following, we consider $\mathcal{X}$ a non-empty subspace of the input space (for example $\mathbb{R}$).

**Definition 2.2.17** (Hilbert space)**.** *A Hilbert Space $\mathcal{H}$ is an inner product space that is complete and separable with respect to the norm defined by the inner product (i.e. Cauchy sequence limits).*

**Definition 2.2.18** (Characterisation of kernels)**.** *A function $k : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ is a kernel if there exists a Hilbert space $\mathcal{H}(k)$ and a feature map $\phi : \mathcal{X} \mapsto \mathcal{H}(k)$ such that, for all $x, x_0 \in \mathcal{X}$, we have*

$$k(x, x_0) := \langle \phi(x), \phi(x_0) \rangle_{\mathcal{H}(k)}. \tag{2.22}$$

*The feature map $\phi$ of every point $x \in \mathcal{X}$ is a function such that $\phi(x) = k(\cdot, x)$. In particular, for any $x, y \in \mathcal{X}$, $k(x, x_0) = \langle k(x,\cdot), k(\cdot, x_0) \rangle_{\mathcal{H}(k)} = \langle \phi(x), \phi(x_0) \rangle_{\mathcal{H}(k)}$.*

A kernel $k$, by definition, satisfies the properties of symmetry and being positive semi-definite. For simplicity, we say that the kernel $k$ is defined on $\mathcal{X}$. In the following, unless specified otherwise, the RKHS $\mathcal{H}(k)$ is denoted simply by $\mathcal{H}$.

**Theorem 2.2.19** (Sum of kernels is kernel)**.** *Let $\alpha > 0$, if $k, k_1$ and $k_2$ are kernels on $\mathcal{X}$, then $\alpha k$ and $k_1 + k_2$ are kernels on $\mathcal{X}$.*

**Theorem 2.2.20** (Product of kernels is kernel)**.** *If $k_1$ and $k_2$ are two kernels defined on $\mathcal{X}$, then the map $k := k_1 \times k_2$ defined on $\mathcal{X}$ by*

$$(k_1 \times k_2)(x, x_0) = k_1(x, x_0) k_2(x, x_0) \tag{2.23}$$

*is a kernel.*

We refer to Schölkopf & Smola (2002) in Chapter 13 for the proof of the previous theorems.

**Theorem 2.2.21** (Tensorised Product of kernels is kernel)**.** *Given two kernels, $k_1$ defined on $\mathcal{X}_1$ and $k_2$ defined on $\mathcal{X}_2$, then the map $k : k_1 \times k_2$ defined on $\mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2$ by*

$$(k_1 \times k_2)(x, x_0) = k_1(x^{(1)}, x_0^{(1)})k_2(x^{(2)}, x_0^{(2)}), \tag{2.24}$$

*where $x = (x^{(1)}, x^{(2)}) \in \mathcal{X}_1 \times \mathcal{X}_2$ and $x_0 = (x_0^{(1)}, x_0^{(2)}) \in \mathcal{X}_1 \times \mathcal{X}_2$, is a kernel. We call it the tensorized product kernel.*

**Definition 2.2.22** (Space of real-valued functions on $\mathcal{X}$)**.** *The space*

$$\mathcal{F}(\mathcal{X}) = \{g : \mathcal{X} \mapsto \mathbb{R} \mid g \text{ is a function}\}$$

*together with the standard scalar multiplication and summation defined for all $\lambda \in \mathbb{R}$, and for all $g, h \in \mathcal{F}(\mathcal{X})$, by:*

$$(\lambda g)(x) := \lambda h(x) \qquad \forall x \in \mathcal{X}$$

*,*

$$(g + h)(x) := g(x) + h(x) \qquad \forall x \in \mathcal{X}$$

*, forms a linear space over $\mathbb{R}$. We call $\mathcal{F}(\mathcal{X})$ the space of real-valued functions on $\mathcal{X}$.*

The Reproducing kernel Hilbert spaces on $\mathcal{X}$, defined below, are well-behaved sub-spaces of $\mathcal{F}(\mathcal{X})$.

**Definition 2.2.23** (Reproducing Kernel Hilbert spaces)**.** *Let $\mathcal{H} \subseteq \mathcal{F}(\mathcal{X})$ be a Hilbert space. Then $\mathcal{H}$ is called a RKHS if there exists a kernel $k$ satisfying:*

- $\forall x \in \mathcal{X} : k(x, \cdot) \in \mathcal{H}$.
- $\forall g \in \mathcal{H}, \forall x \in \mathcal{X} : \langle g, k(x, \cdot)\rangle_{\mathcal{H}} = g(x)$.

The second property is called *the reproducing property* of $k$, we say that $k$ is a reproducing kernel of $\mathcal{H}$.

**Theorem 2.2.24** (Uniqueness of the kernel (Schölkopf & Smola, 2002))**.** *Let $\mathcal{H}$ be an RKHS on $\mathcal{X}$. Assume both $k$ and $\tilde{k}$ are reproducing kernels of $\mathcal{H}$, then $k = \tilde{k}$.*

**Theorem 2.2.25** (Moore-(Aronszajn, 1950))**.** *Let $k : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ be positive definite kernel. There is a **unique** RKHS $\mathcal{H}$ with reproducing kernel $k$.*

**Remark 2.2.26.** *The feature map $\phi$ is not unique. Only kernel $k$ is unique.*

To summarize up the RKHS theory, if $\mathcal{H}$ is a RKHS and $\mathcal{X}$ is non-empty set of points, then ,for each $x \in \mathcal{X}$, there exists, by the Riesz's representation theorem a function (i.e feature map $\phi$) such that $\phi(x) = k(x, \cdot)$ in $\mathcal{H}$ (called representer) with the reproducing property of $\mathcal{F}_x(g) = \langle g, k(x, \cdot)\rangle_{\mathcal{H}} = g(x)$ where $\mathcal{F}_x(g)$ denotes the evaluation application of $g \in \mathcal{H}$ on $x$.

We recall one of the most used stationary kernels in $\mathbb{R}$, the Matérn kernel:

$$k^{\nu}_{\sigma^2,\theta}(x, x') = \sigma^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\sqrt{2\nu}\frac{|x - x'|}{\theta}\right)^{\nu} K_{\nu}\left(\sqrt{2\nu}\frac{|x - x'|}{\theta}\right), \tag{2.25}$$

where

- $\sigma^2 > 0$ is the variance amplitude, the larger $\sigma^2$ is, the larger the scale of the trajectories.
- $\theta > 0$ is the characteristic length-scale. It controls how fast the functions sampled from the Gaussian Processes oscillate.
- $\nu$ is the smoothness hyperparameter that controls the degree of regularity (i.e. differentiability) of the Gaussian Process.
- $\Gamma$ is the complete Gamma function
- $K_\nu$ is the modified Bessel function of the second kind.

For a Gaussian Process with Matérn covariance and smoothness parameter $\nu$, the paths are *almost surely* $\lceil \nu - 1 \rceil$ times differentiable on $\mathbb{R}$. Lower values of $\nu$ correspond to rougher functions, whereas higher values of $\nu$ correspond to smoother functions.

Some particular cases of Matérn kernel are when $\nu = \frac{1}{2}, \frac{3}{2}, \frac{5}{2}$ and $\nu \to \infty$.

- Exponential kernel ($\nu = \frac{1}{2}$): $k_{Exp}(x, x') = \sigma^2 \exp\left(-\frac{|x-x'|}{\theta}\right)$ corresponding to the known Ornstein & Uhlenbeck (1930) process.
- Matérn 3/2 kernel: $k_{Mat3/2}(x, x') = \sigma^2 \left(1 + \sqrt{3}\frac{|x-x'|}{\theta}\right) \exp\left(-\sqrt{3}\frac{|x-x'|}{\theta}\right)$.
- Matérn 5/2 kernel: $k_{Mat5/2}(x, x') = \sigma^2 \left(1 + \sqrt{5}\frac{|x-x'|}{\theta} + \frac{5}{3}\frac{(x-x')^2}{\theta^2}\right) \exp\left(-\sqrt{5}\frac{|x-x'|}{\theta}\right)$.
- Gaussian kernel ($\nu \to \infty$): $k_{Gauss}(x, x') = \sigma^2 \exp\left(-\frac{|x-x'|^2}{2\theta^2}\right)$.

The choice of the covariance function is important as it synthesizes information from the Gaussian Process. For example, the choice of the Gaussian kernel assumes that the function $f$ that we want to learn is very smooth of class $\mathcal{C}^\infty$ (infinitely differentiable). This is often too strict as a condition. A common alternative is the functions Matérn 5/2 or Matérn 3/2 kernel. Some cases of the influence of the covariance parameters can be seen on Figures 2.1, 2.2 and 2.3.

In the case of a Gaussian Process defined on $\mathcal{D} \subseteq \mathbb{R}^d$, the amplitude $\sigma^2$ is defined as one value, but the length-scale $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_d) \in \mathbb{R}_+^d$ is now defined as a vector. When $\theta_i$ is particularly small, then the variable $X_i$ is particularly important, this allows us to get a rank/hierarchy of the input variables $(X_1, \ldots, X_d)$ according to their correlation lengths $\theta_1, \ldots, \theta_d$.

As mentioned in the Subsection 2.2, it is possible to combine the sum and the product of kernels (see Theorems 2.2.19 and 2.2.20). Thus, we can obtain more complex covariance models in $\mathbb{R}^d$ based on classical kernels in $\mathbb{R}$:

- The radial model (anisotropic geometric model) defined by:

$$\boldsymbol{k}_{\sigma^2,\boldsymbol{\theta}}^{\text{radial}}(\boldsymbol{x}, \boldsymbol{x}') = k_{\sigma^2,\boldsymbol{\theta}}^{\nu}\left(\sqrt{\sum_{j=1}^{d} \frac{|x_j - x_j'|^2}{\theta_j^2}}\right). \tag{2.26}$$

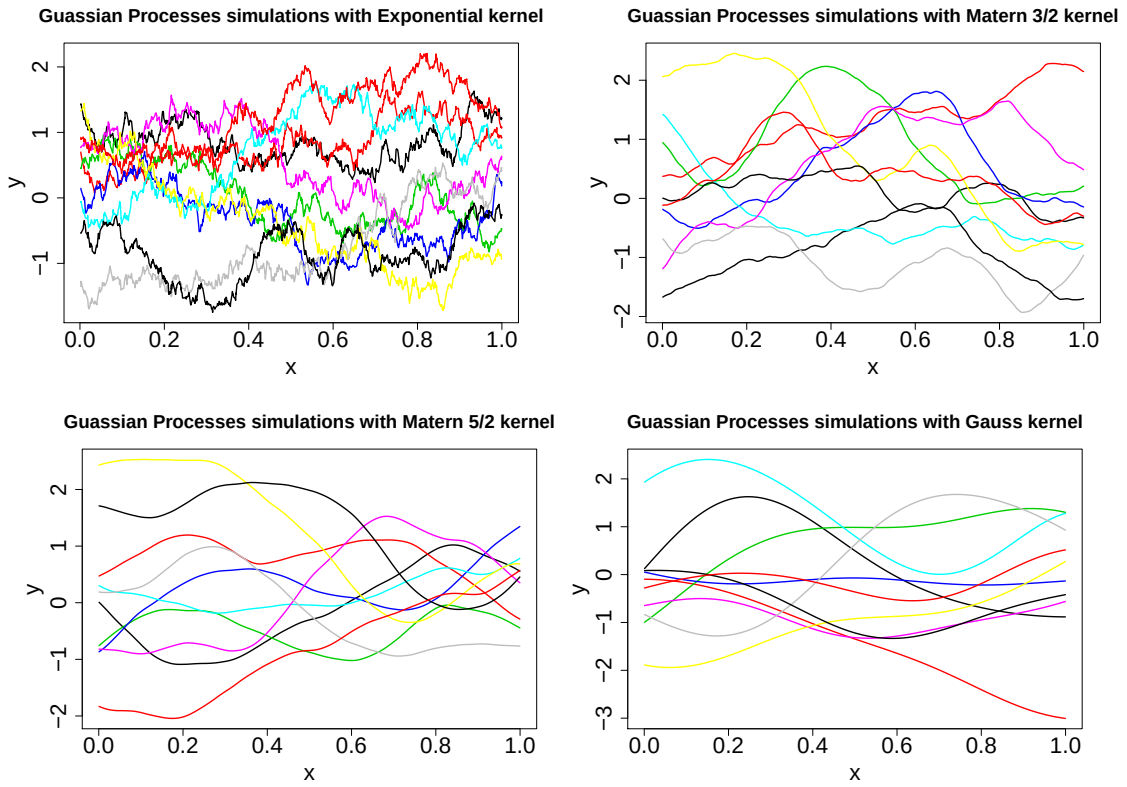



Figure 2.1: Trajectories of Gaussian processes for different covariance functions with $\nu = 1/2$ from the top left to $n \to +\infty$ in the bottom right.

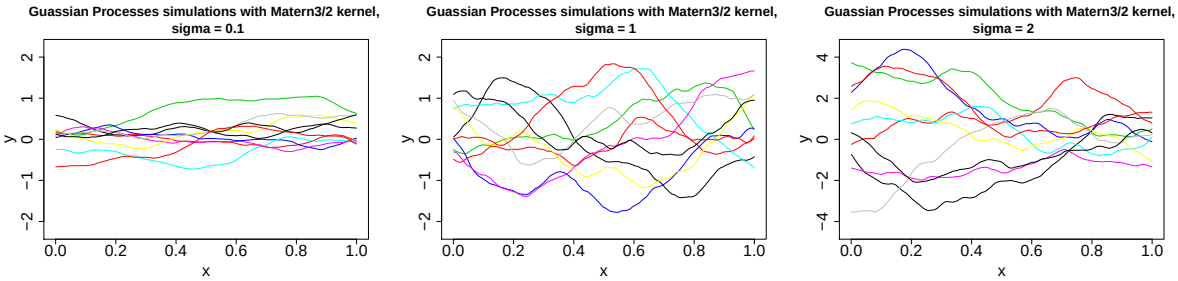



Figure 2.2: The influence of the variance amplitude $\sigma^2$: Trajectories of Gaussian processes with Matérn 3/2 and an amplitude of (from the left to the right) $\sigma^2 = 0.1, 1, 2$.

- The tensorized product model defined by:

$$\boldsymbol{k}_{\sigma^2,\boldsymbol{\theta}}^{\text{TensorProd}}(\boldsymbol{x}, \boldsymbol{x}') = \sigma^2 \bigotimes_{j=1}^{d} k_{1,\theta_i}^{\nu}(x_j, x_j'). \tag{2.27}$$

- The tensorized additive model defined by:

$$\boldsymbol{k}_{\sigma^2,\boldsymbol{\theta}}^{\text{TensorSum}}(\boldsymbol{x}, \boldsymbol{x}') = \bigoplus_{j=1}^{d} k_{\sigma_j,\theta_j}^{\nu}(x_j, x_j'). \tag{2.28}$$

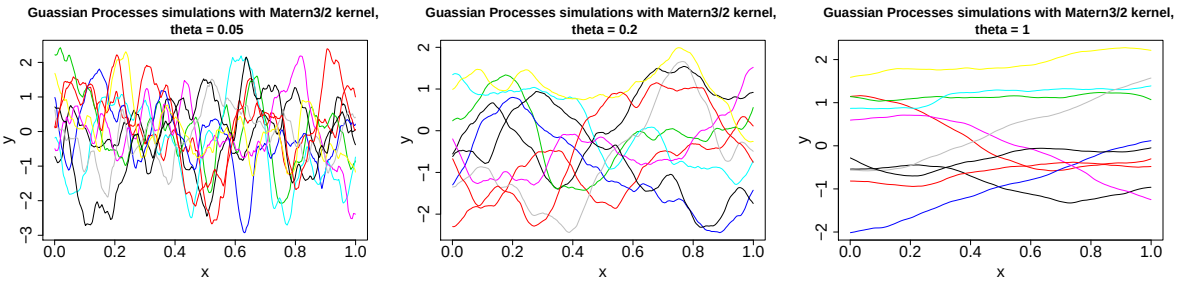



Figure 2.3: The influence of the length-correlation $\theta$: Trajectories of Gaussian processes with Matérn 3/2 and a correlation length of (from left to right) $\theta = 0.05, 0.2, 1$.

Other classical covariance functions can be build, such as the power-exponential by tensorizing the exponential kernel $k_{Exp}$ parameterized also by $0 < p \leq 2$:

$$\boldsymbol{k}_{\sigma^2,\boldsymbol{\theta}}^{\text{PowExp}}(\boldsymbol{x}, \boldsymbol{x}') = \sigma^2 \prod_{j=1}^{d} \exp\left(-\left(\frac{|x_j - x_j'|}{\theta_j}\right)^p\right), \tag{2.29}$$

or the quasi-periodic GP (Tolba et al., 2019) by multiplying a periodic kernel by a non-periodic kernel.

Assume that $\mathcal{X} = \mathbb{R}^d$, we show now the important equivalence between the RKHS of Matérn kernels and Sobolev spaces. We refer the reader to Berlinet & Thomas-Agnan (2004) and Wendland (2004) in Chapter 10 for more details about Sobolev spaces.

**Definition 2.2.27** (Sobolev space)**.** *Let $f \in L^2(\mathbb{R}^d)$ be a squared integrable function defined on $\mathbb{R}^d$. Let $\widehat{f}(\boldsymbol{\xi}) = \int_{\mathbb{R}^d} f(\boldsymbol{x}) \exp(-i\boldsymbol{\xi}^\top \boldsymbol{x}) \mathrm{d}\boldsymbol{x}$ denote the Fourier transform of the function $f$. The Sobolev space $\mathcal{H}_2^s(\mathbb{R}^d)$ of order $s > d/2$ is the Hilbert space*

$$\mathcal{H}_2^s(\mathbb{R}^d) = \{f \in L^2(\mathbb{R}^d) \text{ s.t. } \boldsymbol{\xi} \mapsto \hat{f}(\boldsymbol{\xi})(1 + \|\boldsymbol{\xi}\|^2)^{s/2} \in L^2(\mathbb{R}^d)\}. \tag{2.30}$$

**Remark 2.2.28.** *The assumption $s > d/2$ is required to ensure, by the Sobolev embedding theorem, that every element of $\mathcal{H}_2^s(\mathbb{R}^d)$ is continuous.*

For a stationary kernel of the form $\boldsymbol{k}(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{r}(\boldsymbol{x} - \boldsymbol{y})$ with $\boldsymbol{r} \in L^2(\mathbb{R}^d)$, we assume that its Fourier transform $\hat{\boldsymbol{r}}(\boldsymbol{\xi})$ satisfies

$$C_1(1 + \|\boldsymbol{\xi}\|^2)^{-s} \leq \hat{r}(\boldsymbol{\xi}) \leq C_2(1 + \|\boldsymbol{\xi}\|^2)^{-s}, \tag{2.31}$$

with $s > d/2$ and two positive constants $0 < C_1 \leq C_2$. The Matérn covariance (2.26) with smoothness $\nu$ satisfies this regularity condition with $s = \nu + d/2$ (see Theorem 10.12 of Wendland (2004)). However, Tensor-product covariance functions such as (2.27) do not satisfy this condition (see Ritter (2000) in Chapter 7).

If the stationary kernel satisfies (2.31), then the induced RKHS $\mathcal{H}$ is the Sobolev space $\mathcal{H}^s(\mathbb{R}^d)$ of order $s = \nu + d/2$ and the RKHS norm is equivalent to the Sobolev norm $\|f\|_{\mathcal{H}_2^s} = \|\hat{f}(\cdot)(1 + \|\cdot\|^2)^{s/2}\|_{L^2}$. For this result, we refer to Wendland (2004) in Corollary 10.13.

**Remark 2.2.29.** *The norm-equivalence is useful for inferring and studying the hyperparameters' asymptotic bounds (e.g. in Karvonen (2022)).*

Note that the realizations of a Gaussian process with covariance kernel $k$ do not belong to the RKHS ($\mathbb{P}(Y \in \mathcal{H}) = 0$) by Driscoll's theorem, see Lukić & Beder (2001). It is also possible to introduce Gaussian processes from the general theory of Gaussian measures. The RKHS is then known as the Cameron-Martin space, see Bogachev (1998).

In the following manuscript, we will consider in particular the Matérn anisotropic geometric model $\boldsymbol{k}_{\sigma^2,\boldsymbol{\theta}}^{\text{radial}}$ defined in (2.26), denoted simply by $\boldsymbol{k}$, as we have many theoretical and asymptotic results of Kriging models with anisotropic correlation kernel, we refer to the thesis of Muré (2018) for more details. In addition, this covariance model is available in many packages such ***kergp*** (Deville et al., 2019). Other Matérn covariance functions are also proposed in this package or in ***DiceKriging*** (Roustant et al., 2012).

## Gaussian Process regressor

We recall our initial setting as defined in the introduction. We consider $n$ observations of some unknown function $f$ (physical model, computer code, production system etc.). Each observation of the output corresponds to a $d$-dimensional input vector $\boldsymbol{x} = (x_1, \ldots, x_d) \in \mathcal{D}$. The $n$ points corresponding to the model are called an experimental design and are denoted as $\mathbf{X} = \left(\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(n)}\right)$ where $\boldsymbol{x}^{(i)} = (\boldsymbol{x}_1^{(i)}, \ldots, \boldsymbol{x}_d^{(i)}) \in \mathcal{D}$. The outputs will be denoted as $\boldsymbol{y} = (y_1, \ldots, y_n)$ with $y_i = f(\boldsymbol{x}^{(i)}) + \epsilon_i$.

**Definition 2.2.30** (Gaussian Process model (Rasmussen & Williams, 2005))**.** *The Gaussian Process model is a Bayesian non-parametric regression which assumes a prior distribution over the regression function $f$. In particular, it assumes a Gaussian process prior with some given mean and covariance functions. This prior is updated and converted into a posterior over functions once some data points have been observed.*

In our case, we assume that mean function has the form

$$m(\boldsymbol{x}) = \sum_{j=0}^{p-1} \beta_j f_j(\boldsymbol{x}) = \boldsymbol{f}_{\text{trend}}(\boldsymbol{x})^\top \boldsymbol{\beta}, \tag{2.32}$$

where $f_j, j = 0, \ldots, p-1$ are some predefined functions and $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_{p-1})$ are the regression coefficients.

We assume also that the covariance function satisfies, for $i, j \in \{1, \ldots, n\}$,

$$\boldsymbol{k}(\boldsymbol{x}^{(i)}, \boldsymbol{x}^{(j)}) + \sigma_\epsilon^2 \mathbf{1}_{\{i=j\}}. \tag{2.33}$$

**Remark 2.2.31.** *Definition 2.2.30 is more relevant in the Machine Learning community. Another definition by Sacks et al. (1989) is also commonly used by geostatisticians and computer experiments community. It states that the Gaussian Process modeling treats the response $f(\boldsymbol{x})$ as a realization of a random stochastic process $\xi(\boldsymbol{x})$, for $\boldsymbol{x}$ in $\mathcal{D}$, in the space $(\Omega, \mathcal{F}, \mathbb{P})$ such that*

$$\xi(\boldsymbol{x}) = m(\boldsymbol{x}) + Z(\boldsymbol{x}), \tag{2.34}$$

*where $Z(\boldsymbol{x})$ is a zero-mean stationary Gaussian Process such that*

$$\text{Cov}[Z(\boldsymbol{x}), Z(\boldsymbol{x}')] = \boldsymbol{k}(\boldsymbol{x}, \boldsymbol{x}') + \sigma_\epsilon^2 \mathbf{1}\{\boldsymbol{x} = \boldsymbol{x}'\} \quad \forall \boldsymbol{x}, \boldsymbol{x}' \in \mathcal{D} \times \mathcal{D}. \tag{2.35}$$

$\sigma_\epsilon^2 \geq 0$ is called the nugget effect (Matheron, 1970). It is common to assume that $\sigma_\epsilon^2 = 0$ in computer models because they are assumed to be deterministic, and the output is noise-free. When $f(\boldsymbol{x})$ is observed, repeated simulations at the same point $\boldsymbol{x}$ should produce the same result.

However, in many cases, the assumption of a noise-free model is not feasible. One reason is that the output has an aleatoric uncertainty due to measurement error (as described in Section 2.1). The second reason is the theoretical aspects of smoothness and derivatives of the output. The other reasons are more computational and linked to the numerical stability of zero-nugget models. The presence of the nugget effect in the Gaussian Processes model has been studied in many works (Andrianakis & Challenor, 2012; de Oliveira, 2007; Pepelyshev, 2010), and we refer to these works to understand the effect of the nugget on the likelihood function and the predictions made with the Gaussian Process model. There are three sub-cases of Kriging, depending on the assumption made on the existing knowledge about the model $f$:

- The Simple Kriging: $m$ is assumed to be known, usually null $m = 0$. Equivalently, when working in the simple Kriging framework, we will consider a centered Gaussian process.
- The Ordinary Kriging: $m$ is assumed to be constant but unknown.
- The Universal Kriging: $m$ is assumed to be of the form $\sum_{j=0}^{p-1} \beta_j f_j(x)$, where $f_j$ are predefined (e.g. affine functions $f_0(\boldsymbol{x}) = 1$ or monomial functions of degree less than one $f_j(\boldsymbol{x}) = x_j, j = 1, \ldots, p-1$) and unknown scalar coefficients $\beta_j$.

**Assumption 2.2.32.** *In the case of ordinary or universal kriging, we assume that $n \geq p$, $\mathbf{F}$ is a full rank matrix, and $\mathbf{e} \in \text{Im } \mathbf{F}$ where $\mathbf{e} = (1, \ldots, 1)^\top$.*

Assumption 2.2.32 is reasonable. Indeed, in the Ordinary Kriging, this assumption is always satisfied. In the Universal Kriging, the assumption $\mathbf{e} \in \text{Im } \mathbf{F}$ is satisfied as soon as the constant function $f_0(\boldsymbol{x}) = C$ is included in the chosen family of functions $f_j$.

**Remark 2.2.33.** *We require $\mathbf{F}$ to a full rank matrix in order to ensure that $\mathbf{F}^\top \mathbf{F}$ is non-singular.*

The regression parameters $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_{p-1})$ are subject to an estimation by Generalized Least Squares (GLS), see Section 2.3.

### Joint and conditional predictive distribution

Under the hypothesis of the Gaussian Process model (2.2.30) and given $\boldsymbol{\beta}$ the regression coefficients, $(\sigma^2, \boldsymbol{\theta})$ the hyperparameters of the covariance function $\boldsymbol{k}$ and $\sigma_\epsilon^2$ the nugget effect, then, for all $i = 1, \ldots, n$, the output $Y(\boldsymbol{x}^{(i)})$ corresponding to the point $\boldsymbol{x}^{(i)} \in \mathbf{X}$ is Gaussian

$$Y(\boldsymbol{x}^{(i)}) \mid \boldsymbol{\beta}, \sigma^2, \boldsymbol{\theta}, \sigma_\epsilon^2 \sim \mathcal{N}(\boldsymbol{f}_{\text{trend}}(\boldsymbol{x}^{(i)})^\top \boldsymbol{\beta}, \sigma^2 + \sigma_\epsilon^2), \tag{2.36}$$

where $\boldsymbol{f}_{\text{trend}}(\boldsymbol{x}) = (f_j(\boldsymbol{x}))_{j=0}^{p-1}$, and $\text{Cov}[Y(\boldsymbol{x}^{(i)}), Y(\boldsymbol{x}^{(j)})] = \boldsymbol{k}(\boldsymbol{x}^{(i)}, \boldsymbol{x}^{(j)}) + \mathbf{1}_{\{i=j\}} \sigma_\epsilon^2$ for $i, j = 1, \ldots, n$.

As a result, the *prior* distribution of $\boldsymbol{Y} = \big(Y(\boldsymbol{x}^{(1)}), ..., Y(\boldsymbol{x}^{(n)})\big)$ on the learning sample $\mathbf{X}$ is multivariate Gaussian

$$\boldsymbol{Y} \mid \boldsymbol{\beta}, \sigma^2, \boldsymbol{\theta}, \sigma_\epsilon^2 \sim \mathcal{N}(\mathbf{F}\boldsymbol{\beta}, \mathbf{K}), \tag{2.37}$$

where:

- $\mathbf{F} = (F_{ij}) \in \mathbb{R}^{n \times p}$ is the regression matrix such that $F_{ij} = f_{j-1}(\boldsymbol{x}^{(i)})$.
- $\boldsymbol{\beta} = \{\beta_0, \ldots, \beta_{p-1}\}^\top \in \mathbb{R}^p$ are the regression coefficients when the kriging frame is specified.
- $\mathbf{K} = \left(\boldsymbol{k}(\boldsymbol{x}^{(i)}, \boldsymbol{x}^{(j)})\right)_{1 \leq i,j \leq n} + \sigma_\epsilon^2 \, \mathbf{I}_n \in \mathbb{R}^{n \times n}$ is the covariance matrix of the learning design $\mathbf{X}$.

Using this result, we want to predict $Y_{\text{new}} = Y(\boldsymbol{x}_{\text{new}})$, the output at a new point $\boldsymbol{x}_{\text{new}} = (x_{\text{new},1}, \ldots, x_{\text{new},d}) \in \mathcal{D}$. The joint probability distribution of $(\boldsymbol{Y}, Y_{\text{new}})$ is given by:

$$\begin{bmatrix} \boldsymbol{Y} \\ Y_{\text{new}} \end{bmatrix} \sim \mathcal{N}\left( \begin{bmatrix} \mathbf{F}\boldsymbol{\beta} \\ \boldsymbol{f}_{\text{trend}}(\boldsymbol{x}_{\text{new}})^\top \boldsymbol{\beta} \end{bmatrix}, \begin{bmatrix} \mathbf{K} & \boldsymbol{k}(\mathbf{X}, \boldsymbol{x}_{\text{new}}) \\ \boldsymbol{k}(\mathbf{X}, \boldsymbol{x}_{\text{new}})^\top & \boldsymbol{k}(\boldsymbol{x}_{\text{new}}, \boldsymbol{x}_{\text{new}}) + \sigma_\epsilon^2 \end{bmatrix} \right), \tag{2.38}$$

where $\boldsymbol{k}(\boldsymbol{x}_{\text{new}}, \mathbf{X}) = \left(\boldsymbol{k}(\boldsymbol{x}_{\text{new}}, \boldsymbol{x}^{(i)})\right)_{1 \leq i \leq n} \in \mathbb{R}^n$ is the cross-covariance vector and $\boldsymbol{f}_{\text{trend}}(\boldsymbol{x}_{\text{new}}) = (f_j(\boldsymbol{x}_{\text{new}}))_{j=0}^{p-1}$ the regression trend vector at $\boldsymbol{x}_{\text{new}}$.

By the Gaussian conditioning theorem (2.2.6), it can be shown that the conditional distribution of $Y_{\text{new}}$ is also Gaussian:

$$Y_{\text{new}} = Y(\boldsymbol{x}_{\text{new}}) \mid \mathbf{X}, \boldsymbol{y}, \boldsymbol{\beta}, \sigma^2, \boldsymbol{\theta}, \sigma_\epsilon^2 \sim \mathcal{N}\left(\tilde{y}(\boldsymbol{x}_{\text{new}}), \tilde{\sigma}^2(\boldsymbol{x}_{\text{new}})\right), \tag{2.39}$$

where $\tilde{y}(\boldsymbol{x}_{\text{new}})$ and $\tilde{\sigma}^2(\boldsymbol{x}_{\text{new}})$ are the predictive mean and variance at the new point $\boldsymbol{x}_{\text{new}}$. In the case of Ordinary or Universal Kriging, $\tilde{y}(\boldsymbol{x}_{\text{new}})$ and $\tilde{\sigma}^2(\boldsymbol{x}_{\text{new}})$ are given

$$\tilde{y}_{\sigma^2, \boldsymbol{\theta}, \sigma_\epsilon^2}(\boldsymbol{x}_{\text{new}}) = \boldsymbol{f}_{\text{trend}}(\boldsymbol{x}_{\text{new}})^\top \boldsymbol{\beta} + \boldsymbol{k}(\boldsymbol{x}_{\text{new}}, \mathbf{X})^\top \mathbf{K}^{-1}(\boldsymbol{y} - \mathbf{F}\boldsymbol{\beta}), \tag{2.40}$$

$$\tilde{\sigma}^2_{\sigma^2, \boldsymbol{\theta}, \sigma_\epsilon^2}(\boldsymbol{x}_{\text{new}}) = \boldsymbol{k}(\boldsymbol{x}_{\text{new}}, \boldsymbol{x}_{\text{new}}) + \sigma_\epsilon^2 - \boldsymbol{k}(\boldsymbol{x}_{\text{new}}, \mathbf{X})^\top \mathbf{K}^{-1} \; \boldsymbol{k}(\boldsymbol{x}_{\text{new}}, \mathbf{X}). \tag{2.41}$$

Hence, the Gaussian Process regression is a Bayesian non-parametric regression which assumes a GP prior over the regression functions (Rasmussen & Williams, 2005), which can be converted into a posterior over functions once some data has been observed. It consists in updating the prior distribution over $Y$ using a training set $\mathbf{D}$ of $n$ observations in order to predict $Y(\boldsymbol{x}_{\text{new}})$ at a new point $\boldsymbol{x}_{\text{new}}$.

The predictive mean $\tilde{y}_{\sigma^2, \boldsymbol{\theta}, \sigma_\epsilon^2}$ in (2.40), denoted now by $\tilde{y}$ without specifying its dependence on hyperparameters or the nugget effect, is used as a predictor of the mean value of $Y$ at $\boldsymbol{x}_{\text{new}}$. It has a regression part $\boldsymbol{f}_{\text{trend}}(\boldsymbol{x}_{\text{new}})^\top \boldsymbol{\beta} = \sum_{j=0}^{p-1} \beta_j f_j(\boldsymbol{x}_{\text{new}})$ and a local correction. Thus, it can be written as a linear combination of kernel functions, each one centered on a training point:

$$\begin{aligned} \tilde{y}(\boldsymbol{x}_{\text{new}}) &= f_{\text{trend}}(\boldsymbol{x}_{\text{new}})^\top \boldsymbol{\beta} + \boldsymbol{k}(\boldsymbol{x}_{\text{new}}, \mathbf{X})^\top \mathbf{K}^{-1}(\boldsymbol{y} - \mathbf{F}\boldsymbol{\beta}) \\ &= \sum_{j=0}^{p-1} \beta_j f_j(\boldsymbol{x}_{\text{new}}) + \sum_{i=1}^{n} \alpha_i k(\boldsymbol{x}^{(i)}, \boldsymbol{x}_{\text{new}}), \end{aligned} \tag{2.42}$$

where $\boldsymbol{\alpha} = \mathbf{K}^{-1}(\boldsymbol{y} - \mathbf{F}\boldsymbol{\beta})$. These coefficients $\alpha_i$ are updated each time a new observation is made (as opposed to the parameters of the kernel, referred to as *hyperparameters*), which are not updated once training is over (see Section 2.3).

**Remark 2.2.34.** *The kernel part of prediction function* $\boldsymbol{x}_{\text{new}} \mapsto \sum_{i=1}^{n} \alpha_i \boldsymbol{k}(\boldsymbol{x}^{(i)}, \boldsymbol{x}_{\text{new}})$ *vanishes when* $\boldsymbol{x}_{\text{new}}$ *is far from the observation points* $(\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(n)})$*. Therefore, the kriging model is essentially used for interpolation and prediction.*

**Remark 2.2.35.** *When the model is noise-free* $\sigma_\epsilon^2 = 0$ *and if* $\boldsymbol{x}_{\text{new}} = \boldsymbol{x}^{(i)}$ *for some* $i$*, then* $\tilde{y}(\boldsymbol{x}_{\text{new}}) = y_i$ *and* $\tilde{\sigma}^2(\boldsymbol{x}_{\text{new}}) = 0$*. This result is expected because the prediction mean of an observed value is the value itself and the predictive variance correspond only to the measurement error. We say that the GP model interpolates the experimental design* $\mathbf{X}$*.*

*However, when there is a nugget effect* $\sigma_\epsilon^2 > 0$*, the Gaussian Process model does not interpolate the data* $\boldsymbol{y}$*. It approximates them as good as possible with the Mean Squared Error (MSE) and attaches a positive uncertainty bound around them. (Andrianakis & Challenor, 2012). Furthermore, the leverage of the nugget effect is also investigated by Bostanabad et al. (2018) to train the GP model and estimate the optimal hyperparameters efficiently.*

The variance formula in (2.41) corresponds to the uncertainty of the predictor and is also known as the kriging variance $\tilde{\sigma}^2$. It gives a local indicator of the prediction accuracy.

We note here that the predictive mean and variance as defined in (2.40) and (2.41) assume a complete knowledge about the regression coefficients $\boldsymbol{\beta}$. In other terms, we treat $\boldsymbol{\beta}$ as a deterministic vector, and we plug it in directly in the formulas of predictive mean and variance. However, as we will see in Section 2.3, the regression coefficients $\boldsymbol{\beta}$ are estimated via the GLS method, so they are treated as a random variable with a given mean $\widehat{\boldsymbol{\beta}} = \mathbb{E}(\boldsymbol{\beta}) = \left(\mathbf{F}^\top \mathbf{K}^{-1} \mathbf{F}\right)^{-1} \mathbf{F}^\top \mathbf{K}^{-1} \boldsymbol{y}$ and a given covariance $\text{Cov}(\boldsymbol{\beta}) = \left(\mathbf{F}^\top \mathbf{K}^{-1} \mathbf{F}\right)^{-1}$. We do not present the proofs of this estimation, but we refer to Santner et al. (2003) for further details of $\widehat{\boldsymbol{\beta}}$.

**Definition 2.2.36** (The Best Linear Unbiased Predictor)**.** *Let* $\boldsymbol{Y} = (Y(\boldsymbol{x}^{(1)}), ..., Y(\boldsymbol{x}^{(n)}))$*, we say that* $\widehat{Y}(\cdot)$ *is the Best Linear Unbiased Predictor (BLUP) of* $Y(\cdot)$ *if it satisfies the following:*

- $\widehat{Y}$ *is linear i.e.* $\widehat{Y}(\boldsymbol{x}) = v(\boldsymbol{x})^\top \boldsymbol{Y}$ *for a vector* $v(\boldsymbol{x}) = (v_1(\boldsymbol{x}), \ldots, v_n(\boldsymbol{x}))^\top \in \mathbb{R}^n$*.*
- $\widehat{Y}$ *is unbiased estimator of* $Y$ *i.e.* $\mathbb{E}_{\pi_{Y(\cdot)}}[\widehat{Y}(\boldsymbol{x}) - Y(\boldsymbol{x})] = 0$ *for fixed* $\boldsymbol{x} \in \mathcal{D}$ *where* $\pi_{Y(\cdot)}$ *is the distribution of the process* $Y(\cdot)$*.*
- $\widehat{Y}$ *is the best in the Mean Squared Error sense i.e.* $\widehat{Y}(\boldsymbol{x}) = (v^*(\boldsymbol{x}))^\top \boldsymbol{Y}$ *with* $v^*(\boldsymbol{x}) = \arg\min_{v} \mathbb{E}_{\pi_{Y(\cdot)}}\left[(v(\boldsymbol{x})^\top \boldsymbol{Y} - Y(\boldsymbol{x}))^2\right]$*.*

The BLUP of the Gaussian Process model has been derived by Sacks et al. (1989). The mean prediction of the BLUP is given by:

$$\tilde{y}_{\sigma^2, \boldsymbol{\theta}, \sigma_\epsilon^2}(\boldsymbol{x}_{\text{new}}) = \boldsymbol{f}_{\text{trend}}(\boldsymbol{x}_{\text{new}})^\top \widehat{\boldsymbol{\beta}} + \boldsymbol{k}(\boldsymbol{x}_{\text{new}}, \mathbf{X})^\top \mathbf{K}^{-1} (\boldsymbol{y} - \mathbf{F}\widehat{\boldsymbol{\beta}}). \tag{2.43}$$

The mean square error of the BLUP satisfies:

$$\begin{aligned} \tilde{\sigma}^2_{\sigma^2, \boldsymbol{\theta}, \sigma_\epsilon^2}(\boldsymbol{x}_{\text{new}}) = \boldsymbol{k}(\boldsymbol{x}_{\text{new}}, \boldsymbol{x}_{\text{new}}) + \sigma_\epsilon^2 - \boldsymbol{k}(\boldsymbol{x}_{\text{new}}, \mathbf{X})^\top \mathbf{K}^{-1}\ \boldsymbol{k}(\boldsymbol{x}_{\text{new}}, \mathbf{X}) + (\boldsymbol{f}_{\text{trend}}(\boldsymbol{x}_{\text{new}}) - \\ \mathbf{F}\mathbf{K}^{-1}\boldsymbol{k}(\boldsymbol{x}_{\text{new}}, \mathbf{X})\Big)^\top \left(\mathbf{F}^\top \mathbf{K}^{-1} \mathbf{F}\right)^{-1} \left(\boldsymbol{f}_{\text{trend}}(\boldsymbol{x}_{\text{new}}) - \mathbf{F}\ \mathbf{K}^{-1}\boldsymbol{k}(\boldsymbol{x}_{\text{new}}, \mathbf{X})\right). \end{aligned} \tag{2.44}$$

We refer to Santner et al. (2003) in Chapter 4 for a detailed proof of Equations (2.43) and (2.44). In particular, we note that the predictive variance of the BLUP considers an additional

non-negative term. This term is due to the propagation of the non-informative improper form of the prior distribution on the estimation of $\boldsymbol{\beta}$. We also note that the BLUP of the Gaussian process model does its best to be optimal (in the Mean Squared Error sense) even if we do not assume a Gaussian distribution over $\boldsymbol{y}$.

Given a GP regression model and a point $\boldsymbol{x}_{\text{new}} \in \mathcal{D}$, the *posterior* predictive distribution (2.39) can be standardized into:

$$\tilde{Z}(\boldsymbol{x}_{\text{new}}) = \frac{Y(\boldsymbol{x}_{\text{new}}) - \tilde{y}(\boldsymbol{x}_{\text{new}})}{\tilde{\sigma}(\boldsymbol{x}_{\text{new}})} \mid \mathbf{X}, \boldsymbol{y}, \boldsymbol{\beta}, \sigma^2, \boldsymbol{\theta}, \sigma_\epsilon^2 \sim \mathcal{N}(0,1)\,. \tag{2.45}$$

The variable $\tilde{Z}(\boldsymbol{x}_{\text{new}})$ follows the standardized Gaussian distribution. Therefore, for a given confidence level $1-\alpha$, the Prediction Interval $\mathcal{PI}_{1-\alpha}$ can be build directly by considering the quantiles $q_{1-\alpha/2} = \boldsymbol{\Phi}^{-1}(1-\alpha/2)$ and $q_{\alpha/2} = \boldsymbol{\Phi}^{-1}(\alpha/2) = -q_{1-\alpha/2}$ where $\boldsymbol{\Phi}$ is the CDF of the standard normal distribution

$$\mathcal{PI}_{1-\alpha}(\boldsymbol{x}_{\text{new}}) = \left[\tilde{y}(\boldsymbol{x}_{\text{new}}) - q_{1-\alpha/2} \times \tilde{\sigma}(\boldsymbol{x}_{\text{new}});\ \tilde{y}(\boldsymbol{x}_{\text{new}}) + q_{1-\alpha/2} \times \tilde{\sigma}(\boldsymbol{x}_{\text{new}})\right], \tag{2.46}$$

which gives a natural definition for the mappings $u_{1-\alpha}, l_{1-\alpha} : \mathcal{D} \mapsto \mathbb{R}$ (see Definition 2.1.5) characterizing Prediction Intervals as:

$$l_{1-\alpha}(\boldsymbol{x}_{\text{new}}) = \tilde{y}(\boldsymbol{x}_{\text{new}}) - q_{1-\alpha/2} \times \tilde{\sigma}(\boldsymbol{x}_{\text{new}}), \tag{2.47}$$

$$u_{1-\alpha}(\boldsymbol{x}_{\text{new}}) = \tilde{y}(\boldsymbol{x}_{\text{new}}) + q_{1-\alpha/2} \times \tilde{\sigma}(\boldsymbol{x}_{\text{new}}). \tag{2.48}$$

In particular, for the confidence level $(1-\alpha) = 95\%$, the corresponding Prediction Intervals are:

$$\mathcal{PI}_{1-\alpha}(\boldsymbol{x}_{\text{new}}) = \left[\tilde{y}(\boldsymbol{x}_{\text{new}}) - 1.96 \times \tilde{\sigma}(\boldsymbol{x}_{\text{new}});\ \tilde{y}(\boldsymbol{x}_{\text{new}}) + 1.96 \times \tilde{\sigma}(\boldsymbol{x}_{\text{new}})\right]. \tag{2.49}$$

It follows that this *plug-in* interval is an exact type II Prediction Interval

$$\mathbb{P}_\pi\left(Y(\boldsymbol{x}) \in \mathcal{PI}_{1-\alpha}(\boldsymbol{x}) \mid \mathbf{D}\right) = 1-\alpha, \tag{2.50}$$

where $\pi$ is the posterior distribution of $Y(\boldsymbol{x})$ given $\mathbf{D}$ and $\boldsymbol{x} \in \mathcal{D}$ is a point sampled according to the distribution $\pi_{\boldsymbol{X}}$.

**Remark 2.2.37.** *The Prediction Intervals in (2.46) are in fact type IV Prediction Intervals, that is, for the posterior distribution of $Y(\cdot)$ given $\mathbf{D}$, they satisfy the propriety*

$$\mathbb{P}_{\pi_{Y(\cdot)|\mathbf{D}}}\left(Y(\boldsymbol{X}) \in \mathcal{PI}_{1-\alpha}(\boldsymbol{X}) \mid \mathbf{D}, \boldsymbol{X} = \boldsymbol{x}\right) = 1-\alpha, \tag{2.51}$$

*which is much stronger than Type II Coverage.*

The most outstanding advantage of the GP model compared to other models comes from the previous equations. In fact, Kriging model provides a mathematical formula for the distribution of the output variable at an arbitrary new point $\boldsymbol{x}_{\text{new}}$, given by (2.43), (2.44) and (2.46). This distribution formula can be used in a wide variety of applications such as time series modelling (Roberts et al., 2013), sensitivity analysis (Le Gratiet et al., 2017; Paananen et al., 2019), uncertainty quantification (Teimouri et al., 2017), quantile evaluation (Oakley et al., 2004) as well as the estimation of functional risk Curves Iooss & Le Gratiet (2019). Other possible extensions of GP modelling can also be found in (Currin et al., 1991; Rasmussen & Williams, 2005).

## 2.3 Estimating GP model parameters and hyper-parameters

Defining a GP model and computing the kriging mean and variance as shown in (2.40) and (2.41) requires the estimation of the regression coefficients, the covariance hyperparameters $(\sigma^2, \boldsymbol{\theta})$ as well as the nugget effect $\sigma_\epsilon^2$. In practice, we do not know none of the quantities, and we need to estimate them from the training dataset $\mathbf{D} = \{(\boldsymbol{x}^{(i)}, y_i)\}_{i=1}^n$.

### Estimating the regression coefficients

The regression parameters $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_{p-1})$ are subject to an estimation by Generalized Least Squares (GLS). Given the covariance hyperparameters $(\sigma^2, \boldsymbol{\theta})$ and the nugget effect $\sigma_\epsilon^2$, the generalized least squares regression weights $\widehat{\boldsymbol{\beta}}$ satisfy:

$$\widehat{\boldsymbol{\beta}} = \left(\mathbf{F}^\top \mathbf{K}^{-1} \mathbf{F}\right)^{-1} \mathbf{F}^\top \mathbf{K}^{-1} \boldsymbol{y}. \tag{2.52}$$

We refer to Sacks et al. (1989) and Cox (2004) for the proof of this formula.

### Estimating the covariance hyperparameters by Maximum Likelihood

Given a Gaussian Process model *i.e.* $\boldsymbol{Y} \mid \boldsymbol{\beta}, \sigma^2, \boldsymbol{\theta}, \sigma_\epsilon^2, \sim \mathcal{N}(\mathbf{F}\boldsymbol{\beta}, \mathbf{K})$, the likelihood function of $\boldsymbol{y}$ is given by the probability density function (pdf) of the Multivariate Gaussian distribution (2.18)

$$\ell(\boldsymbol{y} \mid \boldsymbol{\beta}, \sigma^2, \boldsymbol{\theta}, \sigma_\epsilon^2) = (2\pi)^{-n/2} (\det \mathbf{K})^{-1/2} \exp\left(-\frac{1}{2}(\boldsymbol{y} - \mathbf{F}\boldsymbol{\beta})^\top \mathbf{K}^{-1} (\boldsymbol{y} - \mathbf{F}\boldsymbol{\beta})\right). \tag{2.53}$$

The Maximum likelihood estimation (Mardia & Marshall, 1984; Stein, 1999) is a common method used to select the hyperparameters $(\sigma^2, \boldsymbol{\theta})$ within a family of parameterized covariance functions $\mathcal{K} = \{\boldsymbol{k}_{(\sigma^2, \boldsymbol{\theta})}; (\sigma^2, \boldsymbol{\theta}) \in \mathbb{R}^+ \times (0, +\infty)^d\}$. By maximizing the likelihood, this method seeks to find the optimal mean vector $\mathbf{F}\boldsymbol{\beta}$ and covariance matrix $\mathbf{K}$ so that the optimized model produces the observed data with the highest probability.

The negative log-likelihood (Santner et al., 2003; Stein, 1999) of the data $\boldsymbol{y}$ given $(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\theta}, \sigma_\epsilon^2)$ is

$$-\log \ell\left(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\theta}, \sigma_\epsilon^2 \mid \boldsymbol{y}\right) = \frac{n}{2}\log(2\pi) + \frac{1}{2}\log(\det \mathbf{K}) + \frac{1}{2}(\boldsymbol{y} - \mathbf{F}\boldsymbol{\beta})^\top \mathbf{K}^{-1} (\boldsymbol{y} - \mathbf{F}\boldsymbol{\beta}). \tag{2.54}$$

For a given nugget effect $\sigma_\epsilon^2$, if we replace $\boldsymbol{\beta}$ with the GLS formulas $\widehat{\boldsymbol{\beta}}$ and if we use the facts

$$\frac{\partial \mathbf{K}^{-1}}{\partial \cdot} = -\mathbf{K}^{-1} \frac{\partial \mathbf{K}}{\partial \cdot} \mathbf{K}^{-1} \quad \text{and} \quad \frac{\partial \log(\det \mathbf{K})}{\partial \cdot} = \operatorname{Tr}\left(\mathbf{K}^{-1} \frac{\partial \mathbf{K}}{\partial \cdot}\right), \tag{2.55}$$

we get closed-form expressions for the gradient of the negative log-likelihood $-\log \ell$ with respect to $\sigma^2$ and $\boldsymbol{\theta}$

$$\frac{\partial(-\log \ell)}{\partial \cdot} = \frac{1}{2} \operatorname{Tr}\left(\mathbf{K}^{-1} \frac{\partial \mathbf{K}}{\partial \cdot}\right) - \frac{1}{2} \boldsymbol{y}^\top \overline{\mathbf{K}}^{-1} \frac{\partial \mathbf{K}}{\partial \cdot} \overline{\mathbf{K}}^{-1} \boldsymbol{y}, \tag{2.56}$$

where $\overline{\mathbf{K}}$ is the matrix defined by

$$\overline{\mathbf{K}} = \mathbf{K}^{-1} - \mathbf{K}^{-1}\mathbf{F}\left(\mathbf{F}^\top \mathbf{K}^{-1} \mathbf{F}\right)^{-1} \mathbf{F}^\top \mathbf{K}^{-1}. \tag{2.57}$$

The proof can be found in (Mardia & Marshall, 1984) and (Bachoc, 2013).

Unfortunately, setting the gradient to zero gives some expressions that cannot be solved analytically. Thus, the likelihood can be optimized using standard numerical methods. Therefore, the Maximum Likelihood Estimator (MLE) $(\hat{\sigma}^2_{ML}, \hat{\boldsymbol{\theta}}_{ML})$ of $(\sigma^2, \boldsymbol{\theta})$ is given by a numerical optimization of

$$(\hat{\sigma}^2_{ML}, \hat{\boldsymbol{\theta}}_{ML}) \in \operatorname{argmin}_{\sigma^2,\boldsymbol{\theta}} \; \boldsymbol{y}^\top \overline{\mathbf{K}} \boldsymbol{y} + \log\left(\det \mathbf{K}\right). \tag{2.58}$$

Once the covariance hyperparameters $(\hat{\sigma}^2_{ML}, \hat{\boldsymbol{\theta}}_{ML})$ are determined by Maximum Likelihood estimator, the estimator $\widehat{\boldsymbol{\beta}}$ is updated using the Generalized Least Squares formulas as shown in (2.52):

$$\widehat{\boldsymbol{\beta}}_{ML} = \left(\mathbf{F}^\top \mathbf{K}_{ML}^{-1} \mathbf{F}\right)^{-1} \mathbf{F}^\top \mathbf{K}_{ML}^{-1} \boldsymbol{y}, \tag{2.59}$$

where $\mathbf{K}_{ML} = \left(\boldsymbol{k}_{\hat{\sigma}^2_{ML},\hat{\boldsymbol{\theta}}_{ML}}(\boldsymbol{x}^{(i)}, \boldsymbol{x}^{(j)})\right)_{1\leq i,j\leq n} + \sigma_\epsilon^2 \; \mathbf{I}_n$.

**Remark 2.3.1.** *In the absence of the nugget effect, we have $\mathbf{K} = \sigma^2 \mathbf{R}_{\boldsymbol{\theta}}$ where $\mathbf{R}_{\boldsymbol{\theta}}$ is called the auto-correlation matrix. The negative log-likelihood has now the form (Santner et al., 2003):*

$$-\log \ell\left(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\theta} | \boldsymbol{y}\right) = \frac{n}{2}\log(2\pi) + \frac{n}{2}\log(\sigma^2) + \frac{1}{2\sigma^2}(\boldsymbol{y} - \mathbf{F}\boldsymbol{\beta})^\top \mathbf{R}_{\boldsymbol{\theta}}^{-1}(\boldsymbol{y} - \mathbf{F}\boldsymbol{\beta}) + \frac{1}{2}\log(\det \mathbf{R}_{\boldsymbol{\theta}}). \tag{2.60}$$

*We see clearly that the new expression of $-\log \ell$ separates the covariance hyperparameters and makes the maximum likelihood estimator of $\widehat{\sigma}_{ML}$ explicit:*

$$\hat{\sigma}^2_{ML} = \frac{1}{n}(\boldsymbol{y} - \mathbf{F}\widehat{\boldsymbol{\beta}}_{ML})^\top \mathbf{R}_{\boldsymbol{\theta}}^{-1}(\boldsymbol{y} - \widehat{\boldsymbol{\beta}}_{ML}), \tag{2.61}$$

*where*

$$\widehat{\boldsymbol{\beta}}_{ML} = \left(\mathbf{F}^\top \mathbf{R}_{\boldsymbol{\theta}}^{-1} \mathbf{F}\right)^{-1} \mathbf{F}^\top \mathbf{R}_{\boldsymbol{\theta}}^{-1} \boldsymbol{y}. \tag{2.62}$$

*Thereby, as the variance $\widehat{\sigma}^2_{ML}$ and the regression coefficients $\widehat{\boldsymbol{\beta}}_{ML}$ depend now on the correlation length-vector $\boldsymbol{\theta}$, we can substitute them into the negative log-likelihood $-\log \ell$. Thus, maximum likelihood estimation $\hat{\boldsymbol{\theta}}_{ML}$ of $\boldsymbol{\theta}$ consists in numerical optimization of the function*

$$\widehat{\boldsymbol{\theta}}_{ML} \in \operatorname{argmin}_{\boldsymbol{\theta}} -\log \tilde{\ell}(\boldsymbol{\theta}) = \log\left(\hat{\sigma}^2_{ML}(\boldsymbol{\theta})\right) + \frac{1}{n}\log\left(\det \mathbf{R}_{\boldsymbol{\theta}}\right). \tag{2.63}$$

We note that Minimizing function $-\log \ell$ in (2.54) and (2.60) is an heavy optimization problem. The computational cost for calculating a likelihood criterion and its gradient is $O(n^3)$. Some additional difficulties are also raised. The large number of parameters imposes the use of a sequential method of resolution. Moreover, the non-convexity of the objective function requires an exploratory algorithm (stochastic gradient, multi-start etc.) able to explore the domain in an optimal way (Marrel et al., 2008).

It has been shown that the Maximum Likelihood method is optimal when the covariance function is well-specified (Bachoc, 2013). In this case, the predictive posterior distribution in (2.39) is fully characterized for any new point $\boldsymbol{x}_{\text{new}}$. We give below the definition of a well-specified model.

**Definition 2.3.2** (Well-specified model)**.** *Let $\mathcal{K} = \{\boldsymbol{k}_{(\sigma^2,\boldsymbol{\theta})};\ (\sigma^2, \boldsymbol{\theta}) \in \mathbb{R}^+ \times (0, +\infty)^d\}$ be a family of covariance function in $\mathbb{R}^{d+1}$. The model is said to be well-specified if there exists a couple of hyperparameters $(\sigma_0^2, \boldsymbol{\theta}_0)$ such that $\boldsymbol{y}$ comes from a function $f$ that is a realization of a Gaussian Process model with covariance function $\boldsymbol{k}_{(\sigma_0^2,\boldsymbol{\theta}_0)} \in \mathcal{K}$.*

**Remark 2.3.3.** *In reality, we do not look for identifying the exact couple $(\sigma_0^2, \boldsymbol{\theta}_0)$. We rather say that the model is well-specified if the Leave-One-Out residuals are normally distributed (i.e. satisfies the normality assumption) given the obtained hyperparameters by MLE estimation method $(\hat{\sigma}^2_{ML}, \hat{\boldsymbol{\theta}}_{ML})$. However, we note that it is difficult to assess the normality assumption if the sample is too small.*

When the definition of well-specified model 2.3.2 is no more satisfied, we will say that the model is *misspecified*.

The well-posedness of Maximum Likelihood Estimation in the case of misspecified models was investigated in the literature, particularly noiseless data. On the one hand, Zhang (2004) show that the simultaneous estimation of the amplitude $\sigma^2$, length-scale $\theta$, and smoothness parameters $\nu$ of Matérn kernels does not identify the correct solution. Recently, Karvonen & Oates (2022) proves that the simultaneous Maximum Likelihood Estimation of both amplitude $\sigma^2$ and length-scale $\theta$ can be ill-posed. On the other hand, estimating only the amplitude $\sigma^2$ by Maximum Likelihood can provide significant adaptation against misspecification of the Gaussian process model as shown by Karvonen et al. (2020). Similarly, the estimation of the smoothness parameter $\nu$ is also shown to be consistent if the other hyperparameters remain fixed (Chen et al., 2021b).

### Learning the nugget effect

In this subsection, we consider the inference of the nugget effect $\sigma_\epsilon^2$. Indeed, the nugget effect is either known (which is rarely the case) or can be estimated by several approaches, including the Maximum Likelihood or the method proposed in Iooss & Marrel (2017) for instance.

The approach of Iooss & Marrel (2017), known as *GP joint modelling*, consists in a sequential building of two Gaussian Process models to fit the mean $Y_m$ and $Y_d$ dispersion (variance) components. These two components are used to estimate the nugget sequentially by targeting predictions errors. $Y_m$ and $Y_d$ are given by

$$\begin{aligned} Y_m(\boldsymbol{x}) &= \mathbb{E}(Y \mid \boldsymbol{X} = \boldsymbol{x}), \\ Y_d(\boldsymbol{x}) = \mathrm{Var}(Y|\boldsymbol{X} = \boldsymbol{x}) &= \mathbb{E}\left[(Y - Y_m(\boldsymbol{X}))^2 \mid \boldsymbol{X} = \boldsymbol{x}\right]. \end{aligned} \tag{2.64}$$

**Remark 2.3.4.** *In their original paper, Iooss & Marrel (2017) considered a subset $\mathbf{X}_{\text{exp}}$ of influential inputs variables while building these two models to reduce the complexity of the GP models. The subset $\mathbf{X}_{\text{exp}}$ can be obtained using a screening method (the Hilbert-Schmidt Independence Criterion (HSIC) (Gretton et al., 2007), for instance). In our case, we assume that the dimension $d$ is not large so that we can apply the GP joint modelling without screening.*

**Remark 2.3.5.** *The approach of Iooss & Marrel (2017) was mainly designed for heteroscedastic nugget effects. We slightly adapt this approach to homoscedastic nugget (modifications in brown).*

At a given iteration $j$, a first Gaussian Process model, denoted $\mathcal{GP}^j_{m,1}$, is built using the covariance function $\boldsymbol{k}$ with homoscedastic nugget effect (learned by Maximum Likelihood) to

fit $\boldsymbol{y}$ on the mean component. Then a second model, denoted $\mathcal{GP}^j_{d,1}$, is built for the variance component with the same covariance function $\boldsymbol{k}$ to fit the squared residuals $\boldsymbol{y}^2_{d,1} = (\boldsymbol{y} - \tilde{\boldsymbol{y}}_{m,1})^2$ where $\tilde{\boldsymbol{y}}_{m,1}$ is the mean predictions of $\mathcal{GP}^j_{m,1}$. Here, the model $\mathcal{GP}^j_{d,1}$ estimates the dispersion errors $\tilde{\boldsymbol{y}}_{d,1}$ at training points, it can be considered as the value of the heteroscedastic nugget effect and thus is updated in the covariance matrix $\mathbf{K}$. *If we decide to keep the assumption of a homoscedastic nugget effect, then we update the covariance matrix* $\mathbf{K}$ *by adding* $\tilde{\sigma}^2_\epsilon = \mathbb{E}_n(\tilde{\boldsymbol{y}}^2_{d,1})$, *where* $\mathbb{E}_n$ *denotes the empirical mean, to its diagonal terms.*

We repeat the same step by building additional models $\mathcal{GP}^j_{m,2}$ and $\mathcal{GP}^j_{d,2}$ on the mean and dispersion component and updating the estimated (heteroscedastic $\tilde{\boldsymbol{y}}_{d,2}$ or *homoscedastic* $\tilde{\sigma}^2_\epsilon = \mathbb{E}_n(\tilde{\boldsymbol{y}}^2_{d,2})$) nugget effect.

The final model $\mathcal{GP}^j$ is built with the updated nugget effect. Its hyperparameters are optimized by taking hyperparameters obtained at the $(j-1)^{th}$ iteration as starting point.

The *GP joint modelling* procedure of Iooss & Marrel (2017) (with the possible adjustment for homoscedastic nugget effect) can be summarized in the following algorithm:

**Algorithm 1** Sequential procedure of joint modeling

(0) Set $\mathbf{X}$, $\boldsymbol{y}$ and $\mathbf{K} = \boldsymbol{k}(\mathbf{X}, \mathbf{X})$ with default hyperparameters $(\sigma^2, \boldsymbol{\theta}) = (1, \ldots, 1)$ and $\boldsymbol{\beta} = \left(\mathbf{F}^\top \mathbf{K}^{-1} \mathbf{F}\right)^{-1} \mathbf{F}^\top \mathbf{K}^{-1} \boldsymbol{y}$.

**for** $j = 1, \ldots, m$ **do**

(1) Build a GP model $\mathcal{GP}^j_{m,1}$ with $\mathbf{X}$ to fit $\boldsymbol{y}$ and estimate $\tilde{\boldsymbol{y}}_{m,1}$ as the mean prediction of $\mathcal{GP}^j_{m,1}$.

(2) Build a GP model $\mathcal{GP}^j_{d,1}$ with $\mathbf{X}$ to fit $\left(\boldsymbol{y} - \boldsymbol{y}_{m,1}\right)^2$ and estimate $\tilde{\boldsymbol{y}}_{d,1}$ as the mean prediction of $\mathcal{GP}^j_{d,1}$.

(3) Update the covariance matrix $\mathbf{K} \leftarrow \mathbf{K} + \mathrm{Diag}(\tilde{\boldsymbol{y}}_{d,1})$ if assuming heteroscedastic nugget, *or* $\mathbf{K} \leftarrow \mathbf{K} + \tilde{\sigma}^2_\epsilon \mathbf{I}_n$ *if assuming homoscedastic nugget with* $\tilde{\sigma}^2_\epsilon = \mathbb{E}_n(\tilde{\boldsymbol{y}}^2_{d,1})$.

(4) Build a final GP model $\mathcal{GP}^j_{m,2}$ to with $\mathbf{X}$ and the new covariance matrix $\mathbf{K}$ with $\mathbf{X}$ and estimate $\boldsymbol{y}_{m,2}$ as the mean prediction of $\mathcal{GP}^j_{m,2}$.

(5) Repeat (2) and (3) using a GP model $\mathcal{GP}^j_{d,2}$ .

(6) Build a GP model $\mathcal{GP}^j$ with $\mathbf{X}$ to fit $\boldsymbol{y}$.

(7) Estimate the new hyperparameters $(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\theta})_j$ by taking $(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\theta})_{j-1}$ as starting point.

(8) Compute the model accuracy ${Q_j}^2$

$$Q^2_j = 1 - \frac{\sum_{i=1}^{n_{test}} \left(y_i - \hat{y}_i\right)^2}{\sum_{i=1}^{n_{\text{test}}} \left(y_i - \overline{y}\right)^2}.$$

**end for**

**Remark 2.3.6.** *It has been noticed empirically that two iterations* $j = 1, 2$ *are sufficient to estimate the nugget effect.*

**Remark 2.3.7.** *In our work, we do not need to compute the accuracy in step (8); we require only steps (0) to (7) and for* $j = 1$.

### Estimating the covariance hyperparameters by the full-Bayesian approach

In this subsection, we consider the full-Bayesian treatment of GP models (Williams & Barber, 1998). We recall the likelihood function of $\boldsymbol{y}$

$$\ell(\boldsymbol{y} \mid \boldsymbol{\beta}, \sigma^2, \boldsymbol{\theta}, \sigma_\epsilon^2) = (2\pi)^{-n/2} (\det \mathbf{K})^{-1/2} \exp\left(-\frac{1}{2}(\boldsymbol{y} - \mathbf{F}\boldsymbol{\beta})^\top \mathbf{K}^{-1} (\boldsymbol{y} - \mathbf{F}\boldsymbol{\beta})\right). \tag{2.65}$$

It has been shown by de Oliveira (2007) (and Berger et al. (2001) in noise-free case) that the *marginal* likelihood can be written as

$$\begin{aligned} \ell(\boldsymbol{y} \mid \sigma^2, \boldsymbol{\theta}, \sigma_\epsilon^2) &= \int \ell(\boldsymbol{y} \mid \boldsymbol{\beta}, \sigma^2, \boldsymbol{\theta}, \sigma_\epsilon^2)\, \mathrm{d}\boldsymbol{\beta} \\ &\propto |\mathbf{K}|^{-\frac{1}{2}} \left|\mathbf{F}^\top \mathbf{K}^{-1} \mathbf{F}\right|^{-\frac{1}{2}} \left(\widehat{\sigma}^2\right)^{-\left(\frac{n-p}{2}\right)}, \end{aligned} \tag{2.66}$$

where $\widehat{\sigma}^2 = \boldsymbol{y}^\top \overline{\mathbf{K}} \boldsymbol{y}$ and $\widehat{\boldsymbol{\beta}} = \left(\mathbf{F}^\top \mathbf{K}^{-1} \mathbf{F}\right)^{-1} \mathbf{F}^\top \mathbf{K}^{-1} \boldsymbol{y}$.

The nugget effect $\sigma_\epsilon^2$ is assumed to be estimated as described in the previous subsection. Consequently, the expression in (2.66) implies the *marginal* likelihood would have to be estimated jointly with $(\sigma^2, \boldsymbol{\theta})$ or be marginalized with respect to $(\sigma^2, \boldsymbol{\theta})$. Exceptionally in this subsection and for simplicity purposes, we denote the vector of GP hyperparameters $(\sigma^2, \boldsymbol{\theta})$ by $\Theta$ and we omit conditioning on $\sigma_\epsilon^2$.

The full-Bayesian analysis of the hyperparameters integrates the uncertainty and treats $\Theta$ as a random variable. In this method, the hyperparameters are considered as random and their posterior distribution is integrated in the predictive distribution.

We recall the Bayes's rule in Theorem 2.3.8 below:

**Theorem 2.3.8** (Bayes' Rule for parameters distribution)**.** *Let $\Theta$ be a random variable with a given probability distribution that best explains the observations $\boldsymbol{y}$, the Bayes' Rule assumes that:*

$$p(\Theta \mid \boldsymbol{y}) = \frac{p(\boldsymbol{y} \mid \Theta)\; p(\Theta)}{p(\boldsymbol{y})} \quad \textit{i.e.} \quad \textit{Posterior} = \frac{\textit{Likelihood} \times \textit{Prior}}{\textit{Evidence}}. \tag{2.67}$$

The full-Bayesian approach integrates the uncertainty about the unknown hyperparameters and assumes a *prior* on the hyperparameters $\Theta \sim \pi(\Theta)$. Therefore, the hyperparameters' posterior distribution satisfies, by Bayes' rule:

$$p(\Theta \mid \boldsymbol{y}) \propto \pi(\Theta) \ell(\boldsymbol{y} \mid \Theta), \tag{2.68}$$

where $\pi(\Theta)$ is the prior for hyperparameters and $\ell(\boldsymbol{y} \mid \Theta)$ is the *marginal* likelihood of $\boldsymbol{y}$ given the hyperparameters in (2.66).

Consequently, the probability density function of the *posterior* predictive distribution of $Y(\boldsymbol{x}_{\text{new}})$ at a new point $\boldsymbol{x}_{\text{new}}$ can be expressed as an integral over the hyperparameters:

$$p(y_{\text{new}} \mid \boldsymbol{y}) = \int p(y_{\text{new}} \mid \boldsymbol{y}, \Theta) p(\Theta \mid \boldsymbol{y})\, \mathrm{d}\Theta, \tag{2.69}$$

where $p(y_{\text{new}} \mid \boldsymbol{y}, \Theta)$ is the pdf of $Y(\boldsymbol{x}_{\text{new}})$ given the hyperparameters $\Theta$ in (2.39) and $p(\Theta \mid \boldsymbol{y})$ is the hyperparameters' posterior distribution given by (2.68).

The implementation of the full-Bayesian approach requires the evaluation of the previous integral and the posterior $p(\Theta \mid \boldsymbol{y})$, which is known up to a multiplicative constant. It is common to use Markov chain Monte Carlo (MCMC) methods (we refer to Robert & Casella (2004) for more detail about MCMC) for sampling and inference from the posterior distribution of the hyperparameters to overcome this issue, using, in particular, the Metropolis-Hastings (MH) algorithm (Robert & Casella, 2004) or Hamiltonian Monte Carlo (HMC) (Neal, 1993, 1996).

Therefore, the predictive distribution is obtained by Monte Carlo

$$p(y_{\text{new}} \mid \boldsymbol{y}) \simeq \frac{1}{N} \sum_{i=1}^{N} p(y_{\text{new}} \mid \boldsymbol{y}, \Theta_i), \tag{2.70}$$

where $N$ denotes the MCMC sample size and $\Theta_i$ is the $i$-th sample drawn from the posterior distribution $p(\Theta \mid \boldsymbol{y})$.

Finally, one can draw a sample $\left(Y_i(\boldsymbol{x}_{\text{new}})\right)_{i=1}^{N}$ of $Y(\boldsymbol{x}_{\text{new}})$ following the posterior distribution $p(y_{\text{new}} \mid \boldsymbol{y}, \Theta_i)$ as in (2.39) for each $i = 1, \ldots, N$. This sample is used to estimate either the empirical mean prediction $\tilde{y}(\boldsymbol{x}_{\text{new}})$ at $\boldsymbol{x}_{\text{new}}$ or Prediction Intervals $\mathcal{PI}_{1-\alpha}$ by taking the empirical quantiles of order $\alpha/2$ and $1-\alpha/2$ of the sample $\left(Y_i(\boldsymbol{x}_{\text{new}})\right)_{i=1}^{N}$.

**Remark 2.3.9.** *The MLE method can be seen as a plug-in approach that considers (2.69) and replaces $p(\sigma^2, \boldsymbol{\theta} \mid \boldsymbol{y})$ by a Dirac distribution centered on a value such as $(\hat{\sigma}^2_{ML}, \hat{\boldsymbol{\theta}}_{ML})$ that maximizes the likelihood function.*

To conclude, in contrast to *plug-in* approaches, the full-Bayesian considers the uncertainty of the hyperparameters and allows relevant results for estimating Prediction Intervals, but it also comes with a huge computational cost due mainly to the estimation of posterior distribution with MCMC.

### Covariance hyperparameters estimation by Cross-Validation

We have seen before that the Maximum Likelihood method fits well the data $\boldsymbol{y}$ when the model is well-specified (see Definition 2.3.2). However, in most cases, the covariance function is misspecified. The function $f$ is not, or does not seem to be a sample path of a Gaussian Process with covariance function $\boldsymbol{k}_{\hat{\sigma}^2_{ML}, \hat{\boldsymbol{\theta}}_{ML}}$. The Maximum Likelihood estimator may be less robust, and the obtained Gaussian Process model would perform poorly if asked to make new predictions for data it has not already observed. The problem of model misspecification raises the critical importance of an appropriate approach to learn and select optimal covariance hyperparameters that ensure a better point-wise prediction, whether the model is well-specified or not.

The Cross-Validation estimation, therefore, represents an alternative to estimate the hyperparameters $(\sigma^2, \boldsymbol{\theta})$ of the covariance function (Rasmussen & Williams, 2005). Indeed, the Cross-Validation is a practical tool for training models and assessing their predictive quality (Hastie et al. (2009) in chapter 7). It consists of leaving out some points in the dataset at a time and determining how well this data can be estimated from the remaining data for given hyperparameters, then finding the optimal hyperparameters that maximize the point-wise prediction of the Gaussian Process model. It has been shown in particular in (Bachoc, 2013)

that the Cross-Validation method is more efficient and robust when the covariance function is misspecified

In this section, we consider the same learning set of $n$ observations $\mathbf{D} = \{(\boldsymbol{x}^{(i)}, y_i)\}_{i=1}^n$. We assume that the value of the nugget effect $\sigma_\epsilon^2$ is known, and we do not consider the estimation of the regression coefficients $\boldsymbol{\beta}$ by the Cross-Validation method. We use the optimal GLS estimator of $\boldsymbol{\beta}$ in the following. We place us more precisely in the framework of the $n$-Cross-Validation, also known as the Leave-One-Out method. The following propositions and results are already drawn in the paper of Dubrule (1983) and later by Bachoc (2013); Zhang & Wang (2010) for point-wise prediction. For the multi-folds cross-validation, we refer to Dubrule (1983) for the Simple Kriging case, and Ginsbourger & Schärer (2021) for the Universal Kriging case.

For $i \in \{1, \ldots, n\}$, the Leave-One-Out method (i.e. $n$-Cross-Validation) consists in predicting $y_i$ by building a Gaussian Process model, denoted $\mathcal{GP}_{-i}$, when *virtually* removing $(\boldsymbol{x}^{(i)}, y_i)$ from the $\mathbf{D}$. The model $\mathcal{GP}_{-i}$ is trained on $\mathbf{D}_{-i} = \{(\boldsymbol{x}^{(j)}, y_j)\}_{j \in \{1,\ldots,n\} \setminus \{i\}}$. The obtained predictive mean $\tilde{y}_i$ and variance $\tilde{\sigma}_i^2$ at the point $\boldsymbol{x}^{(i)}$ are functions of parameters $(\sigma^2, \boldsymbol{\theta})$ (we recall that $\sigma_\epsilon^2$ is fixed) as shown in (2.43) and (2.44). The Leave-One-Out prediction error at the point $\boldsymbol{x}^{(i)}$ is given by

$$\tilde{\epsilon}_i = y_i - \tilde{y}_i. \tag{2.71}$$

Dubrule (1983) has shown that the Leave-One-Out prediction errors and variance can be calculated directly using the matrix $\overline{\mathbf{K}}$ defined in (2.57). It yields thus a very practical and efficient estimator of the predictive mean and variance. These formulas are known as the Virtual Cross-Validation formulas and are given by:

$$y_i - \tilde{y}_i = \frac{\left(\overline{\mathbf{K}}\boldsymbol{y}\right)_i}{\left(\overline{\mathbf{K}}\right)_{i,i}}, \tag{2.72}$$

and

$$\tilde{\sigma}_i^2 = \frac{1}{\left(\overline{\mathbf{K}}\right)_{i,i}}. \tag{2.73}$$

We refer to Dubrule (1983) (or to Ginsbourger & Schärer (2021) for the generalized case in the Universal Kriging) for detailed proof, which is based on Inverting block matrices and Schur complement.

Since the predictive mean $\tilde{y}_i$ and variance $\tilde{\sigma}_i^2$ imply the diagonal term $\overline{\mathbf{K}}_{i,i}$ in the denominator, we shall make the following assumption:

**Assumption 2.3.10.** *Let $(\mathbf{e}_i)_{i=1}^n$ be the canonical basis of $\mathbb{R}^n$. We assume that $\mathbf{e}_i \notin \mathrm{Im}\mathbf{F}$ for all $i \in \{1, \ldots, n\}$.*

Under this assumption and for all $i \in \{1, \ldots, n\}$, we have $\overline{\mathbf{K}}_{i,i} > 0$. So the Leave-One-Out quantities are well defined. The proof of this result is given in by Lemma A.1.3 in Appendix A.1.

While using the Cross-Validation method, it is common to consider the Mean Squared prediction Error to assess the quality of the point-wise prediction of the obtained Gaussian Process model.

**Definition 2.3.11** (The Leave-One-Out Mean Squared Errors criterion (Zhang & Wang, 2010))**.** *The Leave-One-Out Mean Squared Error criterion is defined by:*

$$\mathcal{LOO}(\sigma^2, \boldsymbol{\theta}) := \frac{1}{n}\tilde{\boldsymbol{\epsilon}}^\top \tilde{\boldsymbol{\epsilon}} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2\,, \tag{2.74}$$

*where, for $1 \leq i \leq n$, $\tilde{y}_i$ is the predictive mean of $y_i$ by a GP model trained on $\mathbf{D}_{-i}$ with covariance hyperparameters of $(\sigma^2, \boldsymbol{\theta})$.*

It has been shown that the Leave-One-Out Mean Squared Error criterion reflects the quality of the point-wise prediction of the GP model (Bachoc, 2013; Zhang & Wang, 2010). Minimizing this criterion, in the case of a stationary noise-free model, has been studied by Bachoc (2013) to address the problem of covariance hyperparameters estimation for a misspecified model.

In our case, it can be showed immediately that $\mathcal{LOO}$ can be also written with explicit quadratic forms as

$$\mathcal{LOO}(\sigma^2, \boldsymbol{\theta}) = \frac{1}{n}\boldsymbol{y}^\top \overline{\mathbf{K}}\,\mathrm{Diag}\left(\overline{\mathbf{K}}\right)^{-2}\overline{\mathbf{K}}\boldsymbol{y}. \tag{2.75}$$

Therefore, the Cross-Validation Mean Squared Error (CV-MSE) estimator of the covariance hyperparameters $(\sigma^2, \boldsymbol{\theta})$ is given by

$$(\hat{\sigma}^2_{MSE}, \hat{\boldsymbol{\theta}}_{MSE}) \in \mathrm{argmin}_{\sigma^2,\boldsymbol{\theta}}\ \ \boldsymbol{y}^\top \overline{\mathbf{K}}\,\mathrm{Diag}\left(\overline{\mathbf{K}}\right)^{-2}\overline{\mathbf{K}}\ \boldsymbol{y}. \tag{2.76}$$

The CV-MSE of the covariance hyperparameters $(\sigma^2, \boldsymbol{\theta})$ has the same computational complexity $O(n^3)$ as Maximum Likelihood, but it has the advantage of being more efficient when the covariance function is misspecified (Bachoc, 2013).

**Remark 2.3.12.** *As already discussed in Bachoc (2013), when there is no nugget effect, the Leave-One-Out Mean Squared Error criterion (2.75) is a function of the length-scale vector $\boldsymbol{\theta}$. Consequently, the CV-MSE estimator in this case is*

$$\hat{\boldsymbol{\theta}}_{MSE} \in \mathrm{argmin}_{\boldsymbol{\theta}}\ \ \frac{1}{n}\boldsymbol{y}^\top \overline{\mathbf{R}}_{\boldsymbol{\theta}}\,\mathrm{Diag}\left(\overline{\mathbf{R}}_{\boldsymbol{\theta}}\right)^{-2}\overline{\mathbf{R}}_{\boldsymbol{\theta}}\boldsymbol{y}, \tag{2.77}$$

*where* $\overline{\mathbf{R}}_{\boldsymbol{\theta}} = \mathbf{R}_{\boldsymbol{\theta}}^{-1} - \mathbf{R}_{\boldsymbol{\theta}}^{-1}\mathbf{F}\left(\mathbf{F}^\top\mathbf{R}_{\boldsymbol{\theta}}^{-1}\mathbf{F}\right)^{-1}\mathbf{F}^\top\mathbf{R}_{\boldsymbol{\theta}}^{-1}$.

*Unfortunately, the previous equation excludes the variance of the model $\sigma^2$ in the estimation procedure. We define another Cross-Validation criterion for this purpose which is:*

$$V_{LOO}(\sigma^2, \hat{\boldsymbol{\theta}}_{MSE}) = \frac{1}{n}\sum_{i=1}^{n}\frac{(y_i - \tilde{y}_i)^2}{\tilde{\sigma}_i^2}, \tag{2.78}$$

*where $\tilde{y}_i$ and $\tilde{\sigma}_i^2$ are the predictive mean and variance obtained using a covariance function with the length-scale vector $\hat{\boldsymbol{\theta}}_{MSE}$. Cressie (1993) in Section 2.6.4 claimed that this variance should be close to 1 if the covariance function is correctly specified. Therefore, enforcing the criterion $V_{LOO}$ to be equal to 1 gives, after direct calculations, a "CV-MSE" estimator of the amplitude $\sigma^2$ given by:*

$$\hat{\sigma}^2_{MSE} = \frac{1}{n}\boldsymbol{y}^\top \overline{\mathbf{R}}_{\hat{\boldsymbol{\theta}}_{MSE}}\,\mathrm{Diag}\left(\overline{\mathbf{R}}_{\hat{\boldsymbol{\theta}}_{MSE}}\right)^{-1}\overline{\mathbf{R}}_{\hat{\boldsymbol{\theta}}_{MSE}}\boldsymbol{y}. \tag{2.79}$$

*Here, the notation of "CV-MSE" is to indicate that the variance's model estimator $\hat{\sigma}^2_{MSE}$ does not minimize in reality the MSE of the model. It is a more reasonable choice given the state-of-the-art.*

## 2.4 Current Kriging-related research

We have seen in this chapter several properties and use of the Gaussian Process model that make it a powerful tool in Machine Learning and Uncertainty Quantification. However, Gaussian Processes belong to a field that is in continuous development.

One major axis of GPs research is their computational cost and the necessity of manipulating large covariance matrices. Recent works aim to approximate GPs in an optimal and efficient to be used in large-scale data (See Liu et al. (2020) for a review about scalable GPs)

GPs have also been proposed to emulate complex problems. A first example is a multidimensional output. Indeed, *multi-fidelity co-Kriging*, which is an extension of ordinary kriging to Multi-output, was originally proposed by Kennedy & O'Hagan (2000) then developed by Forrester et al. (2008). It has been successfully used to emulate *efficiently* hierarchical multi-fidelity codes Le Gratiet (2013) and time-series output (Kerleguer, 2021). The second example is *nested Kriging* (Perrin et al., 2017), where the output of one is the input of the next, called nested codes. Rullière et al. (2018) propose aggregating small Kriging models in the case of large data.

Moreover, the use of quantitative and qualitative inputs in science, engineering and business motivated inputs limits GPs. Roustant et al. (2020) extend Gaussian Processes based methods to categorical inputs (group kernels), Zhang et al. (2021) propose a sparse covariance estimation approach for both numerical and categorical inputs, and Bachoc et al. (2018) develop a theory for Kriging of distributional rather than numerical inputs.

Furthermore, additional knowledge about data can be useful in improving the predictive task of GPs. Veiga & Marrel (2012) introduced a new theoretical framework, with promising results (López-Lopera et al., 2018), called *Constrained Gaussian Processes*. It includes some constraints (e.g. positivity, boundedness, monotonicity and convexity, see Swiler et al. (2020) for a review) while modelling GPs.

Finally, the model misspecification has also been discussed in the literature. Particularly with Bachoc (2013) who introduced the Cross-Validation as an alternative to overcome model misspecification. Wang (2021), and Wynne et al. (2021) also studied the prediction error bounds and convergence guarantees of misspecified Gaussian Process models. This issue of model misspecification is discussed further in the next chapter.

# CHAPTER 3

# Quantifying Prediction Intervals for Gaussian Processes using Cross-Validation method

This chapter is contains passages from the paper (Acharki et al., 2023), to appear in Computational Statistics and Data Analysis Journal.

## 3.1 Introduction

In Chapter 2, we have defined a framework of regression in 2.1 for an output inference. We have reviewed different methods for Uncertainty Quantification. We focused more on the Gaussian Process model as one of the powerful Bayesian nonparametric models. With the Gaussian Process model, we constructed Prediction Intervals in light of the Definition for a given training dataset and confidence level. The upper and lower bounds of these prediction Intervals were fully characterized by the predictive mean and variance of the model.

Recent work has shown that both Maximum Likelihood and the full-Bayesian methods are optimal when the model is well-specified, according to Definition 2.3.2. The mean prediction and the prediction intervals are representative of the uncertainty of the model. In particular, they achieve optimal coverage with respect to Type IV and, consequently, Type II Coverage Probability. Usually, it is preferable to consider the Maximum Likelihood method for computational reasons. Indeed, the Full-Bayesian approach is very complex to implement, typically with a Markov chain Monte Carlo (MCMC) algorithm and can be sensitive to the choice of the *prior* distribution of the hyperparameters.

These results, although promising, are valid only if the obtained covariance function and its hyperparameters fit the assumption of the Gaussian Process on $f$. They also suppose that the set of possible covariance functions with corresponding hyperparameters is given prior to the estimation phase.

The Gaussian process model is misspecified if the observations $\boldsymbol{y}$ do not correspond to a realization of a Gaussian process with a covariance function belonging to this family. Consequently, the Maximum Likelihood approach may fail to fit data. The estimated covariance hyperparameters by Maximum Likelihood do not reflect the uncertainty of the model and Prediction Intervals are no longer reliable as they do not respect the prescribed coverage.

An interesting example is conjectured by Xu & Stein (2017). When modelling the function $f(x) = x^{\gamma}$ on $[0, 1]$ with a Gaussian kernel ($\nu \to +\infty$), the estimated variance can either go to

zero or infinity as the sample size increases to infinity. Prediction Intervals would be either too short with zero coverage or too wide with 100% coverage.

Unfortunately, in many cases and real-world applications, the Gaussian process is misspecified. One cannot know easily what would be the form of the covariance function or to what family it belongs.

When modeling with Gaussian Process, it is common practice to limit the covariance function family to a predefined set of simplified models (for example, the radial model in 2.26). These models aid in maintaining a closed-form expression of the likelihood or MSE criterion and simplifying the optimization procedure. However, they may not be a faithful representation of the latent function $f$, resulting in a weak and unreliable approximation.

Improving the modelling of the covariance function seems to be efficient in overcoming the issue of a misspecified model. Still, it may lead to complex covariance models and severe difficulties in estimating the covariance function's hyperparameters, especially in high dimensions. Moreover, sometimes, it is challenging to find proper modelling without further knowledge of the system and the sources of uncertainty.

The problem of model misspecification is gaining more attention in the Gaussian Process community, and many recent works discuss the properties of the Gaussian process regression given model misspecification. Bachoc (2013) considers the problem of model misspecification to develop a Cross-Validation method for point-wise prediction. Later, Bachoc (2018) shows that, asymptotically, the Maximum Likelihood estimator minimizes the Kullback-Leibler divergence to the misspecified parametric set. Wynne et al. (2021) present error bounds for the mean predictions of misspecified GP models. They demonstrate the sensitivity of the hyperparameter's choice and the experimental design on the error bounds. Wang (2021) provides some insights on explaining the poor coverage of Prediction intervals. The results indicate that, when applying a misspecified model, the prediction interval's reliability and the predictor's optimality cannot be achieved simultaneously.

While most literature emphasizes the difficulty of making an accurate and reliable prediction with misspecified models, the question was whether valid inferences could still be made. The answer turns out to be optimistic at this stage based on Bachoc (2013) work. Indeed, the Cross-Validation method allows for the modification and the selection of the covariance function hyperparameters based on a specific metric (point-wise criterion). However, the Cross-Validation does not correct the model's misspecification; rather, it minimizes the integrated mean squared error, which is insufficient to overcome the main issue. The variance of the MSE Cross-validation model, in particular, may not accurately estimate the true uncertainty of the model. The reason relies mainly on the choice of the $\hat{\sigma}^2$ as explained in Remark 2.3.12. Therefore, the Prediction interval must be carefully constructed to quantify the uncertainties. An example is Luna & Young (2003) who propose to calibrate the *Maximum Likelihood* variance with a bootstrap approach.

In Chapter 3, we propose a method based on Cross-Validation of the Gaussian Process model to address the problem of model misspecification. The goal is to calibrate Prediction Intervals by adjusting the upper and lower bounds. The method gives Prediction Intervals with appropriate coverage probabilities and small widths.

## 3.2 Prediction Intervals estimation with Cross-Validation

In this section, we consider the $n$-Cross-Validation framework as already defined in 2.3, the training dataset is denoted by $\mathbf{D} = \{(\boldsymbol{x}^{(i)}, y_i)\}_{i=1}^n$.

We keep the notations of $\tilde{y}_i$ and $\tilde{\sigma}_i^2$, the predictive mean and variance on $\boldsymbol{x}^{(i)} \in \mathbf{X}$, using the GP model $\mathcal{GP}_{-i}$, trained on the dataset $\mathbf{D}_{-i} = \{(\boldsymbol{x}^{(j)}, y_j)\}_{j\in\{1,\dots,n\}\setminus\{i\}}$. We recall the expression of $\tilde{y}_i$ and $\tilde{\sigma}_i^2$ as given (2.72) and (2.73), given by the Virtual Cross-Validation formulas of Dubrule (1983):

$$y_i - \tilde{y}_i = \frac{\left(\overline{\mathbf{K}}\boldsymbol{y}\right)_i}{\left(\overline{\mathbf{K}}\right)_{i,i}}, \tag{3.1}$$

and

$$\tilde{\sigma}_i^2 = \frac{1}{\left(\overline{\mathbf{K}}\right)_{i,i}}, \tag{3.2}$$

where $\overline{\mathbf{K}} = \mathbf{K}^{-1} - \mathbf{K}^{-1}\mathbf{F}\left(\mathbf{F}^\top\mathbf{K}^{-1}\mathbf{F}\right)^{-1}\mathbf{F}^\top\mathbf{K}^{-1}$ (see (Mardia & Marshall, 1984) or (de Oliveira, 2007) for more details about $\overline{\mathbf{K}}$).

We have seen in Section 2.3 that, using Cross-Validation method, Bachoc (2013) established an estimator of the covariance hyperparameter's $(\hat{\sigma}^2_{MSE}, \hat{\boldsymbol{\theta}}_{MSE})$ based on a point-wise prediction metric. Unfortunately, the obtained model's variance $\sigma^2$ might not represent the model's uncertainty. Consequently, the Prediction Intervals could be shorter or wider and do not respect the required coverage (neither Type II nor Type I).

Based on the Cross-Validation method, our approach proposes Empirical Coverage Probability in 2.1.10 as a metric. We adjust the hyperparameters $(\sigma^2, \boldsymbol{\theta})$ with respect to this metric. The upper $u_{1-\alpha}$ and lower $l_{1-\alpha}$ bounds are calibrated. Doing so will guarantee that the prediction Intervals are well-calibrated and respect Type II coverage.

Let the Leave-One-Out Coverage Probability $\tilde{\mathbb{P}}_{1-\alpha}$ define the Empirical Coverage Probability on $(\boldsymbol{x}^{(i)}, y_i)$ using the dataset $\mathbf{D}_{-i}$:

$$\tilde{\mathbb{P}}_{1-\alpha} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{y_i \in \mathcal{PI}_{1-\alpha}(\boldsymbol{x}^{(i)}; \mathbf{D}_{-i})\}, \tag{3.3}$$

where $\mathcal{PI}_{1-\alpha}(\boldsymbol{x}^{(i)}; \mathbf{D}_{-i})$ are the Prediction Intervals given by the Leave-One-Out method

$$\begin{aligned} \mathcal{PI}_{1-\alpha}(\boldsymbol{x}^{(i)}; \mathbf{D}_{-i}) &= [l_{1-\alpha}(\boldsymbol{x}^{(i)}); u_{1-\alpha}(\boldsymbol{x}^{(i)})] \\ &= [\tilde{y}_i + q_{\alpha/2} \times \tilde{\sigma}_i;\ \tilde{y}_i + q_{1-\alpha/2} \times \tilde{\sigma}_i]. \end{aligned} \tag{3.4}$$

Under some assumptions on the Leave-One-Out predictive mean and variance, Steinberger & Leeb (2018) in Theorem 2.4 provide conditional coverage guarantees of the Leave-One-Out Prediction Intervals. Moreover, these intervals are asymptotically valid *i.e.* the Leave-One-Out Coverage Probability $\tilde{\mathbb{P}}_{1-\alpha}$ converges asymptotically to $1-\alpha$.

When the model is well-specified, the coverage of the Prediction Intervals $\mathcal{PI}_{1-\alpha}$ is optimal, and Leave-One-Out Coverage Probability $\tilde{\mathbb{P}}_{1-\alpha}$ is close to $1-\alpha$. Conversely, if the model is misspecified, this probability is significantly different from $1-\alpha$.

Therefore, the Prediction Intervals or, equivalently, the upper and lower bounds $l_{1-\alpha}, u_{1-\alpha}$ need to be appropriately quantified with respect to Leave-One-Out Coverage Probability $\tilde{\mathbb{P}}_{1-\alpha}$, so it achieves the desired level.

As discussed in remark 2.2.37, despite these intervals being Type IV, we do not intend to calibrate Prediction intervals with respect to Type IV coverage. Unlike the well-specified model case, this coverage is difficult to achieve in the misspecified case without knowing the posterior distribution $Y(\cdot) \mid \mathbf{D}$ or making assumptions about its structure.

The Leave-One-Out Coverage Probability $\tilde{\mathbb{P}}_{1-\alpha}$ can be written as

$$\begin{aligned}\tilde{\mathbb{P}}_{1-\alpha} &= \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{y_i \in \mathcal{PI}_{1-\alpha}(\boldsymbol{x}^{(i)}; \mathbf{D}_{-i})\}, \\ &= \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{\tilde{y}_i + q_{\alpha/2} \times \tilde{\sigma}_i < \ y_i \leq \ \tilde{y}_i + q_{1-\alpha/2} \times \tilde{\sigma}_i\}.\end{aligned} \tag{3.5}$$

We introduce the Heaviside step function $h$

$$h(x) = \mathbf{1}\{x \geq 0\} = \begin{cases} 1 & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases}, \tag{3.6}$$

which allows us to write $\tilde{\mathbb{P}}_{1-\alpha}$ as

$$\tilde{\mathbb{P}}_{1-\alpha} = \frac{1}{n}\sum_{i=1}^{n} h\left(q_{1-\alpha/2} - \frac{y_i - \tilde{y}_i}{\tilde{\sigma}_i}\right) - \frac{1}{n}\sum_{i=1}^{n} h\left(q_{\alpha/2} - \frac{y_i - \tilde{y}_i}{\tilde{\sigma}_i}\right). \tag{3.7}$$

Let $a \in (0, 1/2) \cup (1/2, 1)$ describe a nominal level of quantile. We define the *quasi-Gaussian* proportion $\psi_a$ as a map from $[0, +\infty) \times (0, +\infty)^d$ to $[0, 1]$

$$\psi_a\left(\sigma^2, \boldsymbol{\theta}\right) = \frac{1}{n}\sum_{i=1}^{n} h\left(q_a - \frac{y_i - \tilde{y}_i}{\tilde{\sigma}_i}\right). \tag{3.8}$$

Given the Virtual Cross-Validation formulas (Dubrule, 1983), $\psi_a$ can be written in terms of the covariance matrix $\overline{\mathbf{K}}$

$$\psi_a(\sigma^2, \boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} h\left(q_a - \frac{\left(\overline{\mathbf{K}}\boldsymbol{y}\right)_i}{\sqrt{\left(\overline{\mathbf{K}}\right)_{i,i}}}\right). \tag{3.9}$$

The *quasi-Gaussian* proportion $\psi_a$ describes how close the $a$-quantile $q_a$ of the standardized predictive distribution is to the level $a$ (ideally, it should correspond to $a$).

Since there exists a correspondence between $u_{1-\alpha}$ (respectively, $l_{1-\alpha}$) and $\psi_{1-\alpha/2}$ (respectively, $\psi_{\alpha/2}$), the objective is to fit the hyperparameters $(\sigma^2, \boldsymbol{\theta})$ according to the *quasi-Gaussian* proportions and find two pairs $(\overline{\sigma}^2, \overline{\boldsymbol{\theta}})$ and $(\underline{\sigma}^2, \underline{\boldsymbol{\theta}})$ such that $\psi_{1-\alpha/2}(\overline{\sigma}^2, \overline{\boldsymbol{\theta}}) = 1 - \alpha/2$ and $\psi_{\alpha/2}(\underline{\sigma}^2, \underline{\boldsymbol{\theta}}) = \alpha/2$. This allows us modifying the upper and lower bounds $l_{1-\alpha}, u_{1-\alpha}$ to get the optimal coverage, by setting the Leave-One-Out Coverage to its nominal level, that is $\tilde{\mathbb{P}}_{1-\alpha} = 1 - \alpha$.

### Presence of nugget effect

In this subsection, we assume $\sigma_\epsilon^2 > 0$. The *quasi-Gaussian* proportion $\psi_a$ is, however, piece-wise constant and can take values only in the finite set $\{k/n, k \in \{0, \dots, n\}\}$. We first need to modify the problem $\psi_a\left(\sigma^2, \boldsymbol{\theta}\right) = a$. Let $\delta > 0$, we define the continuous functions $h_\delta^-$ and $h_\delta^+$

$$\begin{aligned} h_\delta^+(x) &= \begin{cases} 1 & \text{if } x > \delta, \\ x/\delta & \text{if } 0 < x \leq \delta, \\ 0 & \text{otherwise.} \end{cases} \\ h_\delta^-(x) &= \begin{cases} 1 & \text{if } x \geq 0, \\ 1 + x/\delta & \text{if } -\delta \leq x < 0, \\ 0 & \text{otherwise.} \end{cases} \end{aligned} \tag{3.10}$$

If $a > 1/2$ we define

$$\psi_a^{(\delta)}\left(\sigma^2, \boldsymbol{\theta}\right) = \frac{1}{n}\sum_{i=1}^{n} h_\delta^+\left(q_a - \frac{\left(\overline{\mathbf{K}}\boldsymbol{y}\right)_i}{\sqrt{\left(\overline{\mathbf{K}}\right)_{i,i}}}\right). \tag{3.11}$$

If $a < 1/2$ we define

$$\psi_a^{(\delta)}\left(\sigma^2, \boldsymbol{\theta}\right) = \frac{1}{n}\sum_{i=1}^{n} h_\delta^-\left(q_a - \frac{\left(\overline{\mathbf{K}}\boldsymbol{y}\right)_i}{\sqrt{\left(\overline{\mathbf{K}}\right)_{i,i}}}\right). \tag{3.12}$$

Let $\delta > 0$ be small enough so that $\delta < q_a$ if $a > 1/2$ (respectively, $\delta < q_{1-a}$ if $a < 1/2$) in such a way that $h_\delta^+(q_a) = 1$ (respectively, $h_\delta^-(q_a) = 0$). We consider the problem

$$\psi_a^{(\delta)}\left(\sigma^2, \boldsymbol{\theta}\right) = a, \tag{3.13}$$

and we denote by $\mathcal{A}_{a,\delta}$ the solution set of the problem (3.13)

$$\mathcal{A}_{a,\delta} := \left\{(\sigma^2, \boldsymbol{\theta}) \in [0, +\infty) \times (0, +\infty)^d,\ \psi_a^{(\delta)}(\sigma^2, \boldsymbol{\theta}) = a\right\}. \tag{3.14}$$

**Assumption 3.2.1.** *Let* $k_\epsilon = \mathrm{Card}\{i \in \{1, \dots, n\},\ \frac{(\boldsymbol{\Pi}\boldsymbol{y})_i}{\sqrt{(\boldsymbol{\Pi})_{ii}}} \leq \sigma_\epsilon q_a\}$ *where* $\boldsymbol{\Pi}$ *is the orthogonal projection matrix on* $(\mathrm{Im}\mathbf{F})^\perp$ *such that* $\boldsymbol{\Pi} = \mathbf{I}_n - \mathbf{F}\left(\mathbf{F}^\top\mathbf{F}\right)^{-1}\mathbf{F}^\top$. *We assume that* $k_\epsilon < na$ *if* $a > 1/2$ *and* $k_\epsilon > na$ *if* $a < 1/2$.

**Remark 3.2.2.** *The assumption 3.2.1 is typically satisfied in Ordinary and Universal Kriging. Indeed,* $\boldsymbol{\Pi}$ *is the projection on the space* $(\mathrm{Im}\mathbf{F})^\perp$ *and is expected to remove the trend of the model. It is reasonable to think that* $(\boldsymbol{\Pi}\boldsymbol{y})$ *is centered and that*

$$\mathrm{Card}\{i \in \{1, \dots, n\},\ (\boldsymbol{\Pi}\boldsymbol{y})_i \leq 0\} \approx \frac{n}{2}. \tag{3.15}$$

*If* $\sigma_\epsilon^2$ *is smaller than* $\sigma^2$, *then we should also have*

$$\mathrm{Card}\{i \in \{1, \dots, n\},\ \frac{(\boldsymbol{\Pi}\boldsymbol{y})_i}{\sqrt{(\boldsymbol{\Pi})_{ii}}} \leq \sigma_\epsilon q_a\} \approx \frac{n}{2}, \tag{3.16}$$

*so that the assumption 3.2.1 should be fulfilled.*

**Proposition 3.2.3.** *Let us assume the assumptions 2.2.32, 2.3.10 and 3.2.1, then $\mathcal{A}_{a,\delta}$ is non-empty.*

*Proof.* In A.1. ■

The challenge now is to identify and choose wisely the optimal solutions $(\sigma^2_{\text{opt}}, \boldsymbol{\theta}_{\text{opt}}) \in \mathcal{A}_{a,\delta}$. In High-Quality principles methods, some authors (Khosravi et al., 2010; Pearce et al., 2018) suggest the mean Prediction Intervals width (MPIW) 2.1.12 of Prediction Intervals $\mathcal{PI}_{1-\alpha}$ as an additional constraint to reduce the set of solutions. The upper and lower bounds are built simultaneously using a Neural Network, which makes computing the MPIW in the loss metric possible.

In our approach, however, this constraint may not be suitable. Indeed, we target the upper and lower bounds separately (the other bound of the corresponding interval would be infinite) and ensure that each bound respects its coverage.

Instead, our strategy consists on comparing these solutions with MLE's solution $(\hat{\sigma}^2_{ML}, \hat{\boldsymbol{\theta}}_{ML})$ (subsection 2.3) or MSE-CV solution $(\hat{\sigma}^2_{MSE}, \hat{\boldsymbol{\theta}}_{MSE})$ (subsection 2.3) and we will take the closest pair $(\sigma^2_{\text{opt}}, \boldsymbol{\theta}_{\text{opt}})$ by using an appropriate notion of similarity between multivariate Gaussian distributions. Ideally, we aim to solve the following problem

$$\text{argmin}_{(\sigma^2,\boldsymbol{\theta})\in\mathcal{A}_{a,\delta}}\, d^2\left((\sigma^2,\boldsymbol{\theta}),(\sigma_0^2,\boldsymbol{\theta}_0)\right), \tag{3.17}$$

where $d$ is a continuous similarity measure of hyperparameters $(\sigma^2, \boldsymbol{\theta})$ operating on the mean $\boldsymbol{m}$ and the covariance matrix $\mathbf{K}$, and $(\sigma_0^2, \boldsymbol{\theta}_0) = (\hat{\sigma}^2_{ML}, \hat{\boldsymbol{\theta}}_{ML})$ or $(\hat{\sigma}^2_{MSE}, \hat{\boldsymbol{\theta}}_{MSE})$ as described in (2.58) or (2.76).

Since the mean $\boldsymbol{m} = \mathbf{F}\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\beta}}$ is a function of the covariance matrix $\mathbf{K}$, the comparison of two covariance functions with given hyperparameters is equivalent somehow to considering similarity measure (i.e. distance) between the covariance matrices.

## 3.3 Similarity measures of covariance matrices

In this subsection, we discuss several distances in the space of symmetric positive semi-definite matrices $\mathbb{S}_n^+(\mathbb{R})$ that can be used to compare covariance matrices. The particularity of this space is that it is non-Euclidean. Thus, non-Euclidean representations are required to compare matrices belonging to it. The logarithm of a matrix, the square root of a matrix and the Cholesky decomposition, which is shown to be unique for positive definite matrices (Golub & Van Loan, 2013), are used for this purpose.

Let $\mathbf{K}_1$ and $\mathbf{K}_2$ be two covariance matrices, symmetric positive definite i.e. $\mathbf{K}_1, \mathbf{K}_2 \in \mathbb{S}_n^{++:}(\mathbb{R})$. Unless specified otherwise, we consider a fixed experimental design $\mathbf{X} = (\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(n)})$ and we assume that $\mathbf{K}_1$ and $\mathbf{K}_2$ are the covariance matrices associated to two covariance models $\boldsymbol{k}_1$ and $\boldsymbol{k}_2$.

We define the logarithm and the square root of $\mathbf{K}_i$, for $i \in \{1, 2\}$ from its spectral decomposition as

$$\begin{aligned} \log(\mathbf{K}_i) &= \mathbf{U}_i \log(\mathbf{D}_i)\mathbf{U}_i^\top, \\ \mathbf{K}_i^{1/2} &= \mathbf{U}_i \mathbf{D}_i^{1/2}\mathbf{U}_i^\top. \end{aligned} \tag{3.18}$$

$\mathbf{U}_i \in \mathcal{O}_n(\mathbb{R})$ is an orthogonal matrix $\mathbf{U}_i\mathbf{U}_i^\top = \mathbf{I}_n$ and $\mathbf{D}_i \in \mathcal{D}_n(\mathbb{R})$ a diagonal matrix containing the eigenvalues of $\mathbf{K}_i$.

The Root-Euclidean distance $d_{\text{Root}}$, the log-Euclidean distance $d_{\log}$ (Arsigny et al., 2007) and the Cholesky distance $d_{\text{Chol}}$ (Zhizhou Wang et al., 2004) for $\mathbf{K}_1$ and $\mathbf{K}_2$ are defined

$$d_{\text{Root}}(\mathbf{K}_1, \mathbf{K}_2) = \|\mathbf{K}_1^{1/2} - \mathbf{K}_2^{1/2}\|, \tag{3.19}$$

$$d_{\log}(\mathbf{K}_1, \mathbf{K}_2) = \|\log(\mathbf{K}_1) - \log(\mathbf{K}_2)\|, \tag{3.20}$$

$$d_{\text{Chol}}(\mathbf{K}_1, \mathbf{K}_2) = \|\text{Chol}(\mathbf{K}_1) - \text{Chol}(\mathbf{K}_2)\|, \tag{3.21}$$

where $\text{Chol}(\mathbf{K}_i)$ is the Cholesky decomposition of $\mathbf{K}_i$.

Förstner & Moonen (2003) and Pennec et al. (2006) proposed a distance, known also as version of the canonical invariant Riemannian metric for matrices

$$d_{\text{Fors}}(\mathbf{K}_1, \mathbf{K}_2) = \sqrt{\text{Tr}\left(\log^2(\mathbf{K}_1^{-1/2}\mathbf{K}_2\mathbf{K}_1^{-1/2})\right)}. \tag{3.22}$$

A distance $d$ is said to be

- Invariant under translation of $\mathbf{K}_i$, if $d(\mathbf{K}_1 + \boldsymbol{t}\boldsymbol{t}^\top, \mathbf{K}_2 + \boldsymbol{t}\boldsymbol{t}^\top) = d(\mathbf{K}_1, \mathbf{K}_2)$ for a translation vector $\boldsymbol{t} \in \mathbb{R}^n$.
- Invariant under simultaneous rotation and reflection of $\mathbf{K}_i$, if $d(\mathbf{U}\mathbf{K}_1\mathbf{U}^\top, \mathbf{U}\mathbf{K}_2\mathbf{U}^\top) = d(\mathbf{K}_1, \mathbf{K}_2)$ for an orthogonal matrix $\mathbf{U} \in \mathcal{O}_n(\mathbb{R})$.
- Invariant under scaling of $\mathbf{K}_i$, if $d(\beta\mathbf{K}_1, \beta\mathbf{K}_2) = d(\mathbf{K}_1, \mathbf{K}_2)$ for $\beta > 0$.
- Affine invariant, if $d(\mathbf{A}\mathbf{K}_1\mathbf{A}^\top, \mathbf{A}\mathbf{K}_2\mathbf{A}^\top) = d(\mathbf{K}_1, \mathbf{K}_2)$ where $\mathbf{A}$ is a general full rank matrix.
- Inverse invariant, if $d(\mathbf{K}_i^{-1}, \mathbf{I}_n) = d(\mathbf{K}_i, \mathbf{I}_n)$.

A review of Dryden et al. (2009) shows that $d_{\text{Chol}}$ is not invariant under simultaneous rotation and reflection of $\mathbf{K}_1$ and $\mathbf{K}_2$, $d_{\text{Root}}$ is not invariant under simultaneous scaling, only $d_{\log}$ and $d_{\text{Fors}}$ are inverse invariant and affine invariant. However, the inverse matrix in the Forstner distance $d_{\text{Fors}}$ may raise some computational issues and lead to unbounded behavior, as we will see below at the end of the subsection.

Pigoli et al. (2014) proposed also a Procrustes size-and-shape distance to compare two positive definite matrices:

$$\Pi(\mathbf{K}_1, \mathbf{K}_2) = \inf_{\mathbf{R} \in \mathcal{O}_n(\mathbb{R})} \|\mathbf{K}_1^{1/2} - \mathbf{K}_2^{1/2}\mathbf{R}\|. \tag{3.23}$$

**Proposition 3.3.1.** *The Procrustes size-and-shape distance* $\Pi$ *is invariant under translation, rotation and reflection.*

In addition, the Procrustes distance $\Pi$ has the advantage of dealing efficiently with deficient rank matrices (Dryden et al., 2009), unlike the invariant Riemannian distance $d_{\text{Fors}}$ which is not valid for this purpose.

An important proposition is shown by Masarotto et al. (2019) proving that the Procrustes distance in (3.23) between two covariance matrices $\mathbf{K}_1$ and $\mathbf{K}_2$ coincides with the second

Wasserstein distance between two Gaussian processes $Y_1 \sim \mathcal{GP}(\boldsymbol{m}_1, \boldsymbol{k}_1)$ and $Y_2 \sim \mathcal{GP}(\boldsymbol{m}_2, \boldsymbol{k}_2)$, given by (Dowson & Landau, 1982)

$$\Pi^2(\mathbf{K}_1, \mathbf{K}_2) = W_2^2(Y_1, Y_2) \tag{3.24}$$

$$= \|\boldsymbol{m}_1 - \boldsymbol{m}_2\|^2 + \operatorname{Tr}\left(\mathbf{K}_1 + \mathbf{K}_2 - 2\sqrt{\mathbf{K}_1^{1/2}\mathbf{K}_2\mathbf{K}_1^{1/2}}\right). \tag{3.25}$$

The Wasserstein distance, widely used in optimal transport problems (see Chapter 6 of Villani (2009) for more details). From now on, given an experimental design of inputs $\mathbf{X}$, the distance $\Pi^2(\mathbf{K}_1, \mathbf{K}_2)$ refers to the second Wasserstein between $\mathcal{GP}(\boldsymbol{m}_1, \boldsymbol{k}_1)$ and $\mathcal{GP}(\boldsymbol{m}_2, \boldsymbol{k}_2)$. When the mean of the two Gaussian processes $Y_1$ and $Y_2$ is constant $\boldsymbol{m}_1 = \boldsymbol{m}_2 = \mathbf{F}\boldsymbol{\beta}$, the second Wasserstein distance considers only the difference associated to the term of $\operatorname{Tr}(\cdot)$, that is,

$$\Pi^2(\mathbf{K}_1, \mathbf{K}_2) = \operatorname{Tr}\left(\mathbf{K}_1 + \mathbf{K}_2 - 2\sqrt{\mathbf{K}_1^{1/2}\mathbf{K}_2\mathbf{K}_1^{1/2}}\right). \tag{3.26}$$

In the following, we will assume a free-noise setting $\sigma_\epsilon^2 = 0$, and we will derive additional properties of the invariant Riemannian and the 2-Wasserstein distance.

**Proposition 3.3.2.** *Let $\mathbf{X}_1$ (respectively, $\mathbf{X}_2$) be an experimental design in $\mathbb{R}^{n\times d}$, and let $\mathbf{K}_1 = \boldsymbol{k}(\mathbf{X}_1, \mathbf{X}_1)$ (respectively, $\mathbf{K}_2 = \boldsymbol{k}(\mathbf{X}_2, \mathbf{X}_2)$) be the Gram matrix by a Matérn anisotropic geometric model $\boldsymbol{k}$ (i.e. the associated covariance matrix), then $d_{\text{Fors}}(\mathbf{K}_1, \mathbf{K}_2)$ is a function of $\boldsymbol{\theta}$ whereas $\Pi(\mathbf{K}_1, \mathbf{K}_2)$ depends on both $\sigma^2$ and $\boldsymbol{\theta}$.*

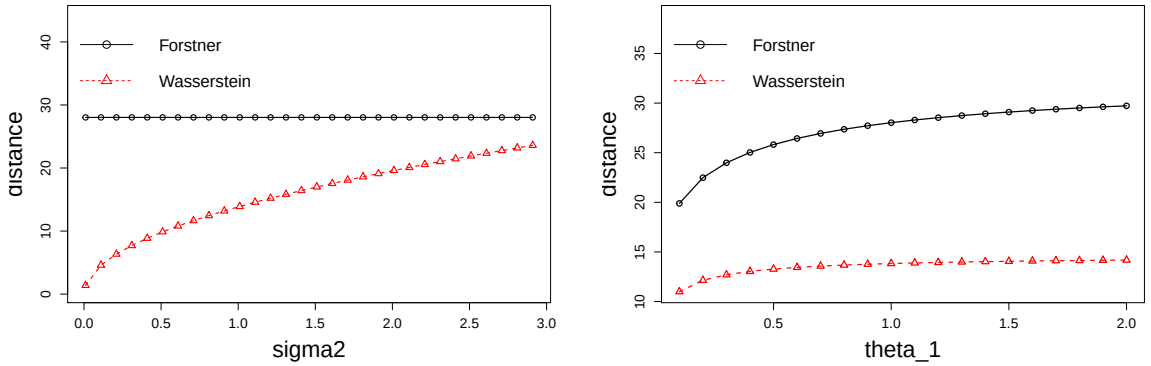
*Proof.* Under the assumption of a radial covariance model as in (2.26), the covariance matrices can be written as $\mathbf{K}_i = \sigma^2\mathbf{R}_{\boldsymbol{\theta},i}$ for $i \in \{1, 2\}$, we obtain by direct calculation :

$$\begin{aligned}
d_{\text{Fors}}^2(\mathbf{K}_1, \mathbf{K}_2) &= \operatorname{Tr}\left(\log^2(\mathbf{K}_1^{-1/2}\mathbf{K}_2\mathbf{K}_1^{-1/2})\right) = \operatorname{Tr}\left(\log^2(\sigma^{-1}\mathbf{R}_{\boldsymbol{\theta},1}^{-1/2}\sigma^2\mathbf{R}_{\boldsymbol{\theta},2}\sigma^{-1}\mathbf{R}_{\boldsymbol{\theta},1}^{-1/2})\right) \\
&= \operatorname{Tr}\left(\log^2(\mathbf{R}_{\boldsymbol{\theta},1}^{-1/2}\mathbf{R}_{\boldsymbol{\theta},2}\mathbf{R}_{\boldsymbol{\theta},1}^{-1/2})\right). \\
\Pi^2(\mathbf{K}_1, \mathbf{K}_2) &= \operatorname{Tr}\left(\mathbf{K}_1 + \mathbf{K}_2 - 2\sqrt{\mathbf{K}_1^{1/2}\mathbf{K}_2\mathbf{K}_1^{1/2}}\right) \\
&= \sigma^2 \operatorname{Tr}\left(\mathbf{R}_{\boldsymbol{\theta}_0}\right) + \sigma^2 \operatorname{Tr}\left(\mathbf{R}_{\boldsymbol{\theta}_0}\right) - 2\sigma^2 \operatorname{Tr}\left(\sqrt{\mathbf{R}_{\boldsymbol{\theta}_0}^{1/2}\mathbf{R}_{\boldsymbol{\theta}_0}\mathbf{R}_{\boldsymbol{\theta}_0}^{1/2}}\right) \\
&= 2\sigma^2\left(n - \operatorname{Tr}\left(\sqrt{\mathbf{R}_{\boldsymbol{\theta},1}^{1/2}\mathbf{R}_{\boldsymbol{\theta},2}\mathbf{R}_{\boldsymbol{\theta},1}^{1/2}}\right)\right).
\end{aligned} \tag{3.27}$$

The amplitude $\sigma^2$ is hence captured with the second Wasserstein distance as a scaling factor. ■

An illustration of Proposition 3.3.2 is shown in Figure 3.1. In this example, we have considered two experimental designs: a Latin-Hypercube-Sample $\mathbf{X}_1$ and a random sample $\mathbf{X}_2$. Both experimental designs are $d = 10$-dimensional and contains $n = 300$ observations. We can clearly see in Figure 3.1a that the Forstner distance $d_{\text{Fors}}(\mathbf{K}_1, \mathbf{K}_2)$ is invariant with respect to $\sigma^2$ and the 2-Wasserstein distance $\Pi(\mathbf{K}_1, \mathbf{K}_2)$ has a squared root curve with respect to $\sigma^2$.

**Proposition 3.3.3.** *For a given length-scale vector $\boldsymbol{\theta}_0 \in \mathbb{R}^d$ and an experimental design $\mathbf{X}$ in $\mathbb{R}^{n\times d}$, we denote $\mathbf{K}_0 = \boldsymbol{k}_0(\mathbf{X}, \mathbf{X})$ (respectively, $\mathbf{K} = \boldsymbol{k}(\mathbf{X}, \mathbf{X})$) the covariance matrix associated to the covariance model $\boldsymbol{k}_0 = \boldsymbol{k}_{\sigma_0^2,\boldsymbol{\theta}_0}$ (respectively, $\boldsymbol{k} = \boldsymbol{k}_{\sigma^2,\boldsymbol{\theta}_0}$). The Forstner and second Wasserstein distances are scale-variant and can be expressed as:*




(a) With respect to $\sigma^2$.

(b) With respect to $\theta_1$.

Figure 3.1: Comparison of Forstner's $d^2_{\text{Fors}}(\mathbf{K}_1, \mathbf{K}_2)$ and the 2-Wasserstein $\Pi(\mathbf{K}_1, \mathbf{K}_2)$ distances, on Latin-Hypercube-Sample $\mathbf{X}_1$ and random sample $\mathbf{X}_2$ of $d = 10$-dimensional and contains $n = 300$ observations, given a Matérn anisotropic geometric model $\boldsymbol{k} = \boldsymbol{k}_{\sigma^2,\boldsymbol{\theta}}$ with smoothness $\nu = 3/2$.

- $d_{\text{Fors}}(\mathbf{K}_0, \mathbf{K}) = 2\sqrt{n}\,|\log\sigma - \log\sigma_0|$
- $\Pi(\mathbf{K}_0, \mathbf{K}) = \sqrt{n}\left|\sqrt{\sigma^2} - \sqrt{\sigma_0^2}\right|$.

We refer to Figure 3.2a for an illustration of the previous proposition.

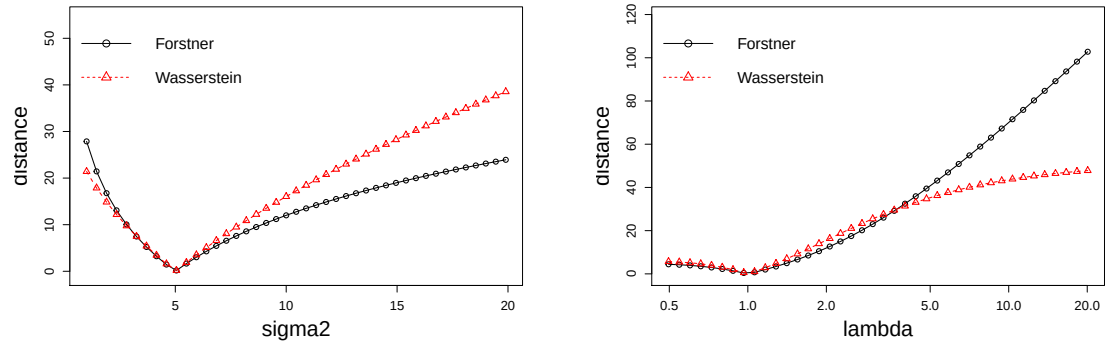
*Proof.* Under the same assumptions and calculations as in the proof of Proposition 3.3.2, we have :

$$\begin{aligned} d^2_{\text{Fors}}(\mathbf{K}_0, \mathbf{K}) &= \operatorname{Tr}\left(\log^2(\sigma_0^{-1}\mathbf{R}_{\boldsymbol{\theta}_0}^{-1/2}\sigma^2\mathbf{R}_{\boldsymbol{\theta}_0}\sigma_0^{-1}\mathbf{R}_{\boldsymbol{\theta}_0}^{-1/2})\right) \\ &= \operatorname{Tr}\left(\log^2(\sigma^2/\sigma_0^2\mathbf{I}_n)\right) = \sum_{i=1}^{n}\log^2(\sigma^2/\sigma_0^2) \\ &= \sum_{i=1}^{n} 4\left[\log(\sigma) - \log(\sigma_0)\right]^2 = 4n\left[\log(\sigma) - \log(\sigma_0)\right]^2. \end{aligned} \tag{3.28}$$

$$\begin{aligned} \Pi^2(\mathbf{K}_0, \mathbf{K}) &= \sigma_0^2\operatorname{Tr}(\mathbf{R}_{\boldsymbol{\theta}_0}) + \sigma^2\operatorname{Tr}(\mathbf{R}_{\boldsymbol{\theta}_0}) - 2\operatorname{Tr}\left(\sqrt{\sigma_0\mathbf{R}_{\boldsymbol{\theta}_0}^{1/2}\sigma^2\mathbf{R}_{\boldsymbol{\theta}_0}\sigma_0\mathbf{R}_{\boldsymbol{\theta}_0}^{1/2}}\right) \\ &= \sigma_0^2\operatorname{Tr}(\mathbf{R}_{\boldsymbol{\theta}_0}) + \sigma^2\operatorname{Tr}(\mathbf{R}_{\boldsymbol{\theta}_0}) - 2\sqrt{\sigma^2}\sqrt{\sigma_0^2}\operatorname{Tr}\left(\sqrt{\mathbf{R}_{\boldsymbol{\theta}_0}^2}\right) \\ &= n\sigma_0^2 + n\sigma^2 - 2n\sqrt{\sigma^2}\sqrt{\sigma_0^2} = n\left[\sqrt{\sigma^2} - \sqrt{\sigma_0^2}\right]^2. \end{aligned} \tag{3.29}$$

The last line holds because because $(\mathbf{R}_{\boldsymbol{\theta}})_{ii} = 1$. ■

The variation of the Forstner and second Wasserstein distances with respect to $\boldsymbol{\theta}$ is more difficult to express as it involves the square root and product of multiples matrices. However, we can simplify it by writing, for $\mathbf{A}$ and $\mathbf{B}$ symmetric positive definite matrices:




(a) With respect to $\sigma^2$.

(b) With respect to $\lambda\boldsymbol{\theta}$.

Figure 3.2: Comparison of Forstner's $d^2_{\text{Fors}}(\mathbf{K}, \mathbf{K}_0)$ and the 2-Wasserstein $\Pi(\mathbf{K}, \mathbf{K}_0)$ distances between two covariance matrices $\mathbf{K}$ and $\mathbf{K}_0$, on a random sample $\mathbf{X}$ of $d = 10$ and $n = 300$ observations, given by two Matérn anisotropic geometric models $\boldsymbol{k} = \boldsymbol{k}_{\sigma^2,\lambda\boldsymbol{\theta}_0}$ and $\boldsymbol{k}_0 = \boldsymbol{k}_{\sigma_0^2,\boldsymbol{\theta}_0}$ with smoothness $\nu = 3/2$ where $\sigma_0^2 = 5$ and $\boldsymbol{\theta}_0 = (1, \ldots, 1)$.

$$\begin{aligned}
\det\left(\mathbf{AB} - \lambda\mathbf{I}_n\right) &= \det\left[\mathbf{A}^{1/2}\left(\mathbf{A}^{1/2}\mathbf{B}\mathbf{A}^{1/2} - \lambda\mathbf{I}_n\right)\mathbf{A}^{-1/2}\right] = \det\left(\mathbf{A}^{1/2}\mathbf{B}\mathbf{A}^{1/2} - \lambda\mathbf{I}_n\right), \\
\det\left(\mathbf{A}^{-1}\mathbf{B} - \lambda\mathbf{I}_n\right) &= \det\left[\mathbf{A}^{-1/2}\left(\mathbf{A}^{-1/2}\mathbf{B}\mathbf{A}^{-1/2} - \lambda\mathbf{I}_n\right)\mathbf{A}^{1/2}\right] = \det\left(\mathbf{A}^{-1/2}\mathbf{B}\mathbf{A}^{-1/2} - \lambda\mathbf{I}_n\right).
\end{aligned} \tag{3.30}$$

As consequence, the characteristic polynomials of $\mathbf{A}^{-1/2}\mathbf{B}\mathbf{A}^{-1/2}$ and $\mathbf{A}^{-1}\mathbf{B}$ coincide on $\mathbb{R}$ and therefore, they have the same eigenvalues. In addition, $\operatorname{Tr}\left(f(\mathbf{A})\right) = \sum_{i=1}^n f(\lambda_i(\mathbf{A}))$ for any function $f$, we can write:

$$\begin{aligned}
\operatorname{Tr}\left(\sqrt{\mathbf{R}_{\boldsymbol{\theta}_0}^{1/2}\mathbf{R}_{\boldsymbol{\theta}}\mathbf{R}_{\boldsymbol{\theta}_0}^{1/2}}\right) &= \sum_{i=1}^n \lambda_i\left(\mathbf{R}_{\boldsymbol{\theta}_0}\mathbf{R}_{\boldsymbol{\theta}}\right), \\
\operatorname{Tr}\left(\log^2(\mathbf{R}_{\boldsymbol{\theta}_0}^{-1/2}\mathbf{R}_{\boldsymbol{\theta}}\mathbf{R}_{\boldsymbol{\theta}_0}^{-1/2})\right) &= \sum_{i=1}^n \log^2\lambda_i\left(\mathbf{R}_{\boldsymbol{\theta}_0}^{-1}\mathbf{R}_{\boldsymbol{\theta}}\right).
\end{aligned} \tag{3.31}$$

Since $\mathbf{R}_{\boldsymbol{\theta}}$ converges to $\mathbf{J} = \boldsymbol{e}\boldsymbol{e}^\top$ the matrix full of ones with $\boldsymbol{e} = (1, \ldots, 1)^\top$, which is singular, we must have at least one eigenvalue of $\mathbf{R}_{\boldsymbol{\theta}_0}^{-1/2}\mathbf{R}_{\boldsymbol{\theta}}$ and $\mathbf{R}_{\boldsymbol{\theta}_0}^{-1}\mathbf{R}_{\boldsymbol{\theta}}$ that converges to 0, this explains Figure 3.2b where the Forstner distance $d_{\text{Fors}}$ diverges to $+\infty$ and $\Pi$ to a finite limit.

As a conclusion of this subsection, considering properties and comparisons above, the final choice of the optimal similarity $d$ goes to the second Wasserstein distance and all its advantages.

### Robust Prediction Intervals Estimation method

From now on, given an experimental design of inputs $\mathbf{X}$, each pair $(\sigma^2, \boldsymbol{\theta})$ is associated to a Gaussian distribution $\mathcal{N}(\boldsymbol{m}, \mathbf{K})$ and we define the similarity measure $d$ as the 2-Wasserstein distance

$$d^2\left((\sigma^2, \boldsymbol{\theta}), (\sigma_0^2, \boldsymbol{\theta}_0)\right) = W_2^2(\mathcal{N}(\boldsymbol{m}, \mathbf{K}), \mathcal{N}(\boldsymbol{m}_0, \mathbf{K}_0)). \tag{3.32}$$

where, $\boldsymbol{m} = \mathbf{F}\widehat{\boldsymbol{\beta}} = \left(\mathbf{F}^\top \mathbf{K}^{-1}\mathbf{F}\right)^{-1}\mathbf{F}^\top\mathbf{K}^{-1}\boldsymbol{y}$, $\boldsymbol{m}_0 = \mathbf{F}\widehat{\boldsymbol{\beta}}_0 = \left(\mathbf{F}^\top \mathbf{K}_0^{-1}\mathbf{F}\right)^{-1}\mathbf{F}^\top\mathbf{K}_0^{-1}\boldsymbol{y}$ and $\mathbf{K}_0$ is the covariance matrix associated to covariance hyperparameters obtained by MLE or MSE-CV methods.

The resolution of Problem (3.17) may be too costly and heavy to solve when the dimension is high, say $d \geq 10$. An alternative is to apply *the relaxation* method where we redefine this optimization problem of $\boldsymbol{\theta}$ from $(0,+\infty)^d$ to $(0,+\infty)$ by shifting the length-scale vector $\boldsymbol{\theta}_0$ by a parameter $\lambda \in (0,+\infty)$.

Let $\boldsymbol{\theta}_0$ denote the correlation-length vector obtained by MLE (2.58) or MSE-CV (2.76) and, for $\lambda \in (0,+\infty)$, let $H_\delta(\lambda)$ denote the subset

$$H_\delta(\lambda) = \{\sigma^2 \in [0,+\infty),\ \psi_a^{(\delta)}(\sigma^2, \lambda\boldsymbol{\theta}_0) = a\}. \tag{3.33}$$

**Assumption 3.3.4.** *The set-valued mapping (the so-called correspondence function) $H_\delta : (0,+\infty) \to \mathcal{P}((0,+\infty))$, where $\mathcal{P}(S)$ denotes the power set of a set $S$, is lower semi-continuous, that is, for all $\lambda \in (0,+\infty)$, for each open set $\mathcal{U}$ with $H_\delta(\lambda) \cap \mathcal{U} \neq \emptyset$, there exists a neighborhood $\mathcal{O}(\lambda)$ such that if $\lambda^* \in \mathcal{O}(\lambda)$ then $H_\delta(\lambda^*) \cap \mathcal{U} \neq \emptyset$.*

In the kriging framework, $\sigma^2$ should be as small as possible to reduce the uncertainty of the model, a natural choice of $\sigma^2_{\text{opt}}$ is

$$\forall \lambda \in (0,+\infty)\ : \sigma^2_{\text{opt}}(\lambda) := \min\{\sigma^2 \in [0,+\infty),\ \psi_a^{(\delta)}(\sigma^2, \lambda\boldsymbol{\theta}_0) = a\}. \tag{3.34}$$

**Proposition 3.3.5.** *The function $\lambda \mapsto \sigma^2_{\text{opt}}(\lambda)$ is well-defined under hypotheses 2.2.32, 2.3.10 and 3.2.1, and continuous on $(0,+\infty)$ under the additional assumption 3.3.4.*

*Proof.* In A.1. ■

The choice of the second Wasserstein distance $W_2$ jointly with $\sigma^2_{\text{opt}}$ makes the Prediction Intervals $\mathcal{PI}_{1-\alpha}$ shorter without the need for an additional metric like the MPIW and without modifying the distribution of the obtained model significantly. We will see in Section 3.4 that, empirically, the bary-centers of Prediction Intervals are not far from the predictive means obtained by MLE or MSE-CV methods.

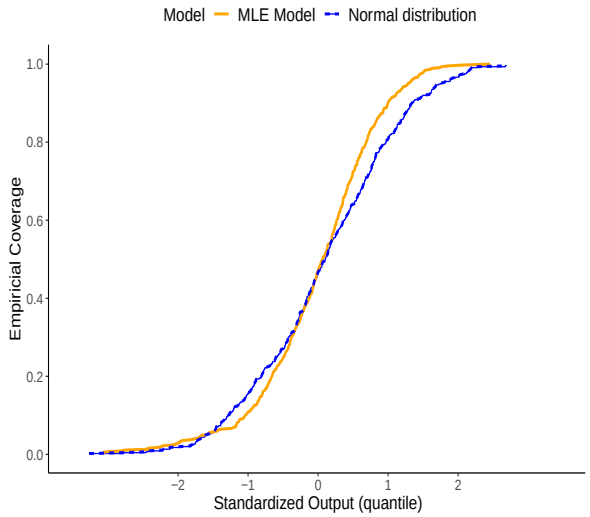
The *relaxed* optimization problem in (3.13) for the Prediction Interval bound's estimation is given by the problem $\mathcal{P}_\lambda$

$$\mathcal{P}_\lambda : \quad \text{argmin}_{\lambda\in(0,+\infty)}\ \mathcal{L}(\lambda) := d^2\left((\sigma^2_{\text{opt}}(\lambda), \lambda\boldsymbol{\theta}_0), (\sigma_0^2, \boldsymbol{\theta}_0)\right). \tag{3.35}$$

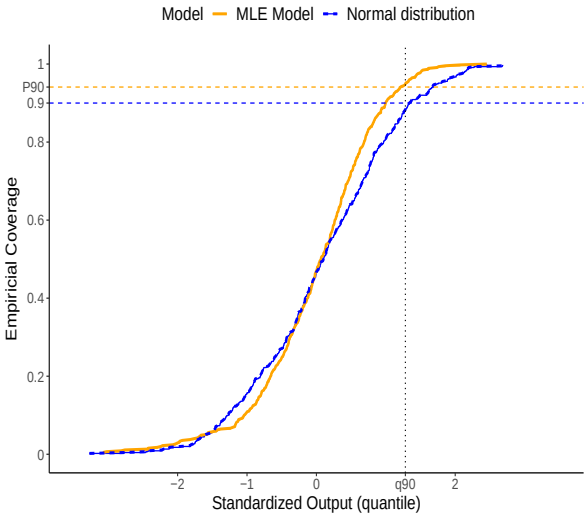
**Proposition 3.3.6.** *Under assumptions 2.2.32 to 3.3.4, the function $\mathcal{L} : (0,+\infty) \to \mathbb{R}^+$ is continuous and coercive on $(0,+\infty)$. The problem $\mathcal{P}_\lambda$ admits at least one global minimizer $\lambda^*$ in $(0,+\infty)$.*

*Proof.* See A.1. ■

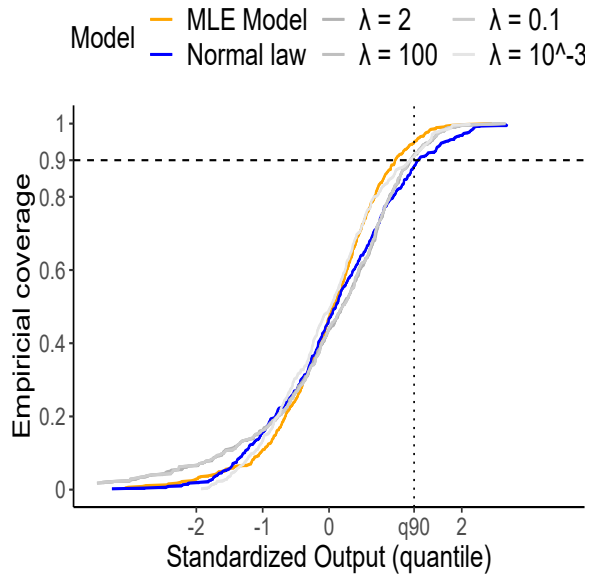
**Remark 3.3.7.** *The coercivity of the function $\mathcal{L}$ is guaranteed by the assumptions 2.2.32 to 3.2.1 (see Appendix A.1). The function $\mathcal{L}$ is also upper semi-continuous (Zhao, 1997). The assumption 3.3.4 ensures that $\mathcal{L}$ is continuous and that a global minimizer exists. This hypothesis is not easy to check. If it does not hold or cannot be checked, then it is possible to solve the problem (3.35) on a regular grid by a grid search method.*



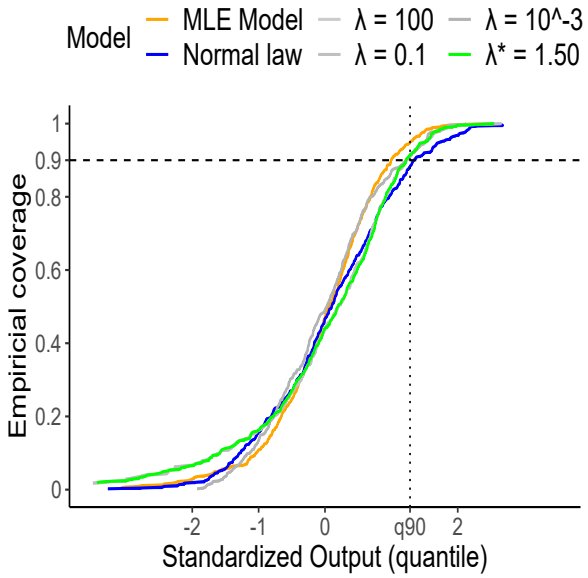

(a) The standardized predictive distribution for the MLE method compared with the standard normal distribution.




(b) Targeting the upper bound with respect to the quantile $q_{90\%}$.




(c) An infinite set of distributions that ensure the coverage of 90% is possible.




(d) The green curve is the optimal distribution with respect to Wasserstein distance.

Figure 3.3: Summary of our approach: In Subfigure 3.3a the model here is misspecified as the standardized predictive distribution with MLE is significantly different from the normal distribution. In Subfigure 3.3b, the upper bound of Prediction Interval with respect to the quantile $q_{90\%}$ is above the coverage of 90%. When trying to ensure the coverage of 90%, we can identify an infinite set of solutions and each solution would give a different distribution as shown in Subfigure 3.3c. With the Wasserstein distance, we manage to choose the closest distribution (green curve in Subfigure 3.3d) to the MLE distribution with the 2-Wasserstein distance.

Let $\widehat{\boldsymbol{\beta}}_{\text{opt}}$ denote the corresponding regression parameter

$$\widehat{\boldsymbol{\beta}}_{\text{opt}}(\lambda^*) = \left(\mathbf{F}^\top \mathbf{K}^{-1}_{\sigma^2_{\text{opt}}(\lambda^*),\lambda^*\boldsymbol{\theta}_0}\mathbf{F}\right)^{-1} \mathbf{F}^\top \mathbf{K}^{-1}_{\sigma^2_{\text{opt}}(\lambda^*),\lambda^*\boldsymbol{\theta}_0)}\boldsymbol{y}. \tag{3.36}$$

The purpose of this resolution is to create a GP model with hyperparameters $(\widehat{\boldsymbol{\beta}}_{\text{opt}}(\lambda^*), \sigma^2_{\text{opt}}(\lambda^*), \lambda^*\boldsymbol{\theta}_0)$ able to predict the quantile $\tilde{y}_a$ such that a proportion $a$ of true values are below $\tilde{y}_a$ with respect to the constraint of *quasi-Gaussian* proportion $\psi_a$ (see Figure 3.3). Finally, the Prediction Intervals $\mathcal{PI}_{1-\alpha}$ will be obtained using two GP models built with the same method, one for the upper bound $u_{1-\alpha} \leftrightarrow \psi_{1-\alpha/2}$ with optimal relaxation parameter $\overline{\lambda}^*$ and the other one for the lower bound $l_{1-\alpha} \leftrightarrow \psi_{\alpha/2}$ with parameter $\underline{\lambda}^*$ (here $\leftrightarrow$ is to denote that there is a correspondence between ... ). The Coverage Probability of $\mathcal{PI}_{1-\alpha}$ is optimal and achieved by respecting the coverage of each bound as shown in (3.7). In the following, we call this method *Robust Prediction Intervals Estimation* (RPIE).

**Remark 3.3.8.** *It is clear that the GP hyperparameters selected by the RPIE method depends on the level a. Given the continuity properties of the different steps of the RPIE method, one may expect, however, that the hyperarameters selected for a specific level a should also give good CP locally for other levels of coverage $a'$ close to a. Nevertheless, this local property is certainly not global. This sensitivity can also be related to the known observation that guarantees for conditional coverage are more challenging to obtain than for marginal coverage (Foygel Barber et al., 2020).*

### Absence of nugget effect

When the nugget effect is null $\sigma^2_\epsilon = 0$, the set of solutions $\mathcal{A}_{a,\delta}$ is still non-empty because one can show that, for $\boldsymbol{\theta}$ in the neighborhood of $\mathbf{0} \in \mathbb{R}^d$, the problem $\psi_a^{(\delta)}(\sigma^2, \boldsymbol{\theta}) = a$ has a solution $\sigma^2 \in (0, +\infty)$ (see A.2). In particular, the correspondence function $H_\delta$ is non-empty valued for $\lambda > 0$ small enough, and it may be empty-valued for some large $\lambda \in (0, +\infty)$. We may think, however, that $H_\delta$ is non-empty valued and that $\sigma^2_{\text{opt}}(\lambda)$ exists for $\lambda$ close to one. Indeed, assume for a while that the model is well-specified, that is, there exist hyperparameters $(\boldsymbol{\beta}_*, \sigma^2_*, \boldsymbol{\theta}_*)$ such that $\boldsymbol{y}$ corresponds to a realization of a random vector $\boldsymbol{Y} \sim \mathcal{N}(\mathbf{F}\boldsymbol{\beta}_*,\ \sigma^2_*\mathbf{R}_{\boldsymbol{\theta}_*})$. The existence of $H_\delta(\lambda)$ and $\sigma^2_{\text{opt}}(\lambda)$ depend on the condition $k_\lambda \leq na$, where $k_\lambda$ is the integer defined by

$$k_\lambda := \text{Card}\left\{i \in \{1, \ldots, n\}, \left(\overline{\mathbf{R}}_{\lambda\boldsymbol{\theta}_0}\boldsymbol{y}\right)_i \leq 0\right\}. \tag{3.37}$$

Since $\overline{\mathbf{R}}_{\boldsymbol{\theta}_*}\boldsymbol{Y}$ is centered, we can anticipate that

$$\text{Card}\left\{i \in \{1, \ldots, n\}, \left(\overline{\mathbf{R}}_{\boldsymbol{\theta}_*}\boldsymbol{y}\right)_i \leq 0\right\} \approx \frac{n}{2}. \tag{3.38}$$

Hence, the condition $n/2 < k_\lambda \leq na$ should be satisfied in a neighborhood of $\lambda = 1$ since $\boldsymbol{\theta}_0$ should be close to $\boldsymbol{\theta}_*$.

For the coercivity, we show in Appendix A.2 that the function $\mathcal{L}$ is coercive for anisotropic geometric Matérn models with smoothness parameter $\nu < 2$. Nevertheless, for Matérn models with smoothness parameter $\nu \geq 2$, the coercivity cannot be satisfied (see A.2 for more discussion), but this should not be an important issue of the method. On the one hand, we may agree that Matérn models with smoothness parameter $\nu \geq 2$ are less robust in Uncertainty Quantification in the free-noise setting. On the other hand, even though the function $\mathcal{L}$ would not be defined

on $(0, +\infty)$, we can solve (3.35) by a grid search method on its domain and pick a minimizer $\lambda^*$, preferably close to 1.

## 3.4 Numerical Results

### Test cases with analytical functions

In this section, we give three numerical examples to illustrate Prediction Intervals estimation by the RPIE method. We show that for the *Wing-Weight* function, the model is well-specified as the CP is optimal for different levels, hence, no robust calibration of Prediction Intervals is required. However, for Zhou (1998) and Morokoff & Caflisch (1995) functions where the model is misspecified and for a given confidence level $\alpha$, we apply the RPIE method as described in section 3.2 to estimate both upper and lower bounds of Predictions Intervals. The following metrics: the Leave-One-Out CP $\tilde{\mathbb{P}}_{1-\alpha}$ defined in (3.7), the Coverage Probability (CP), the mean (MPIW) and standard-deviation (SdPIW) of the Prediction Interval width, and the accuracy $Q^2$ (Kleijnen & Sargent, 2000) are used to assess and compare GP models built by MLE or MSE-CV methods, full Bayesian approach or the RPIE method. They can be used either for point-wise prediction comparison ($Q^2$ will be given in some cases for information, it does not represent the main metric of this section):

$$Q^2 = 1 - \frac{\sum_{i=1}^{n_{test}} \left(y_{test}^{(i)} - \tilde{y}_{i,test}\right)^2}{\sum_{i=1}^{n_{test}} \left(y_{test}^{(i)} - \overline{y}\right)^2}, \tag{3.39}$$

or for quantifying the *goodness* of Prediction Intervals:

$$\tilde{\mathbb{P}}_{1-\alpha} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\{y_i \in \mathcal{PI}_{1-\alpha}(\boldsymbol{x}^{(i)}; \mathbf{D}_{-i})\}, \tag{3.40}$$

$$\mathrm{CP}_{1-\alpha} = \frac{1}{n_{test}} \sum_{i=1}^{n_{test}} \mathbf{1}\{y_{test}^{(i)} \in \mathcal{PI}_{1-\alpha}\left(\boldsymbol{x}_{test}^{(i)}; \mathbf{D}\right)\}, \tag{3.41}$$

$$\mathrm{MPIW}_{1-\alpha} = \frac{1}{n_{test}} \sum_{i=1}^{n_{test}} \left|\mathcal{PI}_{1-\alpha}(\boldsymbol{x}_{test}^{(i)}; \mathbf{D})\right|, \tag{3.42}$$

and,

$$\mathrm{SdPIW}_{1-\alpha} = \sqrt{\frac{1}{n_{test}} \sum_{i=1}^{n_{test}} \left[\left|\mathcal{PI}_{1-\alpha}(\boldsymbol{x}_{test}^{(i)}; \mathbf{D})\right| - \mathrm{MPIW}_{1-\alpha}\right]^2}, \tag{3.43}$$

where $\boldsymbol{y}_{test} = \left(y_{test}^{(1)}, \ldots, y_{test}^{(n_{test})}\right)$ is the vector to predict at $\left(\boldsymbol{x}_{test}^{(1)}, \ldots, \boldsymbol{x}_{test}^{(n_{test})}\right)$, $\mathcal{PI}_{1-\alpha}$ is the $(1-\alpha) \times 100\%$ confidence Prediction Interval delimited by the quantiles $q_{1-\alpha/2}$ and $q_{\alpha/2}$, and $|\mathcal{PI}_{1-\alpha}|$ is the length of the interval.

Note that the $\mathrm{CP}_{1-\alpha}$ may be different from the Leave-One-Out CP $\tilde{\mathbb{P}}_{1-\alpha}$, this case can happen when the distributions of the training and testing sets are different. However, if the Leave-One-Out CP $\tilde{\mathbb{P}}_{1-\alpha}$ is close to $1-\alpha$ and if the assumptions of *i.i.d.* observations and same joint distributions $\pi_{train} = \pi_{test}$ are respected, then $\mathrm{CP}_{1-\alpha}$ should be also close to $1-\alpha$.

This subsection provides results obtained on $d = 10$-dimensional GP with constant mean function (Ordinary Kriging). The value of $\delta$ is fixed at $\delta = 10^{-2}$. We implement our methods using the package *kergp* (Roustant et al., 2020) on R. For the computational time, we use an Intel(R) Core(TM) i5-9400H CPU @ 2.50GHz with a RAM of 32 Go.

### Example 1: Well-specified model - The Wing Weight function

The Wing Weight function is a model in dimension $d = 10$ proposed by Forrester et al. (2008) that estimates the weight of a light aircraft wing. For an input vector $\boldsymbol{x} \in \mathbb{R}^{10}$, the response $y$ is:

$$f(\boldsymbol{x}) = 0.036 x_1^{0.758} x_2^{0.0035} \left(\frac{x_3}{\cos^2(x_4)}\right)^{0.6} x_5^{0.006} x_6^{0.04} \left(\frac{100\ x_7}{\cos(x_4)}\right)^{-0.3} (x_8 x_9)^{0.49} + x_1 x_{10}. \quad (3.44)$$

The components $x_i$ denote some physical and aero-dynamical parameters of the light aircraft wing (e.g. $x_1$ is the wing area in feet squared), see Forrester et al. (2008) and Moon (2010) for details. They are assumed to vary over the ranges given in Table 3.1.

Table 3.1: The input variables $x_j$ and their domain ranges $[a_j; b_j]$.

| Component | Domain | Component | Domain |
|---|---|---|---|
| $x_1$ | $[150; 200]$ | $x_6$ | $[0.5; 1]$ |
| $x_2$ | $[220; 300]$ | $x_7$ | $[0.08; 0.18]$ |
| $x_3$ | $[6, 10]$ | $x_8$ | $[2.5; 6]$ |
| $x_4$ | $[-10; 10]$ | $x_9$ | $[1700, 2500]$ |
| $x_5$ | $[16; 45]$ | $x_{10}$ | $[0.025; 0.08]$ |

We create an experimental design $\mathbf{X}$ of $n = 600$ observations and $d = 10$ variables where observations $\boldsymbol{x}^{(i)} = \left(x_1^{(i)}, \ldots, x_d^{(i)}\right)$ are sampled i.i.d. with uniform distribution over $\bigotimes_{j=1}^{d} [a_j, b_j]$. We generate the response $\boldsymbol{y} = \left(y^{(1)}, \ldots, y^{(n)}\right)$ such that $y_i = f(\boldsymbol{x}^{(i)}) + \epsilon^{(i)}$ with $f$ defined in (3.44) and $\epsilon^{(i)}$ are sampled i.i.d. with the distribution $\mathcal{N}(0, \sigma_\epsilon^2 = 25)$. Here the nugget effect is estimated with the methodology described in Iooss & Marrel (2017) and the covariance kernel is the Matérn 3/2.

The GP model is trained on 75% of the data (25% of data is left for testing). The diagnostics of the model are presented in Table 3.2 with the metrics described above. The accuracy $Q^2$ is moderate for MLE and Full-Bayesian methods. The MSE-CV does much better, an expected result since the MSE-CV method is more adapted for point-wise prediction criterion. However, the Leave-One-Out CP $\tilde{\mathbb{P}}_{1-\alpha}$ for two different levels $\alpha = 5\%, 10\%$ is far from the required level, which means that they were poorly estimated with point-wise prediction criterion. In addition, Table 3.2 shows in particular that the model is well-specified for Matérn 3/2 correlation kernel with the MLE method since the CPs are optimal and close to the required level. This claim is empirical and can be verified either by comparing the standardized predictive distribution with the standard normal distribution as in Figure 3.3 or using Shapiro & Wilk (1965) normality test (in this example, $p$-value $= 0.203$). The Full-Bayesian approach also does well in estimating Prediction Intervals in the case of a well-specified model. Indeed, the hyperparameters' posterior

distribution $p(\sigma^2, \boldsymbol{\theta} \mid \boldsymbol{y})$ is concentrated around the MLE estimator, so the plug-in MLE approach and the Full-Bayesian approach give similar predictive distributions and Prediction Intervals. Morever, the computational cost of the Full-Bayesian approach is extremely long compared to other methods (e.g. 100 times longer than the MLE method). Concerning the RPIE method, one can notice that it provides the optimal coverage at each required level, either on training or testing sets. However, we do not see significant interest in applying it here (except for the MSE-CV solution).

Example 1 is a case of well-specified model in which the CPs obtained by the MLE method satisfy the nominal value and the RPIE method does not bring a significant additional value (at least for the MLE solution).

#### Example 2: Misspecified model with noise - Morokoff & Caflisch function -

We consider the Morokoff & Caflisch (1995) function defined on $[0,1]^d$ by

$$f(\boldsymbol{x}) = \frac{1}{2}\Big(1 + \frac{1}{d}\Big)^d \prod_{i=1}^{d} (x_i)^{1/d}. \tag{3.45}$$

In Example 2, we consider an experimental design $\mathbf{X}$ of $n = 600$ observations and $d = 10$ correlated inputs. Each observation has the form $\boldsymbol{x}^{(i)} = \Big(\Phi(z_1^{(i)}), \ldots, \Phi(z_d^{(i)})\Big) \in \mathbb{R}^d$, $\Phi$ is the CDF of the standard normal distribution, $\boldsymbol{z}^{(i)}$ are sampled from the multivariate distribution

Table 3.2: Performances of methods (MLE, MSE-CV and Full-Bayesian) for Wing Weight function.

| | Before RPIE | | After RPIE | | Full-Bayesian |
|---|---|---|---|---|---|
| | MLE | MSE-CV | MLE | MSE-CV | - |
| $Q^2$ | 0.563 | 0.764 | n.c | n.c | 0.562 |
| $\tilde{\mathbb{P}}_{99\%}$ | 99.1 | 99.8 | 98.9 | 98.9 | 99.1 |
| $\mathrm{CP}_{99\%}$ | 98.7 | 100 | 98.7 | 98.0 | 98.7 |
| $\tilde{\mathbb{P}}_{95\%}$ | 94.0 | 98.9 | 94.9 | 94.9 | 94.2 |
| $\mathrm{CP}_{95\%}$ | 95.3 | 99.3 | 96.7 | 96.0 | 95.3 |
| $\tilde{\mathbb{P}}_{90\%}$ | 90.1 | 96.9 | 90.0 | 90.0 | 90.9 |
| $\mathrm{CP}_{90\%}$ | 91.3 | 96.0 | 89.3 | 90.0 | 91.3 |
| Ct | 2min 12s | 32min 42s | 6min* | 37min* | 4h 39min 27s |

$Q^2$: Accuracy; $\tilde{\mathbb{P}}_{1-\alpha}$: The Leave-One-Out CP in % on the training set; $\mathrm{CP}_{1-\alpha}$: The CP in % on the testing set and Ct: computational time.
*: The approximated cumulative computational time after running the RPIE method for all levels.

$\mathcal{N}(\mathbf{0}, \mathbf{C})$ and $\mathbf{C} \in \mathbb{R}^{d \times d}$ is the following covariance matrix:

$$\mathbf{C} = \begin{bmatrix} 1 & 0.90 & 0 & 0 & 0 & 0.50 & -0.30 & 0 & 0 & 0 \\ 0.90 & 1 & 0 & 0 & 0 & 0 & 0 & 0.10 & 0 & 0 \\ 0 & 0 & 1 & 0 & -0.30 & 0.10 & 0.40 & 0 & 0.05 & 0 \\ 0 & 0 & 0 & 1 & 0.40 & 0 & 0 & -0.35 & 0 & 0 \\ 0 & 0 & -0.30 & 0.40 & 1 & 0 & 0 & 0 & 0.10 & 0 \\ 0.05 & 0 & 0.10 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ -0.30 & 0 & 0.40 & 0 & 0 & 0 & 1 & 0 & 0 & -0.30 \\ 0 & 0.1 & 0 & -0.35 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0.05 & 0 & 0.10 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0. & -0.3 & 0 & 0 & 1 \end{bmatrix}.$$

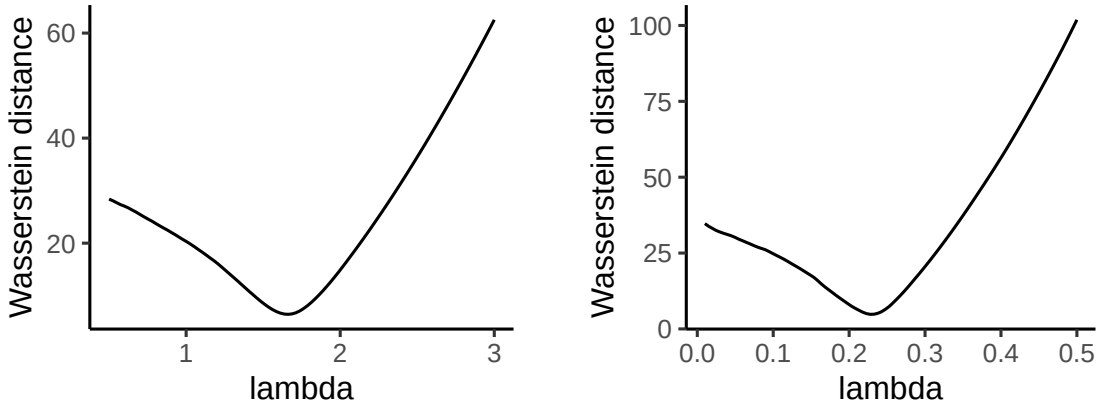
The response vector $\boldsymbol{y}$ is generated as $y_i = f(\boldsymbol{x}^{(i)}) + \epsilon^{(i)}$ with $f$ the *Morokoff & Caflisch* function defined in (3.45) and $\epsilon^{(i)}$ are sampled i.i.d. with the distribution $\mathcal{N}(0, \sigma_\epsilon^2 = 10^{-4})$. We consider the Matérn anisotropic geometric correlation model with smoothness 5/2 as covariance model and we study the Prediction Interval's problem with a nugget effect estimated with the methodology Iooss & Marrel (2017).

Table 3.3: Performances of methods before and after RPIE for Morokoff & Caflisch (1995) function; here $1 - \alpha = 90\%$.

| | Before RPIE | | After RPIE | | Full-Bayesian |
|---|---|---|---|---|---|
| | MLE | MSE-CV | MLE | MSE-CV | - |
| $Q^2$ | 0.892 | 0.895 | n.c | n.c | 0.891 |
| $\tilde{\mathbb{P}}_{1-\alpha}$ | 93.6 | 98.3 | 90.0 | 90.0 | 93.8 |
| $\text{CP}_{1-\alpha}$ | 94.0 | 98.0 | 92.6 | 87.3 | 93.3 |
| $\text{MPIW}_{1-\alpha}$ | $1.68\ 10^{-1}$ | $1.81\ 10^{-1}$ | $5.51\ 10^{-2}$ | $5.78\ 10^{-2}$ | $1.66\ 10^{-1}$ |
| $\text{SdPIW}_{1-\alpha}$ | $9.61\ 10^{-3}$ | $4.16\ 10^{-2}$ | $1.29\ 10^{-2}$ | $1.41\ 10^{-2}$ | $9.27\ 10^{-3}$ |
| Ct | 1min 16s | 24min 18s | 3min 55s | 27min 43s | 4h 43min 38s |

$Q^2$: Accuracy; $\tilde{\mathbb{P}}_{1-\alpha}$: The Leave-One-Out CP in % on the training set; $\text{CP}_{1-\alpha}$: CP in % on the testing set; MPIW: Mean of Prediction Interval widths; SdPIW: standard deviation of Prediction Interval widths and Ct: computational time.

The model is not well-specified as Example 1 and the Shapiro & Wilk (1965) test gives $p$-value $= 1.253\ 10^{-7}$. Table 3.3 summarizes the results of MLE and MSE-CV estimations before and after applying the RPIE, compared with the Full-Bayesian approach. The accuracy $Q^2$ of both models is satisfactory and is slightly improved when using the MSE-CV method. However, before applying the RPIE, the Prediction Intervals are overestimated for both models. The CP does not correspond to the required level of 90%, and the MSE-CV model performs even worse. We note that the Full-Bayesian approach does not improve the quality of estimated Prediction Intervals for the same reason as explained before: the hyperparameters' posterior distribution $p(\sigma^2, \boldsymbol{\theta} \mid \boldsymbol{y})$ is concentrated around the MLE estimator and the performances of both approaches are similar. We will see that this claim is also valid in Example 3.




(a) MLE solution $\boldsymbol{\theta}_0 = \hat{\boldsymbol{\theta}}_{ML}$ (b) MSE-CV solution $\boldsymbol{\theta}_0 = \hat{\boldsymbol{\theta}}_{MSE}$

Figure 3.4: The variation of the *relaxed* Wasserstein distance $\mathcal{L}$ for Morokoff & Caflisch (1995) function; $a = 1 - \alpha/2 = 95\%$.

We now address the problem of Prediction Intervals Estimation for each solution of MLE $\hat{\boldsymbol{\theta}}_{ML}$ and MSE-CV $\hat{\boldsymbol{\theta}}_{MSE}$. We consider the upper and lower bounds $1 - \alpha/2 = 95\%$ and $\alpha/2 = 5\%$ and we apply the RPIE method as described in section 3.2. The optimal values $\overline{\lambda}^*$ and $\underline{\lambda}^*$ obtained from the resolution of the problem (3.35) are used to build two GP models to estimate each bound. Figure 3.4 shows the variation of the function $\mathcal{L}$ for *Morokoff & Caflisch* example while solving the problem (3.35) on the upper bound $1-\alpha/2 = 95\%$, it illustrates the statement of Proposition 3.3.6 : $\mathcal{L}$ is continuous and coercive on $(0, +\infty)$ and reaches a global minimum.

We consider now the Prediction Intervals built according to the RPIE method. In Table 3.3, one observes that these Prediction Intervals are three times shorter than those of MLE, MSE-CV models or Full-Bayesian approaches and have appropriate variances (e.g. more heterogeneous than MLE or Full-Bayesian method's Prediction Intervals). The coverage rate of $1 - \alpha = 90\%$ on the training set is achieved, which is the main objective of the RPIE method, and the CP on the testing set is very close to this level. Concerning the computational time, it appears that applying the RPIE method to MLE or MSE-CV solutions counts for a short computational time (only a few minutes to run in this example). The Full-Bayesian approach is still computationally heavy, as already discussed in the previous example and section 2.3. This represents a competitive advantage of the RPIE method as it delivers good results at a relatively small computational cost compared to the full-Bayesian treatment.

Example 2 is a case of misspecified model with noise in which the CP obtained by MLE, MSE-CV and Full-Bayesian methods are not good. The RPIE method fulfills its purpose: its reduces Prediction Intervals width and improves the robustness of Prediction Intervals in such a way that they achieve the optimal coverage rate.

#### Example 3: Misspecified model without noise - Zhou function -

The Zhou (1998) function, considered initially for the numerical integration of spiky functions, is defined on $[0, 1]^d$ by

$$f(\boldsymbol{x}) = \frac{10^d}{2}\Big[\phi\Big(10(\boldsymbol{x} - \frac{1}{3})\Big) + \phi\Big(10(\boldsymbol{x} - \frac{2}{3})\Big)\Big], \tag{3.46}$$

where

$$\phi(\boldsymbol{x}) = (2\pi)^{-d/2} \exp\left(-0.5\|\boldsymbol{x}\|^2\right). \tag{3.47}$$

In Example 3, we create an experimental design $\mathbf{X}$ similar to Example 1, containing $n = 600$ and $d = 10$ variables where observations $\boldsymbol{x}^{(i)} = \left(x_1^{(i)}, \ldots, x_d^{(i)}\right)$ are sampled independently with uniform distribution over $[0, 1]^d$. As the *Zhou* function in (3.46) takes some high values, we generate the response $\boldsymbol{y}$ by applying a logarithmic transformation:

$$y_i = \log f(\boldsymbol{x}^{(i)})/(d \log 10). \tag{3.48}$$

Note that there is no measurement noise here. We will address two situations: In the first setting, we assume that we know that there is no measurement noise, we impose that there is no nugget effect in the model $\sigma_\epsilon^2 = 0$ and we consider the Exponential anisotropic geometric correlation model ($\nu = 1/2$) as covariance model. In the second setting, we assume that we do not know whether there is measurement noise and we estimate the nugget effect of the model. We consider consequently the Matérn 3/2 anisotropic geometric correlation model ($\nu = 3/2$), a reasonable choice for a smooth covariance model when assuming a nugget effect (See A.2 for further discussion).

Table 3.4: Performances of methods for Zhou (1998) function (3.46) in the first setting ($\sigma_\epsilon^2 = 0$) ; here $1 - \alpha = 90\%$.

| | Before RPIE | | After RPIE | | Full-Bayesian |
|---|---|---|---|---|---|
| | MLE | MSE-CV | MLE | MSE-CV | - |
| $Q^2$ | 0.947 | 0.947 | n.c | n.c | 0.948 |
| $\tilde{\mathbb{P}}_{1-\alpha}$ | 92.0 | 42.1 | 90.0 | 90.0 | 92.0 |
| $\text{CP}_{1-\alpha}$ | 92.7 | 45.3 | 90.0 | 88.0 | 92.9 |
| $\text{MPIW}_{1-\alpha}$ | $4.60\ 10^{-1}$ | $1.46\ 10^{-1}$ | $4.35\ 10^{-1}$ | $4.32\ 10^{-1}$ | $4.59\ 10^{-1}$ |
| $\text{SdPIW}_{1-\alpha}$ | $1.06\ 10^{-1}$ | $3.48\ 10^{-2}$ | $1.00\ 10^{-1}$ | $1.00\ 10^{-1}$ | $1.08\ 10^{-1}$ |
| Ct | 10s | 31min 2s | 2min 31s | 33min 32s | 4h 56min 15s |

$Q^2$: Accuracy; $\tilde{\mathbb{P}}_{1-\alpha}$: The Leave-One-Out CP in % on the training set; $\text{CP}_{1-\alpha}$: The CP in % on the testing set; MPIW: Mean of Prediction Interval widths; SdPIW: standard deviation of Prediction Interval widths and Ct: computational time.

In Table 3.4, the models are good in terms of accuracy $Q^2$ with a small advantage for the Full-Bayesian approach, but none of them satisfies the required level of CP, especially the MSE-CV model with an extremely low CP. As we do not estimate the nugget effect in this setting, the computational time of the MLE method is low (a few seconds) where the RPIE still takes a couple of minutes, as in Example 2. We will notice (also in the industrial application) that the computational time after the RPIE method is generally twice to three times the computational time of MLE method when there is a nugget effect.

When proceeding similarly as Example 2 to build robust Prediction Intervals by the RPIE model, the result is striking in Table 3.4: The estimated Prediction Intervals for the MSE-CV solution $\hat{\boldsymbol{\theta}}_{MSE}$ after RPIE are now four times larger, meaning that the amplitude $\hat{\sigma}^2_{MSE}$ was

largely underestimated. Table 3.4 also shows that the CPs for the testing set are close to their desired value $1-\alpha = 90\%$.

Table 3.5: Performances of methods for Zhou (1998) function (3.46) in the second setting ($\hat{\sigma}_\epsilon^2 = 1.71\ 10^{-2}$) ; here $1-\alpha = 90\%$.

| | Before RPIE | | After RPIE | | Full-Bayesian |
|---|---|---|---|---|---|
| | MLE | MSE-CV | MLE | MSE-CV | - |
| $Q^2$ | 0.941 | 0.944 | n.c | n.c | 0.941 |
| $\tilde{\mathbb{P}}_{1-\alpha}$ | 99.4 | 100 | 90.0 | 90.0 | 99.3 |
| $\text{CP}_{1-\alpha}$ | 99.3 | 100 | 92.0 | 85.3 | 99.6 |
| $\text{MPIW}_{1-\alpha}$ | $6.48\ 10^{-1}$ | 1.19 | $2.26\ 10^{-1}$ | $2.28\ 10^{-1}$ | $6.56\ 10^{-1}$ |
| $\text{SdPIW}_{1-\alpha}$ | $6.88\ 10^{-2}$ | $2.56\ 10^{-1}$ | $4.73\ 10^{-2}$ | $5.27\ 10^{-2}$ | $6.97\ 10^{-2}$ |
| Ct | 1min 20s | 31min 22s | 3min 39s | 33min 37s | 4h 25min 59s |

$Q^2$: Accuracy; $\tilde{\mathbb{P}}_{1-\alpha}$: The Leave-One-Out CP in % on the training set; $\text{CP}_{1-\alpha}$: The CP in % on the testing set; MPIW: Mean of Prediction Interval widths; SdPIW: standard deviation of Prediction Interval widths and Ct: computational time.

In the second setting, the nugget effect is estimated to $\hat{\sigma}_\epsilon^2 = 1.71\ 10^{-2}$ by using Iooss & Marrel (2017). The results of MLE, MSE-CV and Full-Bayesian methods are shown in Table 3.5. The accuracy is still satisfying and similar to the previous setting, but the CP is close to 100%, meaning that the Prediction Intervals of all three methods are overestimated. Table 3.5 shows that, with the RPIE method, we reduce Prediction Intervals width, five times shorter than Prediction Intervals of the MSE-CV solution, and three shorter than Prediction Intervals of the MLE solution. The variances of the obtained Prediction Intervals are between MLE and MSE-CV Prediction Intervals variances. One can notice also a decrease of 50% of the MPIW compared to the first setting, while maintaining an optimal coverage of $1-\alpha = 90\%$.

Example 3 illustrates a case of misspecified model without noise where the RPIE method adjusts Prediction Intervals width and improves the robustness of Prediction Intervals so that the CP is respected. One can also conclude that it is preferable to consider a nugget effect for shorter Prediction Intervals and optimal coverage.

### Application to Gas production for future wells

In this section, we illustrate the interest of the RPIE method in energy production forecasting. It includes many industrial applications such as battery capacity, wind turbine, solar panel performance or, more specifically, unconventional gas wells where a decline in production may be observed. We show that the RPIE can estimate robust Prediction Intervals, covering the lower bounds of level $\alpha/2 = 10\%$ (pessimistic scenario) and the upper bounds of level $1-\alpha/2 = 90\%$ (optimistic scenario).

Indeed, a fundamental challenge of Oil and Gas companies is to predict their assets and their production capacities in the future. It drives both their exploration and development strategy. However, forecasting a well future production is challenging because subsurface reservoirs properties are never fully known. This makes estimating well production with their associated

uncertainty a crucial task. The agencies Securities and Exchange Commission and Society of Petroleum Engineers define specific rules **1P/2P/3P**, known as Petroleum Reserves and Resources Definitions (PRMS), for reserves estimates based on quantile estimates:

- **1P**: 90% of wells produce more than **1P** predictions **(proven)**.
- **2P**: 50% of wells produce more than **2P** predictions **(probable)**.
- **3P**: 10% of wells produce more than **3P** predictions **(possible)**.

These rules are to be disclosed to security investors for publicly traded Oil and Gas companies and aim to provide investors with consistent information and associated value assessments. Many Machine Learning algorithms have shown their efficiency in estimating the median **2P** (e.g. using GP with MLE method, or MSE-CV if interested more in point-wise predictions) but failed to estimate **1P** and **3P**. Thus, the objective of this study is to build a proper estimation of the quantiles $p_{90\%}$ and $p_{10\%}$ by applying the RPIE method described in section 3.2.

Our dataset, *field data*, is derived from unconventional wells localized in the *Utica* shale reservoir, located in the north-east of the United States. It contains approximately $n = 1850$ wells and $d = 12$ variables, including localization, Cumulative Production of natural gas over 12 months in MCFE, completion design and exploitation conditions. The raw dataset can be found at the Ohio Oil & Gas well locator of the Ohio Department of Natural Resources (2022).

Table 3.6: Results obtained for GP model, Random Forest and Gradient Boosting; here $1 - \alpha = 80\%$.

| | MLE | Random Forest | XGBoost |
|---|---|---|---|
| $Q^2$ | 0.872 | 0.870 | 0.885 |
| $\text{CP}_{1-\alpha}$ | 92.8 | 98.1 | 49.8 |
| $\text{MPIW}_{1-\alpha}$ | 1.18 | 1.52 | 0.48 |
| $\text{SdPIW}_{1-\alpha}$ | 0.21 | 0.29 | 0.22 |
| Ct | 14min 37s | 2s | 1min 36s |

$Q^2$: Accuracy; CP: The CP in % on validation set I; MPIW: Mean of Prediction Interval widths; SdPIW: standard deviation of Prediction Interval widths and Ct: computational time.

We standardized the data $(\mathbf{X}, \boldsymbol{y})$, and we divided into a $60\% - 20\% - 20\%$ partition of three datasets: a training set and two validation sets. The response $\boldsymbol{y}$ (Cumulative Production over 12 months in MCFE) is noisy due to the uncertainty of the reservoir parameters in the field. The nugget effect $\sigma_\epsilon^2$ is unknown but estimated to $\widehat{\sigma}_\epsilon^2 = 0.16$ using the method of Iooss & Marrel (2017).

Based on results drawn from the previous subsection and for practical reasons (particularly the computational cost of methods), we will present only the application of the RPIE method on the MLE solution. Table 3.6 shows the performances of the GP model trained by MLE compared with two other statistical models: Random Forest and Gradient Boosting whose Prediction Intervals are estimated using the Bootstrap method. Here we consider the Prediction

Intervals of level $1-\alpha = 80\%$: the lower bound is the 10% quantile ($p_{10\%}$) and the upper bound the 90% quantile ($p_{90\%}$) of the predictive distribution.

The accuracy of the MLE model is 0.873 and has approximately the same accuracy as other models like Random Forest or Gradient Boosting. Furthermore, the CP of the Prediction Intervals of $1-\alpha = 80\%$ is not satisfactory, but it is quite *reasonable* for MLE model compared to Random Forest (overestimated Prediction Intervals) or Gradient Boosting (underestimated Prediction Intervals). Finally, it appears that the GP model requires some computing resources to be built and to estimate its hyperparameters by MLE method.

Table 3.7: Obtained results before and after RPIE method; here $1-\alpha = 80\%$.

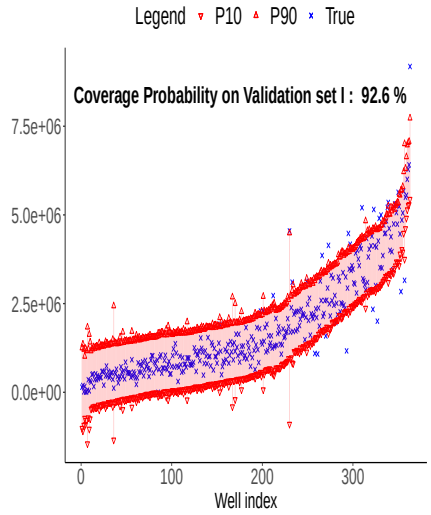
| | MLE before RPIE | MLE after RPIE |
|---|---|---|
| $\tilde{\mathbb{P}}_{1-\alpha}$ | 90.9 | 79.9 |
| $\mathrm{CP}_{1-\alpha}^{\mathrm{Val},1}$ | 92.6 | 81.0 |
| $\mathrm{MPIW}_{1-\alpha}^{\mathrm{Val},1}$ | 1.18 | 1.06 |
| $\mathrm{SdPIW}_{1-\alpha}^{\mathrm{Val},1}$ | $2.09\ 10^{-1}$ | $8.25\ 10^{-3}$ |
| $\mathrm{CP}_{1-\alpha}^{\mathrm{Val},2}$ | 94.1 | 83.2 |
| $\mathrm{MPIW}_{1-\alpha}^{\mathrm{Val},2}$ | 1.17 | 1.06 |
| $\mathrm{SdPIW}_{1-\alpha}^{\mathrm{Val},2}$ | $1.68\ 10^{-1}$ | $7.00\ 10^{-3}$ |
| Ct | 14min 37s | 59min 25s |

$\mathrm{CP}_{1-\alpha}^{\mathrm{Val},1}$ (resp. $\mathrm{CP}_{1-\alpha}^{\mathrm{Val},2}$) : The CP in % on Validation set I (resp. Validation set II); $\mathrm{MPIW}_{1-\alpha}^{\mathrm{Val},1}$ (resp. $\mathrm{MPIW}_{1-\alpha}^{\mathrm{Val},2}$): Mean of Prediction Interval widths on Validation set I (resp. Validation set II); $\mathrm{SdPIW}_{1-\alpha}^{\mathrm{Val},1}$ (resp. $\mathrm{SdPIW}_{1-\alpha}^{\mathrm{Val},2}$): standard deviation of Prediction Interval widths on Validation set I (resp. Validation set II) and Ct: computational time.

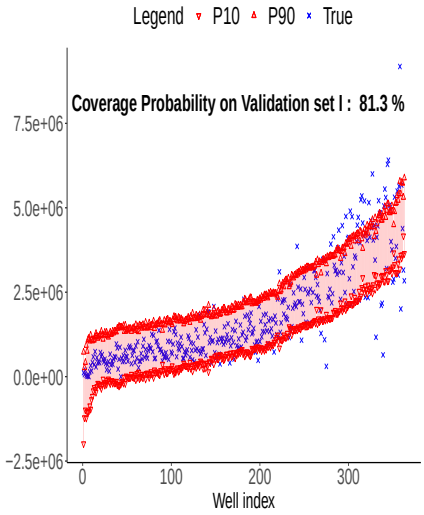
In the following, we define the MLE's solution as reference $\boldsymbol{\theta}_0 = \hat{\boldsymbol{\theta}}_{ML}$ in the optimization problem (3.35) for the quantiles $\alpha/2 = 10\%$ and $1-\alpha/2 = 90\%$ and we build robust Prediction Intervals confidence level $1-\alpha = 80\%$ with the RPIE method. The results are presented in Table 3.7. When considering the estimated Prediction Intervals by the RPIE method, we can see the CP is optimal for the training set and is close to $1-\alpha = 80\%$ for both validation sets. Therefore, we fulfil the objective of estimating the upper and lower bounds, the obtained quantiles $p_{90\%}$ and $p_{10\%}$ respect **1P** and **3P** rules as mentioned above. Finally, in Figure 3.5a, we present the estimated Prediction Intervals defined by the upper bounds $P90$ and lower bounds $P10$ against the true values of $\boldsymbol{y}$ on Validation set I. The x-axis designs well's indices ordered with respect to the barycenters of the Prediction Intervals (engineers choose this representation for interpretation purposes). We can see that the estimated Prediction Intervals by the MLE method are not homogeneous, and some of them are longer. The RPIE method makes them shorter and more homogeneous as it can be seen in Figure 3.5b, and in the evolution of the standard deviation width SdPIW in Table 3.7.

In a second attempt and following the engineers' recommendation, we consider a logarithmic transformation to the raw response $\boldsymbol{y}$ to avoid having non-positive lower bounds and integrate heterogeneity between performant and less performant well. The accuracy of the MLE method decreases now to $Q^2 = 0.615$, the MLE method still overestimates Prediction Intervals as it can be seen in Table 3.8. Most claims of the previous analysis remain true, in particular we

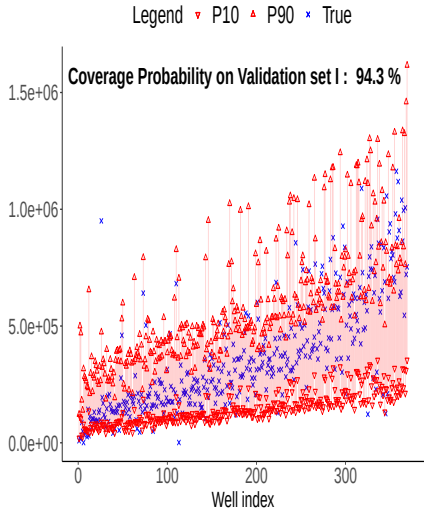
can clearly see (also in Figures 3.5c and 3.5d) that Prediction Intervals obtained by RPIE are shorter and have reduced standard-deviations.



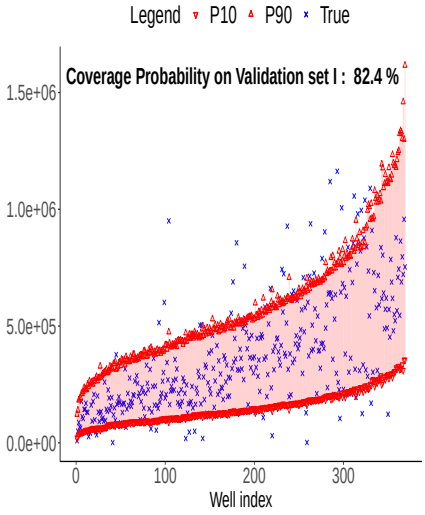

(a) Before the RPIE on standardized output.




(b) After the RPIE on standardized output.




(c) Before the RPIE on log-output.




(d) After the RPIE on log-output

Figure 3.5: Production data after re-scaling: True values vs 80% confidence Prediction Intervals.

### Application to batteries lifetime capacity forecast

The second industrial application is related to Lithium-ion batteries. The original study of Severson et al. (2019) aimed to determine a Lithium-ion battery's cycle lifetime after 100 cycles of discharging. The primary objective is to see if machine learning algorithms are able to accurately predict battery capacity using early life cycle data.

In this subsection, we consider the problem of predicting the battery's cycle lifetime with the associated uncertainty. Indeed, predicting the lower and upper bounds is critical while designing

Table 3.8: Obtained results before and after RPIE method; $1-\alpha = 80\%$. Here the output data are log-transformed.

| | MLE before RPIE | MLE after RPIE |
|---|---|---|
| $\tilde{\mathbb{P}}_{1-\alpha}$ | 91.1 | 79.9 |
| $\text{CP}_{1-\alpha}^{\text{Val},1}$ | 94.3 | 83.2 |
| $\text{MPIW}_{1-\alpha}^{\text{Val},1}$ | 1.53 | 1.40 |
| $\text{SdPIW}_{1-\alpha}^{\text{Val},1}$ | $2.20\ 10^{-1}$ | $1.40\ 10^{-2}$ |
| $\text{CP}_{1-\alpha}^{\text{Val},2}$ | 90.4 | 76.6 |
| $\text{MPIW}_{1-\alpha}^{\text{Val},2}$ | 1.54 | 1.40 |
| $\text{SdPIW}_{1-\alpha}^{\text{Val},2}$ | $1.92\ 10^{-1}$ | $1.42\ 10^{-2}$ |
| Ct | 17min 47s | 53min 21s |

$\text{CP}_{1-\alpha}^{\text{Val},1}$ (resp. $\text{CP}_{1-\alpha}^{\text{Val},2}$) : The CP in % on Validation set I (resp. Validation set II); $\text{MPIW}_{1-\alpha}^{\text{Val},1}$ (resp. $\text{MPIW}_{1-\alpha}^{\text{Val},2}$): Mean of Prediction Interval widths on Validation set I (resp. Validation set II); $\text{SdPIW}_{1-\alpha}^{\text{Val},1}$ (resp. $\text{SdPIW}_{1-\alpha}^{\text{Val},2}$): standard deviation of Prediction Interval widths on Validation set I (resp. Validation set II) and Ct: computational time.

new batteries. It may help companies to cover themselves against earlier failures of batteries (e.g. for maintenance purposes or subscribing to insurance).

The original dataset of Severson et al. (2019) contains 43 cell batteries for training and 42 for validation. However, we combine both datasets in one dataset $\mathbf{D}$ of $n = n_{\text{train}} = 85$ cell batteries to minimize the sensitivity of the Leave-One-Out Empirical Coverage Probability

$$\tilde{\mathbb{P}}_{1-\alpha} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{y_i \in \mathcal{PI}_{1-\alpha}(\boldsymbol{x}^{(i)}; \mathbf{D}_{-i})\} \tag{3.49}$$

to changes whenever there is a point that falls (or not) within a prediction interval. There is no validation set for this application as it will be designed later by engineers.

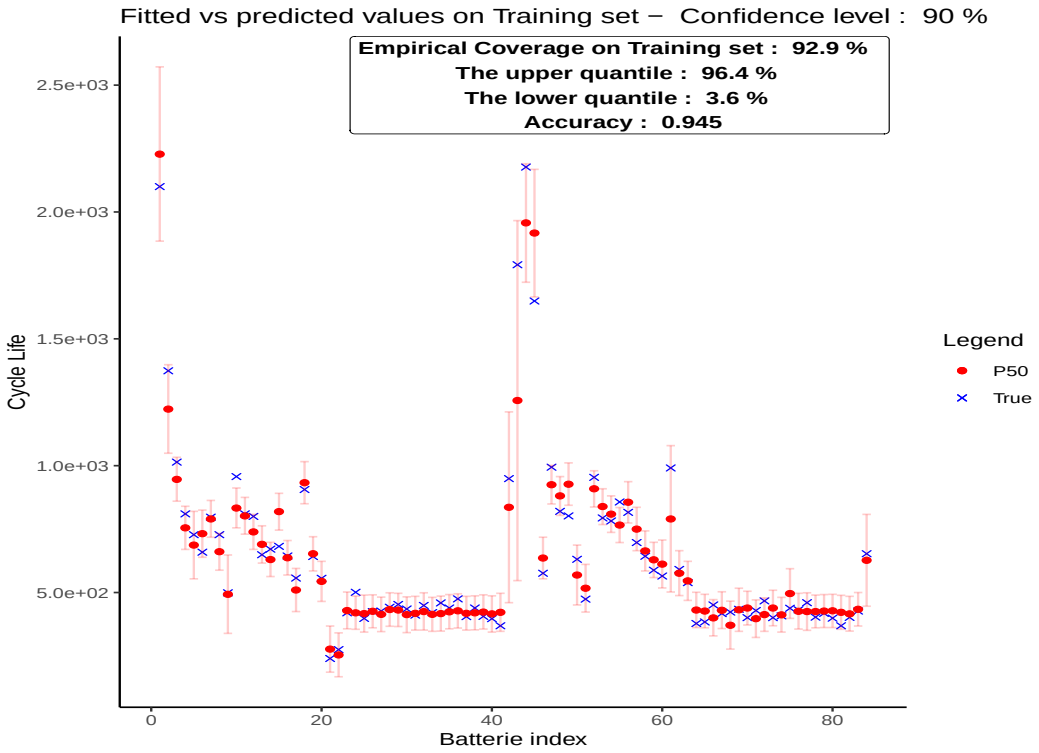
The experimental design $\mathbf{X}$ is $d = 8$-dimensional. It has eight features, including the change in discharge capacity, the discharge capacity fade curve features and other features such as the average charging time, temperature and internal resistance. The output $y$ to predict is the battery's cycle lifetime, corresponding to the number of cycles before 80% of the initial discharge capacity.

We consider a Matérn 3/2 anisotropic geometric model. We assume the existence of a nugget effect $\sigma_\epsilon^2$, and we consider a confidence level of $1-\alpha = 90\%$.

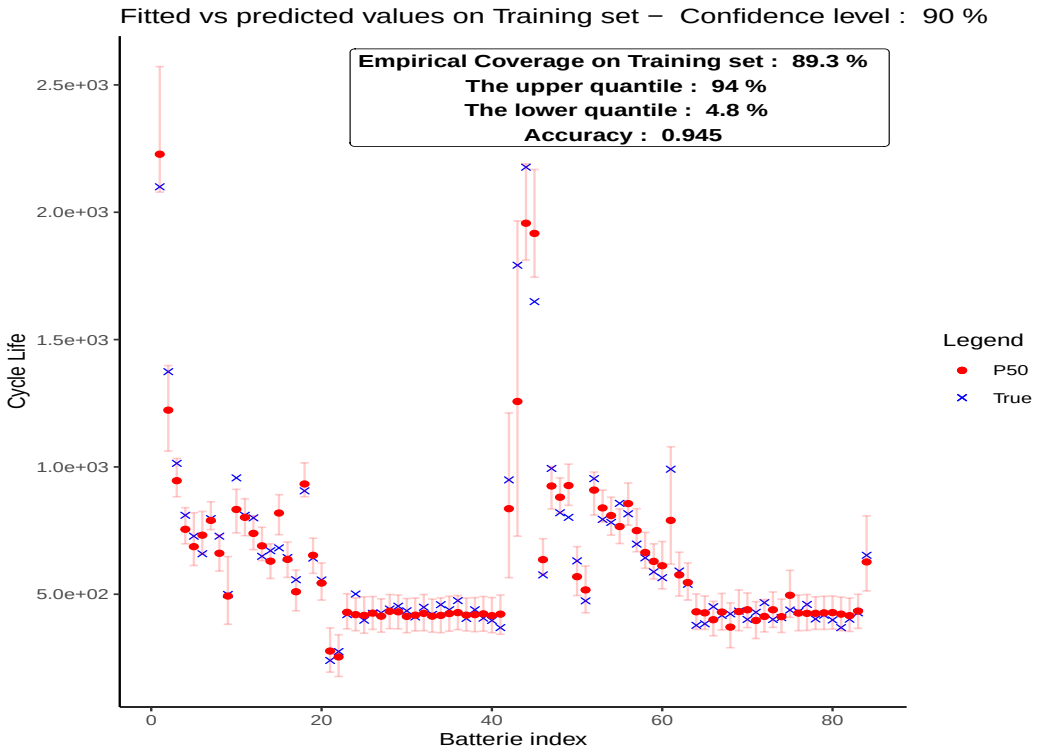
The results of both the Maximum Likelihood and the RPIE methods are shown in Figure 3.6. The accuracy of prediction is put as an informative criterion. One can notice that the initial Prediction Intervals coverage was above the desired level. The upper and lower bounds overestimate the true bounds of Prediction Intervals. With the RPIE method, we reduce their width and achieve a reasonable coverage close to $1-\alpha = 90\%$. We warn the reader that, because of the sample size, the empirical coverage cannot be set to the exact level of $1-\alpha = 90\%$.

## 3.5 Conclusion

In this chapter, we have introduced a new approach for Prediction Intervals estimation based on the Cross-Validation method. We use the Gaussian Processes model because the predictive distribution at a new point is completely characterized by Gaussian distribution. We address an optimization problem for model's hyperparameters estimation by considering the notion of Coverage Probability. The optimal hyperparameters are identified by minimizing the Wasserstein distance between the Gaussian distribution with the hyperparameters determined by Cross-Validation, and the Gaussian distribution with hyperparameters achieving the desired Coverage Probability. This method is relevant when the model is misspecified. It insures an optimal Leave-One-Out Coverage Probability for the training set. It also achieves a reasonable Coverage Probability for the validation set when it is available. The method can be also extended to other statistical models with a predictive distribution, but more detailed work is needed to consider the influence of hyperparameters on Prediction Interval's coverage and solve the optimization problem more efficiently in these cases. Finally, it should be possible to include categorical inputs in the covariance function by using group kernels (Roustant et al., 2020), which would extend the application range of the RPIE method.




(a) Before the RPIE method.




(b) After the RPIE method.

Figure 3.6: Batteries cycle lifetime: True values vs 90% confidence Prediction Intervals.

# PART II

# Causal Inference and estimation of treatment effects

*Data do not give up their secrets easily. They have to be tortured to confess.*

— Thomas C. Redman, *Data Driven*

Predicting an event or outcome is good; providing the associated uncertainty to anticipate risks is even better; however, gaining a deeper understanding of why it would happen is more important.

Most questions raised in the energy industry are not predictive but rather causal. Causal relationships, by definition, are invariant and hold across various circumstances and environments. Causality is thus an exciting tool for overcoming some predictions' current limitations.

# CHAPTER 4

# State of the Art

In this chapter, we begin in Section 4.1 by presenting a historical introduction to Causality and its main two directions in research. In Section 4.2, we introduce the Potential outcome theory and Rubin Causal model In Sections 4.3 and 4.4, we present the framework of estimating average and heterogeneous treatment effects for a binary treatment. Finally, in Section 4.5 we present the ongoing research on the extension of Rubin Causal Model to multiple and continuous treatments.

## 4.1 Introduction to Causality

### A historical and philosophical introduction

Causality refers to the study of cause-and-effect relationships observed during day-to-day experiences. These (relationships) can be related to any field of intellectual, social or political activities of a human being.

The concept of Causality is fundamental and is one of the most important mechanisms by which our mind works—seeking questions such as: *what causes* $Y$? *Why does* $Y$ *occur*? *What would happen to* $Y$ *if ...?* These questions depend entirely on Causality, which results from the mechanism of attaching and connecting the cause to its effect. Specifically, Causality seeks to identify how effects (or events) come to be (or are caused) by their causes.

From the dawn of philosophy, Causality has been a major concern of many philosophers and scientists due to the constant evolution and persistence of the phenomena of the universe within their senses. Causality research is an important area in which philosophy in general, and philosophy of science in particular, have been involved. The philosophical statement of Causality states that every phenomenon, whether physical, social, political or other, has a principle (a cause) which explains its existence (to cause it).

Furthermore, Causality raised many problems at anthropological (existential) and epistemological (cognitive) levels. It led to paradoxes that have caused confusion among scientists and philosophers, especially in the early 20th century. One of these contradictions is the grandfather paradox. When someone travels into the past and kills his grandfather, changing his past, it cancels out the possibility of its existence. Here the paradox shows how that person (the effect) can travel into the past and kill the cause for his existence (the grandfather), and the result becomes a precedent for the cause why it would happen.

Aristotle (ca. 300 BC) is the first philosopher to have considered the theory of the four causes

to understand the human experience of physical nature and answer the question *"because of what?"*. He wrote, *"We think we have knowledge of a thing only when we have grasped its cause"* (Physics, 194 b17–20). Galilei (1600) was the first scientist to consider an interventionist notion of the cause, which is a set of necessary and sufficient conditions for an effect to occur. It can be stated with the condition clause *"If... then ..."*.

In the 18th century, the concept of Causality received revolutionary contributions from David Hume and formed the bedrock of most contemporary studies about Causality. Hume in his book [A Treatise of Human Nature (1739-40)] made an explicit definition of Causality based on the regular succession of event-types: *"We may define a Cause to be an object precedent and contiguous to another, and where all the objects resembling the former are placed in like relations of precedency and contiguity to those objects that resemble the latter"*. This definition implies four required components 1) The constant conjunction of cause and effect; 2) The temporal priority of the cause; 3) The contiguity in space and time and 4) the necessary association between the cause and the effect.

In the contemporary discussions about Causality, three major approaches were proposed:

- Causality as INUS conditions: This notion was introduced by Mackie (1965, 1974) to describe the cause as *"an Insufficient but Necessary part of a condition which is itself Unnecessary but Sufficient for the result"*. Mackie gave the example of a burning house caused by an electrical short circuit to illustrate the INUS concept.

- Causality as probabilistic causation: Suppes (1968) developed a probabilistic framework of Causality. He discovered that many causal relationships could be seen as probable occurrences or chances of events. The cause is an event whose occurrence makes the occurrence of another event more likely to happen than if the first event had not occurred.

- Causality from Counterfactual perspective: Some contemporary philosophers like Lewis (1973, 1986, 2000) see that causal relationships can be understood in terms of counterfactual dependence. That is, 'if $X$ had not occurred, then $Y$ would not have occurred. This definition was introduced but never explored by Hume in Section VII *"if the first object had not been, the second never had existed"*.

Russell (1912) argued that if Causality had more empirical meaning than just one (effect) following another (cause) in time, then it would have been worth appearing in the laws of physics. Indeed, he noticed physical theories are incompatible with causation, as it was understood before because most laws of physics go both ways. Physicists didn't notice what Russell observed, but Statisticians did and raised the need to distinguish causal relationships between two variables. The concept of correlation emerged as an attempt when Galton (1886) decided to do a survey on the relationship between arm length and head size and established a dependence between two variables with abstract numbers. Pearson (1892, 1896), the founder of modern statistics, did not see the importance of the concept of Causality and thought the idea of correlation was enough.

Twenty-five years later, Fisher (1925, 1935) introduced randomization as a critical notion for designing, conducting and analyzing experiments. Randomized Control Trials are popular among statisticians and have been considered (and still are) one of the scientifically proven methods for evaluating causal effects in social and clinical experiments.

However, one of the main reasons that refrained statisticians from the concept of causation is the lack of a formal mathematical language to describe Causality, from both a theoretical and a practical perspective. Pearl & Shafer (1995) realized that the concept of conditional independence was insufficient and argued that a different approach was needed to address the issue of Causality, which was one of his concerns at the beginning of the 21st century. Pearl (2009) was the first to bring a causal mathematical framework and was later joined by others (Pearl & Mackenzie, 2018; Peters et al., 2017) in publishing outstanding works to tackle the problem of Causality in statistics.

To distinguish between tools for associational modelling and causal modelling, Pearl (2019) introduces a 3-level hierarchy based on the kind of information required to answer questions at each level. The three levels are: i) Association, ii) Intervention and iii) Counterfactuals as illustrated in Figure 4.1 with some examples of questions at each level.

The first level corresponds to associational and predictive reasoning from observations. The purpose here is to identify statistical relations (correlation, Odds ratio, dependencies etc.) using exclusively data. Most (but not all) questions at this level can be addressed using classical Machine Learning models.

The second level corresponds to interventional reasoning and predicts what will happen when a system is changed. At this level, we focus on understanding the effects of causes as stated by Holland (1986): *"No causation without manipulation"*. In many frameworks, the intervention can be hypothetical.

The third level corresponds to counterfactual reasoning. Questions at this level are more about what would have happened to the system if circumstances were different not what has happened to the system. Associational or interventional reasoning are not enough to answer them. The counterfactual reasoning is mostly used to reason about the causes of effects but also for the effects of causes from retrospective view.

The second and final layers of Pearl (2019) hierarchy allow to answer many causal questions such that:

- Medicine: Was it the aspirin that stopped my headache? Would I still have had the headache if I did not take Aspirin ?
- Economy: How effective are financial incentives for teachers (Imberman, 2015)?
- Sociology: Did busing programs increase the school achievement of disadvantaged minority youth (Morgan & Winship, 2014)?
- Politics: Do polls influence the electoral choice and behavior of voters (Arceneaux et al., 2006)?
- Industry: What is the effect of a specific efficiency measure (e.g. type of insulation material) on the expected return on investment (EROI) ?

### Survey on Causality methods

Over the last recent decades, several formal frameworks for causality have been proposed and aimed to answer rigorously causal questions. These frameworks had multi-disciplinary applications, especially in epidemiology, economics and statistics. The study of causality is

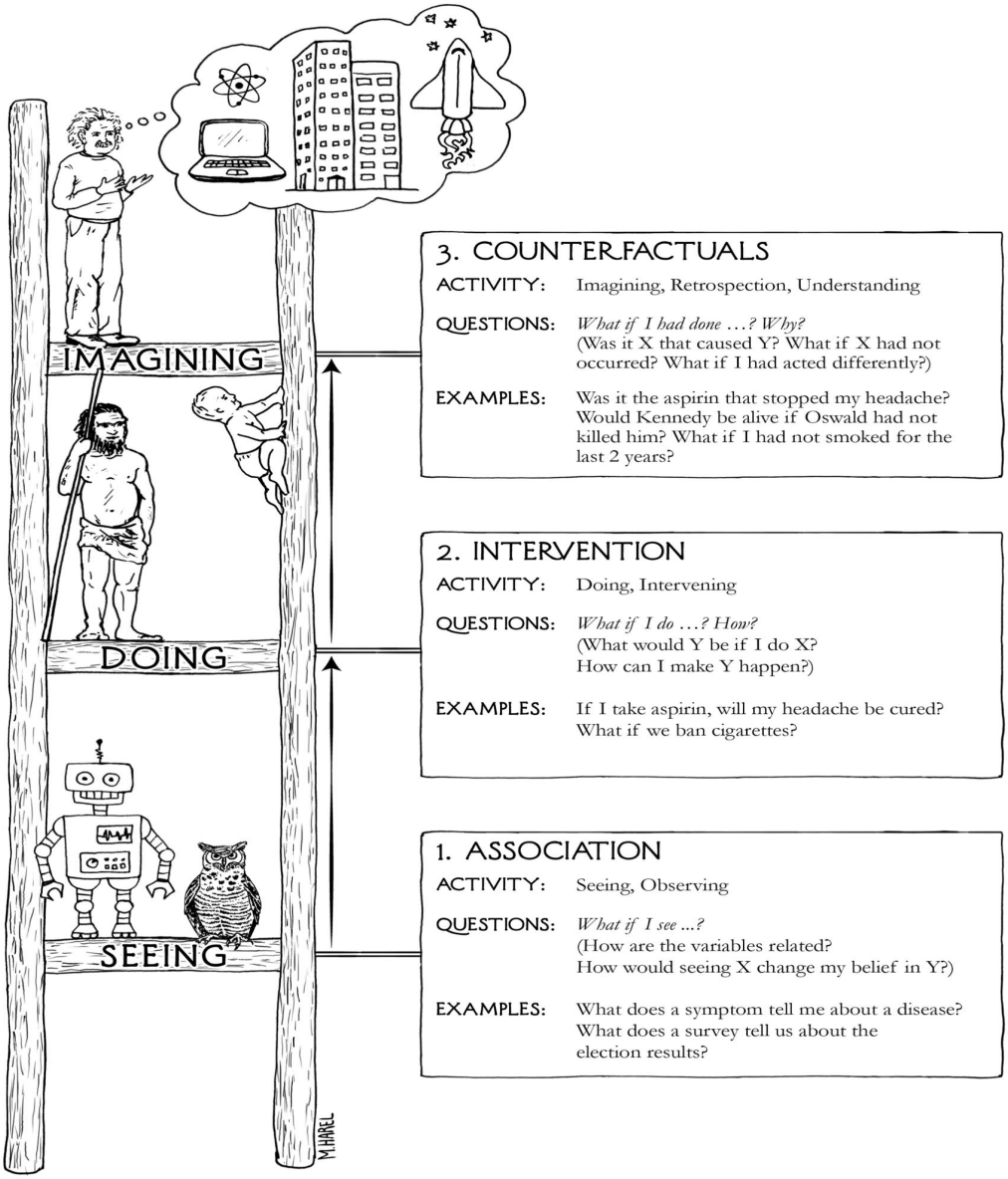



Figure 4.1: The three levels of causal hierarchy (Pearl, 2019).

fragmented into two main complementary tasks: causal discovery and causal inference. From a set of observational data, the first task is to infer the causal relationships between different variables. The second task is to determine and assess the effect of one variable on the other.

### Causal discovery

Causal discovery aims to study and identify causal relationships between a set of variables $\boldsymbol{X}$. The idea is to analyze a given dataset and learn a Directed Acyclic Graph (DAG), called *causal diagram* (Pearl, 1995), that encodes the causal structure of the system described by the dataset. However, learning the true causal diagram is not always possible from observational data. One

would require knowledge or assumptions on the underlying data generating mechanism (Pearl, 2009).

One of the most common assumptions on the underlying true DAG is to assume that the variables $\boldsymbol{X}$ form a Bayesian Network and that the associated DAG satisfies the *global Markov condition*, the *d-separation* and *the Faithfulness assumption*. We refer the reader to Pearl (1995, 2009) for more details of these notions.

Once the structure of a Causal diagram is defined, Causal discovery uses methods to scan the dataset and identify statistical dependencies between variables (Peters et al., 2017; Schölkopf et al., 2012). It is also common to use Structural Causal Models (SCMs) (Pearl, 2009) to identify causal relationships between two variables. These models use functional relationships between ancestor and descendent nodes in DAGs and fit them on the observed dataset.

Often, causal discovery methods (see Guo et al. (2020) for an in-depth review) do not learn only a unique causal graph but a set of candidate causal graphs that may generate the observational distributions. We evaluate the learned causal graphs with the ground-truth causal graph using the concept of the *Markov equivalence class* (Pearl, 1995). Based on this notion, several popular structure learning algorithms have been proposed to estimate this Markov equivalence class from observational data:

- constraint-based algorithms (PC-algorithm, (Spirtes & Glymour, 1991) and FCI-algorithm (Spirtes et al., 2000)) which rely on statistical tests to verify if a candidate graph fits all conditional dependencies from the data and choose the candidate that respects the faithfulness assumption.
- Score-based algorithms relax the faithfulness assumption and use score-based penalties (e.g. Gaussian likelihood penalization or Bayesian Information Criterion) to replace conditional independence tests (the Greedy Equivalence Search algorithm (Chickering, 2002)).
- Functional Causal Models algorithms (LiNGAM (Shimizu, 2014; Shimizu et al., 2006)) estimate SCM of a variable and its direct causes by assuming the non-Gaussianity of the data. They offer a framework to distinguish different DAGs in the same equivalence class.

### Causal Inference

Causal inference is the study of the causal effects of variables. It assumes a relationship among variables and aims to quantify the causal impact of a specific variable over a particular outcome of interest. One can imagine that causal effects can be quantified in two different settings: 1) Through an intervention or a manipulation in the system, given a causal structure that describes it or describes the phenomenon of interest. 2) Via some observational data that can be examined with respect to some causal assumptions when the system's causal structure is unknown.

There are two main frameworks for causal inference. The first framework is the Potential Outcomes theory, which originated from randomized controlled trials (RCTs) in the 1920s by Neyman (1923) and Fisher (1925) and improved formally by Rubin (1974, 1978, 2005, 1990) to infer causal effects from observational data.

The second framework is the *do-calculus* and Directed Acyclic Graphs (DAGs), developed by Pearl (2009) few years ago. The do-calculus is a formalization of causal models that uses *do-operator* which simulates interventions among systems and allows identifying causal effects.

These frameworks are complementary, with different strengths that make them particularly appropriate for different questions—both have shown a considerable interest for statisticians and researchers in various fields. We may refer the reader to Imbens (2020) for in-depth literature about both approaches.

We will consider the potential outcome theory for causal inference in the following.

## 4.2 The potential outcome framework

The potential outcome theory, known as Neyman-Rubin Causal Model, found its origins in the works Neyman (1923) and Fisher (1925). Indeed, in his thesis, Neyman (1923) analyzed randomized hypothetical agricultural experiments. He introduced the potential outcome notation to describe the potential yields of crops associated with distinct plots of land. The potential outcome was developed later by Rubin (1974, 1978, 1990) to perform causal analysis of randomized and non-randomized experiments. It has taken its place as the primary approach in causal inference literature shown its applicability in medicine (Alaa & van der Schaar, 2017; Foster et al., 2011; Robins et al., 2000) economics (Angrist et al., 1996; LaLonde, 1986) and social sciences (Murnane & Willett, 2010; Sobel, 1995). We refer to Rubin (2005) for a detailed history of the Potential Outcome theory and to Imbens & Rubin (2015) for a detailed description of the Rubin Causal Model. Most definitions and notations of this subsection are taken from the same book.

Let $T$ denote the random variable designing the treatment of interest (e.g. drug, policy). We suppose in this section that the treatment is binary $T \in \{0, 1\}$. Let $\boldsymbol{X}$ denote the $d$-vector of pre-treatment covariates (e.g. age, design) and let $Y$ denote the response variable, also called the outcome. To assess the notion of a cause, the treatment $T$ must be manipulable (at least hypothetically), and $Y$ should define the real-valued effect of this cause.

**Definition 4.2.1** (Observed outcomes)**.** *We define the observed outcome $Y_{\text{obs}}$ as the outcome of the treatment that is actually assigned.*

**Definition 4.2.2** (Counterfactual outcomes)**.** *We define the counterfactual outcome $Y_{\text{cf}}$ as the outcome that would have been observed if another treatment had been assigned.*

Let $i$ be a unit (e.g. a person, a system) with covariates $\boldsymbol{x}^{(i)} = (x_1^{(i)}, \dots, x_d^{(i)})^\top \in \mathcal{D}$ subject to the treatment $T$. Let $t_i$ denote the treatment that is actually assigned to the unit $i$. When exposed to this treatment , the unit $i$ responds and shows an outcome $y_i$. In reality, the outcome $y_i$ correspond to the observed outcome $y_i = Y_{\text{obs},i}$ that we have observed in the unit $i$ after assigning the treatment $T = t_i$. Finally, the expression $Y_i(t)$ stands for *potential outcomes*, that we define below:

**Definition 4.2.3** (Potential outcomes)**.** *For a unit $i$, we define the potential outcomes $Y_i(t)$ as the real-valued outcome that would have been observed if the treatment $T$ had been at level $t$.*

*In the case of binary treatment $T \in \{0, 1\}$, the potential outcomes are denoted by $Y_i(0), Y_i(1)$.*

The following assumption is needed to ensure that potential outcome $Y(t)$ is well-defined.

**Assumption 4.2.4** (Stable Unit Treatment Value Assumption (SUTVA))**.** *The potential outcomes for any unit do not vary with the treatments assigned to other units, and, for each unit, there are no different forms or versions of each treatment level, which lead to different potential outcomes.*

Potential outcomes can be expressed using the observed $Y_{\text{obs}}$ and counterfactual $Y_{\text{cf}}$ outcomes as

$$\begin{aligned} Y_{\text{obs}} &= TY(1) + (1-T)Y(0), \\ Y_{\text{cf}} &= (1-T)Y(1) + TY(0). \end{aligned} \tag{4.1}$$

These notations are more relevant when inferring causal effects (Holland & Rubin, 1988).

Suppose now that we have observed a finite sample of $n$ units $\mathbf{D}_{\text{obs}} = (D_{\text{obs},i})_{i=1}^n = (\boldsymbol{x}^{(i)}, t_i, y_i)_{i=1}^n$. Each unit $i$ has a vector covariates denoted $\boldsymbol{x}^{(i)}$ to whom is assigned (actually) a binary treatment $t_i \in \{0,1\}$ and shows an outcome of interest $y_i$.

Following the ideas of Imbens & Rubin (2015), the observed sample $\mathbf{D}_{\text{obs}}$ can be seen as a random sample drawn from an *infinite super-population* with a joint distribution $p_D$ of $D = (\boldsymbol{X}, T, Y(0), Y(1))$. We can assume, therefore, that $(\boldsymbol{x}^{(i)}, t_i, Y_i(0), Y_i(1))$ are *independently and identically distributed (i.i.d.)* from the same distribution $p_D$.

Following this framework and the model (4.1), the SUTVA assumption holds immediately and implies that $Y_{\text{obs},i} = Y_i(t_i) = y_i$ for all units $i$.

In the following, unless otherwise indicated, $\mathbb{P}$ and $\mathbb{E}$ refer to the probability and expectation taken over the joint distribution $p_D$ of $D = (\boldsymbol{X}, T, Y(0), Y(1))$. All causal estimands in the following will be considered with respect to this joint distribution.

**Proposition 4.2.5.** *Let the distribution $p_{Y(0),Y(1),\boldsymbol{X}}$ denote the joint distribution of potential outcomes and $\boldsymbol{X}$, called the model for Science, and let the distribution $p_{T|\boldsymbol{X},Y(0),Y(1)}$ denote the so-called assignment mechanism. For $(\boldsymbol{x}, t, y_{\text{obs}}, y_{\text{cf}}) \in \mathcal{D} \times \{0,1\} \times \mathbb{R} \times \mathbb{R}$, we have*

$$p_{Y_{\text{cf}}|\boldsymbol{X},T,Y_{\text{obs}}}(y_{\text{cf}} \mid \boldsymbol{x}, t, y_{\text{obs}}) \propto p_{\boldsymbol{X},Y(0),Y(1)}(\boldsymbol{x}, y_0, y_1) \times p_{T|\boldsymbol{X},Y(0),Y(1)}(t \mid y_0, y_1, \boldsymbol{x}), \tag{4.2}$$

*where $y_0 = (1-t)y_{\text{obs}} + ty_{\text{cf}}$, $y_1 = ty_{\text{obs}} + (1-t)y_{\text{cf}}$ and $p_{Y_{\text{cf}}|\boldsymbol{X},T,Y_{\text{obs}}}$ is the posterior predictive distribution of the counterfactual outcomes $Y_{\text{cf}}$ given the observed values of $T, \boldsymbol{X}$ and $Y_{\text{obs}}$.*

*Proof.* In Appendix B.1. ■

Therefore, by specifying the assignment mechanism, the model for Science and conditionally on all observed quantities $\boldsymbol{X}, T, Y_{\text{obs}}$, we can address a Bayesian framework that allows predicting the counterfactual outcomes. This framework can be used to infer causal effects based on the notion of counterfactuals. We refer to Rubin (1978, 2005) for more details about the Bayesian framework of causal inference.

The treatment assignment mechanism $p_{T|\boldsymbol{X},Y(0),Y(1)}$, which is a function of the covariates and the potential outcomes, is crucial for causal inference. One must define a probabilistic model for the treatment assignment mechanism when inferring causal effects. A commonly used treatment assignments in experiments is randomization.

**Definition 4.2.6** (Randomized Experiments)**.** *A randomized experiment is an assignment mechanism such that:*

- *The assignment mechanism is ignorable: the assignment mechanism does not depend on the counterfactual outcomes, that is,*

$$\mathbb{P}(T = 1 \mid \boldsymbol{X}, Y(0), Y(1)) = \mathbb{P}(T = 1 \mid \boldsymbol{X}, Y_{\text{obs}}) \tag{4.3}$$

- *The assignment mechanism is probabilistic: the probability of treatment assignment to a unit satisfies*

$$0 < \mathbb{P}(T = 1 \mid \boldsymbol{X}, Y(0), Y(1)) < 1. \tag{4.4}$$

- *The assignment mechanism is a known function of its arguments.*

**Definition 4.2.7** (Randomized Control Trials (RCT))**.** *A randomized controlled trial is a randomized experiment such that*

$$T \perp\!\!\!\perp \{\boldsymbol{X}, Y(0), Y(1)\}. \tag{4.5}$$

**Definition 4.2.8** (Observational studies)**.** *An assignment mechanism corresponds to an observational study if it is an unknown function of its arguments.*

Using the potential outcome theory, we want to infer the effect of the treatment $T$ on the outcome $Y$ from the sample of $n$ units. Causal Effects can be estimated by comparing the potential outcomes of a given treatment assignment.

**Definition 4.2.9** (The Individual Treatment Effect)**.** *The Individual Treatment Effect (ITE) corresponds to the difference between its potential outcomes under treatment and control*

$$\tau_{\text{ITE},i} = Y_i(1) - Y_i(0). \tag{4.6}$$

**Remark 4.2.10.** *$\tau_{\text{ITE},i}$ can be seen as a realization of the (unobserved) random variable $Y(1) - Y(0)$.*

It is impossible to infer this effect directly using $\mathbf{D}_{\text{obs}}$. Indeed, for every unit, by definition of the potential outcomes, we observe only one potential outcome $Y_{\text{obs}}$ corresponding to the potential outcome receiving the treatment $T$, all other potential outcomes $Y_{\text{cf}}$ are missing. This is known as the Fundamental Problem of Causal Inference (Holland, 1986). Hence, causal inference with the Rubin Causal Model can be seen as a missing data problem (Rubin, 2005). Instead, we can target the Average Treatment Effect (ATE) among the observed sample.

**Definition 4.2.11** (The Average Treatment Effect)**.** *The Average Treatment Effect is the treatment effect among the whole sample*

$$\tau = \mathbb{E}[Y(1) - Y(0)]. \tag{4.7}$$

**Remark 4.2.12.** *In the SCM framework, the definition of Average Treatment Effect is equivalent to $\mathbb{E}[Y_{\text{obs}} \mid do(T = 1)] - \mathbb{E}[Y_{\text{obs}} \mid do(T = 0)]$.*

The ATE is much easier to estimate than the ITE because one only needs to compute the means of the marginal distributions of the two potential outcomes. If the treatment is randomly assigned as in RCTs, then $\mathbb{E}[Y_{\text{obs}} \mid T = t] = \mathbb{E}[Y(t)]$ and

$$\tau = \mathbb{E}[Y_{\text{obs}} \mid T = 1] - \mathbb{E}[Y_{\text{obs}} \mid T = 0]. \tag{4.8}$$

However, RCTs are not always conducted, the knowledge of the treatment $T$ and the outcome $Y_{\text{obs}}$ alone does not suffice to identify the true ATE (Hernan & Robins, 2020), this is due to confounding variables.

**Definition 4.2.13** (Confounding variable)**.** *In observational studies, confounding variables, also called confounders, are the variables that influence both the treatment and the outcome.*

We shall make the following assumptions for the identifiability of the causal estimands in observational studies.

**Assumption 4.2.14** (Unconfoundedness)**.** *The potential outcomes $Y(1)$ and $Y(0)$ are independent of the treatment assignment $T$ given the covariates $\boldsymbol{X}$*

$$\{Y(1), Y(0)\} \perp\!\!\!\perp T \mid \boldsymbol{X}. \tag{4.9}$$

This assumption, also called *"strong ignorability"*, assumes that there are no unmeasured confounding variables given the observed covariates $\boldsymbol{X}$. The potential outcomes should be the same for a unit, whether the treatment is assigned or not.

**Assumption 4.2.15** (Overlap)**.** *The probability of receiving the treatment given the observed covariates is positive, that is, there exists $e_{\min} > 0$ such that*

$$e_{\min} < \mathbb{P}(T = 1 \mid \boldsymbol{X} = \boldsymbol{x}) < 1 - e_{\min} \quad \textit{for all } \boldsymbol{x} \in \mathcal{D}. \tag{4.10}$$

The overlap condition is necessary for the identifiability of treatment effects on the support $\mathcal{D}$ because it avoids the degenerate case where all units are either treated or untreated.

The previous assumptions allow to identify the counterfactual response $\mathbb{E}(Y(t) \mid \boldsymbol{X} = \boldsymbol{x})$ for $t \in \{0, 1\}$.

**Proposition 4.2.16.** *Under the assumptions ([4.2.14]-[4.2.15])*

$$\mathbb{E}(Y_{\text{obs}} \mid T = t, \boldsymbol{X} = \boldsymbol{x}) = \mathbb{E}(Y(t) \mid \boldsymbol{X} = \boldsymbol{x}). \tag{4.11}$$

*Proof.*

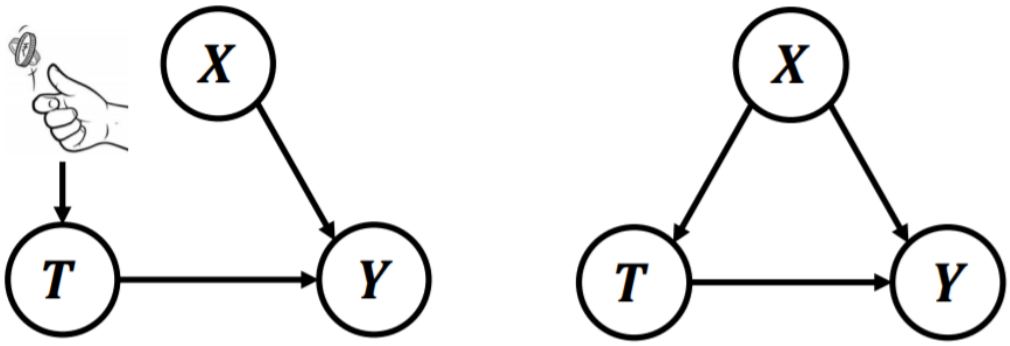
$$\begin{aligned} \mathbb{E}(Y_{\text{obs}} \mid T = t, \boldsymbol{X} = \boldsymbol{x}) &= \mathbb{E}(Y(t) \mid T = t, \boldsymbol{X} = \boldsymbol{x}) && (4.12)\\ &= \mathbb{E}(Y(t) \mid \boldsymbol{X} = \boldsymbol{x}). && (4.13) \end{aligned}$$

■

**Remark 4.2.17.** *The conditional expectation $\mathbb{E}(Y_{\text{obs}} \mid T = t, \tilde{\boldsymbol{X}} = \tilde{\boldsymbol{x}}) \neq \mathbb{E}(Y(t) \mid \tilde{\boldsymbol{X}} = \tilde{\boldsymbol{x}})$ for a subset of covariates $\tilde{\boldsymbol{X}} = (X_{i_1}, \ldots, X_{i_d})$ such that $\{i_1, \ldots, i_d\} \subsetneq \{1, \ldots, d\}$ does not have causal interpretation because the unconfoundedness assumption may not be satisfied by $\tilde{\boldsymbol{X}}$.*

**Definition 4.2.18** (Confounding bias)**.** *In observational studies, the confounding bias refers to the bias responsible for the fact that $\mathbb{E}[Y_{\text{obs}} \mid T = t] \neq \mathbb{E}[Y(t)]$.*

The confounding bias is structural and occurs because of the statistical dependence of the treatment assignment on the confounding variables in observational studies (the confounding variables affect units' treatment choices). It can lead, therefore, to distortion of causal effect,




(a) Randomized Controlled Trials (RCT) (b) Observational Studies

Figure 4.2: Causal structure for RCT and observational studies (Li et al., 2020)

i.e. *spurious correlation* between the treatment $T$ and the outcome $Y$. Indeed, given the causal structure shown in Figure 4.2, conditioning on the treatment $T$ without conditioning on confounding variables is not enough to recover the causal effect on the outcome $Y$.

**Definition 4.2.19** (Selection bias)**.** *In observational studies, given a sample of n units, the selection bias refers to the bias that occurs when directly comparing the observed outcomes of the treated and the untreated units.*

The selection bias is associated with the data-gathering process. It is induced by the preferential selection of units in the sample given their characteristics or the likelihood of being included in the observed data.

To understand the selection bias formally, let us consider the following calculations:

$$\begin{aligned}\mathbb{E}\left[Y_{\text{obs}} \mid T=1\right]-\mathbb{E}\left[Y_{\text{obs}} \mid T=0\right] &= \mathbb{E}\left[Y(1) \mid T=1\right]-\mathbb{E}\left[Y(0) \mid T=0\right] \quad \text{(Consistency)} \\ &= \mathbb{E}\left[Y(1) \mid T=1\right]-\mathbb{E}\left[Y(0) \mid T=1\right]+\mathbb{E}\left[Y(0) \mid T=1\right] \\ &\quad -\mathbb{E}\left[Y(0) \mid T=0\right] \\ &= \underbrace{\mathbb{E}\left[Y(1)-Y(0) \mid T=1\right]}_{\text{ATT}}+\underbrace{\mathbb{E}\left[Y(0) \mid T=1\right]-\mathbb{E}\left[Y(0) \mid T=0\right]}_{\text{Selection Bias}}.\end{aligned} \tag{4.14}$$

These calculations allow us to identify the first term, called the *Average Treatment on the Treated* (ATT), whereas the second term is the *selection bias*. If the randomization holds as in RCTs, it follows that

$$\begin{aligned}\mathbb{E}\left[Y(0) \mid T=0\right]-\mathbb{E}\left[Y(0) \mid T=1\right]=\mathbb{E}\left[Y(0)\right]-\mathbb{E}\left[Y(0)\right]=0, \\ \mathbb{E}\left[Y(1)-Y(0) \mid T=1\right]=\mathbb{E}\left[Y(1)-Y(0)\right].\end{aligned} \tag{4.15}$$

Therefore,

$$\mathbb{E}\left[Y_{\text{obs}} \mid T=1\right]-\mathbb{E}\left[Y_{\text{obs}} \mid T=0\right]=\mathbb{E}\left[Y(1)-Y(0)\right]=\tau. \tag{4.16}$$

In observational studies, both selection and confounding biases are due to a lack of randomization of the treatment assignment, that is $Y(1), Y(0) \not\perp\!\!\!\perp T$. Specifically, the selection bias raises because of *conditioning on common effects* while the confounding bias raises because of *conditioning on common causes* (Hernan & Robins (2020) in Chapter 6).

**Remark 4.2.20.** *It is common in the literature to use the term selection bias to describe both biases. While it may seem to be confusing, this terminology can be understandable. Indeed, Hernan & Robins (2020) explain that the selection bias is a selection of individuals into the analysis while the confounding bias is a selection of individuals into treatment.*

While the presence of both selection and confounding biases harm and lead to biased causal estimands, inferring causal effects is still possible by balancing covariates (Johansson et al., 2016; Shalit et al., 2017) or using the propensity score Curth & van der Schaar (2021b); Hassanpour & Greiner (2019).

**Definition 4.2.21** (Propensity score (Rosenbaum & Rubin, 1983))**.** *The propensity score $e$ is defined as the probability of receiving the treatment given the observed covariates $\boldsymbol{x} \in \mathcal{D}$:*

$$e(\boldsymbol{x}) := \mathbb{P}(T = 1 \mid \boldsymbol{X} = \boldsymbol{x}). \tag{4.17}$$

The propensity score $e$, initially introduced in the causality literature by Rosenbaum & Rubin (1983), have been used to match, stratify or re-weight the samples from the treatment and control groups in observational studies (Rosenbaum & Rubin, 1984). With the propensity score, one can recover the randomized setting where both $\boldsymbol{X}$ and $T$ are independent and obtain similar distributions of observed covariates $\boldsymbol{X}$ across the treatment and control groups. It is useful with the balancing property (see proposition below) to handle both *confounding* and *selection bias* in observational studies.

**Proposition 4.2.22** (Balancing property (Rosenbaum & Rubin, 1983))**.** *The treatment $T$ and the covariates $\boldsymbol{X}$ are independent given the propensity score $e(\boldsymbol{X})$*

$$\boldsymbol{X} \perp\!\!\!\perp T \mid e(\boldsymbol{X}). \tag{4.18}$$

Logistic regression models has been widely used to estimate propensity score $e$ (Austin, 2011; Cepeda et al., 2003). They have the advantage of being a simple parametric approach but it does not offer any guarantees of the goodness of the estimated Propensity score $\widehat{e}$. Some studies show that Machine Learning models are more efficient than logistic regression especially in terms of predictions (Lee et al., 2010; McCaffrey et al., 2004) and bias reduction with iterative variables section but they lack of interprebability or require sometimes additional work on model selection.

As already described previously in Section 4.1, the Potential Outcomes theory is not the only approach used in causal inference. There is also the do-calculus and directed acyclic graphs (DAGs) proposed by Pearl (2009). Richardson & Robins (2013) made an attempt to reconcile and connect these two approaches. Some other works (Pearl, 2011, 2015) show that potential outcomes can be seen as a special case of the do-calculus under some conditions. Indeed, the DAG associated to Potential Outcomes is assumed to have the form in Figure 4.2 (on the right). In this graph, there is no collider nor mediator on $\boldsymbol{X}$. Therefore, intervening on the covariates $do(\boldsymbol{X} = \boldsymbol{x})$ is equivalent to conditioning $\boldsymbol{X} = \boldsymbol{x}$.

Finally, the potential outcomes theory and its assumptions have received different criticisms by Dawid (2000). In the same paper, Dawid (2000) suggested another framework for causal inference without counterfactuals, but it did not gain popularity among causal inference community.

## 4.3 Average treatment Effect

In this section, we present several methods and approaches to estimate the Average Treatment Effect (ATE) using the observational sample $\mathbf{D}_{\text{obs}}$. While presenting different estimators, we do not explicitly distinguish the RCT case to the non-randomized case (with confounding). We refer to Imbens (2004) and Yao et al. (2021) for detailed review of the literature of existing methods to estimate the average treatment effect.

### The naive estimator

The first and the naive estimator of the ATE is the *difference in means* estimator. Given the observational the observational sample $\mathbf{D}_{\text{obs}}$ and the three causal assumptions 4.2.14, 4.2.4 and 4.2.15, the *difference in means* estimator is given by:

$$\hat{\tau}_{\text{naive}} = \sum_{i=1}^{n} \frac{t_i y_i}{\sum_{i=1}^{n} t_i} - \sum_{i=1}^{n} \frac{(1-t_i) y_i}{\sum_{i=1}^{n} (1-t_i)}. \tag{4.19}$$

Since $TY_{\text{obs}} = TY(1)$ and $(1-T)Y_{\text{obs}} = (1-T)Y(0)$, the naive estimator $\hat{\tau}_{\text{naive}}$ satisfies

$$\hat{\tau}_{\text{naive}} = \frac{\sum_{i=1}^{n} t_i Y_i(1)}{\sum_{i=1}^{n} t_i} - \frac{\sum_{i=1}^{n} (1-t_i) Y_i(0)}{\sum_{i=1}^{n} (1-t_i)}. \tag{4.20}$$

Since $(\boldsymbol{x}^{(i)}, Y_i(1), Y_i(0), t_i)$ are *i.i.d.*, the observations $(t_i Y_i(1), (1-t_i)Y_i(0), t_i)$ are also *i.i.d.* drawn from the distribution of $(TY(1), (1-T)Y(0), T)$.By the strong law of large numbers, one obtains

$$\begin{aligned} \hat{\tau}_{\text{naive}} \overset{n \to +\infty}{\longrightarrow} \tau_{\text{lim}} &= \frac{\mathbb{E}[Y(1)T]}{\mathbb{E}[T]} - \frac{\mathbb{E}[(1-T)Y(0)]}{\mathbb{E}[1-T]}, \\ &= \mathbb{E}[Y(1) \mid T=1] - \mathbb{E}[Y(0) \mid T=0]. \end{aligned} \tag{4.21}$$

The naive estimator is simple to construct and has sound theoretical guarantees. Indeed, by the Central Limit Theorem (CLT) and Delta method, one finds that

$$\sqrt{n}(\hat{\tau}_{\text{naive}} - \tau_{\text{lim}}) \to \mathcal{N}(0, V_{\text{naive}}), \tag{4.22}$$

where $V_{\text{naive}}$, that we do not explicitly compute, is the variance of the naive estimator $\hat{\tau}_{\text{naive}}$.

In the RCT framework, the naive estimator $\hat{\tau}_{\text{naive}}$ is strongly convergent and its limit satisfies

$$\tau_{\text{lim}} = \mathbb{E}[Y(1)] - \mathbb{E}[Y(0)] = \tau. \tag{4.23}$$

In the observational framework, we may have $\tau_{\text{lim}} \neq \tau$.

### Propensity score Matching

Propensity Score Matching (PSM) is a statistical matching technique. It aims to mimic randomization and establish the independence between the covariates $\boldsymbol{X}$ and the treatment $T$ by matching treated and control units with similar covariates. PSM attempts to reduce the *selection bias* in observational studies and provide an unbiased estimation of treatment effects.

Propensity Score Matching (PSM) builds its fundamentals on the Balancing property of Rosenbaum & Rubin (1983). Indeed, adjusting units with respect to their propensity score is sufficient to eliminate confounding bias.

Compared to other matching methods that require a specific metric and compare all covariates (see Stuart (2010) for an exhaustive review about matching methods in causal inference), PSM has the advantage of reducing the dimensionality of matching to a single dimension.

In the ideal matching scenario, each treated unit would be matched with one or more control observations with the same values on all the covariates and/or vice versa. However, this situation does not always occur: treated and control units may not be perfectly balanced, and some treated units may differ significantly from other control units. Instead, one can prefer matching the nearest neighbour. We refer the reader to Abadie & Imbens (2016) for a detailed review of existing propensity score matching algorithms (e.g. one-to-one exact, exact matching).

Once the matching is done, it is necessary to assess its quality and check the balancing between treated and control units. The resulting balance quality can be assessed in different ways. Ideally, one compares the distribution of the joint covariates in both groups after matching, but this becomes challenging in high-dimensional settings, or one can use summary statistics such as the Kolmogorov-Smirnov test or multivariate standardized bias (Rosenbaum & Rubin, 1985).

### Propensity score Stratification

Stratification (Angrist, 1998) is an alternative method to adjust *selection bias* due to confounders in observational studies. The idea of stratification is to split the entire sample into homogeneous subgroups and compare treatment effects among them. It generalizes matching to treated and control subgroups, called substrata, with similar covariates distributions.

Ideally, the treated and the control groups in each substratum have similar distributions. The units in the same substratum can be viewed as sampled from the data under Randomized Controlled Trials. Therefore, the treatment effect within each substratum can be calculated directly using the *difference in means* estimator. After computing the treatment effect within each substratum, the average treatment effect can be obtained by combining/averaging the treatment effects of all substrata.

Formally, the idea of stratification consists on dividing the sample $\mathbf{D}_{\text{obs}}$ into $M$ substrata $(\mathcal{S}_m)_{m=1}^{M}$ for a given criterion (propensity score, covariates etc.), then build an estimator $\hat{\tau}_{\text{strat}}$ of the ATE such that:

$$\hat{\tau}_{\text{strat}} = \sum_{m=1}^{M} \frac{n_m}{n} \big(\overline{y}_1(\mathcal{S}_m) - \overline{y}_0(\mathcal{S}_m)\big), \tag{4.24}$$

where $M$ is the number of substrata, $n_m = \text{Card}(\mathcal{S}_m)$ is the number of units in each substratum, $\overline{y}_1(\mathcal{S}_m) = 1/n_m^{(1)} \sum_{i \in \mathcal{S}_m, t_i=1} Y_{\text{obs},i}$ and $\overline{y}_0(\mathcal{S}_m) = 1/n_m^{(0)} \sum_{i \in \mathcal{S}_m, t_i=0} Y_{\text{obs},i}$ are the average of the treated and control outcomes in the $m$-th substratum $\mathcal{S}_m$.

It has also been shown that stratification effectively decreases the bias of ATE estimation compared with the *difference in means* estimator (Yao et al., 2021). However, $\hat{\tau}_{\text{strat}}$ may be biased due to the remaining heterogeneity within strata and due the reduced sample size in each stratum.

### Propensity weighting estimators

*Propensity score re-weighting methods* is a class of estimators used in observational studies to estimate treatment effects. Using the observed sample $\mathbf{D}_{\text{obs}}$, these methods seek to reduce *selection bias* by incorporating the probability of being assigned to the treatment given its covariates in the sampling procedure. In these methods, we associate some weights on the covariates to the sample in order to make treated and control units equate.

Inverse Propensity Weighting (IPW), originally proposed by Horvitz & Thompson (1952), has been proposed in the context of non-randomized studies by Rosenbaum (1987) as a form of model-based direct standardization to estimate treatment effects. In particular, the Inverse propensity weighting (IPW) estimator $\hat{\tau}_{\text{IPW}}$ of the ATE $\tau$ is given by:

$$\hat{\tau}_{\text{IPW}} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{t_i y_i}{\hat{e}(\boldsymbol{x}^{(i)})} - \frac{(1-t_i)\,y_i}{1-\hat{e}(\boldsymbol{x}^{(i)})}\right), \tag{4.25}$$

where $\hat{e}$ is an estimator of the propensity score $e$.

Another normalized version of the IPW estimator $\hat{\tau}_{\text{NIPW}}$, known as Hájek estimator, is introduced by Imbens (2004):

$$\hat{\tau}_{\text{NIPW}} = \left(\sum_{i=1}^{n}\frac{t_i y_i}{\hat{e}(\boldsymbol{x}^{(i)})}\right) \times \left(\sum_{i=1}^{n}\frac{t_i}{\hat{e}(\boldsymbol{x}^{(i)})}\right)^{-1} - \left(\sum_{i=1}^{n}\frac{(1-t_i)\,y_i}{1-\hat{e}(\boldsymbol{x}^{(i)})}\right) \times \left(\sum_{i=1}^{n}\frac{1-t_i}{1-\hat{e}(\boldsymbol{x}^{(i)})}\right)^{-1}. \tag{4.26}$$

Kang & Schafer (2007) show that the precision of this estimator is generally improved compared the standard IPW estimator when weighting the averages of the two groups.

In practice, the correctness of the propensity score estimation is critical and highly impacts the correctness of the IPW estimator. Furthermore, since the propensity score $e$ is present in the denominator, slightly misspecification of propensity scores would increase ATE estimation error dramatically Imai & Ratkovic (2014).

The Augmented Inverse Propensity Weighting (AIPW) has been proposed by Robins et al. (1994) to handle the problem of propensity score misspecifications. We define the AIPW estimator $\hat{\tau}_{\text{AIPW}}$ as:

$$\hat{\tau}_{\text{AIPW}} = \frac{1}{n}\sum_{i=1}^{n}\hat{\mu}_1(\boldsymbol{x}^{(i)}) - \hat{\mu}_0(\boldsymbol{x}^{(i)}) + t_i\frac{Y_{\text{obs},i} - \hat{\mu}_1(\boldsymbol{x}^{(i)})}{\hat{e}(\boldsymbol{x}^{(i)})} - (1-t_i)\frac{Y_{\text{obs},i} - \hat{\mu}_0(\boldsymbol{x}^{(i)})}{1-\hat{e}(\boldsymbol{x}^{(i)})}, \tag{4.27}$$

where, for $j \in \{0,1\}$, $\hat{\mu}_j$ is an estimator of $\mu_j(\boldsymbol{x}) = \mathbb{E}[Y_{\text{obs}} \mid \boldsymbol{X} = \boldsymbol{x}, T = j]$

We can see in (4.27) that the original IPW estimator is augmented using two regression estimators $\hat{\mu}_j$, which allows more flexible modelling. These regression models do not have any causal interpretation and are only used for prediction. The most important property of the AIPW estimator is its *doubly-robustness*, that is, $\hat{\tau}_{\text{AIPW}}$ is consistent and asymptotically unbiased if either the propensity score estimator $\hat{e}$ or the outcomes model estimators $\hat{\mu}_j$ are well-specified (Robins et al., 1994).

### Covariates Balancing Propensity Score

The Covariate Balancing Propensity Score (CBPS) is proposed by Imai & Ratkovic (2014) to overcome the drawback of misspecified propensity score. In contrast to other estimation

methods that use Maximum Likelihood to estimate the propensity score (e.g. logistic regression), the CBPS creates a parametric solution that focuses on achieving a good balance between the treated and control groups.

Indeed, the CBPS estimates propensity scores with respect to a parametric form $e = e(\cdot, \theta)$ by solving the following problem:

$$\mathbb{E}\left[\frac{T\tilde{\boldsymbol{X}}}{e(\tilde{\boldsymbol{X}};\theta)} - \frac{(1-T)\tilde{\boldsymbol{X}}}{1-e(\tilde{\boldsymbol{X}};\theta)}\right] = 0, \tag{4.28}$$

where $\tilde{\boldsymbol{X}} = f(\boldsymbol{X})$ for a measurable function $f$ and the expectation $\mathbb{E}$ is over the joint distribution $p_D$.

By taking the empirical instead of the expectation and by solving the corresponding minimization problem, the CBPS directly constructs the covariate balancing score from the estimated parametric propensity score, which increases its robustness to propensity score model's misspecification. In addition, It can improve the accuracy of estimated treatment effects over parametric models even if the model is well specified (Wyss et al., 2014).

### Regression adjustment

Another common way to estimate the ATE is the *regression adjustment*. In this method, we assume that the ATE is as parameter of a regression model on the outcome model $\mu_t(\boldsymbol{x}) = \mathbb{E}[Y_{\text{obs}} \mid \boldsymbol{X} = \boldsymbol{x}, T = t]$. More precisely, for $\boldsymbol{x} \in \mathcal{D}$ and $t \in \{0, 1\}$, we assume a linear functional form on the outcome model:

$$\mu_t(\boldsymbol{x}) = \beta_0 + \boldsymbol{\beta}^\top \boldsymbol{x} + \tau\, t, \tag{4.29}$$

where $\beta_0$, $\boldsymbol{\beta} \in \mathbb{R}^d$ are some regression coefficients and $\tau$ is the quantity of interest (ATE).

With the previous model, it can be shown easily that:

$$\tau = \mathbb{E}\big[\mu_1(\boldsymbol{X}) - \mu_0(\boldsymbol{X})\big]. \tag{4.30}$$

Therefore, one can build an estimator of the outcome model $\mu_\cdot$, denoted by $\hat{\mu}_\cdot$, then target the ATE by averaging over the empirical distributions of the covariates $\boldsymbol{X}$ for both treated and control units such that:

$$\hat{\tau}_{\text{reg}} = \frac{1}{n}\sum_{i=1}^{n}\big[\hat{\mu}_1(\boldsymbol{x}^{(i)}) - \hat{\mu}_0(\boldsymbol{x}^{(i)})\big]. \tag{4.31}$$

Note that other regression models, such as kernel regression and generalized linear or additive models (see Hastie et al. (2009) for a review of different regression models), can be used and offer more choice without relying on the parametric and linear forms of the outcome model.

Despite the efficiency of the *regression adjustment* to reduce bias and increase precision in estimating the ATE, Rubin (1979) points out that regression adjustments are sensitive to model misspecification when there is insufficient overlap between treated and control units. It can lead, unfortunately, to more bias when the functional form of the outcome model is misspecified.

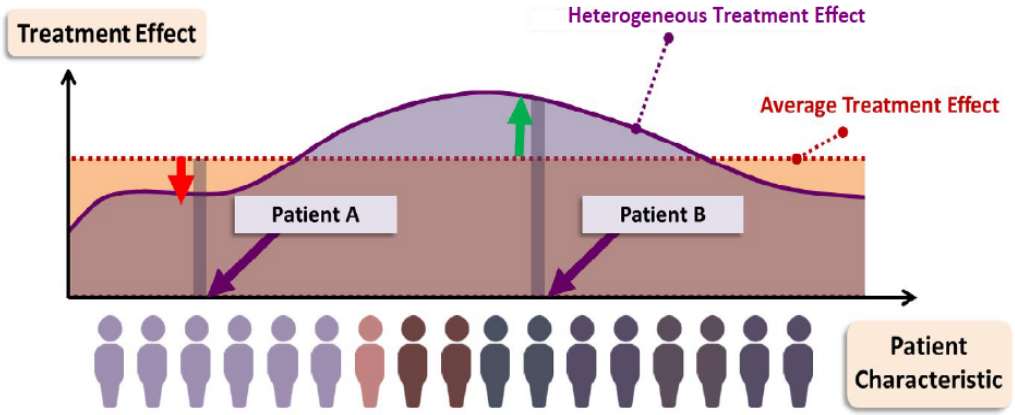



Figure 4.3: Illustration of the difference between the Average Treatment Effect and Individualized Treatment Effects (Bica et al., 2021).

## 4.4 Heterogeneity of treatment effects

In many situations, researchers are not interested in aggregated causal effects but would known how the treatment would affect units with particular covariates. Indeed, it may happen that the treatment has a no effect in average (i.e. the ATE satisfies $\tau = 0$) for the population but differs significantly among some subgroups (e.g. being positive for some units and negative for others). In causal inference literature, varied causal effects for individuals with varied characteristics are called heterogeneous treatment effects. Figure 4.3 illustrates an example of Treatment effect heterogeneity where treatment effects are above the average in some regions and below for some other regions.

Treatment effect heterogeneity is an important topic in many fields, especially in the medical sciences, economics, public policy. The heterogeneity of treatment effect offer more valuable information and allow them to adjust and personalize the treatment/policy for different subgroups of the population.

Crump et al. (2008) develop nonparametric tests for the null of no treatment effect heterogeneity, which bypass the multiple testing problem but fail to specify exactly which subgroups have heterogeneity. This has motivated many recent works to estimate heterogeneous treatment effects and identify subgroups of interest (Alaa & van der Schaar, 2017; Hill, 2011; Imai & Ratkovic, 2014; Johansson et al., 2016; Knaus et al., 2020b; Nie & Wager, 2020).

In causal Inference framework, estimating heterogeneous treatment effects is equivalent to estimate the average treatment effect for different subgroups. The subgroups are defined by specific covariates $\boldsymbol{X} = \boldsymbol{x}$ and the average treatment effect within a subgroup is commonly known as the conditional average treatment effect (CATE).

**Definition 4.4.1** (Conditional Average Treatment Effect (CATE))**.** *For a given vector of covariates $\boldsymbol{x} \in \mathcal{D}$, we define the Conditional Average Treatment Effect (CATE) function by:*

$$\tau(\boldsymbol{x}) = \mathbb{E}(Y(1) - Y(0) \mid \boldsymbol{X} = \boldsymbol{x}). \tag{4.32}$$

To address the problem of estimating the CATE, several methods and models have been

proposed by researchers. Some of them incorporate Machine-Learning through modified models, while others, known as *meta-learners*, do not require a specific Machine-Learning method.

Caron et al. (2022a) and Knaus et al. (2020a) for provide an in-depth literature review of the most recent and popular methods for CATE estimation, particularly for Machine-Learning based models. In the following, we will present quickly Machine-Learning based models for the CATE's estimation, but we refer the reader to the same references for a detailed review about these models, and we focus more on the Meta-Learners framework for estimating the CATE.

### Machine-Learning based models

The recent interest and advances in CATEs estimation have led to the development of numerous algorithms and methods based on Machine-Learning models (e.g. tree ensembles, gradient boosting methods, neural networks). These models use the observational data to construct consistent estimators of the CATE.

The first contribution for estimating the CATE belongs to Hill (2011). The idea consisted in assuming the following functional form of the potential outcomes:

$$Y(t) = f(\boldsymbol{X}, t) + \epsilon, \tag{4.33}$$

where $f$ is an unknown function and $\epsilon$ is additive noise. Hill (2011) proposed to learn $f$ using Bayesian Additive Regression Trees (BART) and infer the CATE from the learnt function.

Since then, a wide variety of tree-based and, more generally, ensemble methods have been developed by the Causal Inference community to derive efficient and consistent CATE estimators. These methods include Causal tree (Athey & Imbens, 2016) Causal Forests (Lechner, 2018; Wager & Athey, 2018), support vector machines (Imai & Ratkovic, 2013), causal boosting and causal multivariate adaptive regression splines (MARS) (Powers et al., 2018), LASSO regression (Zhao et al., 2022), non-parametric kernel smoothing (Fan et al., 2022; Zimmert & Lechner, 2019), Bayesian Causal Forests (Caron et al., 2022b; Hahn et al., 2020) to handle the problem of confounding and multi-task learning approach using Gaussian Processes (Alaa & van der Schaar, 2017).

In the deep learning field, several models employing deep neural networks have been proposed to learn balanced representations and handle confounding. Among them, we can cite Balancing Counterfactual Regression (BCR) (Johansson et al., 2016), Treatment Agnostic Representation Networks (TARNET) (Shalit et al., 2017), Causal Effect Variational AutoEncoder (CEVAE) (Louizos et al., 2017), Generative Adversarial Nets for inference of Individualised Treatment Effects (GANITE) (Yoon et al., 2018), Similarity Individual Treatment Effect method (SITE)(Yao et al., 2018) and Dragonnet (Shi et al., 2019). We may refer the reader to Dorie et al. (2019) for a review of (hybrid) Machine-Learning models for causal inference.

Generally, the previous cited methods are built either upon a single model or upon two distinct models. They fall into the S- or T-learners class, which we will emphasize further in the following subsection.

### Meta-learners for Heterogeneous treatment Effects estimation

One possible framework to tackle the problem of estimating the CATE are meta-learners as initially introduced and discussed by Künzel et al. (2019). Meta-learners derive consistent

estimators of the CATE in both Randomized Controlled Trials (RCT) and Observational studies.

**Definition 4.4.2** (Meta-learner (Künzel et al., 2019)). *A Meta-learner is a statistical framework that models and estimates the CATE model such that*

$$\tau(\boldsymbol{x}) = \mathbb{E}[Y(1) - Y(0) \mid \boldsymbol{X} = \boldsymbol{x}]. \tag{4.34}$$

The advantage of meta-learners is that they do not require a specific Machine Learning method. They can support any supervised regression parametric or nonparametric method (e.g. random forest, gradient boosting methods). These methods are called *base-learners* when applied to a meta-learner.

All meta-learners fall in a taxonomy of CATE's estimators given by Curth & van der Schaar (2021a); Knaus et al. (2020a). Namely, direct plug-in (one step) meta-learners (T- and S-learners), pseudo-outcome (two-step) meta-learners (X-, M- and DR-learners) and Neyman-Orthogonality based learners (R-learner).

### T-learner

From the definition of CATE in (4.32), the first meta-learner to be considered is the *T-learner*, where $T$ refers to *two*-models procedure. This meta-learner builds a CATE estimator using two models:

- Regress $Y(j)$ separately on the covariates $\boldsymbol{X}$ using $(D_{\text{obs},i})_{i\in\mathbf{S}_j}$ where $\mathbf{S}_j = \{i, t_i = j\}$ for $j \in \{0, 1\}$ to build estimators $\widehat{\mu}_j$ of $\mu_j(\boldsymbol{x}) = \mathbb{E}(Y(j) \mid \boldsymbol{X} = \boldsymbol{x})$.
- Estimate the CATE as $\widehat{\tau}_T(\boldsymbol{x}) = \widehat{\mu}_1(\boldsymbol{x}) - \widehat{\mu}_0(\boldsymbol{x})$.

Some authors (Curth & van der Schaar, 2021b; Künzel et al., 2019) claim that the main drawback of the T-learning approach is that it does not take the interaction between treatment $T$ and the outcome $Y$ and that it may suffer from *confounding bias*. This problem occurs typically while sampling $(D_{\text{obs},i})_{i\in\mathbf{S}_j}$ for $j \in \{0, 1\}$ at the first stage of regression procedure and the outcome models $\mu_j$ are, therefore, estimated with respect to the wrong distribution of the training sample, that is,

$$\mathbb{E}_{\boldsymbol{X}\sim p(\cdot)}\big[(\widehat{\mu}_j(\boldsymbol{X}) - \mu_j(\boldsymbol{X}))^2\big] \neq \mathbb{E}_{\boldsymbol{X}\sim p(\cdot|T=j)}\big[(\widehat{\mu}_j(\boldsymbol{X}) - \mu_j(\boldsymbol{X}))^2\big], \tag{4.35}$$

where $p(\cdot)$ denotes the marginal distribution of $\boldsymbol{X}$ and $p(\cdot \mid T = j)$ denotes the conditional distribution of $\boldsymbol{X}$ given $T = j$.

Therefore, the optimal $\widehat{\mu}_j$ for $j \in \{0, 1\}$ should be fitted on the sample $(D_{\text{obs},i})_{i\in\mathbf{S}_j}$ by considering a weight while minimizing the expected (integrated) error (Curth & van der Schaar, 2021a):

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{X}\sim p(\cdot)}\big[(\widehat{\mu}_1(\boldsymbol{X}) - \mu_1(\boldsymbol{X}))^2\big] &= \mathbb{E}_{\boldsymbol{X}\sim p(\cdot|T=1)}\big[\frac{p(T=1)}{e(\boldsymbol{X})}(\widehat{\mu}_1(\boldsymbol{X}) - \mu_1(\boldsymbol{X}))^2\big], \\
\mathbb{E}_{\boldsymbol{X}\sim p(\cdot)}\big[(\widehat{\mu}_0(\boldsymbol{X}) - \mu_0(\boldsymbol{X}))^2\big] &= \mathbb{E}_{\boldsymbol{X}\sim p(\cdot|T=0)}\big[\frac{\mathbb{P}(T=0)}{1-e(\boldsymbol{X})}(\widehat{\mu}_0(\boldsymbol{X}) - \mu_0(\boldsymbol{X}))^2\big].
\end{aligned} \tag{4.36}$$

### S-learner

The second meta-learner to be defined is the S-learner where S refers to *single*. It is based on Proposition 4.2.16 of the identifiability of the counterfactual response, indeed

$$\tau(\boldsymbol{x}) = \mathbb{E}[Y_{\text{obs}} \mid T = 1, \boldsymbol{X} = \boldsymbol{x}] - \mathbb{E}[Y_{\text{obs}} \mid T = 0, \boldsymbol{X} = \boldsymbol{x}]. \tag{4.37}$$

Therefore, one can take the treatment $T$ as a feature similar to all the other covariates and build as follows :

- Regress $Y$ on the treatment $T$ and the covariates $\boldsymbol{X}$ by a single model $\widehat{\mu}$ using $\mathbf{D}_{\text{obs}}$.
- Estimate the CATE as $\widehat{\tau}_S(\boldsymbol{x}) = \widehat{\mu}(\boldsymbol{x}, 1) - \widehat{\mu}(\boldsymbol{x}, 0)$.

**Remark 4.4.3.** *The T-Learner and the S-Learner may not produce the same result as the regression procedure is different for each learner.*

Using the propensity score $e$, we may define additional meta-learning algorithms whose objective is to estimate the CATE in (4.32) more efficiently.

### X-learner

The X-learner, where $X$ refers to the *cross*-learning approach of the algorithm (Künzel et al., 2019), has been proposed to overcome the problem of unbalancing groups, which adopts information from the control group to give a better estimator on the treated group and vice versa.

Let us consider the two random variables $D^{(1)} := Y(1) - \mu_0(\boldsymbol{X})$ and $D^{(0)} := \mu_1(\boldsymbol{X}) - Y(0)$. We have

$$\begin{aligned} \mathbb{E}(D^{(1)} \mid \boldsymbol{X} = \boldsymbol{x}) &= \mathbb{E}(Y(1) - \mu_0(\boldsymbol{X}) \mid \boldsymbol{X} = \boldsymbol{x}) \\ &= \mathbb{E}[Y(1) - \mathbb{E}(Y(0) \mid \boldsymbol{X}) \mid \boldsymbol{X} = \boldsymbol{x}] \\ &= \mathbb{E}(Y(1) - Y(0) \mid \boldsymbol{X} = \boldsymbol{x}) \mid = \tau(\boldsymbol{x}), \end{aligned} \tag{4.38}$$

and

$$\begin{aligned} \mathbb{E}(D^{(0)} \mid \boldsymbol{X} = \boldsymbol{x}) &= \mathbb{E}(\mu_1(\boldsymbol{X}) - Y(0) \mid \boldsymbol{X} = \boldsymbol{x}) \\ &= \mathbb{E}[\mathbb{E}(Y(1) \mid \boldsymbol{X}) - Y(0) - \mid \boldsymbol{X} = \boldsymbol{x}] \\ &= \mathbb{E}(Y(1) - Y(0) \mid \boldsymbol{X} = \boldsymbol{x}) \mid = \tau(\boldsymbol{x}). \end{aligned} \tag{4.39}$$

The X-Learner can be built from the sample $\mathbf{D}_{\text{obs}}$ as follows :

- Similarly to T-Learner, regress $Y(j)$ on the covariates $\boldsymbol{X}$ using the subsets $(D_{\text{obs},i})_{i \in \mathbf{S}_j}$ for $j \in \{0, 1\}$ to build estimators $\widehat{\mu}_j$ of $\mu_j(\boldsymbol{x}) = \mathbb{E}(Y(j) \mid \boldsymbol{X} = \boldsymbol{x})$.
- Estimate the missing potential outcomes $\tilde{D}_i^{(1)} = Y_{\text{obs},i} - \widehat{\mu}_0(\boldsymbol{x}^{(i)})$ if $i \in \mathbf{S}_1$ and $\tilde{D}_i^{(0)} = \widehat{\mu}_1(\boldsymbol{x}^{(i)}) - Y_{\text{obs},i}$ if $i \in \mathbf{S}_0$.
- Regress $D^{(1)}$ and $D^{(0)}$ on the covariates $\boldsymbol{X}$ by two models $\widehat{\tau}_1$ and $\widehat{\tau}_0$ using the subsets $(\boldsymbol{x}^{(i)}, \tilde{D}_i^{(0)})_{i \in \mathbf{S}_0}$ and $(\boldsymbol{x}^{(i)}, \tilde{D}_i^{(1)})_{i \in \mathbf{S}_1}$.
- Estimate the CATE by a weighted average function $g$ (e.g. propensity score $e$) of the estimated models such that $\widehat{\tau}_X(\boldsymbol{x}) = g(\boldsymbol{x})\widehat{\tau}_0(\boldsymbol{x}) + (1 - g(\boldsymbol{x}))\widehat{\tau}_1(\boldsymbol{x})$.

**Remark 4.4.4.** *$\widehat{\tau}_1$ and $\widehat{\tau}_0$ are both estimators for CATE model $\tau$, while $g$ is chosen to combine these estimators to an improved estimator $\widehat{\tau}_X$.*

The choice of the weighting function $g$ is crucial and affects the final estimation of the CATE $\tau$ (Curth & van der Schaar, 2021a). The same authors suggest as alternative the regression adjustment learning strategy: a two-steps cross procedure (instead of four as in the original X-learner) that does not require any weighting function.

**Proposition 4.4.5.** *If the assumptions ([4.2.14]-[4.2.15]) hold, we define the regression adjustment pseudo-outcome $Z_{\mathrm{RA}}$ as*

$$Z_{\mathrm{RA}} = T(Y_{\mathrm{obs}} - \mu_0(\boldsymbol{X})) + (1-T)(\mu_1(\boldsymbol{X}) - Y_{\mathrm{obs}}), \tag{4.40}$$

*then*

$$\mathbb{E}(Z_{\mathrm{RA}} \mid \boldsymbol{X} = \boldsymbol{x}) = \tau(\boldsymbol{x}). \tag{4.41}$$

*Proof.*

$$\begin{aligned}
\mathbb{E}(Z_{\mathrm{RA}} \mid \boldsymbol{X} = \boldsymbol{x}) &= \mathbb{E}\left[T(Y_{\mathrm{obs}} - \mu_0(\boldsymbol{X})) + (1-T)(\mu_1(\boldsymbol{X}) - Y_{\mathrm{obs}}) \mid \boldsymbol{X} = \boldsymbol{x}\right] && (4.42)\\
&= \mathbb{E}\left[TY_{\mathrm{obs}} \mid \boldsymbol{X} = \boldsymbol{x}\right] - e(\boldsymbol{x})\mu_0(\boldsymbol{x}) + (1-e(\boldsymbol{x}))\mu_1(\boldsymbol{x}) && (4.43)\\
&\quad - \mathbb{E}\left[(1-T)Y_{\mathrm{obs}} \mid \boldsymbol{X} = \boldsymbol{x}\right]. && (4.44)
\end{aligned}$$

Since $Y_{\mathrm{obs}} = Y(T) = TY(1) + (1-T)Y(0)$, we have $TY = T^2Y(1) + T(1-T)Y(0) = TY(1)$ and $(1-T)Y_{\mathrm{obs}} = T(1-T)Y(1) + (1-T)^2Y(0) = (1-T)Y(0)$ and thus

$$\mathbb{E}(Z_{RA} \mid \boldsymbol{X} = \boldsymbol{x}) = \mathbb{E}\left[TY(1) \mid \boldsymbol{X} = \boldsymbol{x}\right] - e(\boldsymbol{x})\mu_0(\boldsymbol{x}) + (1-e(\boldsymbol{x}))\mu_0(\boldsymbol{x}) - \mathbb{E}\left[(1-T)Y(0) \mid \boldsymbol{X} = \boldsymbol{x}\right]. \tag{4.45}$$

Therefore, since $\mathbb{E}[T \mid \boldsymbol{X}] = \mathbb{P}[T = 1 \mid \boldsymbol{X}]$ and if the assumption [4.2.14] holds, then

$$\begin{aligned}
\mathbb{E}(Z_{\mathrm{RA}} \mid \boldsymbol{X} = \boldsymbol{x}) &= e(\boldsymbol{x})\mu_1(\boldsymbol{x}) - e(\boldsymbol{x})\mu_0(\boldsymbol{x}) + (1-e(\boldsymbol{x}))\mu_0(\boldsymbol{x}) - (1-e(\boldsymbol{x}))\mu_0(\boldsymbol{x}) && (4.46)\\
&= (e(\boldsymbol{x}) + 1 - e(\boldsymbol{x}))(\mu_1(\boldsymbol{x}) - \mu_0(\boldsymbol{x})) = \tau(\boldsymbol{x}). && (4.47)
\end{aligned}$$

■

The improved X-Learner can be built as follows:

- Similarly to T-Learner, regress $Y(j)$ on the covariates $\boldsymbol{X}$ using the subsets $(D_{\mathrm{obs},i})_{i \in \mathbf{S}_j}$ for $j \in \{0,1\}$ to build estimators $\widehat{\mu}_j$ of $\mu_j(\boldsymbol{x}) = \mathbb{E}(Y(j) \mid \boldsymbol{X} = \boldsymbol{x})$.
- Estimate the CATE by regressing the regression-adjustment pseudo-outcome $\tilde{\boldsymbol{z}}_{\mathrm{RA}} = (\tilde{z}_{\mathrm{RA},i})_{i=1}^n$ on the covariates $\boldsymbol{X}$ using $\mathbf{D}_{\mathrm{obs}}$, where

$$\tilde{z}_{\mathrm{RA},i} = t_i(Y_{\mathrm{obs},i} - \widehat{\mu}_0(\boldsymbol{x}^{(i)})) + (1-t_i)(\widehat{\mu}_1(\boldsymbol{x}^{(i)}) - Y_{\mathrm{obs},i}). \tag{4.48}$$

### M-learner

The *M-learner* (Athey & Imbens, 2016), where M refers to the *modified* learned pseudo-outcome in the algorithm, is inspired from the Inverse Propensity Weighting (IPW) transformation as proposed by (Rosenbaum, 1987) for estimating the ATE.

**Proposition 4.4.6.** *If the assumptions ([4.2.14]-[4.2.15]) hold, we define the modified pseudo-outcome $Z_{\text{IPW}}$ by IPW as*

$$Z_{\text{IPW}} = \frac{T}{e(\boldsymbol{X})} Y_{\text{obs}} - \frac{1-T}{1-e(\boldsymbol{X})} Y_{\text{obs}}, \tag{4.49}$$

*then*

$$\mathbb{E}(Z_{\text{IPW}} \mid \boldsymbol{X} = \boldsymbol{x}) = \tau(\boldsymbol{x}). \tag{4.50}$$

*Proof.*

$$\begin{aligned}
\mathbb{E}(Z_{\text{IPW}} \mid \boldsymbol{X} = \boldsymbol{x}) &= \mathbb{E}\left[\frac{T}{e(\boldsymbol{X})} Y_{\text{obs}} - \frac{1-T}{1-e(\boldsymbol{X})} Y_{\text{obs}} \mid \boldsymbol{X} = \boldsymbol{x}\right] && (4.51)\\
&= \mathbb{E}\left[\frac{T}{e(\boldsymbol{X})} Y_{\text{obs}} \mid \boldsymbol{X} = \boldsymbol{x}\right] - \mathbb{E}\left[\frac{1-T}{1-e(\boldsymbol{X})} Y_{\text{obs}} \mid \boldsymbol{X} = \boldsymbol{x}\right] && (4.52)\\
&= \frac{1}{e(\boldsymbol{x})} \mathbb{E}\left[T Y_{\text{obs}} \mid \boldsymbol{X} = \boldsymbol{x}\right] - \frac{1}{1-e(\boldsymbol{x})} \mathbb{E}\left[(1-T) Y_{\text{obs}} \mid \boldsymbol{X} = \boldsymbol{x}\right]. && (4.53)
\end{aligned}$$

Thus,

$$\mathbb{E}(Z_{\text{IPW}} \mid \boldsymbol{X} = \boldsymbol{x}) = \frac{1}{e(\boldsymbol{x})} \mathbb{E}\left[T Y(1) \mid \boldsymbol{X} = \boldsymbol{x}\right] - \frac{1}{1-e(\boldsymbol{x})} \mathbb{E}\left[(1-T) Y(0) \mid \boldsymbol{X} = \boldsymbol{x}\right]. \tag{4.54}$$

Therefore, by assumption [4.2.14]

$$\begin{aligned}
\mathbb{E}(Z_{IPW} \mid \boldsymbol{X} = \boldsymbol{x}) &= \frac{1}{e(\boldsymbol{x})} \mathbb{E}[T \mid \boldsymbol{X} = \boldsymbol{x}] \mu_1(\boldsymbol{x}) - \frac{1}{1-e(\boldsymbol{x})} \mathbb{E}[(1-T) \mid \boldsymbol{X} = \boldsymbol{x}] \mu_0(\boldsymbol{x}) && (4.55)\\
&= \mu_1(\boldsymbol{x}) - \mu_0(\boldsymbol{x}) = \tau(\boldsymbol{x}) 0 && (4.56)
\end{aligned}$$

■

Hence, the M-learner is built in two stages :

- Estimate the propensity score $e$ by regressing $T$ on the covariates $\boldsymbol{X}$ using $\mathbf{D}_{\text{obs}}$ and denote $\widehat{e}$ the obtained model.
- Estimate the CATE by regressing the IPW pseudo-outcome $\tilde{\boldsymbol{z}}_{\text{IPW}} = (\tilde{z}_{\text{IPW},i})_{i=1}^{n}$ on the covariates $\boldsymbol{X}$ using $\mathbf{D}_{\text{obs}}$, where

$$\tilde{z}_{\text{IPW},i} = \frac{t_i}{\widehat{e}(\boldsymbol{x}^{(i)})} Y_{\text{obs},i} - \frac{1-t_i}{1-\widehat{e}(\boldsymbol{x}^{(i)})} Y_{\text{obs},i}. \tag{4.57}$$

However, one needs the consistency of the propensity score estimator $\widehat{e}$ to get a correct estimation of the CATE.

### DR-learner

As explained in the previous section, the *doubly-robust* method was suggested by Robins et al. (1994) to overcome the problem of model misspecification. It tries to estimate two components: the outcome model and the propensity score. The consistency of the causal effect estimator is achieved if at least one of these components is well specified, that is the estimation of either the outcome model or the propensity score consistent. Under the assumptions ([4.2.14]-[4.2.15]), we define the augmented inverse probability weighting (AIPW) pseudo-outcome by

$$Z_{\text{AIPW}} = \mu_1(\boldsymbol{X}) - \mu_0(\boldsymbol{X}) + T\frac{Y_{\text{obs}} - \mu_1(\boldsymbol{X})}{e(\boldsymbol{X})} - (1-T)\frac{Y_{\text{obs}} - \mu_0(\boldsymbol{X})}{1-e(\boldsymbol{X})}. \tag{4.58}$$

**Proposition 4.4.7.** *Let $Z_{\text{AIPW}}$ be the AIPW pseudo-outcome defined previously, then under the assumptions ([4.2.14] - [4.2.15])*

$$\mathbb{E}(Z_{\text{AIPW}} \mid \boldsymbol{X} = \boldsymbol{x}) = \tau(\boldsymbol{x}). \tag{4.59}$$

*Proof.*

$$\mathbb{E}(Z_{\text{AIPW}} \mid \boldsymbol{X} = \boldsymbol{x}) = \mathbb{E}\left[\mu_1(\boldsymbol{X}) - \mu_0(\boldsymbol{X}) + T\frac{Y_{\text{obs}} - \mu_1(\boldsymbol{X})}{e(\boldsymbol{X})} - (1-T)\frac{Y_{\text{obs}} - \mu_0(\boldsymbol{X})}{1-e(\boldsymbol{X})} \mid \boldsymbol{X} = \boldsymbol{x}\right] \tag{4.60}$$

$$= \mu_1(\boldsymbol{x}) - \mu_0(\boldsymbol{x}) + \frac{\mathbb{E}[T(Y_{\text{obs}} - \mu_1(\boldsymbol{X})) \mid \boldsymbol{X} = \boldsymbol{x}]}{e(\boldsymbol{X})} - \frac{\mathbb{E}[(1-T)(Y_{\text{obs}} - \mu_0(\boldsymbol{X})) \mid \boldsymbol{X} = \boldsymbol{x}]}{1-e(\boldsymbol{X})}. \tag{4.61}$$

By uncounfoundedness Assumption [4.2.14], we have

$$\mathbb{E}(Z_{\text{AIPW}} \mid \boldsymbol{X} = \boldsymbol{x}) = \mu_1(\boldsymbol{x}) - \mu_0(\boldsymbol{x}) + e(\boldsymbol{x})\frac{\mathbb{E}(Y(1) \mid \boldsymbol{X} = \boldsymbol{x}) - \mu_1(\boldsymbol{x})}{e(\boldsymbol{x})} \tag{4.62}$$

$$- (1-e(\boldsymbol{x}))\frac{\mathbb{E}(Y(0) \mid \boldsymbol{X} = \boldsymbol{x}) - \mu_0(\boldsymbol{x}) - \mu_0(\boldsymbol{x})}{1-e(\boldsymbol{x})} \tag{4.63}$$

$$= \mu_1(\boldsymbol{x}) - \mu_0(\boldsymbol{x}) = \tau(\boldsymbol{x}). \tag{4.64}$$

■

However, the definition of $Z_{\text{AIPW}}$ implies estimating both the outcome models $\mu_0, \mu_1$ and the propensity score $e$. We denote $\overline{\mu}_0, \overline{\mu}_1$ and $\overline{e}$ some arbitrary models of the three previous models, we define the *doubly-robust* (DR) pseudo-outcome $Z_{\text{DR}}$ below in (4.65) and we show its doubly-robust property

$$Z_{\text{DR}}(\overline{\mu}_0, \overline{\mu}_1, \overline{e}) = \overline{\mu}_1(\boldsymbol{X}) + T\frac{Y_{\text{obs}} - \overline{\mu}_1(\boldsymbol{X})}{\overline{e}(\boldsymbol{X})} - \overline{\mu}_0(\boldsymbol{X}) - (1-T)\frac{Y_{\text{obs}} - \overline{\mu}_0(\boldsymbol{X})}{1-\overline{e}(\boldsymbol{X})}, \tag{4.65}$$

where $\overline{e}$ is also assumed to satisfied the assumption [4.2.15], that is, $0 < e_{min} < \overline{e}(\boldsymbol{x}) < e_{max} < 1$ for all $\boldsymbol{x} \in \mathcal{D}$.

**Proposition 4.4.8.** *Let $Z_{\text{DR}}(\overline{\mu}_0, \overline{\mu}_1, \overline{e})$ be the Doubly-Robust pseudo-outcome defined previously, then under the assumptions ([4.2.14] - [4.2.15])*

$$\mathbb{E}(Z_{\text{DR}}(\overline{\mu}_0, \overline{\mu}_1, \overline{e}) \mid \boldsymbol{X} = \boldsymbol{x}) = \tau(\boldsymbol{x}), \tag{4.66}$$

*if the outcome models or the propensity model is well-specified, i.e.* $\overline{e}(\boldsymbol{X}) = e(\boldsymbol{X})$ *or* $(\overline{\mu}_0(\boldsymbol{X}), \overline{\mu}_1(\boldsymbol{X})) = (\mu_0(\boldsymbol{X}), \mu_1(\boldsymbol{X}))$ *almost surely.*

*Proof.* We show now the *Doubly-Robust* behaviour of the DR pseudo-outcome,

$$\mathbb{E}(Z_{\mathrm{DR}}(\overline{\mu}_0, \overline{\mu}_1, \overline{e}) \mid \boldsymbol{X} = \boldsymbol{x}) = \mathbb{E}\Big[\overline{\mu}_1(\boldsymbol{X}) + T\frac{Y_{\mathrm{obs}} - \overline{\mu}_1(\boldsymbol{X})}{\overline{e}(\boldsymbol{X})} - \overline{\mu}_0(\boldsymbol{X}) - (1-T)\frac{Y_{\mathrm{obs}} - \overline{\mu}_0(\boldsymbol{X})}{1-\overline{e}(\boldsymbol{X})} \mid \boldsymbol{X} = \boldsymbol{x}\Big] \tag{4.67}$$

$$= \mathbb{E}\Big[\overline{\mu}_1(\boldsymbol{X}) + T\frac{Y(1) - \overline{\mu}_1(\boldsymbol{X})}{\overline{e}(\boldsymbol{X})} - \overline{\mu}_0(\boldsymbol{X}) - (1-T)\frac{Y(0) - \overline{\mu}_0(\boldsymbol{X})}{1-\overline{e}(\boldsymbol{X})} \mid \boldsymbol{X} = \boldsymbol{x}\Big] \tag{4.68}$$

$$= \mathbb{E}\Big[Y(1) + \overline{\mu}_1(\boldsymbol{X}) - Y(1) + T\frac{Y(1) - \overline{\mu}_1(\boldsymbol{X})}{\overline{e}(\boldsymbol{X})} \tag{4.69}$$

$$- Y(0) - \overline{\mu}_0(\boldsymbol{X}) + Y(0) - (1-T)\frac{Y(0) - \overline{\mu}_0(\boldsymbol{X})}{1-\overline{e}(\boldsymbol{X})} \mid \boldsymbol{X} = \boldsymbol{x}\Big] \tag{4.70}$$

$$= \mathbb{E}\Big[Y(1) + \left(\frac{T}{\overline{e}(\boldsymbol{X})} - 1\right)(Y(1) - \overline{\mu}_1(\boldsymbol{X})) \mid \boldsymbol{X} = \boldsymbol{x}\Big] \tag{4.71}$$

$$- \mathbb{E}\Big[Y(0) + \left(\frac{1-T}{1-\overline{e}(\boldsymbol{X})} - 1\right)(Y(0) - \overline{\mu}_0(\boldsymbol{X})) \mid \boldsymbol{X} = \boldsymbol{x}\Big] \tag{4.72}$$

$$= \mathbb{E}(Y(1) - Y(0) \mid \boldsymbol{X} = \boldsymbol{x}) + \eta_1(\boldsymbol{x}) - \eta_0(\boldsymbol{x}), \tag{4.73}$$

where $\eta_1(\boldsymbol{x}) = \mathbb{E}\Big[Y(1) + \left(\frac{T}{\overline{e}(\boldsymbol{X})} - 1\right)(Y(1) - \overline{\mu}_1(\boldsymbol{X})) \mid \boldsymbol{X} = \boldsymbol{x}\Big]$ and $\eta_0(\boldsymbol{x}) = \mathbb{E}\Big[Y(0) + \left(\frac{1-T}{1-\overline{e}(\boldsymbol{X})} - 1\right)(Y(0) - \overline{\mu}_0(\boldsymbol{X})) \mid \boldsymbol{X} = \boldsymbol{x}\Big]$.

- If the propensity score $\overline{e}$ is correctly specified (i.e. $\overline{e}(\boldsymbol{X}) = e(\boldsymbol{X})$ *almost surely*) but the outcome model is misspecified, we would have

$$\eta_1(\boldsymbol{x}) = \mathbb{E}\Big[\left(\frac{T}{e(\boldsymbol{X})} - 1\right)(Y(1) - \overline{\mu}_1(\boldsymbol{X}))|\boldsymbol{X} = \boldsymbol{x}\Big] \tag{4.74}$$

$$= \mathbb{E}\Big[\mathbb{E}\Big[\left(\frac{T}{e(\boldsymbol{X})} - 1\right)(Y(1) - \overline{\mu}_1(\boldsymbol{X}))|\boldsymbol{X}, Y(1)\Big]|\boldsymbol{X} = \boldsymbol{x}\Big] \tag{4.75}$$

$$= \mathbb{E}\Big[(Y(1) - \overline{\mu}_1(\boldsymbol{X}))\mathbb{E}\Big[\left(\frac{T}{e(\boldsymbol{X})} - 1\right)|\boldsymbol{X}, Y(1)\Big]|\boldsymbol{X} = \boldsymbol{x}\Big]. \tag{4.76}$$

  Thus, by the assumption of unconfoundedness [4.2.14]

$$\eta_1(\boldsymbol{x}) = \mathbb{E}\Big[(Y(1) - \overline{\mu}_1(\boldsymbol{X}))\mathbb{E}\Big[\left(\frac{T}{e(\boldsymbol{X})} - 1\right)\|\boldsymbol{X}\Big]|\boldsymbol{X} = \boldsymbol{x}\Big] \tag{4.77}$$

$$= \mathbb{E}\Big[(Y(1) - \overline{\mu}_1(\boldsymbol{X}))\left(\frac{\mathbb{E}(T \mid \boldsymbol{X})}{e(\boldsymbol{X})} - 1\right)|\boldsymbol{X} = \boldsymbol{x}\Big] \tag{4.78}$$

$$= 0, \tag{4.79}$$

  where the last line holds by the definition of the propensity score $e$.

- If the propensity model is misspecified but the outcome models are correctly specified (i.e. $\overline{\mu}_1 = \mu_1$ *almost surely*), we would have

$$\eta_1(\boldsymbol{x}) = \mathbb{E}\Big[\left(\frac{T}{\overline{e}(\boldsymbol{X})} - 1\right)(Y(1) - \mu_1(\boldsymbol{X}))|\boldsymbol{X} = \boldsymbol{x}\Big] \tag{4.80}$$

$$= \mathbb{E}\Big[\mathbb{E}\Big[\left(\frac{T}{\overline{e}(\boldsymbol{X})} - 1\right)(Y(1) - \mu_1(\boldsymbol{X}))|T, \boldsymbol{X}\Big]|\boldsymbol{X} = \boldsymbol{x}\Big] \tag{4.81}$$

$$= \mathbb{E}\Big[\left(\frac{T}{\overline{e}(\boldsymbol{X})} - 1\right)(\mathbb{E}[Y(1) \mid T, \boldsymbol{X}] - \mu_1(\boldsymbol{X})|\boldsymbol{X} = \boldsymbol{x}\Big] \tag{4.82}$$

  Thus, by assumption [4.2.14]

$$\eta_1(\boldsymbol{x}) = \mathbb{E}\Big[\left(\frac{T}{\overline{e}(\boldsymbol{X})} - 1\right)(\mathbb{E}[Y(1) \mid \boldsymbol{X}] - \mu_1(\boldsymbol{X})|\boldsymbol{X} = \boldsymbol{x}\Big] \tag{4.83}$$

$$= \mathbb{E}\Big[\left(\frac{T}{\overline{e}(\boldsymbol{X})} - 1\right)(\mu_1(\boldsymbol{X}) - \mu_1(\boldsymbol{X})|\boldsymbol{X} = \boldsymbol{x}\Big] \tag{4.84}$$

$$= 0. \tag{4.85}$$

Analogously, we prove that $\eta_0(\boldsymbol{x}) = 0$ if one of the models is misspecified and we end the proof. ■

Hence, the *DR-learner*, where *DR* refers to the *Doubly-Robust* learned outcome in the algorithm, is built from observed data $(\boldsymbol{x}^{(i)}, t_i, Y_{\text{obs},i})_{1\leq i\leq n}$ in three stages:

- Estimate by regression the propensity score $e$ and the outcome models $\mu_1$ and $\mu_0$ using $\mathbf{D}_{\text{obs}}$, denote $\widehat{e}$, $\widehat{\mu}_0$ and $\widehat{\mu}_1$ the obtained estimators.
- Estimate CATE by regressing the DR pseudo-outcome $\tilde{\boldsymbol{z}}_{\text{DR}} = (\tilde{z}_{\text{DR},i})_{i=1}^n$ on the covariates $\boldsymbol{X}$ with the correspondent estimators ($\widehat{\mu}_0, \widehat{\mu}_1$ and $\widehat{e}$) using

$$\tilde{z}_{\text{DR},i} = \widehat{\mu}_1(\boldsymbol{x}^{(i)}) + t_i\frac{Y_{\text{obs},i} - \widehat{\mu}_1(\boldsymbol{x}^{(i)})}{\widehat{e}(\boldsymbol{x}^{(i)})} - \widehat{\mu}_0(\boldsymbol{x}^{(i)}) - (1 - t_i)\frac{Y_{\text{obs},i} - \widehat{\mu}_0(\boldsymbol{x}^{(i)})}{1 - \widehat{e}(\boldsymbol{x}^{(i)})}. \tag{4.86}$$

### R-learner

The *R-learner* is an approach of meta-learning for estimating Heterogeneous Treatment Effects, based on the Robinson Robinson (1988) decomposition in partially linear models. Let $\epsilon$ be the random variable defined by

$$\epsilon = Y_{\text{obs}} - T\mu_1(\boldsymbol{X}) - (1 - T)\mu_0(\boldsymbol{X}). \tag{4.87}$$

**Proposition 4.4.9.** *Let $\epsilon$ be the outcome model error defined in (4.87), then $\mathbb{E}(\epsilon \mid T, \boldsymbol{X}) = 0$*

*Proof.* For $t \in \{0, 1\}$ and $\boldsymbol{x} \in \mathcal{D}$

$$\mathbb{E}(\epsilon \mid T = t, \boldsymbol{X} = \boldsymbol{x}) = \mathbb{E}\left[Y_{\text{obs}} - T\mu_1(\boldsymbol{X}) - (1 - T)\mu_0(\boldsymbol{X}) \mid T = t, \boldsymbol{X} = \boldsymbol{x}\right] \tag{4.88}$$

$$= \mathbb{E}\left[Y(t) \mid \boldsymbol{X} = \boldsymbol{x}\right] - t\mu_1(\boldsymbol{x}) - (1 - t)\mu_0(\boldsymbol{x}) \tag{4.89}$$

$$= \begin{cases} \mathbb{E}(Y(0) \mid \boldsymbol{X} = \boldsymbol{x}) - \mu_0(\boldsymbol{x}) = 0 & \text{if } t = 0, \\ \mathbb{E}(Y(1) \mid \boldsymbol{X} = \boldsymbol{x}) - \mu_1(\boldsymbol{x}) = 0 & \text{if } t = 1. \end{cases} \tag{4.90}$$

■

In binary case, the Robinson decomposition can be written as

$$\epsilon = Y_{\text{obs}} - m(\boldsymbol{X}) - (T - e(\boldsymbol{X}))\,\tau(\boldsymbol{X}), \tag{4.91}$$

where $m(\boldsymbol{x}) = \mathbb{E}(Y_{\text{obs}} \mid \boldsymbol{X} = \boldsymbol{x})$ and $e(\boldsymbol{x}) = \mathbb{E}(T \mid \boldsymbol{X} = \boldsymbol{x})$.

*Proof.* For $\boldsymbol{x} \in \mathcal{D}$, we have

$$\begin{aligned}
\mathbb{E}(Y_{\text{obs}} \mid \boldsymbol{X} = \boldsymbol{x}) &= \mathbb{E}\left[\epsilon + T\mu_1(\boldsymbol{X}) + (1-T)\mu_0(\boldsymbol{X}) \mid \boldsymbol{X} = \boldsymbol{x}\right] && (4.92)\\
&= \mathbb{E}[\mathbb{E}(\epsilon \mid T, \boldsymbol{X} = \boldsymbol{x})] + \mathbb{E}\left(\mu_0(\boldsymbol{X}) + T\ \tau(\boldsymbol{X}) \mid \boldsymbol{X} = \boldsymbol{x}\right) && (4.93)\\
&= \mu_0(\boldsymbol{x}) + \mathbb{E}(T \mid \boldsymbol{X} = \boldsymbol{x})\tau(\boldsymbol{x}) && (4.94)\\
&= \mu_0(\boldsymbol{x}) + e(\boldsymbol{x})\tau(\boldsymbol{x}). && (4.95)
\end{aligned}$$

Hence, $\mu_0(\boldsymbol{x}) = m(\boldsymbol{x}) - e(\boldsymbol{x})\tau(\boldsymbol{x})$ which leads finally to the Robinson decomposition

$$\epsilon = Y_{\text{obs}} - m(\boldsymbol{X}) - (T - e(\boldsymbol{X}))\,\tau(\boldsymbol{X}). \tag{4.96}$$

■

The representation above (4.91) has been studied by Nie & Wager (2020) to develop a flexible Meta-Learner, called the *R-Learner*. The goal of this representation is to form a squared error loss based on orthogonalization with respect to both observed outcome and propensity score estimate. Nie & Wager (2020) show that minimizing this loss function captures the CATE efficiently and use it to obtain a Quasi-Oracle estimator $\widehat{\tau}(\cdot)$ of the CATE in two steps:

- Estimate the outcome model $m$ and the propensity score $e$ using $\mathbf{D}_{\text{obs}}$ and denote $\widehat{m}$ and $\widehat{e}$ the obtained models.
- Find the optimal model $\widehat{\tau}_R$ within a family $\mathcal{F}$ of parametric or non-parametric candidate base-learner models such that

$$\begin{aligned}
\widehat{\tau}_R(\cdot) = \text{argmin}_{\tau(\cdot)\in\mathcal{F}} \Bigg(\frac{1}{n}\sum_{i=1}^{n} \Big[\left(Y_{\text{obs},i} - \widehat{m}(\boldsymbol{x}^{(i)})\right) - \left(t_i - \widehat{e}(\boldsymbol{x}^{(i)})\right)\tau(\boldsymbol{x}^{(i)})\Big]^2 \\
+ \Lambda_n\big[\tau(\cdot)\big]\Bigg),
\end{aligned} \tag{4.97}$$

  where $\Lambda_n[\tau(\cdot)]$ is a regularization term of the function $\tau(\cdot)$.

## 4.5 Extension to multi-valued and continuous treatment

The goal of this section is to infer causal effects when the treatment is no more binary but takes at least three possible values. We follow the extension of the Potential outcome theory to multiple and continuous treatments regime as developed by Frölich (2002); Imai & Dyk (2004); Imbens (2000); Lechner (2001) and Galagate (2016).

### The multi-valued treatment regime

Let $\mathcal{T} = \{t^{(0)}, t^{(1)}, \dots, t^{(K)}\}$ (do not confuse with $t_1, \dots, t_n$ corresponding to the treatment assigned to units) be the treatment support for $K+1$ ordered possible treatment levels of $T$. We suppose that we observe always an *i.i.d.* sample of $n$ units $\mathbf{D}_{\text{obs}} = (D_{\text{obs},i})_{i=1}^n = (\boldsymbol{x}^{(i)}, t_i, y_i)_{i=1}^n$ where $\boldsymbol{x}^{(i)}$ denotes a vector of covariates with values in $\mathcal{D}$, $t_i$ denotes the assigned treatment to unit $i$ with possible values in $\mathcal{T}$ and $y_i$ denotes the outcome of the unit $i$ ater .

Following the potential outcome framework, we suppose the existence of $Y(t)$, the real-valued counterfactual outcome that would have been observed under treatment level $t \in \mathcal{T} = \{t^{(0)}, \dots, t^{(K)}\}$. We suppose also that $\mathbf{D}_{\text{obs}}$ is a random sample drawn *i.i.d.* from a joint distribution $p_D$ where $D = (\boldsymbol{X}, T, (Y(t))_{t\in\mathcal{T}})$.

Similarly to the binary setting, for a unit $i$, the observed outcome $Y_{\text{obs},i}$ can be written as a function of the potential outcomes:

$$Y_{\text{obs},i} = \sum_{t\in\mathcal{T}} \mathbf{1}\{t_i = t\} Y_i(t), \tag{4.98}$$

where $\mathbf{1}\{T = t\}$ is the indicator function of the event $T = t$.

The actual model is a generalization of the Rubin causal model Rubin (1974, 1978, 1990) used in the causal inference of a binary treatment effect. We may refer to Lopez & Gutman (2017) for a review of the extension of causal effects estimation in multiple treatments. The consistency assumption $Y_{\text{obs}} = Y(T)$ holds directly with (4.98), the other assumptions and properties of this model remain valid when dealing with multiple treatments:

**Assumption 4.5.1** (Unconfoundedness)**.** *Given the observed covariates $\boldsymbol{X}$, the treatment mechanism is unconfounded for all treatment levels*

$$\forall t \in \mathcal{T} : Y(t) \perp\!\!\!\perp \mathbf{1}\{T = t\} \mid \boldsymbol{X}. \tag{4.99}$$

The previous assumption is a weak version of the unconfoundedness. Some authors in the literature may claim the joint conditional independence of the treatment $T$ and all potential outcomes $(Y(t))_{t\in\mathcal{T}}$ given the covariates $\boldsymbol{X}$.

**Assumption 4.5.2** (Overlap)**.** *The probability of receiving the treatment $T$ given observed covariates $\boldsymbol{X} = \boldsymbol{x}$ is positive, i.e. there exists $r_{\min} > 0$ such that*

$$\forall \boldsymbol{x} \in \mathcal{D}, \forall t \in \mathcal{T} : r_{\min} \leq r(t, \boldsymbol{x}) := \mathbb{P}(T = t \mid \boldsymbol{X} = \boldsymbol{x}). \tag{4.100}$$

$r$ is called the Generalized Propensity Score (GPS) (Imbens, 2000) and extends the classical Propensity Score from $e$ to the multiple treatment setting. It has the same balancing property as the classical Propensity Score, that is (Imbens, 2004):

$$\boldsymbol{X} \perp\!\!\!\perp \mathbf{1}\{T = t\} \mid r(t, \boldsymbol{X}). \tag{4.101}$$

Under the previous assumptions, causal effects can be identified and the counterfactual response satisfies:

$$\mathbb{E}(Y(t) \mid \boldsymbol{X} = \boldsymbol{x}) = \mathbb{E}(Y_{\text{obs}} \mid T = t, \boldsymbol{X} = \boldsymbol{x}). \tag{4.102}$$

In the multiple treatments regime, one can estimate the Average Dose-Response Function (ADRF)

$$\mu(t) = \mathbb{E}[Y(t)]. \tag{4.103}$$

It usually describes the magnitude of the outcome of the population when exposed to a specific level of the treatment (amount of smoked cigarettes, quantity of exposed pollution etc.).

One also can consider the Average Treatment Effect (ATE) between two levels $t$ and $s$, usually $s = t_0$ the baseline treatment value

$$\mu_{t,s} = \mathbb{E}[Y(t) - Y(s)], \tag{4.104}$$

The ATE between levels $t$ and $s$ can be inferred directly if one has already estimated the ADRF. The final causal estimands are heterogeneous treatment effects, given by the Conditional Average Treatment Effect (CATE) between two levels $t$ and $s$:

$$\tau_{t,s}(\boldsymbol{x}) = \mathbb{E}[Y(t) - Y(s) \mid \boldsymbol{X} = \boldsymbol{x}]. \tag{4.105}$$

The ADRF estimation, also known as exposure-response modelling, was considered (though less than the ATE in the binary case) and successfully applied in many domains, including in medicine, economics (Dominici et al., 2002; Flores, 2007; Hu et al., 2020; Lin et al., 2019; Saini et al., 2019). The CATE's estimation, however, remains less prominently studied in the literature. Hill (2011) proposes to model the counterfactual response surface $\mathbb{E}(Y(t) \mid \boldsymbol{X})$ by Bayesian Additive Regression Trees (BART) but did not go further for continuous treatments. Later, Hu et al. (2020) consider the same model and studied it further for the estimation of counterfactual response and causal effects. Harada & Kashima (2021); Schwab et al. (2020) applied neural networks and representations learning to estimate counterfactual response curves for multiple and continuous treatments (more precisely for graph-structured treatments) and Kaddour et al. (2021) proposed Structured Intervention Networks (SIN) for estimating CATEs of structured treatments.

Most methods for estimating the ADRF function use approaches similar to the binary case. Namely, the propensity score weighting (Feng et al., 2012; Imbens, 2000; Mccaffrey et al., 2013), matching and sub-classification (Yang et al., 2016) and vector matching (VM) (Lopez & Gutman, 2017). However, this approach appears challenging to implement and costly when the number of treatments is too large, or the sample is too small. Other methods include regression adjustment using BART (Hu et al., 2020), Generalized Linear models (GLM) (Guardabascio & Ventura, 2014), Generalized Additive models (GAM) (Zhang et al., 2016), stratification on the GPS are also possible. We refer the reader to Zhang et al. (2016) and Galagate (2016) for a review of existing methods for estimating the ADRF.

Given the importance of using the GPS in most methods to estimate the ADRF, special attention should be given while evaluating the ARDF. The following subsection reviews some existing methods for GPS estimation.

### Estimating the Generalized Propensity Score (GPS)

The first method of estimating Generalized Propensity Scores (GPS) to appear required some assumptions on the conditional density of $T$ given $\boldsymbol{X}$ (Imai & Van Dyk, 2004; Imbens, 2004) and do not offer practical guidance to estimate GPS in general cases. Still, some recent studies have

proposed parametric estimation of the propensity score via the multinomial logistic regression (Spreeuwenberg et al., 2010) or ordinal logistic regression model with an iterative approach (Zanutto et al., 2005), ensemble methods to estimate propensity score (Yan et al., 2019). In the following, we present three methods for estimating the GPS that we will consider later in Chapter 5 and in Appendix B.4 (Figures B.1 and B.2), but we refer to Lin et al. (2019) for a detailed review of existing methods.

The estimation of the propensity score $r$ can be seen as a particular application of the multi-class classification problem. In the following paragraphs, we consider a problem of multi-class classification. The covariates $\boldsymbol{X}$ are the inputs. The response is the treatment $T$ with $K+1$ possible values in $\mathcal{T}$, each class corresponds to a treatment level $t^{(k)}$, we aim to build learning model $f$ able to estimate $\mathbb{P}(T = t \mid \boldsymbol{X})$ from the learning sample $\mathbf{D}_{\text{obs}}$.

**Generalized Linear Models (GLM)** Generalized linear modelling (GLM) is a framework for statistical analysis introduced by Nelder & Wedderburn (1972) and developed by McCullagh & Nelder (1989) to overcome the linear modelling framework issues and to deal with non-normally distributed response variables, with the condition of belonging to the *exponential family*. It models in particular the conditional expectation $\mathbb{E}(\boldsymbol{Y} \mid \boldsymbol{X})$ of the multi-variate response $\boldsymbol{Y} = \left(\mathbf{1}\{T = t^{(0)}\}, \ldots, \mathbf{1}\{T = t^{(K)}\}\right)^\top \in \mathbb{R}^K$ given covariates $\boldsymbol{X}$ by a linear model through a link function.

In our setting, for given $\boldsymbol{x} \in \mathcal{D}$, we are interested into $\boldsymbol{\pi}(\boldsymbol{x}) = (\pi_k(\boldsymbol{x}))_{0 \leq k \leq K}$, where $\pi_k(\boldsymbol{x}) = \mathbb{P}(T = t^{(k)} \mid \boldsymbol{X} = \boldsymbol{x})$ is the conditional probability of getting the treatment value $t^{(k)}$ given $\boldsymbol{x}$ satisfying $\sum_{k=0}^{K} \pi_k(\boldsymbol{x}) = 1$. With the previous condition, it is sufficient to estimate only $(\pi_k(\boldsymbol{x}))_{1 \leq k \leq K}$ and deduce immediately $\pi_0(\boldsymbol{x})$.

Let $\boldsymbol{Y} = \left(\mathbf{1}\{T = t^{(K)}\}, \ldots, \mathbf{1}\{T = t^{(K)}\}\right)^\top \in \mathbb{R}^K$ satisfying

$$\mathbb{E}(\boldsymbol{Y} \mid \boldsymbol{X} = \boldsymbol{x}) = \begin{bmatrix} \mathbb{E}(\mathbf{1}\{T = t^{(1)}\} \mid \boldsymbol{X} = \boldsymbol{x}) \\ \vdots \\ \mathbb{E}(\mathbf{1}\{T = t^{(K)}\} \mid \boldsymbol{X} = \boldsymbol{x}) \end{bmatrix} = \begin{bmatrix} \pi_1(\boldsymbol{x}) \\ \vdots \\ \pi_K(\boldsymbol{x}) \end{bmatrix} = \boldsymbol{\pi}(\boldsymbol{x}). \tag{4.106}$$

We assume that the covariates $\boldsymbol{X}$ are related to the multivariate response $\boldsymbol{Y}$ by a continuous invertible mapping $\boldsymbol{g} : \mathbb{R}^K \to \mathbb{R}^K$, called the *link function*, through a vector linear predictor $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_K)$ where $\eta_k(\boldsymbol{x}) = \beta_k^{(0)} + \boldsymbol{\beta}_k^\top \boldsymbol{x}$ and $\boldsymbol{\beta}_k$ is the regression coefficients vector,

Under these assumptions, $\boldsymbol{\pi}(\boldsymbol{x})$ is fully characterized by GLM once the link function $\boldsymbol{g}$ is specified. Indeed,

$$\boldsymbol{\pi}(\boldsymbol{x}) = \begin{bmatrix} \pi_1(\boldsymbol{x}) \\ \vdots \\ \pi_K(\boldsymbol{x}) \end{bmatrix} = \begin{bmatrix} g_1^{-1}(\eta_1(\boldsymbol{x})) \\ \vdots \\ g_K^{-1}(\eta_k(\boldsymbol{x})) \end{bmatrix} = \boldsymbol{g}^{-1}(\boldsymbol{\eta}(\boldsymbol{x})). \tag{4.107}$$

Here, we consider the canonical link function $\boldsymbol{g} : \mathcal{M} \to \mathbb{R}^K$,

$$g_k(\boldsymbol{\pi}) = \log\left(\frac{\pi_k}{1 - \sum_{k=1}^{K} \pi_j}\right) \quad k = 1, \ldots, K, \tag{4.108}$$

where

$$\mathcal{M} = \{\boldsymbol{\pi} = (\pi_k)_{1 \leq k \leq K} \in (0,1)^K; \sum_{k=1}^{K} \pi_k < 1\}. \tag{4.109}$$

Hence, under the assumption of GLM, for all $k = 1, \ldots, K$

$$\begin{aligned} \pi_k(\boldsymbol{x}) &= g_k^{-1}\left(\beta_k^{(0)} + \boldsymbol{\beta}_k^\top \boldsymbol{x}\right) \\ &= \frac{\exp\left(\beta_k^{(0)} + \boldsymbol{\beta}_k^\top \boldsymbol{x}\right)}{1 + \sum_{j=1}^K \exp\left(\beta_j^{(0)} + \boldsymbol{\beta}_j^\top \boldsymbol{x}\right)}. \end{aligned} \tag{4.110}$$

Therefore, we obtain the *multinomial logistic regression* formulas for propensity score's estimation, for $t^{(k)} \in \mathcal{T}\backslash\{t^{(0)}\}$,

$$r(t^{(k)}, \boldsymbol{x}^{(i)}) = \mathbb{P}(T = t^{(k)} \mid \boldsymbol{X} = \boldsymbol{x}^{(i)}) = \frac{\exp\left(\beta_k^{(0)} + \boldsymbol{\beta}_k^\top \boldsymbol{x}^{(i)}\right)}{1 + \sum_{j=1}^K \exp\left(\beta_j^{(0)} + \boldsymbol{\beta}_j^\top \boldsymbol{x}^{(i)}\right)}, \tag{4.111}$$

and,

$$r(t^{(0)}, \boldsymbol{x}^{(i)}) = \mathbb{P}(T = t^{(0)} \mid \boldsymbol{X} = \boldsymbol{x}^{(i)}) = \frac{1}{1 + \sum_{j=1}^K \exp\left(\beta_j^{(0)} + \boldsymbol{\beta}_j^\top \boldsymbol{x}^{(i)}\right)}. \tag{4.112}$$

Finally, the regression coefficients $(\boldsymbol{\beta}_k)_{1\leq k\leq K}$ are estimated from the data $\mathbf{D}_{\text{obs}}$, by maximizing the log-likelihood given parameters $(\boldsymbol{\beta}_k)_{1\leq k\leq K}$ using Newton-Raphson/Fisher's scoring algorithm (Nelder & Wedderburn, 1972).

**Random Forest models** Random forests (Breiman, 2001) is a popular tree-based algorithm using a substantial version of *bagging* (Bootstrap aggregating) to build a large collection of decorrelated decision trees (see (Breiman et al., 1984) for more details about decision trees) to capture complex nonlinear interaction and to deal efficiently with high-variance and low-bias cases (Hastie et al., 2001).

Consider the same framework of multi-class classification as defined previously. A decision tree is a directed graph consisting of nodes and edges. The nodes are either internal (non-terminal) containing some attribute test conditions to split on covariates $\boldsymbol{X}$, or leaf (terminal) corresponding to a class label $t^{(k)}$.

Given the learning data $\mathbf{D}_{\text{obs}}$, the decision tree is built in such a way that its attributes split the data so that each terminal node is as *pure* as possible, that is, each leaf in the tree contains units of a single class.

The *impurity* of the nodes can be computed by using

- The Gini index:

$$Gini(L) = 1 - \sum_{k=0}^{K} \hat{p}_k^2. \tag{4.113}$$

- The Entropy information:

$$Ent(L) = -\sum_{k=0}^{K} \hat{p}_k \log(\hat{p}_k). \tag{4.114}$$

where $\hat{p}_k$ is the relative proportion of class $t^{(k)}$ in the leaf $L$.

Once a decision tree $f_b$ is trained by minimizing the *impurity* function over all its leaves, it predicts the most probable class $t^{(k)}$ among training observations that fall into the same leaf $L(\boldsymbol{x})$ as $\boldsymbol{x}$, using a measure called *vote*, defined by

$$\text{vote}(f_b(\boldsymbol{x})) = t^{(k)}. \tag{4.115}$$

However, constructing a decision tree may be computationally unfeasible and decision trees are usually high variance and prone to over-fit the data. Random forests are created to cope with these drawbacks by growing a multitude of decision trees where each tree is trained on different samples and covariates selected randomly. Indeed, we generate $B$ bootstrap samples from the original sample with (or without) replacement. Then we build a decision tree $f_b$ for $b = 1, \ldots, B$ using each bootstrapped sample. At each node of $f_b$, a split is performed by minimizing the *impurity* criterion as in (4.113) and (4.114). While forming the best split of $f_b$'s nodes, a random sample of $m \leq d$ covariates are chosen as split candidates between $d$ covariates $\boldsymbol{X}$ (In classification, we generally choose $m = \sqrt{d}$ (James et al., 2014) and the tree $f_b$ is grown until the minimum node size $N_L$ is achieved in each leaf). The parameter node size $N_L$ controls the complexity of each tree. If it is not specified, then the tree $f_b$ is expanded until all its leaves are pure with respect to Gini (4.113) or entropy (4.114) measures.

For a point $\boldsymbol{x} \in \mathcal{D}$, as the leaf $L_b(\boldsymbol{x})$ containing $\boldsymbol{x}$ in each decision tree $f_b$ is not necessarily *pure*, $f_b$ predicts now the class that occurs most frequently in $L_b(\boldsymbol{x})$

$$\text{vote}(f_b(\boldsymbol{x})) = \underset{0 \leq k \leq K}{\arg\max}\ \widehat{p}_{kb}(\boldsymbol{x}), \qquad b = 1, \ldots, B, \tag{4.116}$$

where $\widehat{p}_{kb}(\boldsymbol{x})$ is the proportion of $k$-th class observations when $\boldsymbol{x}$ is falling in the $f_b$'s tree terminal node $L_b(\boldsymbol{x})$ containing $\mid L_b(\boldsymbol{x}) \mid$ observations

$$\widehat{p}_{kb}(\boldsymbol{x}) = \frac{1}{\mid L_b(\boldsymbol{x}) \mid} \sum_{i, \boldsymbol{x}^{(i)} \in L_b(\boldsymbol{x})} \mathbf{1}\{t_i = t^{(k)}\}, \qquad b = 1, \ldots, B, \tag{4.117}$$

When the quantity of interest is the class-probability, the built forest of $B$ trees aggregates class proportion's predictions in the terminal node of each decision tree $f_b$ and the GPS $\widehat{r}$ is estimated as

$$\widehat{r}(t^{(k)}, \boldsymbol{x}) = \frac{1}{B} \sum_{b=1}^{B} \widehat{p}_{kb}(\boldsymbol{x}), \qquad k = 0, \ldots, K. \tag{4.118}$$

The optimal random forest can be obtained by tuning its hyperparameters (the number of trees $B$, the maximum number of selected covariates $m$ and the minimum node size $N_L$) using a grid search combined with a cross-validation method Probst et al. (2019).

**Generalized boosted models** Generalized Boosted Models (GBM) are a set of automated data-adaptive algorithms based on a set of standard weak base learners. From these weak learners, we aim to build a strong learner able to predict more precisely real outcomes and capture/nonlinear interactive effects of the covariates. The statistical framework of GBM has been developed by Friedman (2001) for estimating and predicting a function subject to minimizing a loss function or an empirical risk.

In multi-class classification, we seek to learn a multivariate predictive model $\boldsymbol{\phi} : \mathbb{R}^d \to \mathbb{R}^{K+1}$ where $\boldsymbol{\phi}(\boldsymbol{x}) = (\phi_k(\boldsymbol{x}))_{0 \leq k \leq K} \in (0,1)^{K+1}$ designs the empirical class-probability vector of

assigning treatment $T = t^{(k)}$ by the model $\boldsymbol{\phi}$ given covariates $\boldsymbol{x}$ ($\boldsymbol{\phi}$ is in the end no more than a predictor of the GPS $r$). Note that each $\phi_k$ can be linked to $F_k : \mathbb{R}^d \to \mathbb{R}$ by the *softmax* function

$$\phi_k(\boldsymbol{x}) = \frac{\exp(F_k(\boldsymbol{x}))}{\sum_{j=0}^{K} \exp(F_j(\boldsymbol{x}))}, \quad k = 0, \ldots, K. \tag{4.119}$$

The loss function to be considered is the *Cross-Entropy* loss $L : \mathbb{R} \times \mathbb{R}^{K+1} \to \mathbb{R}$ defined as

$$L(y, \boldsymbol{\phi}(\boldsymbol{x})) = -\sum_{k=0}^{K} \mathbf{1}\{y = t^{(k)}\} \log(\phi_k(\boldsymbol{x})), \quad y \in \mathbb{R},\ \boldsymbol{x} \in \mathbb{R}^d. \tag{4.120}$$

In the boosting framework, each $F_k$ is assumed to be sum of additive functions (base-learners) belonging to a functional space $\mathcal{F}_{\text{bl}}$, typically the space of decision trees (CART) $\mathcal{F}_{\text{CART}} = \{f_1, \ldots, f_M, f_m : \mathbb{R}^d \to \mathbb{R}\}$ where each $f_m$ corresponds to a decision tree with a structure (characterized by the maximum depth or number of leaves for example), and $M$ is the number of decision trees in the space $\mathcal{F}_{\text{CART}}$.

Unlike the random forest, which involves bootstrap aggregating (*Bagging*), GBM grows base-learners sequentially (*Boosting*), and each base-learner $f_m$ is fitted on a re-weighted version of the original data. The new base-learner is chosen to provide the best fit to the residuals on the loss function $L$ of the previous base-learner model. When adding the new base-learner, the contribution of each new base-leaser is scaled by a factor $0 \leq \eta < 1$ to improve the smoothness of the resulting model and the final model fit (Friedman, 2001; James et al., 2014).

The Gradient Boosting algorithm is originally described by Friedman (2001) and uses negative gradients to optimize the loss function. However, as the original algorithm is stagewise, adding base-learners members one after the other may randomly influence the loss function, making the optimization procedure harder and unfeasible. Some variants of gradient boosting like XGBoost have been developed to optimize the loss function efficiently, scalable end-to-end tree boosting method by Chen & Guestrin (2016). In the XGBoost framework, the empirical loss function to minimize integrates a regularization term $\Omega$ penalizing the complexity of the base-learner models

$$\mathcal{L}((F_k)_{0 \leq k \leq K}) = \sum_{i=1}^{n} L\left(t_i, \boldsymbol{\phi}(\boldsymbol{x}^{(i)})\right) + \sum_{m=1}^{M} \Omega\left(f_m\right) \quad k = 0, \ldots, K. \tag{4.121}$$

where $\Omega(f_m) = \gamma J + 1/2\lambda_0 \sum_{j=1}^{J} \omega_j^2$ for a decision tree $f_m$, $J$ is the number of leaves in the tree $f_m$ and $\omega_j$ is the score of the $j-$th leaf of the tree. The regularization term $\Omega(f)$ penalizes several leaf nodes and avoids over-fitting by selecting simple and predictive regression trees into the final model $F_M$. In most cases, the regularization parameters take the default values $\lambda_0 = 1$ and $\gamma = 0$.

Some other variants like the Cyclic Gradient Boosting (Zhang et al., 2019) or BOOMER (Rapp et al., 2020) can also be used in learning multi-class classification with gradient boosting. In related work, Mccaffrey et al. (2013) used GBM to estimate initial GPS models $(\widehat{p}_k)_{0 \leq k \leq K}$ then fit these models by defining a specific criterion to assess covariates balancing in the optimization procedure.

### The continuous treatment regime

Suppose now that the treatment assignment variable $T$ is continuous with a support $\mathcal{T} = [t_{\min}, t_{\max}] \subseteq \mathbb{R}$. Following the Neyman (1923) potential outcomes framework and the generalization of the Rubin (1974, 1978, 1979, 1990) Causal Model, we suppose the existence of $Y(t)$, the real-valued counterfactual outcome that would have been observed under a treatment level $t \in \mathcal{T}$. As for binary and multi-valued treatments, we consider $D = (\boldsymbol{X}, T, Y(t)_{t\in\mathcal{T}})$ with joint distribution $p_D$. We suppose that we observe an *i.i.d.* sample of $n$ units $D_{\text{obs},i} = \left(\boldsymbol{x}^{(i)}, t_i, y_i\right)$ drawn from $p_D$ and that $y_i = Y_{\text{obs},i} = Y_i(t_i)$ (consistency assumption).

The assumptions of unconfoundedness and common support are still necessary to make the causal inference in continuous treatments.

**Assumption 4.5.3** (Unconfoundedness)**.** *The treatment mechanism is unconfounded given the observed covariates* $Y(t) \perp\!\!\!\perp T \mid \boldsymbol{X}$ *for all* $t \in \mathcal{T}$.

**Assumption 4.5.4** (Overlap)**.** *The conditional density* $f_{T|\boldsymbol{X}}$ *is uniformly bounded from* 0 *i.e. there exists* $r_{\min} > 0$ *such that* $r_{\min} \leq f_{T|\boldsymbol{X}}(t \mid \boldsymbol{x})$ *for all* $t \in \mathcal{T}$ *and* $\boldsymbol{x} \in \mathcal{D}$.

The conditional density $f_{T|\boldsymbol{X}}$ is also called the generalized propensity score $r$ (Imbens, 2004) such that $r(t, \boldsymbol{x}) = f_{T|\boldsymbol{X}}(t \mid \boldsymbol{x})$. It generalizes the classical propensity score and the multi-valued generalized propensity score to continuous treatments.

Using different terminology, Imai & Dyk (2004) proposed a generalization of the propensity score for continuous treatments, called the propensity function (P-Function). In their seminal work, Imai & Dyk (2004) made an extra assumption to uniquely parameterize the P-Function, that is, For almost every $\boldsymbol{x} \in \mathcal{D}$, $(f_{T|\boldsymbol{X}}(t \mid \boldsymbol{x}))_{t\in\mathcal{T}}$ is characterized by $\Theta(\boldsymbol{x})$, where $\boldsymbol{x} \in \mathcal{D} \mapsto \Theta(\boldsymbol{x}) \in \mathbb{R}^q$ is a measurable map.

The main difference between the GPS and the P-Function lies in the fact the GPS evaluate the conditional density $f_{T|\boldsymbol{X}}$ at the observed covariate, whereas the P-Function, under the assumption above, focuses on uniquely parameterizing it (Zhao et al., 2020).

Both the GPS and the P-Function can be used to eliminate selection and confounding biases. They are useful with the following properties:

- For the P-Function:

$$f_{T|\boldsymbol{X}} = f_{T|\Theta(\boldsymbol{X})}, \tag{4.122}$$

$$\boldsymbol{X} \perp\!\!\!\perp T \mid \Theta(\boldsymbol{X}), \tag{4.123}$$

$$\forall t \in \mathcal{T} : \; Y(t) \perp\!\!\!\perp T \mid \Theta(\boldsymbol{X}). \tag{4.124}$$

- For the GPS:

$$\text{For almost every } t \in \mathcal{T} : \; f_T(t \mid r(t, \boldsymbol{X}), Y(t)) = f_T(t \mid r(t, \boldsymbol{X})). \tag{4.125}$$

*Proof.* The proofs of (4.122-4.124) are in Appendix B.1. The proof of 4.125 can be found in Imbens (2004) in Theorem 1. ■

The advantage of the P-Function over the GPS is that $Y(t)$ and $T$ are conditionally independent given the low-dimensional score (Zhao et al., 2020). This is an interesting *balancing property*

and ensures the randomization between balanced units. However, the assumption of the unique characterization of $(f_{T|\boldsymbol{X}}(t \mid \boldsymbol{x}))_{t\in\mathcal{T}}$ might be too restrictive.

The previous assumptions allow the identification of causal effects. Indeed, the counterfactual response satisfies:

$$\mathbb{E}(Y(t) \mid \boldsymbol{X} = \boldsymbol{x}) = \mathbb{E}(Y_{\text{obs}} \mid T = t, \boldsymbol{X} = \boldsymbol{x}). \tag{4.126}$$

Under continuous treatments, we are interested in estimating the Average Dose-Response Function (ADRF)

$$\mu(t) = \mathbb{E}[Y(t)], \tag{4.127}$$

or the finite difference average treatment effect (ATE),

$$\tau_{t,s} = \mathbb{E}[Y(t) - Y(s)], \tag{4.128}$$

for any two levels of treatment of interest $t, s \in \mathcal{T}$, or Conditional Average Treatment Effects (CATEs) between two levels $t$ and $s$:

$$\tau_{t,s}(\boldsymbol{x}) = \mathbb{E}[Y(t) - Y(s) \mid \boldsymbol{X} = \boldsymbol{x}]. \tag{4.129}$$

Most work in the literature, or maybe all of them, focus only on the estimation of the ARDF modelling (Colangelo & Lee, 2020; Galagate, 2016; Galvao & Wang, 2015; Imbens, 2004; Kennedy et al., 2017; Zhao et al., 2020). On the one hand, to the best of our knowledge and excepting the work of Zhang et al. (2022), the estimation of heterogeneous effects so far is not studied in its theoretical and practical aspects for continuous, and the existing approaches such as (Kaddour et al., 2021; Schwab et al., 2020) are more based on learning representations. On the other hand, the ARDF estimation methods use the generalization (but not the direct extension because this implies handling the indicator function by kernels methods) of methods already present for the binary and multi-valued setting. This includes regression adjustment, propensity-score weighting, matching (Wu et al., 2018), covariates balancing and procedures based on machine learning algorithms. We may refer the reader to Galagate (2016) thesis for a review of some of these methods.

### Estimating the conditional density

The enormous difficulty of estimating the conditional densities dramatically impacts the causal inference under continuous treatments.

However, recent advances have been made in the literature, and various approaches and methods are proposed to estimate the GPS (following the terminology of Imbens (2004)). The developed estimators can use parametric methods such as the kernel density estimator or non-parametric and Machine Learning methods such as the Lasso regression and (Su et al., 2019), artificial neural networks (Chen & White, 1999), random forest (Colangelo & Lee, 2020) and generalized boosting models (Zhu et al., 2015).

One interesting idea is to follow the scheme of Belloni et al. (2019). Indeed, if we assume that the treatment $T$ has a density and that

$$T = m(\boldsymbol{X}) + \epsilon, \tag{4.130}$$

where $m$ is a given function, and $\epsilon$ is the model's error $m$ assumed to be independent of $\boldsymbol{X}$.

**Proposition 4.5.5.** *Under the previous model, if the assumption 4.5.4 holds and if $\epsilon$ has a density $f_\epsilon$, then estimating the GPS is equivalent to estimating the density $f_\epsilon$ of $\epsilon$.*

*Proof.*

$$\begin{aligned} R(t, \boldsymbol{x}) &= \mathbb{P}(T \leq t \mid \boldsymbol{X} = \boldsymbol{x}) = \mathbb{P}(m(\boldsymbol{X}) + \epsilon \leq t \mid \boldsymbol{X} = \boldsymbol{x}) \\ &= \mathbb{P}(\epsilon \leq t - m(\boldsymbol{x}) \mid \boldsymbol{X} = \boldsymbol{x}) \\ &= \mathbb{P}(\epsilon \leq t - m(\boldsymbol{x})) \\ &= F_\epsilon(t - m(\boldsymbol{x})), \end{aligned} \tag{4.131}$$

where $F_\epsilon$ is the Cumulative Distribution Function of $\epsilon$.

Let $\Delta t > 0$ be small enough, then

$$\begin{aligned} 2r(t, \boldsymbol{x})\Delta t &\approx \mathbb{P}(T \in [t - \Delta t, t + \Delta t] \mid \boldsymbol{X} = \boldsymbol{x}) \\ &= \mathbb{P}(T \leq t + \Delta t \mid \boldsymbol{X} = \boldsymbol{x}) - \mathbb{P}(T \leq t - \Delta t \mid \boldsymbol{X} = \boldsymbol{x}) \\ &= F_\epsilon(t + \Delta t - m(\boldsymbol{x})) - F_\epsilon(t - \Delta t - m(\boldsymbol{x})). \end{aligned} \tag{4.132}$$

Therefore,

$$r(t, \boldsymbol{x}) = \frac{F_\epsilon(t + \Delta t - m(\boldsymbol{x})) - F_\epsilon(t - \Delta t - m(\boldsymbol{x}))}{2\Delta t} \xrightarrow{\Delta t \to 0} f_\epsilon(t - m(\boldsymbol{x})), \tag{4.133}$$

which ends the proof. ■

Under this proposition, one can estimate the GPS $r$ from observed data $\mathbf{D}_{\text{obs}}$ in three-steps procedure: First, estimate the treatment model $\widehat{m}$ by regressing $T$ on $\boldsymbol{X}$. Secondly, estimate the CDF $\widehat{F}_\epsilon$ from the observed residuals $(\epsilon_i)_{i=1}^n$. Finally, compute the estimated GPS $\widehat{r}(t, \boldsymbol{x})$ with the previous proposition.

### Estimating the ADRF

**Outcome modelling** The outcome regression modelling of the ARDF consists of assuming some parametric or non-parametric form on the outcome, then performing an estimation procedure using the observed data $\mathbf{D}_{\text{obs}}$. In addition to BART (Hill, 2011), we present two other methods.

Imbens (2004) is the first to propose a method for estimating the ARDF, called the *Y-model*. For $t \in \mathcal{T}$ and $r \in (0, 1)$, we define $\eta(t, r)$ by:

$$\eta(t, r) = \mathbb{E}[Y(t) \mid r(t, \boldsymbol{X}) = r]. \tag{4.134}$$

Hence,

$$\mu(t) = \mathbb{E}[Y(t)] = \mathbb{E}[\mathbb{E}[Y(t) \mid r(t, \boldsymbol{X})]] = \mathbb{E}\left[\eta(t, r(t, \boldsymbol{X}))\right]. \tag{4.135}$$

Moreover, using the balancing property (4.125) of the GPS $\eta(t, r) = \mathbb{E}[Y_{\text{obs}} \mid T = t, r(T, \boldsymbol{X}) = r]$. Therefore, given an estimator $\hat{r}$ of the GPS, one can regress the observed outcome $Y_{\text{obs}}$ on $(t_i, \hat{r}(t_i, \boldsymbol{x}^{(i)}))_{i=1}^n$ or model it given a parametric form and get an estimator $\hat{\eta}$ of the conditional expectation $\eta$. Finally, for each $t \in \mathcal{T}$, one estimate the ADRF function $\hat{\mu}(t)$ as

$$\hat{\mu}(t) = \frac{1}{n} \sum_{i=1}^{n} \hat{\eta}(t, \hat{r}(t, \boldsymbol{x}^{(i)}). \tag{4.136}$$

In a similar approach, Imai & Van Dyk (2004) proposed a *T-model*, which was modified by Zhao et al. (2020), that uses the P-Function to learn and estimate ADRF. Indeed, by (4.124)

$$\begin{aligned}\mu(t) &= \mathbb{E}[\mathbb{E}[Y(t) \mid \Theta(\boldsymbol{X})]] = \mathbb{E}[\mathbb{E}[Y(t) \mid \Theta(\boldsymbol{X}), T = t]] \\ &= \int \mathbb{E}[Y_{\text{obs}} \mid T = t, \Theta(\boldsymbol{X}) = \Theta(\boldsymbol{x})] p_{\boldsymbol{X}}(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} \\ &= \int \mathbb{E}[Y_{\text{obs}} \mid T = t, \Theta(\boldsymbol{X}) = \boldsymbol{\theta}] p_{\Theta(\boldsymbol{X})}(\boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{\theta} \\ &= \int \gamma(t, \boldsymbol{\theta}) p_{\Theta(\boldsymbol{X})}(\boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{\theta},\end{aligned} \tag{4.137}$$

where $p_{\Theta(\boldsymbol{X})}$ denotes the distribution of $\Theta(\boldsymbol{X})$ and $\gamma(t, \boldsymbol{\theta}) = \mathbb{E}[Y_{\text{obs}} \mid T = t, \Theta(\boldsymbol{X}) = \boldsymbol{\theta}]$.

The estimation procedure follows the same ideas as the *Y-model*. First, we estimate $\Theta(\boldsymbol{X})$ from observed data $\mathbf{D}_{\text{obs}}$ to get $\hat{\Theta}(\boldsymbol{X})$. Second, we fit a smooth-coefficient model on $Y_{\text{obs}} \mid T, \hat{\Theta}(\boldsymbol{X})$ and estimate the model $\gamma$. Finally, for each treatment level $t \in \mathcal{T}$ and by considering $\hat{\boldsymbol{\theta}}_i = \hat{\Theta}(\boldsymbol{x}^{(i)})$, we estimate the ADRF function $\mu(t)$ as follows (Zhao et al., 2020):

$$\hat{\mu}(t) = \frac{1}{n} \sum_{i=1}^{n} \hat{\gamma}(t, \hat{\boldsymbol{\theta}}_i). \tag{4.138}$$

**Inverse Propensity Weighting** The Inverse Propensity Weighting is proposed by Flores et al. (2012) for the ADRF estimation. It can be seen as a version of Horvitz & Thompson (1952) weighting applied to continuous treatments. If the assumption 4.5.1 holds, then one can overcome the problem of indicator function (since $\mathbb{P}(T = t) = 0$ when the variable $T$ has a continuous density) by introducing a kernel $K_h$, where $h$ is the bandwidth and the Nadaraya-Watson estimator of the ARDF:

$$\hat{\mu}_{\text{NW}}(t) = \frac{\sum_{i=1}^{n} \tilde{K}_h(t_i - t) y_i}{\sum_{i=1}^{n} \tilde{K}_h(t_i - t)}, \tag{4.139}$$

where, for a given estimator $\hat{r}$ of the GPS,

$$\tilde{K}_h(T_i - t) = \frac{K_h(t_i - t)}{\hat{r}(t, \boldsymbol{x}^{(i)})}. \tag{4.140}$$

The GPS estimator $\hat{r}$ is used to weight the kernel $K_h$ (Flores et al., 2012).

Flores et al. (2012) perform a local linear regression to propose an estimator with the form

$$\hat{\mu}_{\text{IPW}}(t) = \frac{D_0(t) S_2(t) - D_1(t) S_1(t)}{S_0(t) S_2(t) - S_1^2(t)}, \tag{4.141}$$

where, for $j = 0, 1, 2$ and a weighted kernel function $\tilde{K}_h$

$$S_j(t) = \sum_{i=1}^{n} \tilde{K}_h(t_i - t)(t_i - t)^j, \tag{4.142}$$

$$D_j(t) = \sum_{i=1}^{n} \tilde{K}_h(t_i - t)(t_i - t)^j y_i. \tag{4.143}$$

Another method was also proposed by Galagate (2016) as an extension of IPW with second-moment. For a unit $i$ and $\boldsymbol{b}(t) = (b_1(t), \dots, b_p(t))^\top$ a $p$-vector of known basis functions, we assume a linear relationship between the treatment and the outcome,

$$Y_i(t) = \sum_{j=1}^{p} \theta_{ij} b_j(t) = \boldsymbol{\theta}_i^\top \boldsymbol{b}(t). \tag{4.144}$$

Under the assumed functional form of the outcome, the Inverse Second-Moment Weighting (ISMW) generates sample weight matrix $\mathbf{W}$ using the random vector $\boldsymbol{B} = \boldsymbol{b}(T)$ such that $\boldsymbol{W}_i = \left(\mathbb{E}\left[\boldsymbol{B}\boldsymbol{B}^\top \mid \boldsymbol{X} = \boldsymbol{x}^{(i)}\right]\right)^{-1}$. The ISMW estimator $\hat{\mu}_{\text{ISMW}}$ of the ADRF is given by

$$\hat{\mu}_{\text{ISMW}}(t) = \hat{\boldsymbol{\xi}}^\top \boldsymbol{b}(t), \tag{4.145}$$

where $\boldsymbol{B}_i = \boldsymbol{b}(T_i)$ for $i \in \{1, \dots, n\}$ and

$$\hat{\boldsymbol{\xi}} = \left(\sum_{i=1}^{n} \boldsymbol{W}_i \boldsymbol{B}_i \boldsymbol{B}_i^\top\right)^{-1} \left(\sum_{i=1}^{n} \boldsymbol{W}_i \boldsymbol{B}_i y_i\right). \tag{4.146}$$

The last method is the *doubly robust* estimation (Kennedy et al., 2017). It has the same properties of double robustness against model misspecification as already discussed in section 4.4.

For a given arbitrary estimators $\overline{\mu}, \overline{r}$ of $\mu(t, \boldsymbol{x}) = \mathbb{E}(Y_{\text{obs}} \mid T = t, \boldsymbol{X} = \boldsymbol{x})$ and $r(t, \boldsymbol{x}) = f_{T|\boldsymbol{X}}(t \mid \boldsymbol{x})$ Kennedy et al. (2017) consider the following pseudo-outcome:

$$Z_{\text{DR}}(\overline{\mu}, \overline{r}) = \frac{Y_{\text{obs}} - \overline{\mu}(\boldsymbol{X}, T)}{\overline{r}(T, \boldsymbol{X})} \int_{\mathcal{D}} \overline{r}(T, \boldsymbol{x}) p_{\boldsymbol{X}}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} + \int_{\mathcal{D}} \overline{\mu}(T, \boldsymbol{x}) p_{\boldsymbol{X}}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}. \tag{4.147}$$

The pseudo-outcome $Z_{\text{DR}}(\overline{\mu}, \overline{r})$ provides a consistent estimator of the ADRF $\mathbb{E}[Z_{\text{DR}}(\overline{\mu}, \overline{r}) \mid T = t] = \mu(t)$ if either $\overline{\mu} = \mu$ or $\overline{r} = r$. Bonvini & Kennedy (2022) show that the best convergence rate attainable would be using the Doubly robust method and non-parametric regression.

**Covariate Balancing methods** Similarly to binary treatments, covariate balancing methods are extended to continuous treatments to address the misspecification of a conditional density model $r$. The idea of covariate balancing methods is to form modified weights and then solve them through various optimization criteria. The goal is to ensure that optimal weights satisfy the balancing condition, that is, the treatment $T$ and the covariates $\boldsymbol{X}$ are independent.

The existing approaches in the literature of covariates balancing under continuous treatments include the Generalized Covariate Balancing Propensity Score (GCBPS) approach (Fong et al., 2018), Covariates association eliminating weights (Yiu & Su, 2018), entropy balancing for continuous treatments (Tübbicke, 2022; Vegetabile et al., 2021), end-to-end balancing (E2B) based on Generalized Stable Weights (Bahadori et al., 2022) and Generative Adversarial Deconfounding (GAD) (Li et al., 2020). We do not present these methods in detail, but we refer the reader to the related papers.

# CHAPTER 5

# Meta-learners for multi-valued treatments

Some sections and passages of this chapter are taken from the paper (Acharki et al., 2022),

## 5.1 Introduction

With the rapid development of Machine Learning and its efficiency in predicting outcomes, the question of counterfactual prediction *"what would happen if ?"* arises. Engineers may want to know how the outcome (e.g. production) would be affected when a parameter is changed to a specific value. It will help them personalize the parameter at efficient levels and optimize the outcome. Recently, many companies have relied on supervised machine learning models to find the optimal intervention strategy. Yet, the results are not satisfactory. Indeed, these models do not account for other impacting effects (One-At-a-Time approach) and cannot distinguish between correlations and causal relationships in the data.

In Chapter 4, we have seen that, based on the Potential Outcomes theory (Neyman, 1923; Rubin, 1974), epidemiologists and statisticians developed a set of statistical tools to make causal inference and estimate the effects of a treatment on the outcome whether on average among the whole population or inside different sub-groups. They have been successfully applied in many fields such as medicine, economics, public policy and advertising/marketing. Nevertheless, they are still unfamiliar and seldom used in industrial applications.

Further, most existing methods and studies are limited to the setting of a binary treatment, whereas in many real-world applications, the treatment variable can take multiple values. In some cases, it would be helpful to give an in-depth analysis of the impact of the treatment across its possible levels (doses) instead of just considering a binary scenario where the treatment is either assigned or not. In addition, the heterogeneity of effects may provide valuable information regarding the effectiveness of this treatment and help companies or governments to personalize their policies and strategies. Unfortunately, Heiler & Knaus (2021) show that binarizing multi-treatments can lead to a misleading estimation of heterogeneous effects across different levels. Consequently, a detailed study of heterogeneous treatment effects is required under multi-valued treatments.

Finally, randomized controlled trials (RCTs) are not always conducted, and the ground truth of treatment effects cannot be observed and is rarely available. This fact makes heterogeneous treatment effects estimation different from a standard supervised learning problem (Alaa & van der Schaar, 2018). Therefore, it is challenging to assess treatment effect estimators' performances and select the best model with standard point-wise error metrics such as Mean

Squared Errors.

In Chapter 5, we study the problem of estimating Heterogeneous Treatment Effects, also known as Conditional Average Treatment Effects (CATEs), when the treatment is multi-valued. In Section 5.2, following the same taxonomy as Curth & van der Schaar (2021a); Knaus et al. (2020a), we establish *meta-learners* for Conditional Average Treatment Effects under multi-valued treatment. In Section 5.3, we analyze the error bounds of pseudo-outcome meta-learners and show the advantage of the X-learner. In Section 5.4, we present a semi-synthetic dataset that will serve to validate Causal Inference methods. We end this chapter by presenting some numerical studies and experiments showing the performances of the X-learner in the multi-valued setting.

## 5.2 Meta-learners in the multi-treatment regime

We recall the multi-treatment setting as defined in Section 4.5: we consider a treatment variable $T$ that can takes $K+1$ ordered possible levels in $\mathcal{T} = \{t^{(0)}, t^{(1)}, \dots, t^{(K)}\}$. We suppose that we have observe *i.i.d* sample of $n$ units $\mathbf{D}_{\text{obs}} = (D_{\text{obs},i})_{i=1}^n = (\boldsymbol{x}^{(i)}, t_i, y_i)_{i=1}^n$ where $\boldsymbol{x}^{(i)}$ denotes a vector of covariates with values in $\mathcal{D}$, $t_i$ denotes the assigned treatment to unit $i$ with possible values in $\mathcal{T}$ and $y_i$ denotes the outcome of the unit $i$. We suppose finally the existence of potential outcomes $(Y(t))_{t\in\mathcal{T}}$ and the causal assumptions (4.5.1-4.5.2). We are interested in the estimation of the Conditional Average Treatment Effect (CATE) between two levels $t$ and $s$:

$$\tau_{t,s}(\boldsymbol{x}) = \mathbb{E}[Y(t) - Y(s) \mid \boldsymbol{X} = \boldsymbol{x}]. \tag{5.1}$$

To make notations more simple in the following, we consider CATEs $(\tau_k)_{k=1}^K$ estimation problem between $t^{(k)}$ and $t^{(0)}$ for $k = 1, \dots, K$ such that

$$\tau_k(\boldsymbol{x}) = \mathbb{E}[Y(t^{(k)}) - Y(t^{(0)}) \mid \boldsymbol{X} = \boldsymbol{x}]. \tag{5.2}$$

To tackle the problem of estimating CATEs under multi-valued treatment, we generalize the notion of meta-learners to derive consistent estimators of the CATE. This task can be achieved either by modelling the CATE directly in one step or two steps: by decomposing it into regularized regression problems or by addressing a minimization problem with respect to an appropriate loss function. Moreover, all considered meta-learners below, except the R-learner, can support any supervised regression Machine Learning method (e.g. random forest, gradient boosting methods).

In the following, we follow a similar taxonomy of CATEs estimators as Curth & van der Schaar (2021a); Knaus et al. (2020a). Namely, direct plug-in (one-step) meta-learners, pseudo-outcome (two-step) meta-learners and Neyman-Orthogonality based learners (R-learner).

### Direct plug-in meta-learners

In this subsection, we present direct plug-in meta-learners, also known as *one-step* learners that estimate the CATE in (5.1) by targeting directly the observed data $\mathbf{D}_{\text{obs}}$. They are the naive extension of the T- and S-learners in the binary case.

**T-learner with multiple treatments.** T-learner is a naive approach to estimating CATEs. It consists on estimating the *two* conditional response surfaces $\mu_t(\boldsymbol{x}) = \mathbb{E}(Y(t) \mid \boldsymbol{X} = \boldsymbol{x})$ using $\mathbf{S}_t = \{i, t_i = t\}$ for $t \in \{t^{(k)}, t^{(0)}\}$ as in the binary case. The T-learning approach does not account for the interaction between treatment $T$ and the outcome $Y$ and creates different models for different treatments.

Despite its naivety, the T-learning approach may suffer from selection bias (Curth & van der Schaar, 2021b), that is, when the outcome models $\mu_t$ are estimated with respect to the wrong distribution of the training sample. To overcome this issue in the estimation of $\mu_t$ while sampling $(D_{\text{obs},i})_{i\in\mathbf{S}_t}$, we use Importance Sampling (Hassanpour & Greiner, 2019), and we show the following proposition.

**Proposition 5.2.1.** *For a treatment level $t \in \mathcal{T}$, the expected squared error of the estimator $\widehat{\mu}_t$ on the outcome surface $\mu_t$ satisfies:*

$$\mathbb{E}_{\boldsymbol{X}\sim p(\cdot)}\left[\left(\widehat{\mu}_t(\boldsymbol{X}) - \mu_t(\boldsymbol{X})\right)^2\right] = \mathbb{E}_{\boldsymbol{X}\sim p(\cdot|T=t)}\left[\frac{\mathbb{P}(T=t)}{r(t,\boldsymbol{X})}\left(\widehat{\mu}_t(\boldsymbol{X}) - \mu_t(\boldsymbol{X})\right)^2\right]. \tag{5.3}$$

*where $p(\cdot)$ is the marginal distribution of $\boldsymbol{X}$ and $p(\cdot \mid T = t)$ is the conditional distribution of $\boldsymbol{X}$ given $T = t$.*

*Proof.* In Appendix B.1. ■

The proposition 5.2.1 highlights the fact that $\mu_t$ should be estimated by minimizing the expected squared error on the nominal *weighted* distribution.

Therefore, the T-learner in the multi-treatment setting can be built as follows

- For $t \in \{t^{(k)}, t^{(0)}\}$, consider the sample $(D_{\text{obs},i})_{i\in\mathbf{S}_t}$ and estimate the conditional response $\widehat{\mu}_t$ by minimizing the expected squared error of the estimator $\widehat{\mu}_t$.
- Compute the CATE between two treatment levels $t^{(k)}$ and $t^{(0)}$ by:

$$\widehat{\tau}_k^{(\mathrm{T})}(\boldsymbol{x}) = \widehat{\mu}_{t^{(k)}}(\boldsymbol{x}) - \widehat{\mu}_{t^{(0)}}(\boldsymbol{x}). \tag{5.4}$$

**S-learner with multiple treatments.** Using the identification of the CATE by assumptions (4.5.1)-(4.5.2), we can write:

$$\tau_k(\boldsymbol{x}) = \mathbb{E}(Y_{\text{obs}} \mid T = t^{(k)}, \boldsymbol{X} = \boldsymbol{x}) - \mathbb{E}(Y_{\text{obs}} \mid T = t^{(0)}, \boldsymbol{X} = \boldsymbol{x}). \tag{5.5}$$

Therefore, instead of splitting the dataset and building separate models as in T-learning, one can consider a *single* model built from the whole dataset and naturally define the S-learner in case of the multi-treatment setting as

- Regress $Y_{\text{obs}}$ on the treatment $T$ and the covariates $\boldsymbol{X}$ by a single model $\widehat{\mu}$ using $\mathbf{D}_{\text{obs}}$.
- Estimate the CATE between two treatment levels $t^{(k)}$ and $t^{(0)}$ by:

$$\widehat{\tau}_k^{(\mathrm{S})}(\boldsymbol{x}) = \widehat{\mu}(\boldsymbol{x}, t^{(k)}) - \widehat{\mu}(\boldsymbol{x}, t^{(0)}). \tag{5.6}$$

Obviously, including the treatment $T$ as an input feature and sharing some information between covariates $\boldsymbol{X}$ and $T$ may provide better predictions. However, this result is conditioned by the ability of the regression model to capture and distinguish contributions of both $\boldsymbol{X}$ and $T$ on $Y_{\text{obs}}$.

In the binary cases, the S-learner is usually considered a good choice Curth & van der Schaar (2021b); Künzel et al. (2019) and has shown its performance. Although, as we will see in Section 5.5, its results are very sensitive to the base learner, particularly for random forests, because it cannot capture the correct effect of the treatment variable.

Note that the S-learning approach may also suffer from confounding and regularization biases (Chernozhukov et al., 2018; Hahn et al., 2020) when estimating the counterfactual response model $\widehat{\mu}$.

### Pseudo-outcome meta-learners

Despite Proposition 5.2.1 for overcoming selection bias, it implies learning in small samples, which may harm the quality of the meta-learner when $\mathbf{S}_t$ becomes small for a certain $t$. This is all the more critical as the number $K$ of treatments increases. An alternative (and usual) possibility for mitigating this bias is to consider some specific representations of the observed outcome $Y_{\text{obs}}$, called *pseudo-outcome*. These representations incorporate *nuisance components* that generally include valuable information such as the dependence between covariates $\boldsymbol{X}$ and $T$ (i.e. the GPS) and the occurrence of a particular treatment assignment. Further, regressing the pseudo-outcome produces a new *regularized* estimator that predicts the *right* treatment effect instead of predicting a biased effect while keeping the same sample size as $Y_{\text{obs}}$.

**M-learner with multiple treatments.** Similarly to the binary case, the M-learner is inspired from the Inverse Propensity Weighting (IPW) transformation to estimate causal effects by standardizing the outcome on the GPS.

In the multi-valued setting, for $k = 1, \dots, K$, we define the *modified* pseudo-outcome $Z_k^M$ in the multi-treatment regime using the IPW representation as:

$$Z_k^M = \frac{\mathbf{1}\{T = t^{(k)}\}}{r(t^{(k)}, \boldsymbol{X})} Y_{\text{obs}} - \frac{\mathbf{1}\{T = t^{(0)}\}}{r(t^{(0)}, \boldsymbol{X})} Y_{\text{obs}}, \tag{5.7}$$

where $r(t, \boldsymbol{x}) = \mathbb{P}(T = t \mid \boldsymbol{X} = \boldsymbol{x})$ is the GPS.

**Proposition 5.2.2.** *Under the assumptions (4.5.1)-(4.5.2)*

$$\mathbb{E}(Z_k^M \mid \boldsymbol{X} = \boldsymbol{x}) = \tau_k(\boldsymbol{x}). \tag{5.8}$$

*Proof.* For $t \in \mathcal{T}$, we consider $Y_t^M$ the modified IPW representation of $Y_{\text{obs}}$ in such way that $Z_k^M = Y_{t^{(k)}}^M - Y_{t^{(0)}}^M$. By noticing that $\mathbf{1}\{T = t\} Y_{\text{obs}} = \mathbf{1}\{T = t\} Y(t)$, we have for $\boldsymbol{x} \in \mathcal{D}$:

$$\begin{aligned}
\mathbb{E}(Y_t^M \mid \boldsymbol{X} = \boldsymbol{x}) &= \mathbb{E}\left[\frac{\mathbf{1}\{T = t\}}{r(t, \boldsymbol{X})} Y_{\text{obs}} \mid \boldsymbol{X} = \boldsymbol{x}\right] \\
&= \frac{1}{r(t, \boldsymbol{x})} \mathbb{E}\left[\mathbf{1}\{T = t\} Y(t) \mid \boldsymbol{X} = \boldsymbol{x}\right] \\
&= \frac{1}{r(t, \boldsymbol{x})} \mathbb{E}\big[\mathbf{1}\{T = t\} \mid \boldsymbol{X} = \boldsymbol{x}\big] \mathbb{E}\left[Y(t) \mid \boldsymbol{X} = \boldsymbol{x}\right] \qquad \text{(by Assumption 4.5.1)} \\
&= \mathbb{E}(Y(t) \mid \boldsymbol{X} = \boldsymbol{x}) = \mu_t(\boldsymbol{x}).
\end{aligned} \tag{5.9}$$

Thus $\mathbb{E}(Z_k^M \mid \boldsymbol{X} = \boldsymbol{x}) = \mu_{t^{(k)}}(\boldsymbol{x}) - \mu_{t^{(0)}}(\boldsymbol{x})$ and we get the desired result. ■

Unfortunately, the M-learner is very sensitive to the estimation of the GPS and suffers from high variance, even when the propensity score is correctly specified or known and constant (Curth & van der Schaar, 2021a). Moreover, the *modified* pseudo-outcome can often be null, leading to an over-fitting problem as the base-learner may try to predict zero instead of $\tau_k$. Again, this becomes more critical as the number $K$ of treatments increases as some values of the GPS $r$ can be smaller than $1/K$.

**DR-learner with multiple treatments.** Requiring the consistency of the GPS estimator may be hard to get a correct estimation of CATEs. The *Doubly Robust* (DR) method (Kennedy, 2020; Kennedy et al., 2017; Robins et al., 1994) is helpful in overcoming the problem of the model's misspecification by estimating two components, the outcome model $\mu_t$ and the GPS $r$, instead of relying on the correctness of one (and the only) parameter.

Let $\overline{\mu}$ denote an arbitrary model of the outcome $\mu$, let $\overline{r}$ denote also an arbitrary model of the GPS $r$, we assume that $\overline{r}$ respects also Assumption (4.5.2). For $k = 1, \dots, K$, we define *doubly-robust* pseudo-outcome $Z^{DR}_{\overline{\mu},\overline{r},k}$ as

$$\begin{aligned} Z^{DR}_{\overline{\mu},\overline{r},k} = \frac{Y_{\text{obs}} - \overline{\mu}_T(\boldsymbol{X})}{\overline{r}(t^{(k)}, \boldsymbol{X})} \mathbf{1}\{T = t^{(k)}\} - \frac{Y_{\text{obs}} - \overline{\mu}_T(\boldsymbol{X})}{\overline{r}(t^{(0)}, \boldsymbol{X})} \mathbf{1}\{T = t^{(0)}\} \\ + \overline{\mu}_{t^{(k)}}(\boldsymbol{X}) - \overline{\mu}_{t^{(0)}}(\boldsymbol{X}). \end{aligned} \tag{5.10}$$

**Proposition 5.2.3.** *Let $Z^{DR}_{\overline{\mu},\overline{r},k}$ be the Doubly-Robust pseudo-outcome defined in (5.10), then under the assumptions (4.5.1)-(4.5.2)*

$$\mathbb{E}(Z^{DR}_{\overline{\mu},\overline{r},k} \mid \boldsymbol{X} = \boldsymbol{x}) = \tau_k(\boldsymbol{X}), \tag{5.11}$$

*if either the outcome model or the propensity model is well-specified, i.e. $\overline{\mu}_t(\boldsymbol{X}) = \mu_t(\boldsymbol{X})$ and $\overline{\mu}_{t^{(0)}}(\boldsymbol{X}) = \mu_{t^{(0)}}(\boldsymbol{X})$ almost surely, or $\overline{r}(T, \boldsymbol{X}) = r(T, \boldsymbol{X})$ almost surely.*

*Proof.* Let $\overline{\mu}$ denote an arbitrary model of the outcome $\mu$, and let $\overline{r}$ also denote an arbitrary model of the GPS $r$ satisfying the overlap assumption 4.5.2. Similarly to the previous proof, we consider $Y^{DR}_t$ the AIPW representation of $Y_{\text{obs}}$ such that $Z^{DR}_{\overline{\mu},\overline{r},k} = Y^{DR}_{\overline{\mu},\overline{r},t^{(k)}} - Y^{DR}_{\overline{\mu},\overline{r},t^{(0)}}$, and we show that

$$\begin{aligned} \mathbb{E}(Y^{DR}_{\overline{\mu},\overline{r},t} \mid \boldsymbol{X} = \boldsymbol{x}) &= \mathbb{E}\left[\frac{Y_{\text{obs}} - \overline{\mu}_T(\boldsymbol{X})}{\overline{r}(t, \boldsymbol{X})} \mathbf{1}\{T = t\} + \overline{\mu}_t(\boldsymbol{X}) \mid \boldsymbol{X} = \boldsymbol{x}\right] \\ &= \mathbb{E}\left[\frac{Y(t) - \overline{\mu}_t(\boldsymbol{X})}{\overline{r}(t, X)} \mathbf{1}\{T = t\} + \overline{\mu}_t(\boldsymbol{X}) \mid \boldsymbol{X} = \boldsymbol{x}\right] \\ &= \mathbb{E}\left[Y(t) \mid \boldsymbol{X} = \boldsymbol{x}\right] + \mathbb{E}\left[\frac{Y(t) - \overline{\mu}_t(\boldsymbol{X})}{\overline{r}(t, X)} \mathbf{1}\{T = t\} - Y(t) + \overline{\mu}_t(X) \mid \boldsymbol{X} = \boldsymbol{x}\right] \\ &= \mu_t(\boldsymbol{x}) + \eta_t(\boldsymbol{x}), \end{aligned} \tag{5.12}$$

with $\eta_t(\boldsymbol{x}) = \mathbb{E}\left[\frac{\mathbf{1}\{T=t\} - \overline{r}(t,X)}{\overline{r}(t,X)}(Y(t) - \overline{\mu}_t(X)) \mid \boldsymbol{X} = \boldsymbol{x}\right]$.

We show that the second term $\eta_t$ is null under the double robustness of the model, that is, if one of the nuisance components is consistent.

- If the propensity model $\overline{r}$ is correctly specified (i.e. $\overline{r}(T, \boldsymbol{X}) = r(T, \boldsymbol{X})$ *almost surely*) but the outcome model is misspecified, we would have

$$\begin{aligned}
\eta_t(\boldsymbol{x}) &= \mathbb{E}\left[\frac{\mathbf{1}\{T=t\} - \overline{r}(t, \boldsymbol{X})}{\overline{r}(t, \boldsymbol{X})}(Y(t) - \overline{\mu}_t(\boldsymbol{X}))|\boldsymbol{X} = \boldsymbol{x}\right] \\
&= \mathbb{E}\left[\mathbb{E}\Big[\frac{\mathbf{1}\{T=t\} - \overline{r}(t, \boldsymbol{X})}{\overline{r}(t, \boldsymbol{X})}(Y(t) - \overline{\mu}_t(\boldsymbol{X})) \mid Y(t), \boldsymbol{X}\Big]|\boldsymbol{X} = \boldsymbol{x}\right] \\
&= \mathbb{E}\left[(Y(t) - \overline{\mu}_t(\boldsymbol{X}))\mathbb{E}\Big[\frac{\mathbf{1}\{T=t\} - \overline{r}(t, \boldsymbol{X})}{\overline{r}(t, \boldsymbol{X})} \mid Y(t), \boldsymbol{X}\Big]|\boldsymbol{X} = \boldsymbol{x}\right] \\
&= \mathbb{E}\left[(Y(t) - \overline{\mu}_t(\boldsymbol{X}))\frac{\mathbb{E}[\mathbf{1}\{T=t\} \mid Y(t), \boldsymbol{X}] - \overline{r}(t, \boldsymbol{X})}{\overline{r}(t, \boldsymbol{X})}|\boldsymbol{X} = \boldsymbol{x}\right] \\
&= \mathbb{E}\left[(Y(t) - \overline{\mu}_t(\boldsymbol{X}))\frac{\mathbb{E}[\mathbf{1}\{T=t\} \mid \boldsymbol{X}] - \overline{r}(t, \boldsymbol{X})}{\overline{r}(t, \boldsymbol{X})}|\boldsymbol{X} = \boldsymbol{x}\right] \quad \text{(by Assumption 4.5.1)} \\
&= \mathbb{E}\left[(Y(t) - \overline{\mu}_t(\boldsymbol{X}))\frac{r(t, \boldsymbol{X}) - \overline{r}(t, \boldsymbol{X})}{\overline{r}(t, \boldsymbol{X})}|\boldsymbol{X} = \boldsymbol{x}\right] = 0,
\end{aligned} \tag{5.13}$$

  where the last line holds by the definition of the Generalized Propensity Score $r(t, \boldsymbol{x})$.

- If the propensity model is misspecified, but the outcome model is correctly specified (i.e. $\overline{\mu}(T, \boldsymbol{X}) = \mu(T, \boldsymbol{X}) = \mathbb{E}(Y_{\text{obs}} \mid T, \boldsymbol{X})$ *almost surely*), we would have

$$\begin{aligned}
\eta_t(\boldsymbol{x}) &= \mathbb{E}\left[\frac{\mathbf{1}\{T=t\} - \overline{r}(T, X)}{\overline{r}(T, X)}(Y(t) - \mathbb{E}(Y_{\text{obs}} \mid T=t, \boldsymbol{X})) \mid \boldsymbol{X} = \boldsymbol{x}\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\frac{\mathbf{1}\{T=t\} - \overline{r}(T, \boldsymbol{X})}{\overline{r}(T, \boldsymbol{X})}(Y(t) - \mathbb{E}(Y_{\text{obs}} \mid T=t, \boldsymbol{X}))|T, \boldsymbol{X}\right] \mid \boldsymbol{X} = \boldsymbol{x}\right] \\
&= \mathbb{E}\left[\frac{\mathbf{1}\{T=t\} - \overline{r}(t, \boldsymbol{X})}{\overline{r}(t, \boldsymbol{X})}(\mathbb{E}\left[Y(t) \mid T, \boldsymbol{X}\right] - \mathbb{E}(Y_{\text{obs}} \mid T=t, \boldsymbol{X})) \mid \boldsymbol{X} = \boldsymbol{x}\right] \\
&= \mathbb{E}\left[\frac{\mathbf{1}\{T=t\} - \overline{r}(t, \boldsymbol{X})}{\overline{r}(t, \boldsymbol{X})}\mathbb{E}(\left[Y_{\text{obs}}|T=t, \boldsymbol{X}\right] - \mathbb{E}[Y_{\text{obs}} \mid T=t, \boldsymbol{X}])|\boldsymbol{X} = \boldsymbol{x}\right] \\
&= 0.
\end{aligned} \tag{5.14}$$

Note that assuming $\overline{\mu}_t = \mu_t = \mathbb{E}(Y_{\text{obs}} \mid T=t, \boldsymbol{X})$ is sufficient to prove that $\eta(\boldsymbol{x}) = 0$. The result holds similarly for $Y^{DR}_{\overline{\mu}, \overline{r}, t^{(0)}}$. Therefore, the consistency of the DR-learner is achieved if the propensity score is well-specified or if the potential outcome model is well-specified (at least for $t^{(k)}$ and $t^{(0)}$). ■

Therefore, the consistency of the DR-learner is achieved if at least one of the components (the propensity score model or outcome models) is well-specified. It also has the advantage of having a small asymptotic variance compared to the M-learner when the propensity score model is correct, as it will be shown in Section 5.5.

**X-learner with multiple treatments.** The X-learner Künzel et al. (2019), also known as *Regression-Adjustment* (RA)-learning in a developed version by Curth & van der Schaar (2021a), has been proposed as an alternative to T-learning in the case where one treatment group is over-represented. The idea consists of a *cross* procedure of estimation between observations $Y_{\text{obs}}$ and outcome models when one of the treatments occurs.

In the multi-treatment regime, for $k = 1, \dots, K$, we define the *Regression-Adjustment* pseudo-outcome $Z_k^X$ as

$$\begin{aligned} Z_k^X = \mathbf{1}\{T = t^{(k)}\}(Y_{\text{obs}} - \mu_{t^{(0)}}(\boldsymbol{X})) + \sum_{l \neq k} \mathbf{1}\{T = t^{(l)}\} \times \\ (\mu_{t^{(k)}}(\boldsymbol{X}) - Y_{\text{obs}}) + \sum_{l \neq k} \mathbf{1}\{T = t^{(l)}\}(\mu_{t^{(l)}}(\boldsymbol{X}) - \mu_{t^{(0)}}(\boldsymbol{X})). \end{aligned} \tag{5.15}$$

**Proposition 5.2.4.** *Under the assumptions (4.5.1)-(4.5.2)*

$$\mathbb{E}(Z_k^X \mid \boldsymbol{X} = \boldsymbol{x}) = \tau_k(\boldsymbol{x}). \tag{5.16}$$

*Proof.* By direct calculations, we show that

$$\mathbb{E}(Z_k^X \mid \boldsymbol{X} = \boldsymbol{x}) = \mathbb{E}\left[\mathbf{1}\{T = t^{(k)}\}Y(t^{(k)}) \mid \boldsymbol{X} = \boldsymbol{x}\right] - r(t^{(k)}, \boldsymbol{x})\mu_{t^{(0)}}(\boldsymbol{x}) + \sum_{l \neq k} r(t^{(l)}, \boldsymbol{x})\Big(\mu_{t^{(k)}}(\boldsymbol{x}) \tag{5.17}$$

$$- \mathbb{E}\left[\mathbf{1}\{T = t^{(l)}\}Y(t^{(l)}) \mid \boldsymbol{X} = \boldsymbol{x}\right]\Big) + \sum_{l \neq k} r(t^{(l)}, \boldsymbol{x})(\mu_{t^{(l)}}(\boldsymbol{x}) - \mu_{t^{(0)}}(\boldsymbol{x})) \tag{5.18}$$

$$= r(t^{(k)}, \boldsymbol{x})\mu_{t^{(k)}}(\boldsymbol{x}) - r(t, \boldsymbol{x})\mu_{t^{(0)}}(\boldsymbol{x}) + \sum_{l \neq k} \left(r(t^{(l)}, \boldsymbol{x})\mu_t(\boldsymbol{x}) - r(t^{(l)}, \boldsymbol{x})\mu_{t'}(\boldsymbol{x})\right) \tag{5.19}$$

$$+ \sum_{l \neq k} r(t^{(l)}, \boldsymbol{x})(\mu_{t^{(l)}}(\boldsymbol{x}) - \mu_{t^{(0)}}(\boldsymbol{x})) \qquad \text{(by Assumption 4.5.1)} \tag{5.20}$$

$$= r(t^{(k)}, \boldsymbol{x})\mu_{t^{(k)}}(\boldsymbol{x}) - r(t^{(k)}, \boldsymbol{x})\mu_{t^{(0)}}(\boldsymbol{x}) + \sum_{l \neq k} r(t^{(l)}, \boldsymbol{x})\mu_{t^{(k)}}(\boldsymbol{x}) - \sum_{l \neq k} r(t^{(l)}, \boldsymbol{x})\mu_{t^{(0)}}(\boldsymbol{x}) \tag{5.21}$$

$$= (\mu_{t^{(k)}}(\boldsymbol{x}) - \mu_{t^{(0)}}(\boldsymbol{x}))\Big(r(t^{(k)}, \boldsymbol{x}) + \sum_{l \neq k} r(t^{(l)}, \boldsymbol{x})\Big) \tag{5.22}$$

$$= \mu_{t^{(k)}}(\boldsymbol{x}) - \mu_{t^{(0)}}(\boldsymbol{x}) = \tau_k(\boldsymbol{x}). \tag{5.23}$$

■

**Remark 5.2.5.** *The X-learning approach provides also a new method for estimating the difference of Average Dose-Response Function (ARDF) $\eta(t) = \mathbb{E}(Y(t) - Y(t^{(0)}))$.*

In opposition to the DR-learner, the pseudo-outcome $Z_k^X$ incorporates only potential outcome models and does not imply the GPS $r$. Consequently, the X-learner is likely to have the smallest variance compared to other meta-learners when the GPS takes some extreme values (i.e. the overlap assumption (4.5.2) is not sufficiently respected). However, it requires the consistency of all components $(\widehat{\mu}_t)_{t \in \mathcal{T}}$ to estimate the CATE correctly.

The algorithm 2 summarizes CATEs estimation using the previous meta-learners. The "Transformation" function stands for the pseudo-outcome modification that has been applied to $Y_{\text{obs}}$ for the M-, DR- and X-learning approaches.

**Algorithm 2** Pseudo-outcome meta-learning estimation

**Input:** data $(\boldsymbol{x}^{(i)}, t_i, y_i)$, level $t$, model $(\overline{\tau}_k)_{k=1}^K$, Components $\widehat{r}, \widehat{\mu}_\cdot$.
**if** Components not provided **then**
  Estimate $\widehat{r}$ by regressing $T$ on $\boldsymbol{X}$.
  Estimate $\widehat{\mu}_\cdot$ by T-learning or S-learning
**end if**
**for** $k = 1, \ldots, K$ **do**
  $Z_{k,i} = \text{Transformation}(t^{(k)}, \boldsymbol{x}^{(i)}, t_i, y_i, \widehat{r}, \widehat{\mu}_\cdot)$
  Regress $Z_k$ on $\boldsymbol{X}$ using $\overline{\tau}_k$.
**end for**
**Output:** Learned model $(\widehat{\tau}_k)_{k=1}^K$.

In the estimation phase, three main approaches are possible to learn the nuisance components ($r$ and $\mu$) and then estimate the $\tau_k$, namely, Full-Sample, Sample-Split and Cross-Fit methods (Okasa, 2022). This chapter does not discuss estimation procedures and adopts the Full-Sample strategy.

## R-learning approach

The R-learner is based mainly on the Robinson (1988) decomposition to provide a flexible estimator avoiding regularization bias, with strong convergence rates. Principally, the R-learner achieves approximately asymptotic error rates as an *oracle* learner knowing the nuisance parameters perfectly.

The following proposition, which is slightly different from the work of Kaddour et al. (2021), aims to generalize the Robinson (1988) representation in the multi-treatment setting without assuming Product Decomposition of $Y_{\text{obs}}$.

**Proposition 5.2.6.** *In the multi-treatment regime, let $\epsilon$ be the outcome model error*

$$\epsilon = Y_{\text{obs}} - \sum_{t \in \mathcal{T}} \mathbf{1}\{T = t\}\mu_t(\boldsymbol{X}) = Y_{\text{obs}} - \mu_T(\boldsymbol{X}). \tag{5.24}$$

*Then $\epsilon$ satisfies $\mathbb{E}(\epsilon \mid T, \boldsymbol{X}) = 0$ (Neyman Orthogonality) and the decomposition*

$$\epsilon = Y_{\text{obs}} - m(\boldsymbol{X}) - \sum_{k=1}^{K} \left(\mathbf{1}\{T = t^{(k)}\} - r(t^{(k)}, \boldsymbol{X})\right)\tau_k(\boldsymbol{X}), \tag{5.25}$$

*where $m(\boldsymbol{x}) = \mathbb{E}(Y_{\text{obs}} \mid \boldsymbol{X} = \boldsymbol{x})$ is the observed outcome model and $r(t, \boldsymbol{x}) = \mathbb{P}(T = t \mid \boldsymbol{X} = \boldsymbol{x})$ is the GPS.*

*Proof.* We show first the Neyman-Orthogonality propriety, i.e. $\mathbb{E}(\epsilon \mid T, \boldsymbol{X}) = 0$. Indeed, for $t \in \mathcal{T}$ and $\boldsymbol{x} \in \mathcal{D}$, we have

$$\begin{aligned} \mathbb{E}[\epsilon \mid T = t, \boldsymbol{X} = \boldsymbol{x}] &= \mathbb{E}[Y_{\text{obs}} - \mu_T(\boldsymbol{X}) \mid T = t, \boldsymbol{X} = \boldsymbol{x}] \\ &= \mathbb{E}[Y(t) - \mu_T(\boldsymbol{X}) \mid T = t, \boldsymbol{X} = \boldsymbol{x}] \\ &= \mu_t(\boldsymbol{x}) - \mu_t(\boldsymbol{x}) = 0. \end{aligned} \tag{5.26}$$

Thus, the observed outcome model satisfies:

$$\begin{aligned}
\mathbb{E}(Y_{\text{obs}} \mid \boldsymbol{X} = \boldsymbol{x}) &= \mathbb{E}\big[\epsilon + \sum_{k=0}^{K} \mathbf{1}\{T = t^{(k)}\}\mu_{t^{(k)}}(\boldsymbol{X}) \mid \boldsymbol{X} = \boldsymbol{x}\big] \\
&= \mathbb{E}\big[\mathbb{E}[\epsilon \mid T, \boldsymbol{X}] \mid \boldsymbol{X} = \boldsymbol{x}\big] + \sum_{k=0}^{K} \mathbb{E}\big[\mathbf{1}\{T = t^{(k)}\} \mid \boldsymbol{X} = \boldsymbol{x}\big]\mu_{t^{(k)}}(\boldsymbol{x}) \\
&= \sum_{k=0}^{K} \mu_{t^{(k)}}(\boldsymbol{x}) r(t^{(k)}, \boldsymbol{x}) = \mu_{t^{(0)}}(\boldsymbol{x}) r(t^{(0)}, \boldsymbol{x}) + \sum_{k=1}^{K} \mu_{t^{(k)}}(\boldsymbol{x}) r(t^{(k)}, \boldsymbol{x}) \\
&= \mu_{t^{(0)}}(\boldsymbol{x})\big[1 - \sum_{k=1}^{K} r(t^{(k)}, \boldsymbol{x})\big] + \sum_{k=1}^{K} \mu_{t^{(k)}}(\boldsymbol{x}) r(t^{(k)}, \boldsymbol{x}) \\
&= \mu_{t^{(0)}}(\boldsymbol{x}) + \sum_{k=1}^{K} r(t, \boldsymbol{x}) \left[\mu_{t^{(k)}}(\boldsymbol{x}) - \mu_{t^{(0)}}(\boldsymbol{x})\right] \\
&= \mu_{t^{(0)}}(\boldsymbol{x}) + \sum_{k=1}^{K} r(t^{(k)}, \boldsymbol{x}) \tau_k(\boldsymbol{x}) = m(\boldsymbol{x}).
\end{aligned} \tag{5.27}$$

By gathering both quantities :

$$\begin{aligned}
Y_{\text{obs}} - m(\boldsymbol{X}) &= \sum_{k=0}^{K} \mathbf{1}\{T = t^{(k)}\}\mu_{t^{(k)}}(\boldsymbol{X}) - \mu_{t^{(0)}}(\boldsymbol{X}) - \sum_{k=1}^{K} r(t^{(k)}, \boldsymbol{X})\tau_k(\boldsymbol{X}) + \epsilon \\
&= \mathbf{1}\{T = t^{(0)}\}\mu_{t^{(0)}}(\boldsymbol{X}) + \sum_{k=1}^{K} \mathbf{1}\{T = t^{(k)}\}\mu_{t^{(k)}}(\boldsymbol{X}) - \mu_{t^{(0)}}(\boldsymbol{X}) - \sum_{k=1}^{K} r(t^{(k)}, \boldsymbol{X})\tau_k(\boldsymbol{X}) + \epsilon \\
&= (\mathbf{1}\{T = t^{(0)}\} - 1)\mu_{t^{(0)}}(\boldsymbol{X}) + \sum_{k=1}^{K} (\mathbf{1}\{T = t^{(k)}\}\mu_{t^{(k)}}(\boldsymbol{X}) - r(t^{(k)}, \boldsymbol{X})\tau_k(\boldsymbol{X})) + \epsilon \\
&= \sum_{k=1}^{K} (\mathbf{1}\{T = t^{(k)}\}\mu_{t^{(k)}}(\boldsymbol{X}) - r(t^{(k)}, \boldsymbol{X})\tau_k(\boldsymbol{X})) - \sum_{k=1}^{K} \mathbf{1}\{T = t^{(k)}\}\mu_{t^{(0)}}(\boldsymbol{X}) + \epsilon \\
&= \sum_{k=1}^{K} (\mathbf{1}\{T = t^{(k)}\}\mu_{t^{(k)}}(\boldsymbol{X}) - \mathbf{1}\{T = t^{(k)}\}\mu_{t^{(0)}}(\boldsymbol{X}) - r(t^{(k)}, \boldsymbol{X})\tau_k(\boldsymbol{X})) + \epsilon \\
&= \sum_{k=1}^{K} \left[\mathbf{1}\{T = t^{(k)}\} - r(t^{(k)}, \boldsymbol{X})\right]\tau_k(\boldsymbol{X}) + \epsilon.
\end{aligned} \tag{5.28}$$

Therefore, we obtain the generalized Robinson decomposition for the multi-treatment regime. ■

As described in the original paper of Nie & Wager (2020), the main interest of the previous decomposition relies on forming a pseudo-outcome error, implying only the regression of observed quantities on $\boldsymbol{X}$ (i.e. the observed outcome model $m$ and the GPS $r$), that isolates CATEs $\tau_k$ for all $k = 1, \ldots, K$. The generalized Robinson decomposition is relevant for two reasons. Firstly, setting up an error to minimize allows us to target CATEs models $\tau_k$ directly Kaddour et al. (2021). Secondly, requiring the observed outcome model is less restrictive than requiring potential outcome models $\mu_{\cdot}$ as in the DR- and X- pseudo-outcomes.

In the multi-treatment regime, considering the mean squared error of $\epsilon$ as a loss function and minimizing it implies estimating $K$ models $\{\widehat{\tau}_k^{(\mathrm{R})}\}_{k=1}^K$ simultaneously such that

$$\{\widehat{\tau}_k^{(\mathrm{R})}\}_{k=1}^K = \mathrm{argmin}_{\{\overline{\tau}_k\}_{k=1}^K \in \mathcal{F}} \frac{1}{n}\sum_{i=1}^n \Big[ \Big(y_i - \widehat{m}(\boldsymbol{x}^{(i)})\Big) - \sum_{k=1}^K \big(\mathbf{1}\{t_i = t^{(k)}\} - \widehat{r}(t^{(k)}, \boldsymbol{x}^{(i)})\big)\overline{\tau}_k(\boldsymbol{x}^{(i)})\Big]^2, \tag{5.29}$$

where $\widehat{m}$ (respectively, $\widehat{r}$) is an estimator of $m$ (respectively, $r$) and $\mathcal{F}$ is the space of candidate models $[\{\overline{\tau}_k\}_{k=1}^K]$.

Still, the major difficulty with our R-learning approach in the multi-treatment regime comes from the fact that Problem (5.29) cannot be written similarly as *weighted* supervised learning problem with a specific pseudo-outcome. Therefore, only parametric families $\mathcal{F}$ can be considered in the multi-treatment regime.

**Proposition 5.2.7.** *Let us assume that $\overline{\tau}_k$ belongs to the family of linear regression models. Then Problem (5.29) admits at least a solution given by the Ordinary Least Squares estimator.*

*Proof.* For $k = 1, \ldots, K$, we assume that $\overline{\tau}_k$ belongs to the family of linear regression models such that:

$$\mathcal{F} = \Big\{ \{\overline{\tau}_k(\boldsymbol{x}) := \beta_{k,0} + \sum_{j=1}^{p-1} \beta_{k,j} f_j(\boldsymbol{x})\}_{k=1}^K \ / \ \boldsymbol{\beta}_k = (\beta_{k,0}, \ldots, \beta_{k,p-1})^\top \in \mathbb{R}^p \Big\}. \tag{5.30}$$

$f_j$ are predefined functions (e.g. polynomial functions). It is also possible to use a matrix notation and write $\overline{\tau}_k(\mathbf{X}) = \mathbf{F}\boldsymbol{\beta}_k$ where $\mathbf{F} = (f_j(\boldsymbol{x}^{(i)})) \in \mathbb{R}^{n \times p}$ assumed to be full rank matrix $rank(\mathbf{F}) = p \leq n$.

Let $\overline{Y} = (\overline{Y}_i)_{i=1}^n$ and $\overline{T}_k = (\overline{T}_{i,k})_{i=1}^n$ such that $\overline{Y}_i = y_i - \widehat{m}(\boldsymbol{x}^{(i)})$ and $\overline{T}_{i,k} = \mathbf{1}\{t_i = t^{(k)}\} - \widehat{r}(t^{(k)}, \boldsymbol{x}^{(i)})$. Let $\epsilon = (\epsilon_i)_{i=1}^n$ denote the vector of errors $\epsilon$ obtained for the generalized Robinson (1988) decomposition in Proposition 5.2.6.

We show immediately that $\mathcal{L}$, the loss function associated with the mean squared error of $\epsilon$, is quadratic with respect to $\boldsymbol{\beta}$. Indeed,

$$\begin{aligned}
\mathcal{L}(\{\overline{\tau}_k\}_{t \neq t^{(0)}}) &= \frac{1}{n}\epsilon^\top \epsilon = \frac{1}{n}\Big(\overline{Y} - \sum_{k=1}^K \overline{T}_k \odot (\mathbf{F}\boldsymbol{\beta}_k)\Big)^\top \Big(\overline{Y} - \sum_{k=1}^K \overline{T}_k \odot (\mathbf{F}\boldsymbol{\beta}_k)\Big) \\
&= \frac{1}{n}\left[\overline{Y}^\top\overline{Y} - 2\sum_{k=1}^K \overline{Y}^\top(\overline{T}_k \odot (\mathbf{F}\boldsymbol{\beta}_k)) + \sum_{k,k'=1}^K (\overline{T}_k \odot (\mathbf{F}\boldsymbol{\beta}_k))^\top \ (\overline{T}_{k'} \odot (\mathbf{F}\boldsymbol{\beta}_{k'}))\right] \\
&= \frac{1}{n}\Big(\overline{Y}^\top\overline{Y} - 2\sum_{k=1}^K \overline{Y}^\top \mathbf{D}_{\overline{T}_k}\mathbf{F}\boldsymbol{\beta}_k + \sum_{k,k'=1}^K \boldsymbol{\beta}_k^\top \mathbf{F}^\top \mathbf{D}_{\overline{T}_k}\mathbf{D}_{\overline{T}_{k'}}\mathbf{F}\boldsymbol{\beta}_{k'}\Big),
\end{aligned} \tag{5.31}$$

where $\odot$ is the Hadamard product (element-wise product). The last line holds because $\overline{T}_k \odot (\mathbf{F}\boldsymbol{\beta}_k) = \mathbf{D}_{\overline{T}_k}\mathbf{F}\boldsymbol{\beta}_k$ with $\mathbf{D}_{\overline{T}_k}$ is the diagonal matrix of the vector $\overline{T}_k = (\overline{T}_{i,k})_{i=1}^n$

By differentiating $\partial\mathcal{L}/\partial\boldsymbol{\beta}_k = 0$ for $k = 1, \ldots, K$ :

$$\begin{cases} -\boldsymbol{a}_1 + \mathbf{B}_1\widehat{\boldsymbol{\beta}}_1 + \sum_{k=2}^{K} \mathbf{C}_{1k}\widehat{\boldsymbol{\beta}}_k = 0 \\ \quad\vdots \qquad\quad \vdots \qquad\quad \vdots \qquad = 0 \\ -\boldsymbol{a}_K + \sum_{k=1}^{K} \mathbf{C}_{Kk}\widehat{\boldsymbol{\beta}}_k + \mathbf{B}_K\widehat{\boldsymbol{\beta}}_K = 0 \end{cases} \tag{5.32}$$

$$\iff \begin{bmatrix} \mathbf{B}_1 & \mathbf{C}_{12} & \cdots & \mathbf{C}_{1K} \\ \mathbf{C}_{21} & \mathbf{B}_2 & \cdots & \mathbf{C}_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{C}_{K1} & \mathbf{C}_{K2} & \cdots & \mathbf{B}_K \end{bmatrix} \begin{bmatrix} \widehat{\boldsymbol{\beta}}_1 \\ \widehat{\boldsymbol{\beta}}_2 \\ \vdots \\ \widehat{\boldsymbol{\beta}}_K \end{bmatrix} = \begin{bmatrix} \boldsymbol{a}_1 \\ \boldsymbol{a}_2 \\ \vdots \\ \boldsymbol{a}_K \end{bmatrix}, \tag{5.33}$$

where

$$\boldsymbol{a}_j = \frac{1}{n}\mathbf{F}^\top \mathbf{D}_{\overline{T}_j}\overline{Y} \in \mathbb{R}^p, \tag{5.34}$$

$$\mathbf{B}_j = \frac{1}{n}\mathbf{F}^\top \mathbf{D}_{\overline{T}_j}^2\ \mathbf{F} \in \mathbb{R}^{p\times p}, \tag{5.35}$$

$$\mathbf{C}_{ij} = \frac{1}{n}\mathbf{F}^\top \mathbf{D}_{\overline{T}_i}\mathbf{D}_{\overline{T}_j}\mathbf{F} \in \mathbb{R}^{p\times p}. \tag{5.36}$$

Let $\boldsymbol{\beta} = \left(\boldsymbol{\beta}_1^\top, \ldots, \boldsymbol{\beta}_K^\top\right)^\top \in \mathbb{R}^{K\times p}$ and consider the block matrix $\mathbf{A}$ defined as.

$$\mathbf{A} = \begin{bmatrix} \mathbf{B}_1 & \mathbf{C}_{12} & \cdots & \mathbf{C}_{1K} \\ \mathbf{C}_{21} & \mathbf{B}_2 & \cdots & \mathbf{C}_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{C}_{K1} & \mathbf{C}_{K2} & \cdots & \mathbf{B}_K \end{bmatrix}. \tag{5.37}$$

The matrix $\mathbf{A}$ is real symmetric and satisfies:

$$\begin{aligned} \boldsymbol{\beta}^\top\mathbf{A}\boldsymbol{\beta} &= \sum_{1\le k,l\le K} \boldsymbol{\beta}_k^\top \mathbf{F}^\top \mathbf{D}_{\overline{T}_k}\mathbf{D}_{\overline{T}_l}\mathbf{F}\boldsymbol{\beta}_l \\ &= \left\| \sum_{k=1}^{K} \mathbf{D}_{\overline{T}_k}\mathbf{F}\boldsymbol{\beta}_k \right\|^2 \ge 0. \end{aligned} \tag{5.38}$$

This result shows that $\mathbf{A}$ is positive semi-definite, all its eigenvalues are nonnegative and also proves the existence of a minimizer $\widehat{\boldsymbol{\beta}}$ to the loss function $\mathcal{L}$.

The optimal solution $\widehat{\boldsymbol{\beta}}$ to Problem (5.29) can be given by

$$\widehat{\boldsymbol{\beta}} = \mathbf{A}^+\boldsymbol{a}, \tag{5.39}$$

where $\mathbf{A}^+$ is the Moore–Penrose inverse of $\mathbf{A}$ and $\boldsymbol{a} = \left(\boldsymbol{a}_1^\top, \ldots, \boldsymbol{a}_K^\top\right)^\top$.

**Remark 5.2.8.** *If $\mathbf{D}_{\overline{T}_k}\boldsymbol{\beta}_k \notin \mathrm{Im}(\mathbf{F})^\perp$ for all $k \in \{1, \ldots, K\}$, then $\sum_{k=1}^{K} \mathbf{D}_{\overline{T}_k}\boldsymbol{\beta}_k \notin \mathrm{Im}(\mathbf{F})^\perp = \mathrm{Ker}(\mathbf{F}^\top)$ which is sufficient to prove that $\mathbf{A}$ is positive definite. In this case, the system in (5.32) admits a unique solution such that*

$$\widehat{\boldsymbol{\beta}} = \begin{bmatrix} \widehat{\boldsymbol{\beta}}_1 \\ \widehat{\boldsymbol{\beta}}_2 \\ \vdots \\ \widehat{\boldsymbol{\beta}}_K \end{bmatrix} = \begin{bmatrix} \mathbf{B}_1 & \mathbf{C}_{12} & \cdots & \mathbf{C}_{1K} \\ \mathbf{C}_{21} & \mathbf{B}_2 & \cdots & \mathbf{C}_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{C}_{K1} & \mathbf{C}_{K2} & \cdots & \mathbf{B}_K \end{bmatrix}^{-1} \begin{bmatrix} \boldsymbol{a}_1 \\ \boldsymbol{a}_2 \\ \vdots \\ \boldsymbol{a}_K \end{bmatrix}. \tag{5.40}$$

■

**Proposition 5.2.9.** *Let us assume that $\overline{\tau}_k$ belongs to the Reproducing Kernel Hilbert Space (RKHS) with a reproducing kernel $\boldsymbol{k}$ and hyperparameters $(\sigma^2, \boldsymbol{\theta})$; then Problem (5.29) admits at least a solution, whose regression coefficients are given by Ordinary Least Squares estimator and optimal hyperparameters are solved numerically.*

*Proof.* In this proof, we introduce the Kernel regression framework as developed by Schölkopf & Smola (2002). This framework is based on considering the Reproducing Kernel Hilbert Spaces (See Subsection 2.2 in Chapter 2).

The Hilbert space $\mathcal{H}$ is defined as a Reproducing Kernel Hilbert Space (RKHS) Berlinet & Thomas-Agnan (2004) with reproducing Kernel $\boldsymbol{k}$ because it verifies, for any $f \in \mathcal{H}$ and $\boldsymbol{x} \in \mathcal{D}$,

$$\langle f, \boldsymbol{k}(\boldsymbol{x}, \cdot) \rangle_{\mathcal{H}} = f(\boldsymbol{x}), \tag{5.41}$$

where $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ is the dot product associated to the Hilbert space $\mathcal{H}$.

It is shown by the Representer theorem (Schölkopf & Smola, 2002) that any minimizer to the empirical risk of the function $f \in \mathcal{H}$ admits a representation of the form

$$\hat{f}(\boldsymbol{x}) = \sum_{i=1}^{n} \alpha_i \boldsymbol{k}(\boldsymbol{x}^{(i)}, \boldsymbol{x}), \tag{5.42}$$

where $\boldsymbol{\alpha} \in \mathbb{R}^n$.

In the following, we consider the Matérn anisotropic geometric kernel $\boldsymbol{k}_{\sigma^2,\boldsymbol{\theta}} = \sigma^2 \boldsymbol{r}_{\boldsymbol{\theta}}$ as defined in (2.26) and we assume that, for $k = 1, \ldots, K$, each $\overline{\tau}_k$ belongs to $\mathcal{H}$, the RKHS with reproducing kernel $\boldsymbol{k}_{\sigma^2,\boldsymbol{\theta}}$ in such way that

$$\mathcal{F} = \left\{ \{ \overline{\tau}_k(\boldsymbol{x}) = \sum_{i=1}^{n} \alpha_{k,i} \boldsymbol{k}_{\sigma^2,\boldsymbol{\theta}}(\boldsymbol{x}^{(i)}, \boldsymbol{x}) \}_{k=1}^{K} \ / \ \boldsymbol{\alpha}_k = (\alpha_{k,1}, \ldots, \alpha_{k,n})^\top \in \mathbb{R}^n \right\}. \tag{5.43}$$

Similarly to linear regression models, it is possible to use a matrix notation $\tau_{t_k}(\mathbf{X}) = \mathbf{K}\boldsymbol{\alpha}_k$ where $\mathbf{K} = (\boldsymbol{k}_{\sigma^2,\boldsymbol{\theta}}(\boldsymbol{x}^{(i)}, \boldsymbol{x}^{(j)}))_{1 \leq i,j \leq n}$ is the Gram matrix of $\boldsymbol{k}_{\sigma^2,\boldsymbol{\theta}}$.

For a fixed hyperparameter $(\sigma^2, \boldsymbol{\theta})$, we prove immediately that the R-learning problem in (5.29) is similar to a linear regression problem. Therefore, by Proposition 5.2.7, the coefficients $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_K)$ satisfy

$$\widehat{\boldsymbol{\alpha}}_{\sigma^2,\boldsymbol{\theta}} = \mathbf{A}^{+}_{\sigma^2,\boldsymbol{\theta}} \boldsymbol{a}_{\sigma^2,\boldsymbol{\theta}}, \tag{5.44}$$

where $\mathbf{A}^{+}_{\sigma^2,\boldsymbol{\theta}}$ is the Moore–Penrose inverse of $\mathbf{A}_{\sigma^2,\boldsymbol{\theta}}$ and $\boldsymbol{a}_{\sigma^2,\boldsymbol{\theta}} = \left( \boldsymbol{a}_1^\top, \ldots, \boldsymbol{a}_K^\top \right)^\top$ such that

$$(\boldsymbol{a}_{\sigma^2,\boldsymbol{\theta}})_j = \frac{1}{n} \mathbf{K}^\top \mathbf{D}_{\overline{T}_j} \overline{Y} \in \mathbb{R}^n, \tag{5.45}$$

$$(\mathbf{B}_{\sigma^2,\boldsymbol{\theta}})_j = \frac{1}{n}\mathbf{K}^\top \mathbf{D}^2_{\overline{T}_j}\ \mathbf{K} \in \mathbb{R}^{n\times n}, \tag{5.46}$$

$$(\mathbf{C}_{\sigma^2,\boldsymbol{\theta}})_{ij} = \frac{1}{n}\mathbf{K}^\top \mathbf{D}_{\overline{T}_i}\mathbf{D}_{\overline{T}_j}\mathbf{K} \in \mathbb{R}^{n\times n}. \tag{5.47}$$

Finally, by considering $\widehat{\tau}_{k,\sigma^2,\boldsymbol{\theta}}(\boldsymbol{x}) = \sum_{i=1}^n (\widehat{\boldsymbol{\alpha}}_{k,\sigma^2,\boldsymbol{\theta}})_i\ \boldsymbol{k}_{\sigma^2,\boldsymbol{\theta}}(\boldsymbol{x}^{(i)},\boldsymbol{x})$, one can obtain the optimal hyperparameters $(\sigma^2,\boldsymbol{\theta})$ by solving the problem:

$$(\widehat{\sigma}^2,\widehat{\boldsymbol{\theta}}) = \underset{(\sigma^2,\boldsymbol{\theta})}{\arg\min}\Big\{\frac{1}{n}\sum_{i=1}^n\Big[\Big(y_i - \widehat{m}(\boldsymbol{x}^{(i)})\Big) - \sum_{k=1}^K (\mathbf{1}\{t_i = t^{(k)}\} - \widehat{r}(t^{(k)},\boldsymbol{x}^{(i)}))\widehat{\tau}_k(\boldsymbol{x}^{(i)})\Big]^2. \tag{5.48}$$

This problem admits an explicit solution for $\widehat{\sigma}^2$ by direct calculations as in the proof of Proposition 5.2.7 such that

$$\widehat{\sigma}^2(\boldsymbol{\theta}) = \frac{\sum_{k=1}^K \overline{Y}^\top(\mathbf{D}_{\overline{T}_k}\mathbf{R}_{\boldsymbol{\theta}})\widehat{\boldsymbol{\alpha}}_{k,\sigma^2,\boldsymbol{\theta}}}{\sum_{k,k'=1}\widehat{\boldsymbol{\alpha}}^\top_{k,\sigma^2,\boldsymbol{\theta}}(\mathbf{R}^\top_{\boldsymbol{\theta}}\mathbf{D}_{\overline{T}_k}\mathbf{D}_{\overline{T}_{k'}}\mathbf{R}_{\boldsymbol{\theta}})\widehat{\boldsymbol{\alpha}}_{k',\sigma^2,\boldsymbol{\theta}}}, \tag{5.49}$$

where $\mathbf{D}_{\overline{T}_k}$ is the diagonal matrix of the vector $\overline{T}_k = (\overline{T}_{i,k})_{i=1}^n$ and $\mathbf{R}_{\boldsymbol{\theta}} = (\boldsymbol{r}_{\boldsymbol{\theta}}(\boldsymbol{x}^{(i)},\boldsymbol{x}^{(j)}))_{1\leq i,j\leq n}$.

The optimal length-scale vector $\widehat{\boldsymbol{\theta}}$ can be obtained numerically by running, for example, a multistart gradient descent algorithm or multistart BFGS method.

$$\widehat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\arg\min}\Big\{\frac{1}{n}\sum_{i=1}^n\Big[\Big(y_i - \widehat{m}(\boldsymbol{x}^{(i)})\Big) - \sum_{k=1}^K (\mathbf{1}\{t_i = t^{(k)}\} - \widehat{r}(t^{(k)},\boldsymbol{x}^{(i)}))\widehat{\tau}_k(\boldsymbol{x}^{(i)})\Big]^2. \tag{5.50}$$

■

We note that the kernel regression method is heavy to solve (cost of $O(n^3K^3)$ at each iteration). Thus, we do not present its results in Section 5.5 and limit ourselves only to R-learners derived from linear regression.

In recent work, Zhang et al. (2022) demonstrate that the generalized R-learner suffers from the non-identifiability of the generalized R-loss function in (5.29). In other words, minimizing the loss function does not uniquely identify CATEs models $(\tau_k)_{k=1}^K$ and leads to poor estimation performance. This statement is observed for continuous treatments but seems to hold for multi-treatments. Zhang et al. (2022) suggest therefore T-identification, based on Tikhonov et al. (1995) regularization, to get over this problem of identification. We did not consider this regularization but our numerical results in B.4 confirm that, in the majority of simulations, the R-learner fails to estimate CATEs $(\tau_k)_{k=1}^K$.

## 5.3 Error estimation of pseudo-outcome meta-learners.

Given their nature, pseudo-outcome meta-learners need to estimate component parameters on the same data $\mathbf{D}_{\text{obs}}$. Unfortunately, some pseudo-outcomes representations may lead to higher variance (i.e. expected squared error) and poor performance in how these components intervene.

In this section, we propose to analyze the (upper bounds) error estimation of each pseudo-outcome. To do so, we will make the assumptions below:

**Assumption 5.3.1.** *We assume that the outcomes $Y(t)$ are generated from a function $f : \mathbb{R} \times \mathbb{R}^d \to \mathbb{R}$ respecting the causal assumptions (4.5.1-4.5.2) such that*

$$Y(t) = f(t, \boldsymbol{X}) + \epsilon, \tag{5.51}$$

*where $\epsilon \sim \mathcal{N}(0, \sigma^2)$ is an additive noise.*

**Assumption 5.3.2.** *We assume the existence of $\boldsymbol{\beta}_t^* \in \mathbb{R}^p$ such that, for all $t \in \mathcal{T}$ and $\boldsymbol{x} \in \mathcal{D}$*

$$f(t, \boldsymbol{x}) = \sum_{j=0}^{p-1} \beta_{t,j}^* f_j(\boldsymbol{x}). \tag{5.52}$$

*where $f_j$ are some predefined basis functions (e.g. polynomial functions $f_j(\boldsymbol{x}) = (\boldsymbol{x}_k^j)_{1 \leq k \leq d}$). We assume in addition that, for all $j \in \{1, \dots, p\}$, $f_j(\boldsymbol{X})$ has all possible finite moments, i.e. $f_j(\boldsymbol{X}) \in L^a$ for $a > 1$.*

**Assumption 5.3.3.** *We assume that the function $f$ is bounded, i.e. there exists $C > 0$ such that*

$$\forall t \in \mathcal{T}, \forall \boldsymbol{x} \in \mathcal{D} : \ |f(t, \boldsymbol{x})| \leq C. \tag{5.53}$$

Under these three assumptions, the CATE $\tau_k$ can be written as:

$$\tau_k(\boldsymbol{x}) = \sum_{j=0}^{p-1} (\beta_{t^{(k)},j}^* - \beta_{t^{(0)},j}^*) f_j(\boldsymbol{x}) = \sum_{j=0}^{p-1} \beta_{k,j}^* f_j(\boldsymbol{x}), \tag{5.54}$$

where $\boldsymbol{\beta}_k^* = (\beta_{k,j}^*)_{j=0}^{p-1} = \boldsymbol{\beta}_{t^{(k)}}^* - \boldsymbol{\beta}_{t^{(0)}}^* \in \mathbb{R}^p$.

When investigating the pseudo-outcomes $Z_k$ that we have considered for the M-, DR- and X-learners, one can see that these pseudo-outcomes have a linear form with respect to $Y_{\text{obs}}$. Therefore, for $k = 1, \dots, K$, we write $Z_k$ as

$$Z_k = A_{t^{(k)}}(T, \boldsymbol{X}) Y_{\text{obs}} + B_{t^{(k)}}(T, \boldsymbol{X}), \tag{5.55}$$

where $A_{t^{(k)}}(T, \boldsymbol{X})$ and $B_{t^{(k)}}(T, \boldsymbol{X})$ are given for each pseudo-outcome meta-learner.

The regression coefficients $\widehat{\boldsymbol{\beta}}_k$ are given by the Ordinary Least Squares (OLS) method

$$\widehat{\boldsymbol{\beta}}_k = (\mathbf{H}^\top \mathbf{H})^{-1} \mathbf{H}^\top \boldsymbol{z}_k, \tag{5.56}$$

where $\boldsymbol{z}_k = (Z_{k,i})_{1 \leq i \leq n}$ and $\mathbf{H} = (\mathbf{H}_{ij}) \in \mathbb{R}^{n \times p}$ is the regression matrix.

**Theorem 5.3.4.** *Under Assumptions (5.3.1-5.3.3), the OLS estimator $\widehat{\boldsymbol{\beta}}_k$ has a bias $\mathbb{B}(\widehat{\boldsymbol{\beta}}_k) = \mathbb{E}(\widehat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k^*)$ that is null if the nuisance parameters are well-specified, and a covariance matrix $\mathbb{V}(\widehat{\boldsymbol{\beta}}_k) = \mathbf{C}/n$, whose terms $\mathbf{C}_{ij}$, for all $\epsilon > 0$, are bounded by:*

$$|\mathbf{C}_{ij}| \leq \begin{cases} \mathcal{E}^M = \mathcal{O}\left(\frac{1}{r_{\min}^{1+\epsilon}}\right) \text{ for the M-learner,} \\ \mathcal{E}^{DR} = \mathcal{O}\left(\frac{\text{err}(\widehat{\mu}_{t^{(k)}}) + \text{err}(\widehat{\mu}_{t^{(0)}})}{r_{\min}^{1+\epsilon}}\right) \text{ for the DR-learner,} \\ \mathcal{E}^X = \mathcal{O}\left(K^2 \sum_{l \neq k} \text{err}(\widehat{\mu}_{t^{(l)}})\right) \text{ for the X-learner,} \end{cases} \tag{5.57}$$

*where $\text{err}(\widehat{\mu}_t) = \mathbb{E}\left[\left(f(t, \boldsymbol{X}) - \widehat{\mu}_t(\boldsymbol{X})\right)^2\right]$ is the expected mean squared error of $\widehat{\mu}_t$.*

*Proof.* In Appendix B.2. ■

Theorem 5.3.4 is valuable because it allows comparing error estimation of pseudo-outcome meta-learners under different scenarios. Following this theorem, it appears that:

**M-learner.** Without surprise, the M-learner has the largest variance, and its error upper bound is constant.

**M- and DR-learners.** As the GPS is present in the denominator of the upper bounds of both M-learners and DR-learners, the variance is likely to be high when there is a lack of overlap in the propensity score, that is, $r_{\min}$ is close to 0. Besides, when the number of treatments $K$ increases, $r_{\min}$ becomes more and more smaller by a mechanical effect of $r_{\min} \leq 1/K$. One can expect consequently that the performances of M- and DR-learners decrease for larger $K$.

**X-learner.** The upper bounds of the X-learner and DR-learner depend on the quality of the estimated potential outcomes models $\widehat{\mu}$. One can expect that the more precise outcome models, the lower the variance.

**M-learner vs DR-learner.** If the potential outcome models are well-specified, the variance's upper bound is expected to be lower for the DR-learner. Controversially, suppose the outcome models are misspecified (but the propensity score is well-specified). In that case, there is no guarantee that the DR-learner would perform better than M-learner, and it may perform even worse, as we will see in some numerical results in Table B.6 in Appendix B.4.

**X-learner vs M-learner.** The X-learner is likely to have low variance if the expected squared error of all outcome models $\widehat{\mu}_\cdot$ is small enough and if some conditions on $K$ and $r_{\min}$ hold. More precisely, the idea is to take both errors' upper bounds and obtain properly conditions under which the X-learner may perform less than the M-learner.

**X-learner vs DR-learner.** It is difficult to anticipate which meta-learner would perform better. This depends mainly on the expected squared error of $\widehat{\mu}$, $K$ and $r_{\min}$, whom, in some cases, make the X-learner have less error than the DR-learner, and the opposite in other cases. Still, numerical results in Appendix B.4 (Tables B.1, B.6, B.8 and B.10) show that the X-learner outperforms the DR-learner when the nuisance components are well-estimated.

### Comparison of meta-learners

We end this subsection by presenting a summary table 5.1 of different meta-learners with their main advantages and drawback:

## 5.4 A semi-synthetic dataset for causal inference: simulating Enhanced Geothermal System with physics-based models

### Motivation

The difficulty in evaluating a causal model's performance in real-world applications motivates the need to create a semi-synthetic dataset. In this subsection, we consider a multistage

Table 5.1: Summary table of multi-treatments meta-learners.

| Meta-learner | Advantages | Disadvantages |
|---|---|---|
| T-learner | ✓ Simple approach | ✗ Selection bias<br>✗ Low sample regime |
| S-learner | ✓ Simple approach | ✗ Confounding effects<br>✗ Regularization bias |
| M-learner | ✓ Consistency | ✗ High variance |
| DR-learner | ✓ Consistency<br>✓ Doubly Robust | ✗ High variance |
| X-learner | ✓ Consistency<br>✓ Low variance | ✗ Complex expression |
| R-learner | ✓ Flexible representation | ✗ Heavy problem<br>✗ Non-identifiability |

fracturing Enhanced Geothermal System (EGS).

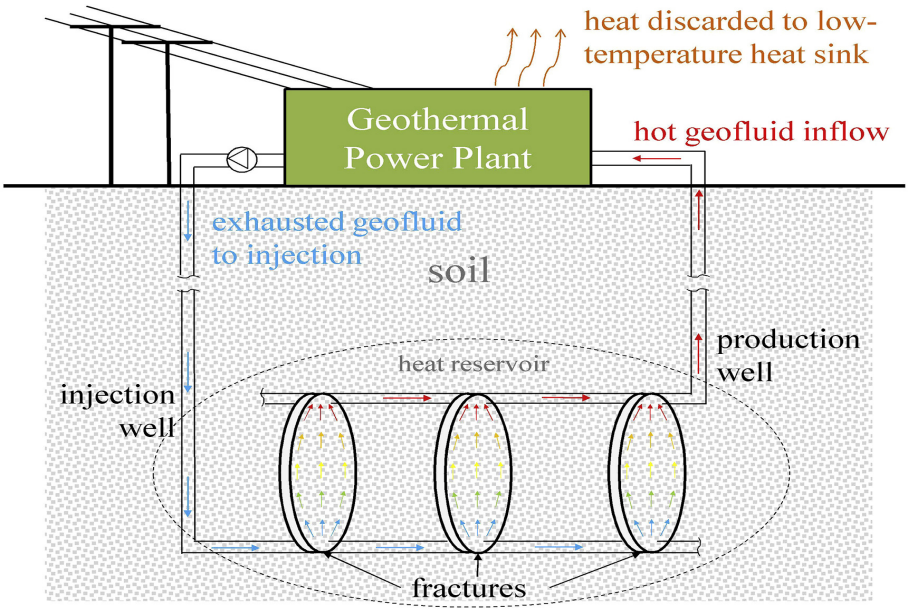



Figure 5.1: Schematic diagram of an EGS system (Li & Lior, 2015).

Enhanced Geothermal Systems (EGS) are geothermal wells that generate geothermal energy by creating fluid connectivity in low permeability conductive rocks through hydraulic, thermal, or chemical stimulation. The EGS concept (See Figure 5.1) involves extracting heat by constructing a subsurface fracture system to which water can be added via injection wells [Geothermal Technologies Office]. Indeed, rocks are permeable due to slight fractures and pore spaces between mineral grains, and the injected water is heated by contact with the rock and returns

to the surface through production wells. Moreover, Enhanced geothermal systems (EGS) have a high potential for developing and supplying renewable energy sources that are more efficient and cheaper than traditional hydrocarbon resources (Bhatia, 2014).

For energy companies, the goal is to optimize the design of the geothermal well (fracture spacing, Lateral Length etc.) to generate the maximum geothermal energy. However, some economic and operational problems present challenges: On the one hand, if the fractures are too small or too few, rocks will not be exploited sufficiently. On the other hand, if the number of fractures in a given rock is too high, the fractures may cool down faster. We would have a costly design that will not maximize the extracted heat.

We assume that the heat extraction performance of the EGS satisfies the following physical model:

$$Q_{well} = Q_{fracture} \times \ell_L / d \times \eta_d, \tag{5.58}$$

where $Q_{well}$ is heat extraction performance delivered by the well (output), $Q_{fracture}$ is the *unknown* heat extraction performance from a single fracture that can be generated using a complex seven-parameter model, including reservoir characteristics and fracture design, $\ell_L$ is the Lateral Length of the well, $d$ is the average spacing between two fractures and $\eta_d$ is the stage efficiency penalizing the individual contribution when fractures are close to each other. We refer to Figure 5.2 for a graphical description of the EGS and its inputs/output.

Finally, the model in (5.58) respects the unconfoundedness assumption (4.5.1), and we can control all its variables in the simulations. We note that, in practice, all inputs are continuous with a given density. However, we discretize these variables in their input space to create a full factorial design.

## Description of the data-set

This section describes the data generating process of our semi-synthetic dataset simulating the heat delivered by a multistage fracturing EGS. The process involved the creation of a conceptual reservoir model and modelling multiple wells' completion scenarios. The output (heat extraction performance) obtained from physics-based simulation experiments was tabulated with inputs in the semi-synthetic dataset.

The input data for the model were fabricated to ease confidentiality and non-disclosure information issues. However, data has been selected from reliable sources such as field observations, journals and books to be within the range of interest. Doing so allowed the building of a plain but representative reservoir model that would provide realistic results of an EGS.

The heat extraction performance from a single fracture ($Q_{fracture}$) is determined using fracture length, fracture height, fracture width, fracture permeability, reservoir porosity, reservoir permeability and pore pressure. Modelling and simulation work were done using preprocessor and reservoir simulation tools PETREL [Schlumberger] and ECLIPSE [Schlumberger].

The four physical parameters of the fracture were investigated, and the list of values used for each parameter can be observed in Table 5.2. In the end, $10 \times 10 \times 2 \times 3 = 600$ fracture's simulation cases have been realized.

To emulate distinct reservoir schemes, it was decided to vary three main parameters; porosity, permeability and pore pressure. For porosity and permeability, the simulator takes the minimum

Table 5.2: Fracture parameters and their range of variation for simulations.

| Variable | Range of variation |
|---|---|
| Fracture length (ft) | $[100, 1000]$ by a step of 100 ft |
| Fracture height (ft) | $[50, 500]$ by a step of 50 ft |
| Fracture width (in) | $\{0.1, 0.2\}$ |
| Fracture Permeability (md) | $\{30000, 85000, 19000\}$ |

and maximum values and estimates the physical properties across the reservoir. Three different multipliers were applied to define three (Low, Base and High) scenarios. Concerning pore pressure, three specific values were defined to simulate under-normal, normal (base) and overpressure (high) gradient conditions. Therefore, $3 \times 3 \times 3 = 27$ possible scenarios were defined. Table 5.3 displays the range of minimum and maximum values for the three reservoir parameters to be varied.

Table 5.3: Reservoir parameters and their range of variation for simulations.

| Variable | Range of variation |
|---|---|
| $(\mathrm{K_{min}}, \mathrm{K_{max}})$ (md) | $\{(0.0054, 0.0157), (0.054, 0.157), (0.109, 0.314)\}$ |
| $(\mathrm{Por_{min}}, \mathrm{Por_{max}})$ (dec) | $\{(0.0054, 0.0157), (0.054, 0.157), (0.109, 0.314)\}$ |
| Pore pressure (psi) | $\{5000, 7000, 9000\}$ |

By combining different reservoir scenarios with single fracture simulations, we obtained a single dataset with 16,200 possible cases for a fracture in a reservoir then we simulated the heat extraction performance for each experiment. Simulation's results were tabulated in the dataset "*Single_Fracture_Simulation_Cases_16200.csv*".

The next step is to define well characteristics (lateral lengths and fracture spacing) to evaluate the heat extraction performance of the well when reservoir and fracture properties are not changed.

Table 5.4: Well parameters and their range of variation.

| Variable | Range of variation |
|---|---|
| Lateral length (ft) | $[2000, 14000]$ by a step of 1000 ft |
| Fracture spacing (ft) | $[100, 500]$ by a step of 100 ft |

Regarding the spacing efficiency coefficient, this coefficient was used to model interactions between fractures and penalize the heat extraction performance of a single fracture in the presence of other close fractures, that is, when the spacing between two fractures is small. Indeed, if the fractures are spaced too close, there may not be enough thermal energy in the rock to heat the water, which decreases the heat extraction efficiency. Modelling this efficiency led to the efficiency table "*Fracture_Efficency.csv*" that describes what would be the well's heat performance behavior with respect to the fracture spacing selected. Based on this table, one

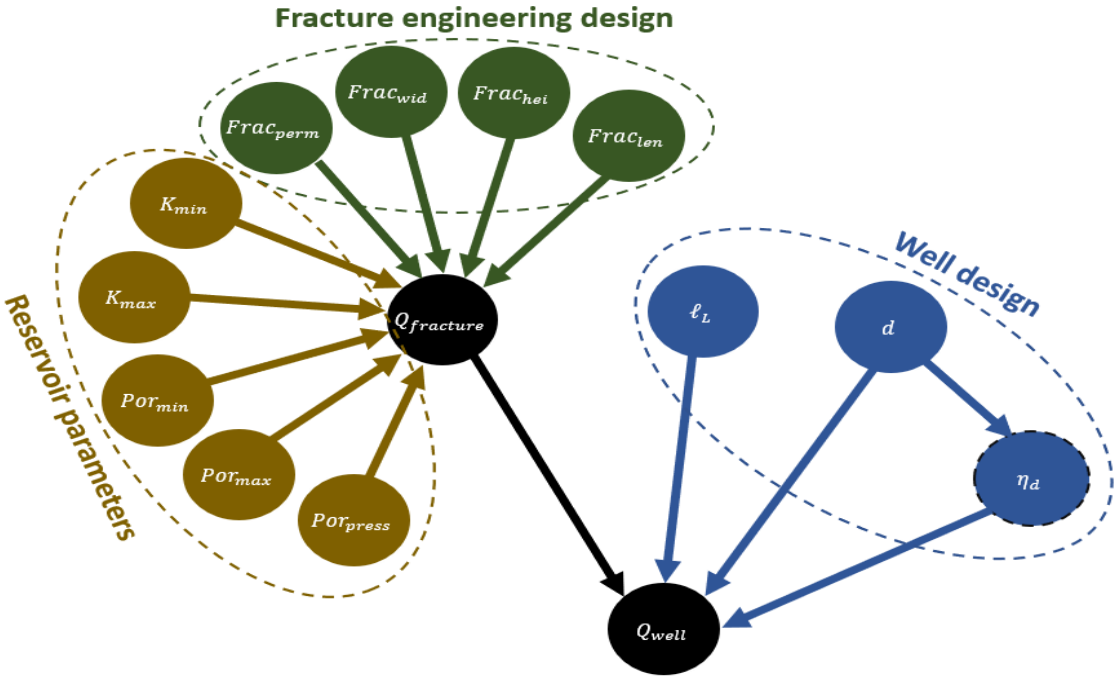



Figure 5.2: The Causal DAG associated with the multistage EGS. Nodes in yellowish brown represent the reservoir characteristics, they can only be simulated, but in reality, we cannot intervene in these variables. Nodes in Dark green represent the fracture design. Engineers control them, and intervening in them is possible whenever there is a need to make a new fracture in the well. Nodes in blue represent a well's design and can be chosen arbitrarily by engineers or statisticians. Nodes in black denote the outputs. $Q_{fracture}$ is only given by the simulator, whereas $Q_{well}$ is given by the physical model in (5.58). Note that this graph contains nine nodes, but both $K_{\min}$ and $K_{\max}$ represent the same physical parameter $K$, and the same remark is valid for $Por_{\min}$ and $Por_{\max}$.

can interpolate the efficiency to draw the curve (see Figure 5.3) and thus obtain the spacing efficiency coefficient for any desired value fracture spacing.

The final generation of the semi-synthetic dataset "*Main_Dataset.csv*" was achieved by combining two main tables created using the R programming language. This table allows calculating the heat performance of a well for any lateral length and fracture spacing between 500 ft and 100 ft with the associated spacing efficiency coefficient defined in the efficiency table, following the physical model in (5.58).

The three datasets are available in the zip file in Supplementary Materials "*Semi-synthetic-EGS.zip*". They will also be shared in the following repository for public use.

Finally, we emphasize that the designed methodology applied for this study focused on generating a semi-synthetic dataset using reservoir numerical simulation and creating a new benchmarking dataset for comparing and validating causal inference methods. Indeed, following the last step

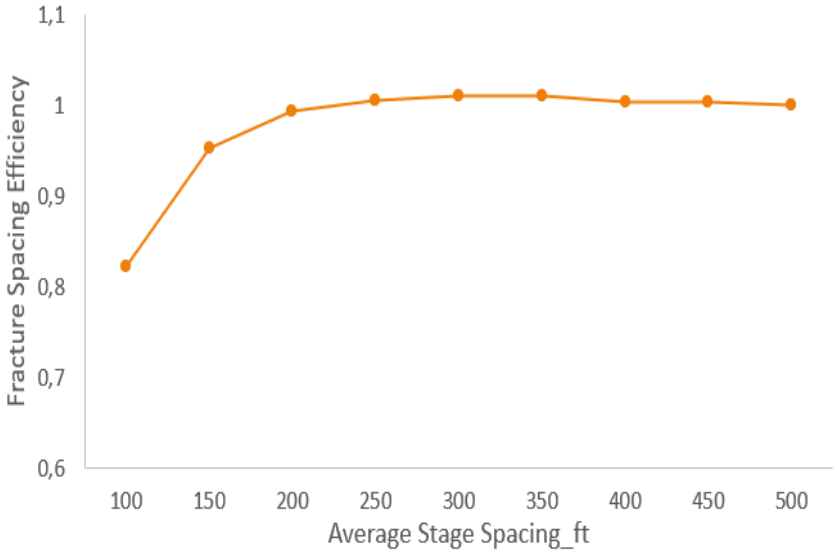



Figure 5.3: Cross plot between fracture spacing efficiency and average stage spacing.

of creating the final dataset "*Main_Dataset.csv*", any user can define different distributions (with different values) on lateral lengths in the range $[2000, 14000]$ and fracture spacing in range $[100, 500]$, pick-up the corresponding spacing efficiency coefficients using the curve drawn in Figure 5.3 and generate a new semi-synthetic dataset by extrapolating them with "*Single_Fracture_Simulation_Cases_16200.csv*" dataset.

### The creation of a non-randomized biased dataset.

The idea of this step was to create a collection of biased data from the main semi-synthetic dataset to emulate observational data found in real-world situations. For example, geothermal wells with larger lateral lengths are likely to have more fractures (expensive wells are located in better geological areas). The opposite is seen for smaller wells that tend to be associated with fewer fractures. This situation creates a discrepancy between what engineers expect from physical models and what they observe in the field data. The biased data, with 9,992 observations, was generated by following the *preferential selection* strategy from the main dataset. Figure 5.4 shows the difference between the *real* heat extraction performance of the EGS and the observed heat extraction performance on the field: low (under-estimated) heat performance for small wells and high (over-estimated) heat performance for large wells.

## 5.5 Experiments and numerical results

We remind that our main goal is to build models able to estimate CATEs as precise as possible for the in-sample counterfactual prediction (i.e. for the same observed covariates $\boldsymbol{X}$ but different treatment level $T$) but also, ideally, for out-sample counterfactual prediction for decision-making

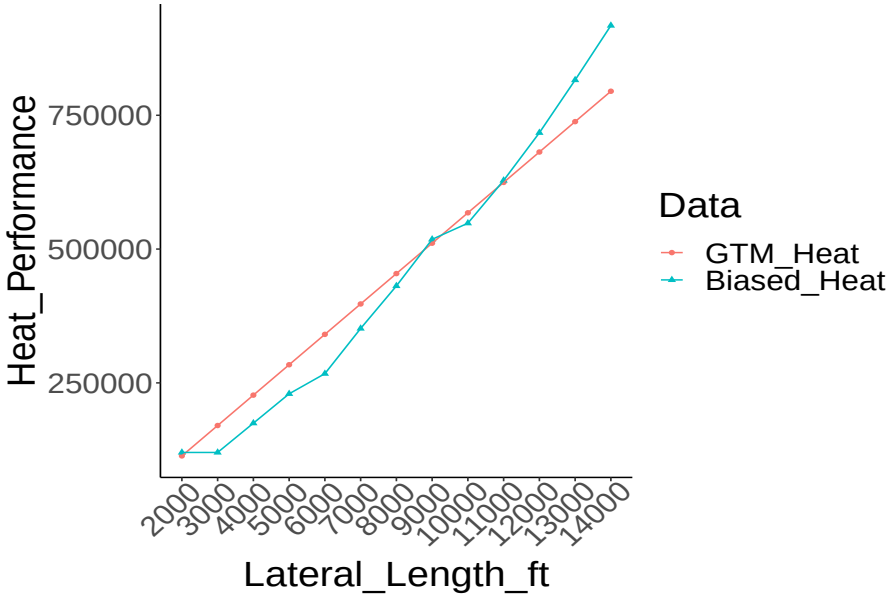



Figure 5.4: An illustration of selection bias on the heat performance. Red line: The heat extraction performance on the main dataset (i.e. Ground Truth Model). Blue line: The heat performance on the biased dataset (i.e. observed response).

purposes. However, as mentioned in Section 5.2, even the task of in-sample prediction is still tricky as realizations of the true CATE $\tau_k$ are not observable. Therefore, training our models on sample $\mathbf{D}_{\text{obs}}$ and predicting on the same sample is quite different from *standard* in-sample prediction and seems somehow as an out-sample prediction if compared to *classical* supervised regression problem.

**Metric.** In the examples where the potential outcome functions and/or CATEs are *a priori* known, the error in estimation is given by **mPEHE**, the mean of the Precision in Estimation of Heterogeneous Effect (PEHE) (Hill, 2011; Shalit et al., 2017) defined as the mean squared error in the estimation of the treatment effect $\widehat{\tau}_k$, over all possible treatment levels $t^{(k)}$ for $k = 1, \ldots, K$:

$$\mathbf{mPEHE} = \frac{1}{K} \sum_{k=1}^{K} PEHE(\widehat{\tau}_k), \tag{5.59}$$

where $PEHE(\widehat{\tau}_k) = \sqrt{\frac{1}{n} \sum_{i=1}^{n} \left(\widehat{\tau}_k(\boldsymbol{x}^{(i)}) - \tau_k(\boldsymbol{x}^{(i)})\right)^2}$ .

This metric will be used to compare and identify conditions (sample size $n$, number of possible treatments $K$, the correctness of nuisance parameters and base-learners) under which we can precisely estimate CATEs. We do not consider here model-fitting of base-learners. More specifically, all hyperparameters (e.g. number of trees, depth etc.) are fixed to their default values during all experiments.

### Synthetic datasets: analytical functions in randomized and non-randomized studies

In this subsection, we begin by empirically evaluating the performances of our meta-learners when the treatment $T$ is taking $K+1=10$ possible values in $[0,1]$ in a RCT setting where the outcome is a linear model and satisfies:

$$Y(t) \mid X \sim \mathcal{N}\big((1+t)X, \sigma^2\big),\ X \sim \mathcal{U}[0,1], \tag{5.60}$$

then, we evaluate meta-learners on the hazard rate outcome:

$$Y(t) \mid \boldsymbol{X} \sim \mathcal{N}\big(t + \|\boldsymbol{X}\| \exp\left(-t\|\boldsymbol{X}\|\right), \sigma^2\big), \tag{5.61}$$

for $\boldsymbol{X} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_5)$ in a non-randomized setting.

Table 5.5: **mPEHE** for XGBoost and RandomForest; linear model (5.60) in RCT setting with $n = 2000$ units.

| Meta-learner | XGBoost | RandomForest |
|---|---|---|
| T-Learner | 0.061 | **0.037** |
| S-Learner | **0.029** | 0.040 |
| M-Learner | 1.23 | 1.15 |
| DR-Learner | 0.063 - 0.063 | 0.060 - 0.060 |
| X-Learner | 0.059 - 0.030 | 0.041 - 0.079 |
| RLin-Learner | 0.122 | 0.112 |

For the DR and X-learners: $\mu_t$ are estimated by T-learning (left value) or S-learning (right value).

Table 5.6: **mPEHE** for XGBoost and RandomForest. Hazard rate model (5.61) in observational setting with $n = 10000$ units.

| Meta-learner | XGboost | RandomForest |
|---|---|---|
| T-Learner | 0.184 | 0.251 |
| *Reg*T-Learner | 0.158 | 0.253 |
| S-Learner | 0.166 | 0.269 |
| M-Learner | 1.56 | 1.55 |
| DR-Learner | 0.151 - 0.171 | 0.275 - 0.288 |
| X-Learner | **0.149** - 0.162 | 0.270 - 0.286 |
| RLin-Learner | 0.235 | **0.178** |

For the DR and X-learners: $\mu_t$ are estimated by T-learning (left value) or S-learning (right value).

To simulate observational data, instead of removing some rows, we propose to create a selection bias in the data by selecting preferentially only observations with specific characteristics (see subsection B.3 in Appendix B). This strategy comes in line with the findings and

recommendations of Curth et al. (2021) about creating a biased sub-sample and evaluating CATE estimators.

The GPS is estimated using gradient boosting models (XGBoost), and the outcome models $\mu_t$ are either estimated by the T-learning or S-learning approaches. In the following tables and Appendix B.4, RLin-learner denotes the R-learner with linear regression models in Proposition 5.2.7 with $p = 2$, the bold font is to indicate the best meta-learner (row) per base-learner (column).

In Tables 5.5 and 5.6, we find that, as expected, the M-learner predicts poorly. The T-learner gives better predictions for Random Forest, whereas the S-learner gives better results for XGBoost. Regularizing T-learner (*Reg*T-Learner) against selection bias (Proposition 5.2.1) increases its performances. The X- and DR-learners improve the predictions of the S-learner for XGBoost, but this improvement is not always observable for Random Forests. Unfortunately, the actual results (and also additional numerical experiments in Appendix B.4) confirm the statement of Zhang et al. (2022): The RLin-learner fails to identify CATEs optimally. Surprisingly, the RLin-learner outperforms when combined with Random Forests for the Hazard rate model.

Despite these satisfying results, we highlight the problem of over-fitted gradient boosting models and Random Forest by comparing them with the linear model in Appendix B.4. This problem should be taken further while estimating CATEs. We think that using out-sample prediction supervised models (e.g. Neural Networks) might solve this problem.

Finally, on the one hand, when $K$ increases, the RLin-learner becomes more effective for CATEs prediction, but the performance of the T-learner becomes compromised, with a slight impact on other learners. Therefore, we recommend the S-learner's estimated potential outcome model when $K \geq 10$ for pseudo-outcome meta-learners. On the other hand, having a large sample size $n$ improves the performances of all meta-learners (except for the M-learner, we do not have any explanation for this behavior). To conclude, two-step meta-learners are robust when applying gradient boosting models as base-learner. In particular, the X-learner improves the quality of one-step meta-learners; when it does not, the differences are very small.

Table 5.7: **mPEHE** for XGBoost and RandomForest. Heat Extraction model (5.58) in observational setting.

| Meta-learner | XGBoost | RandomForest |
|---|---|---|
| T-learner | 0.167 | 0.154 |
| *Reg*T-Learner | 0.153 | **0.153** |
| S-learner | 0.101 | 0.216 |
| M-learner | 1.05 | 0.907 |
| DR-learner | 0.146 - 0.100 | 0.162 - 0.199 |
| X-learner | 0.140 - **0.095** | 0.175 - 0.209 |
| RLin-learner | 0.336 | 0.338 |

For the DR and X-learners: $\mu_t$ are estimated by T-learning (left value) or S-learning (right value).

### Semi-synthetic dataset: estimating heterogeneous treatment effects on the non-randomized biased dataset.

We consider the Lateral Length as treatment $T$ with $K+1 = 13$ possible values and the covariates $\boldsymbol{X} \in \mathbb{R}^{11}$ are the remaining variables. We also consider a logarithmic transformation of the heat performance for a meaningful **mPEHE**, and we normalize the treatment $T$. Following the *preferential selection*, we sample $n = 10000$ units such that wells with high lateral length are likely to have larger fractures and vice versa. The GPS is estimated using gradient boosting models. Table 5.7 resumes the **mPEHE** for different meta-learners. Most findings of subsection 5.5 remain valid: XGBoost model is generally a better choice than Random Forests (except for T-learning); The X-learner, followed by DR-learner, outperforms all other learners.

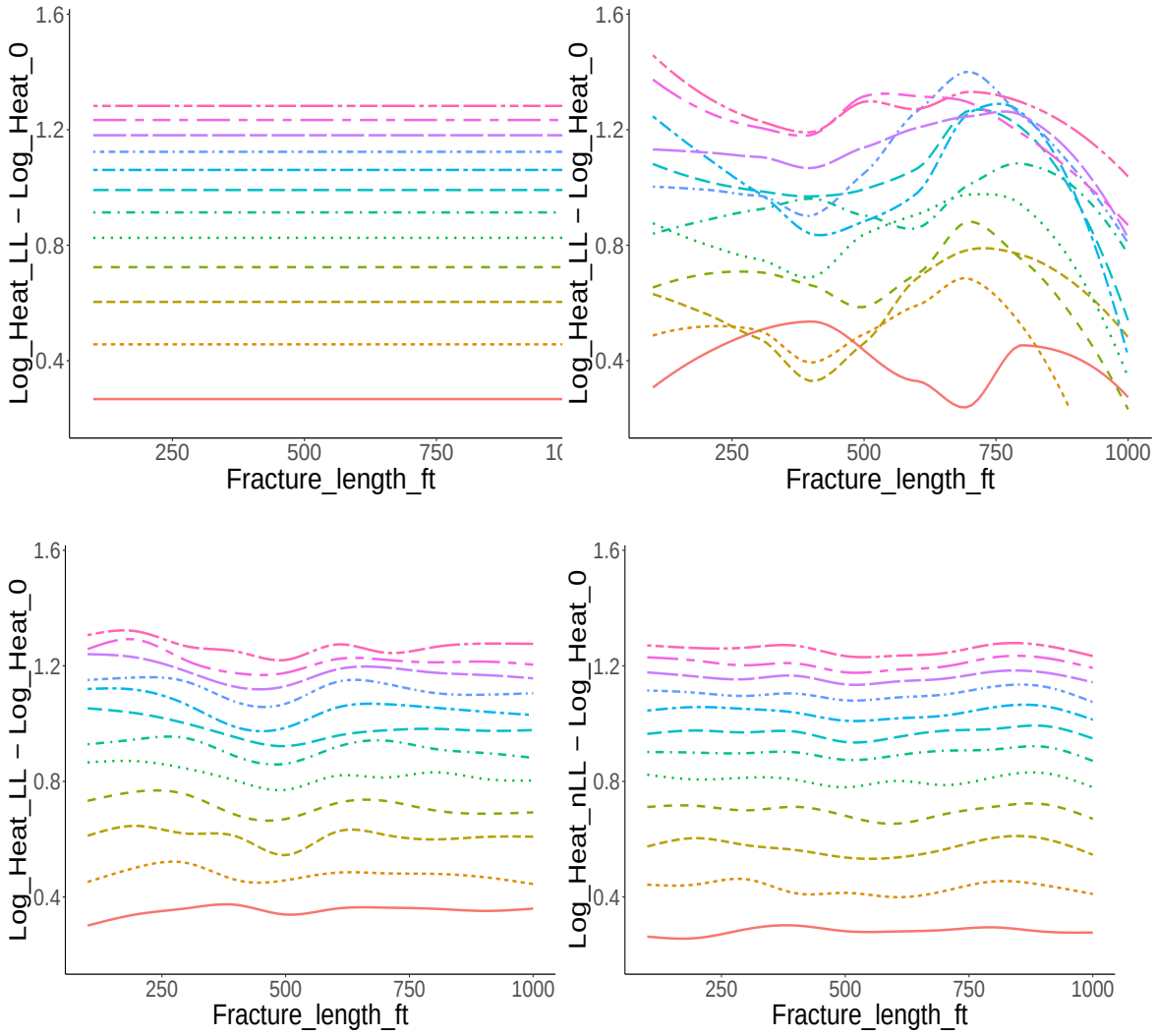



Figure 5.5: CATEs estimation on the semi-synthetic dataset. Each line represents $\tau_k$ for $k = 1, \ldots, K$. a: The ground truth model; b: A biased estimation of CATEs by regressing on Fracture_length_ft; c: T-learner estimation; d: X-learner estimation.

Finally, Figure 5.5 shows the ground truth model, what one would obtain by regressing only

on fracture length (correlation) and T-, X-learner's estimation. It demonstrates the ability of meta-learners, in particular the X-learner, to rebuild the ground truth.

## 5.6 Conclusion

In this chapter, we investigated heterogeneous treatment effects estimation under multi-valued treatment. In addition to standard plug-in meta-learners, we considered representations to build pseudo-outcome meta-learners, and we proposed the generalized Robinson decomposition to build the R-learner. Using the bias-variance analysis, we conducted an in-depth analysis of the error's upper bound of pseudo-outcomes meta-learners. Thanks to this analysis, we were able to address the advantages and limits of each pseudo-outcome meta-learner. In particular, we have identified the impact of $K$ on the X-learner and the lower bound $r_{\min}$ on both M- and DR-learners. Through synthetic and semi-synthetic industrial datasets, we assessed the performances of different meta-learners in a non-randomized case where some covariates are confounded with the treatment. We showed, in particular, the ability of the X-learner to reconstruct the ground truth model. We also highlighted how the choice of base-learner can affect the quality of CATEs estimation. Precisely, it is recommended to choose gradient boosting machines rather than random forests.

# CHAPTER 6

# Heterogeneous treatment effects estimation: Theoretical aspects for continuous treatments

## 6.1 Introduction

Not all causal questions about Causal Inference are binary. Sometimes, answering such questions implies going further and considering a continuous treatment. The challenge now is to estimate the treatment effect (response) for each possible level (dose) of the treatment. This is relevant in many fields (e.g. modelling dose-response in healthcare, evaluating the impact of price increase on demand or return-in investment etc.) because it allows us to identify the optimal intervention policy and personalize it for each unit or subgroup of units.

The state-of-the-art of Causal inference (Section 4.5 with continuous treatments) points out the lack of theoretical and practical guarantees about estimating the dose-response function. In particular, the heterogeneity of the dose-response (treatment effects) is still unpopular in the literature on Causal Inference. The majority of works focus more on learning representations (Harada & Kashima, 2021; Kaddour et al., 2021; Schwab et al., 2020) for graph-structured treatments or on Machine Learning-based models (Hill, 2011). Furthermore, the notion of meta-learners is still (except for the contribution of Zhang et al. (2022) to the R-learner) unknown for continuous treatments.

From a theoretical point of view, these limitations can be justified for many reasons. Indeed, the causal assumptions for continuous treatments are more restrictive. The unconfoundedness assumption, for example, requires the conditional independence of all potential outcomes $(Y(t))_{t\in\mathcal{T}}$ to the treatment $T$ whereas only the conditional independence of $Y(t)$ to the indicator function $\mathbf{1}\{T = t\}$ was required for the multi-treatment setting. Another example is the overlap assumption: assuming the conditional density is uniformly bounded away from zero restricts heavily the set of accepted densities that satisfy this condition (the Gaussian distribution would be excluded, for example, if the support is $\mathbb{R}$). From a practical point of view, the fundamental problem of Causal inference (Holland, 1986) would imply an infinite-dimensional missing data problem. In addition, adjusting selection/confounding bias is extremely difficult for continuous treatments.

In this chapter, we propose to discuss the extension of meta-learners to continuous treatments: The T-learning is meaningless for continuous treatments, and the theoretical properties of the R-learner were already addressed in the paper of Zhang et al. (2022). The focus of the chapter will be more on pseudo-outcome meta-learners (M-, DR- and X-learners), and we aim to answer

the following question: *"Are pseudo-outcome representations worthy for continuous treatment?"*. According to our results, the answer seems to be negative, and our conclusion leans towards the use of a regularized and deconfounded S-learner (Super S-learner) for the estimation of treatment effects under continuous treatments.

In Section 6.2, we recall the framework of heterogeneous treatment effects estimation with continuous treatments and the properties of kernels. In section 6.3, using kernel methods, we propose the extension of pseudo-outcome meta-learners (M-, DR- and X-learners), and we show their consistency. In section 6.4, we conduct a bias-variance analysis of these meta-learners and compare their efficiency with a super S-learner. In Section 6.5, we review the main drawback of generalized R-learner as discussed in the paper of Zhang et al. (2022). Finally, we draw our conclusion in Section 6.6.

## 6.2 Heterogeneous treatment effects estimation under continuous treatments: Set-up

We suppose that we have observed an *i.i.d.* sample of $n$ units $\mathbf{D}_{\text{obs}} = (D_{\text{obs},i})_{i=1}^n = (\boldsymbol{x}^{(i)}, t_i, y_i)_{i=1}^n$ where $\boldsymbol{x}^{(i)}$ denotes a vector of covariates with values in $\mathcal{D}$, $t_i$ denotes the assigned treatment to unit $i$ with possible values in $\mathcal{T}$ and $y_i$ denotes the outcome of the unit $i$. We assume that the treatment assignment variable $T$ is continuous with a support $\mathcal{T} = [t_{\min}, t_{\max}] \subseteq \mathbb{R}$. Following the Neyman (1923) potential outcomes framework and the generalization of the Rubin (1974, 1978, 1979, 1990) Causal Model, we suppose the existence of $Y(t)$, the real-valued counterfactual outcome that would have been observed under a treatment level $t \in \mathcal{T}$. We suppose in addition the causal assumptions (4.5.3-4.5.4) and that $y_i = Y_{\text{obs},i} = Y_i(t_i)$ (consistency assumption). We are interested in the estimation of the Conditional Average Treatment Effect (CATE) between two levels $t$ and $t_0$ in $\mathcal{T}$:

$$\tau_t(\boldsymbol{x}) = \mathbb{E}[Y(t) - Y(t_0) \mid \boldsymbol{X} = \boldsymbol{x}]. \tag{6.1}$$

Following the overlap assumption 4.5.4 and to avoid any possible confusion with the GPS $r$, we refer to the conditional density function by $f_{T|\boldsymbol{X}}$ in the whole chapter.

For direct plug-in meta-learners, as discussed in the previous work, the T-learning approach is unfeasible when the treatment variable is continuous since $(D_{\text{obs},i})_{i \in \mathbf{S}_t}$ $\mathbf{S}_t = \{i, t_i = t\}$ is empty for almost every $t$ and therefore does not contain enough points to estimate the conditional response surfaces (counterfactual predictions surfaces) $\mu_t(\boldsymbol{x}) = \mathbb{E}(Y(t) \mid \boldsymbol{X} = \boldsymbol{x})$. Therefore, only the S-learning approach is considered as a direct plug-in estimator of CATEs.

We remind that the S-learner considers a *single* model built from the whole dataset $\mathbf{D}_{\text{obs}} = (D_{\text{obs},i})_{i=1}^n$ and estimate the CATE in (6.1) as follows:

- Regress $Y_{\text{obs}}$ on the treatment $T$ and the covariates $\boldsymbol{X}$ by a single model $\widehat{\mu}$ using $\mathbf{D}_{\text{obs}}$.
- Estimate the CATE between two treatment levels $t$ and $t_0$ by $\widehat{\tau}_t^{(\text{S})}(\boldsymbol{x}) = \widehat{\mu}(\boldsymbol{x}, t) - \widehat{\mu}(\boldsymbol{x}, t_0)$.

In the following, whenever is mentioned, the conditional response surface will be denoted as $\mu_t(\boldsymbol{x}) = \mathbb{E}(Y_{\text{obs}} \mid T = t, \boldsymbol{X} = \boldsymbol{x})$ and will be estimated using the S-learning approach.

For pseudo-outcome meta-learners, we have seen in Chapter 5 that these estimators incorporate the indicator function $\mathbf{1}\{T = t\}$ whose probability of being equal to one is zero when $T$ is

continuous. Kernel Density Estimation methods seem to be a natural choice and carry over treatment effects estimation for continuous treatments. Indeed, when building the pseudo-outcome vector $\boldsymbol{z}_t = (Z_{t,i})_{i=1}^n$ for each unit $i$, the idea is to replace $\mathbf{1}\{t_i = t\}$ by $K_h(t_i - t)$ where $K_h$ is a weighted kernel with bandwidth $h$ (historically known for approximating smoothly the Dirac delta function) such that:

$$K_h(t_i - t) = \frac{1}{h} K\left(\frac{t_i - t}{h}\right), \tag{6.2}$$

where $K$ is a kernel function.

We consider a kernel function $K$ satisfying the following properties:

- $K$ is non-negative i.e. for all $u \in \mathbb{R} : K(u) \geq 0$.
- $K$ has the density property i.e. $\int_{\mathbb{R}} K(u)\mathrm{d}u = 1$.
- The roughness of the kernel $K$, i.e. $R(K) = \int_{\mathbb{R}} K^2(u)\mathrm{d}u$ exists and is finite.
- The second moment of $K$ i.e. $\kappa_2(K) = \int_{\mathbb{R}} u^2 K(u)\mathrm{d}u$, and the second moment of $K^2$ i.e. $\int_{\mathbb{R}} u^2 K^2(u)\mathrm{d}u)$ are finite.
- $K$ is even, which implies $\kappa_1(K) = \int_{\mathbb{R}} uK(u)\mathrm{d}u = 0$ and $\kappa_1(K^2) = \int_{\mathbb{R}} uK^2(u)\mathrm{d}u = 0$.

## 6.3 Generalization of pseudo-outcome meta-Learners to continuous treatments

In this section, we propose to extend pseudo-outcome meta-learners to the continuous treatment regime. The M- and DR-learners are naturally generalizable to continuous treatments as they use known propensity re-weighting methods. However, the extension X-learner is not trivial for two reasons: Firstly, it requires more reasoning to correct the confounding effect between the treatment $T$ and the covariates $\boldsymbol{X}$. Secondly, unlike binary or multi-treatments scenarios, we cannot easily isolate the level $t$ when $T$ has a density. These two facts would imply significant changes in the expression of the X-learner.

In this section, to guarantee the consistency of all pseudo-outcome meta-learners, the following assumptions on $f_{T|\boldsymbol{X}}$ and $\mu$ are necessary:

**Assumption 6.3.1.** *The conditional density $f_{T|\boldsymbol{X}}$ is continuous and uniformly bounded away from 0 and $+\infty$ i.e. there exists $r_{\min}, r_{\max} > 0$ such that*

$$\forall t \in \mathcal{T}, \forall \boldsymbol{x} \in \mathcal{D} : r_{\min} \leq f_{T|\boldsymbol{X}}(t \mid \boldsymbol{x}) \leq r_{\max}. \tag{6.3}$$

**Assumption 6.3.2.** *The conditional response surface $\mu_t(\boldsymbol{x}) = \mathbb{E}[Y_{\text{obs}} \mid \boldsymbol{X} = \boldsymbol{x}, T = t]$ is continuous on $\mathcal{T} \times \mathcal{D}$.*

A consequence of the assumption 6.3.2 is that, for a fixed $\boldsymbol{x} \in \mathcal{D}$, $\mu_t(\boldsymbol{x})$ is bounded for all $t \in \mathcal{T}$. In other terms, there exists $C_{\boldsymbol{x}} > 0$ such that

$$\forall t \in \mathcal{T} : \ |\mu_t(\boldsymbol{x})| \leq C_{\boldsymbol{x}}. \tag{6.4}$$

### The M-learner in the continuous treatment setting.

Let $t \in \mathcal{T}$ be a treatment level, we define the *modified* pseudo-outcome $Z_t^M$ in multiple treatment regime using the Inverse Propensity Weighting representation as

$$Z_{t,h}^M = \frac{K_h(T-t)}{f_{T|\boldsymbol{X}}(t \mid \boldsymbol{X})} - \frac{K_h(T-t_0)}{f_{T|\boldsymbol{X}}(t_0 \mid \boldsymbol{X})} Y_{\text{obs}}, \tag{6.5}$$

where $f_{T|\boldsymbol{X}}$ is the conditional density, and $K_h$ is the weighted kernel described previously.

**Proposition 6.3.3.** *Under the assumptions (4.5.3)-(4.5.4)*

$$\mathbb{E}(Z_{t,h}^M \mid \boldsymbol{X} = \boldsymbol{x}) \xrightarrow{h \to 0} \tau_t(\boldsymbol{x}). \tag{6.6}$$

*Proof.* We consider $Y_{t,h}^M$ the modified IPW representation of $Y_{\text{obs}}$ in such way that $Z_{t,h}^M = Y_{t,h}^M - Y_{t_0,h}^M$. We have for $\boldsymbol{x} \in \mathcal{D}$:

$$\begin{aligned}
\mathbb{E}(Y_{t,h}^M \mid \boldsymbol{X} = \boldsymbol{x}) &= \mathbb{E}\left[\frac{K_h(T-t)}{f_{T|\boldsymbol{X}}(t \mid \boldsymbol{X})} Y_{\text{obs}} \mid \boldsymbol{X} = \boldsymbol{x}\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\frac{K_h(T-t)}{f_{T|\boldsymbol{X}}(t \mid \boldsymbol{X})} Y_{\text{obs}} \mid \boldsymbol{X}, T\right] \mid \boldsymbol{X} = \boldsymbol{x}\right] \\
&= \mathbb{E}\left[\frac{K_h(T-t)}{f_{T|\boldsymbol{X}}(t \mid \boldsymbol{X})} \mathbb{E}\left[Y_{\text{obs}} \mid \boldsymbol{X}, T\right] \mid \boldsymbol{X} = \boldsymbol{x}\right] \\
&= \int \frac{K_h(s-t)}{f_{T|\boldsymbol{X}}(t \mid \boldsymbol{x})} \mathbb{E}(Y_{\text{obs}} \mid T = s, \boldsymbol{X} = \boldsymbol{x}) f_{T|\boldsymbol{X}}(s \mid \boldsymbol{x}) \, \mathrm{d}s \\
&= \int \frac{K_h(s-t)}{f_{T|\boldsymbol{X}}(t \mid \boldsymbol{x})} \mathbb{E}(Y(s) \mid \boldsymbol{X} = \boldsymbol{x}) f_{T|\boldsymbol{X}}(s \mid \boldsymbol{x}) \, \mathrm{d}s \quad \text{(by Assumption 4.5.3)} \\
&= \int \frac{K_h(s-t)}{f_{T|\boldsymbol{X}}(t \mid \boldsymbol{x})} \mu_s(\boldsymbol{x}) f_{T|\boldsymbol{X}}(s \mid \boldsymbol{x}) \mathbf{1}\{t_{\min} \leq s \leq t_{\max}\} \, \mathrm{d}s \\
&\overset{u=(s-t)/h}{=} \int_{\mathbb{R}} \frac{K(u)}{f_{T|\boldsymbol{X}}(t \mid \boldsymbol{x})} \mu_{t+uh}(\boldsymbol{x}) f_{T|\boldsymbol{X}}(t + uh \mid \boldsymbol{x}) \\
&\qquad \mathbf{1}\{\frac{t_{\min} - t}{h} \leq u \leq \frac{t_{\max} - t}{h}\} \, \mathrm{d}u.
\end{aligned} \tag{6.7}$$

For $u \in \mathbb{R}$ and given the assumptions (6.3.1-6.3.2), we have

$$\Big| \frac{K(u)}{f_{T|\boldsymbol{X}}(t \mid \boldsymbol{x})} \mu_{t+uh}(\boldsymbol{x}) f_{T|\boldsymbol{X}}(t + uh \mid \boldsymbol{x}) \mathbf{1}\{\frac{t_{\min} - t}{h} \leq u \leq \frac{t_{\max} - t}{h}\} \Big| \leq \frac{r_{\max}}{r_{\min}} K(u) C_{\boldsymbol{x}}. \tag{6.8}$$

The function $u \mapsto K(u)$ is integrable by the properties given in Section 6.2. Therefore, by the dominated convergence theorem:

$$\begin{aligned}
\mathbb{E}(Y_{t,h}^M \mid \boldsymbol{X} = \boldsymbol{x}) &\xrightarrow{h \to 0} \int_{\mathbb{R}} \frac{K(u)}{f_{T|\boldsymbol{X}}(t \mid \boldsymbol{x})} \mu_t(\boldsymbol{x}) \, f_{T|\boldsymbol{X}}(t \mid \boldsymbol{x}) \mathrm{d}u. \\
&= \int_{\mathbb{R}} K(u) \mu_t(\boldsymbol{x}) \, \mathrm{d}u = \mu_t(\boldsymbol{x}).
\end{aligned} \tag{6.9}$$

Thus, $\mathbb{E}(Z_{t,h}^M \mid \boldsymbol{X} = \boldsymbol{x}) \xrightarrow{h \to 0} \mu_t(\boldsymbol{x}) - \mu_{t_0}(\boldsymbol{x})$ and we get the desired result. ■

### The DR-learner in the continuous treatment setting.

Similarly to the binary and multi-valued treatment regimes, the DR-learner with continuous treatments is defined using on the Augmented Inverse Propensity Weighting (AIPW) representation (Robins et al., 1994). Let $\widehat{\mu}$ denote an arbitrary estimator of the outcome $\mu$ and let $\widehat{f}_{T|\boldsymbol{X}}$ denote also an arbitrary estimator of the conditional density $f_{T|\boldsymbol{X}}$. We assume that $\widehat{f}_{T|\boldsymbol{X}}$ and $\widehat{\mu}$ respect also Assumptions (6.3.1)-(6.3.2). For $t \in \mathcal{T}$, we define *doubly-robust* pseudo-outcome $Z^{DR}_{\widehat{\mu},\widehat{f}_{T|\boldsymbol{X}},t}$ as

$$Z^{DR}_{\widehat{\mu},\widehat{f}_{T|\boldsymbol{X}},t,h} = \frac{Y_{\text{obs}} - \widehat{\mu}_t(\boldsymbol{X})}{\widehat{f}_{T|\boldsymbol{X}}(t \mid \boldsymbol{X})} K_h(T-t) - \frac{Y_{\text{obs}} - \widehat{\mu}_t(\boldsymbol{X})}{\widehat{f}_{T|\boldsymbol{X}}(t_0 \mid \boldsymbol{X})} K_h(T-t_0) + \widehat{\mu}_t(\boldsymbol{X}) - \widehat{\mu}_{t_0}(\boldsymbol{X}). \tag{6.10}$$

**Proposition 6.3.4.** *Let $Z^{DR}_{\widehat{\mu},\widehat{f}_{T|\boldsymbol{X}},t}$ be the Doubly-Robust pseudo-outcome defined in (6.10), then under the causal assumptions (4.5.3)-(4.5.4) and (6.3.1)-(6.3.2)*

$$\mathbb{E}(Z^{DR}_{\widehat{\mu},\widehat{f}_{T|\boldsymbol{X}},t,h} \mid \boldsymbol{X} = \boldsymbol{x}) \xrightarrow{h \to 0} \tau_t(\boldsymbol{x}), \tag{6.11}$$

*if the outcome models or the propensity model is well-specified, i.e. $\widehat{\mu} = \mu$ almost surely, or $\widehat{f}_{T|\boldsymbol{X}} = f_{T|\boldsymbol{X}}$ almost surely.*

*Proof.* Similarly to the previous proof, we consider $Y^{DR}_{t,h}$ the AIPW representation of $Y_{\text{obs}}$ such that $Z^{DR}_{\widehat{\mu},\widehat{f}_{T|\boldsymbol{X}},t,h} = Y^{DR}_{\widehat{\mu},\widehat{f}_{T|\boldsymbol{X}},t,h} - Y^{DR}_{\widehat{\mu},\widehat{f}_{T|\boldsymbol{X}},t_0,h}$, and we show that

$$\begin{aligned}
\mathbb{E}(Y^{DR}_{\widehat{\mu},\widehat{f}_{T|\boldsymbol{X}},t,h} \mid \boldsymbol{X} = \boldsymbol{x}) &= \mathbb{E}\left[\frac{Y_{\text{obs}} - \widehat{\mu}_t(\boldsymbol{X})}{\widehat{f}_{T|\boldsymbol{X}}(t \mid \boldsymbol{X})} K_h(T-t) + \widehat{\mu}_t(\boldsymbol{X}) \mid \boldsymbol{X} = \boldsymbol{x}\right] \\
&= \widehat{\mu}_t(\boldsymbol{x}) + \mathbb{E}\left[\frac{Y_{\text{obs}} - \widehat{\mu}_t(\boldsymbol{X})}{\widehat{f}_{T|\boldsymbol{X}}(t \mid \boldsymbol{X})} K_h(T-t) \mid \boldsymbol{X} = \boldsymbol{x}\right] \\
&= \widehat{\mu}_t(\boldsymbol{x}) + \mathbb{E}\left[\mathbb{E}\left[\frac{Y_{\text{obs}} - \widehat{\mu}_t(\boldsymbol{X})}{\widehat{f}_{T|\boldsymbol{X}}(t \mid \boldsymbol{X})} K_h(T-t) \mid \boldsymbol{X}, T\right] \mid \boldsymbol{X} = \boldsymbol{x}\right].
\end{aligned} \tag{6.12}$$

- If the propensity model $\widehat{f}_{T|\boldsymbol{X}}$ is correctly specified (i.e. $\widehat{f}_{T|\boldsymbol{X}} = f_{T|\boldsymbol{X}}$ *almost surely*) but

the outcome model is misspecified, we would have

$$\begin{aligned}
\mathbb{E}(Y^{DR}_{\widehat{\mu},\widehat{f}_{T|\boldsymbol{X}},t,h} \mid \boldsymbol{X}=\boldsymbol{x}) &= \widehat{\mu}_t(\boldsymbol{x}) + \mathbb{E}\left[\mathbb{E}\left[\frac{Y_{\text{obs}}-\widehat{\mu}_t(\boldsymbol{X})}{f_{T|\boldsymbol{X}}(t \mid \boldsymbol{X})}K_h(T-t) \mid \boldsymbol{X}, T\right] \mid \boldsymbol{X}=\boldsymbol{x}\right] \\
&= \widehat{\mu}_t(\boldsymbol{x}) + \int \frac{\mathbb{E}(Y_{\text{obs}} \mid T=s, \boldsymbol{X}=\boldsymbol{x}) - \widehat{\mu}_t(\boldsymbol{x})}{f_{T|\boldsymbol{X}}(t \mid \boldsymbol{x})}K_h(s-t)f_{T|\boldsymbol{X}}(s \mid \boldsymbol{x})\,\mathrm{d}s \\
&= \widehat{\mu}_t(\boldsymbol{x}) + \int \frac{\mathbb{E}(Y(s) \mid \boldsymbol{X}=\boldsymbol{x}) - \widehat{\mu}_t(\boldsymbol{x})}{f_{T|\boldsymbol{X}}(t \mid \boldsymbol{x})}K_h(s-t)f_{T|\boldsymbol{X}}(s \mid \boldsymbol{x})\,\mathrm{d}s \\
&\qquad \text{(by Assumption 4.5.3)} \\
&= \widehat{\mu}_t(\boldsymbol{x}) + \int \frac{\mu_s(\boldsymbol{x}) - \widehat{\mu}_t(\boldsymbol{x})}{f_{T|\boldsymbol{X}}(t \mid \boldsymbol{x})}K_h(s-t)f_{T|\boldsymbol{X}}(s \mid \boldsymbol{x})\mathbf{1}\{t_{\min}\le s\le t_{\max}\}\,\mathrm{d}s \\
&\overset{u=(s-t)/h}{=} \widehat{\mu}_t(\boldsymbol{x}) + \int_{\mathbb{R}} \frac{\mu_{t+uh}(\boldsymbol{x}) - \widehat{\mu}_t(\boldsymbol{x})}{f_{T|\boldsymbol{X}}(t \mid \boldsymbol{x})}K(u)f_{T|\boldsymbol{X}}(t+uh \mid \boldsymbol{x}) \\
&\qquad \mathbf{1}\{\frac{t_{\min}-t}{h}\le u\le \frac{t_{\max}-t}{h}\}\,\mathrm{d}u \\
&\overset{h\to 0}{\longrightarrow} \widehat{\mu}_t(\boldsymbol{x}) + \int_{\mathbb{R}} \frac{\mu_t(\boldsymbol{x}) - \widehat{\mu}_t(\boldsymbol{x})}{f_{T|\boldsymbol{X}}(t \mid \boldsymbol{x})}K(u)f_{T|\boldsymbol{X}}(t \mid \boldsymbol{x})\,\mathrm{d}u \\
&\qquad \text{(by the dominated convergence theorem)} \\
&= \widehat{\mu}_t(\boldsymbol{x}) + (\mu_t(\boldsymbol{x}) - \widehat{\mu}_t(\boldsymbol{x}))\int_{\mathbb{R}} K(u)\,\mathrm{d}u \\
&= \widehat{\mu}_t(\boldsymbol{x}) + \mu_t(\boldsymbol{x}) - \widehat{\mu}_t(\boldsymbol{x}) = \mu_t(\boldsymbol{x}).
\end{aligned} \tag{6.13}$$

- If the propensity model is misspecified, but the outcome model is correctly specified (i.e. $\widehat{\mu}=\mu$ *almost surely*), we would have

$$\begin{aligned}
\mathbb{E}(Y^{DR}_{\widehat{\mu},\widehat{f}_{T|\boldsymbol{X}},t,h} \mid \boldsymbol{X}=\boldsymbol{x}) &= \widehat{\mu}_t(\boldsymbol{x}) + \mathbb{E}\left[\mathbb{E}\left[\frac{Y_{\text{obs}}-\mu_t(\boldsymbol{X})}{\widehat{f}_{T|\boldsymbol{X}}(t \mid \boldsymbol{X})}K_h(T-t)|T, \boldsymbol{X}\right] \mid \boldsymbol{X}=\boldsymbol{x}\right] \\
&= \mu_t(\boldsymbol{x}) + \int \frac{\mathbb{E}(Y_{\text{obs}} \mid T=s, \boldsymbol{X}=\boldsymbol{x}) - \mu_t(\boldsymbol{x})}{\widehat{f}_{T|\boldsymbol{X}}(t \mid \boldsymbol{x})}K_h(s-t)f_{T|\boldsymbol{X}}(s \mid \boldsymbol{x})\,\mathrm{d}s \\
&= \mu_t(\boldsymbol{x}) + \int \frac{\mu_s(\boldsymbol{x}) - \mu_t(\boldsymbol{x})}{\widehat{f}_{T|\boldsymbol{X}}(t \mid \boldsymbol{x})}K_h(s-t)f_{T|\boldsymbol{X}}(s \mid \boldsymbol{x})\mathbf{1}\{t_{\min}\le s\le t_{\max}\}\,\mathrm{d}s \\
&\overset{u=(s-t)/h}{=} \mu_t(\boldsymbol{x}) + \int_{\mathbb{R}} \frac{\mu_{t+uh}(\boldsymbol{x}) - \mu_t(\boldsymbol{x})}{\widehat{f}_{T|\boldsymbol{X}}(t \mid \boldsymbol{x})}K(u)f_{T|\boldsymbol{X}}(t+uh \mid \boldsymbol{x}) \\
&\qquad \mathbf{1}\{\frac{t_{\min}-t}{h}\le u\le \frac{t_{\max}-t}{h}\}\,\mathrm{d}u \\
&\overset{h\to 0}{\longrightarrow} \mu_t(\boldsymbol{x}) + \int_{\mathbb{R}} \frac{\mu_t(\boldsymbol{x}) - \mu_t(\boldsymbol{x})}{\widehat{f}_{T|\boldsymbol{X}}(t \mid \boldsymbol{x})}K(u)f_{T|\boldsymbol{X}}(t \mid \boldsymbol{x})\,\mathrm{d}u \\
&\qquad \text{(by the dominated convergence theorem)} \\
&= \mu_t(\boldsymbol{x}).
\end{aligned} \tag{6.14}$$

The result holds similarly for $Y^{DR}_{\widehat{\mu},\widehat{f}_{T|\boldsymbol{X}},t_0}$. Therefore, the consistency of the DR-learner is achieved if either the propensity score or the outcome model is well-specified. ■

It is essential to discuss the well-correctness of the generalized propensity score function. Indeed, since the purpose is to estimate the conditional density $f_{T|\boldsymbol{X}}$ in the M- and DR-learners correctly, the GPS of Imbens (2004) would not be sufficient to achieve the correctness of the conditional density $f_{T|\boldsymbol{X}}$. The P-Function of Imai & Van Dyk (2004) is more adapted for this purpose as it uniquely characterizes the conditional density and ensures the estimation correctness. However, the use of the P-Function implies the extra assumption of Imai & Van Dyk (2004):

**Assumption 6.3.5.** *For almost every $\boldsymbol{x} \in \mathcal{D}$, $(f_{T|\boldsymbol{X}}(t \mid \boldsymbol{x}))_{t\in\mathcal{T}}$ is characterized by $\Theta(\boldsymbol{x})$, where $\boldsymbol{x} \in \mathcal{D} \mapsto \Theta(\boldsymbol{x}) \in \mathbb{R}^q$ is a measurable map.*

Although restrictive, the previous assumption is necessary for the M-learner and the DR-learner if the outcome model is misspecified.

### The X-learner in continuous treatment setting.

In the binary and multi-treatment setting, the X-learner Künzel et al. (2019), also known as *Regression-Adjustment*, consists in a *cross*-procedure of estimation between observations $Y_{\text{obs}}$ and the outcome model when one of the treatments occurs. We remind that the main purpose of the X-learner (in multi-treatment regime) is to learn the CATE $\tau_t$ using all other treatments $t' \neq t$, instead of learning directly as $\tau_t(\boldsymbol{x}) = \mu_t(\boldsymbol{x}) - \mu_{t_0}(\boldsymbol{x})$, which is much easier to do with an S-learner. The extension of the X-learner to continuous treatments should proceed in a similar way: learn the CATE $\tau_t$ using other treatments. The issue of modelling the event that the treatment occurs can be handled by kernel methods. However, as mentioned at the beginning of the section, when $T$ is continuous and has density, we cannot separate and isolate a specific treatment level $t$ from other treatments $t' \neq t \in \mathcal{T}$. Therefore, we propose an adapted *Regression-Adjustment* formula that isolates the treatment $t$ in a local neighbourhood and cross-estimate treatments effects over the treatment support $\mathcal{T}$.

In continuous treatments, for $h > 0$ and for $t \neq t_0 \in \mathcal{T}$, we consider the map $\epsilon : h \mapsto \epsilon(h) > 0$ and the *Regression-Adjustment* pseudo-outcome $Z^X_{t,h}$ such that

$$\begin{aligned} Z^X_{t,h} = 2\epsilon(h)\, K_h(T-t)(Y_{\text{obs}} - \mu_{t_0}(\boldsymbol{X})) + \int_{t_{\min}}^{t-\epsilon(h)} K_h(T-t')(\mu_t(\boldsymbol{X}) - Y_{\text{obs}})\,\mathrm{d}t' + \\ \int_{t+\epsilon(h)}^{t_{\max}} K_h(T-t')(\mu_t(\boldsymbol{X}) - Y_{\text{obs}})\,\mathrm{d}t' + \int_{t_{\min}}^{t-\epsilon(h)} K_h(T-t')(\mu_{t'}(\boldsymbol{X}) - \mu_{t_0}(\boldsymbol{X}))\,\mathrm{d}t' \\ + \int_{t+\epsilon(h)}^{t_{\max}} K_h(T-t')(\mu_{t'}(\boldsymbol{X}) - \mu_{t_0}(\boldsymbol{X}))\,\mathrm{d}t'. \end{aligned} \tag{6.15}$$

**Proposition 6.3.6.** *Under the assumptions (4.5.3)-(4.5.4)*

$$\mathbb{E}(Z^X_{t,h} \mid \boldsymbol{X} = \boldsymbol{x}) \xrightarrow{h\to 0} \tau_t(\boldsymbol{x}). \tag{6.16}$$

*Proof.* By direct calculations, we show that

$$\begin{aligned}\mathbb{E}(Z^X_{t,h} \mid \boldsymbol{X}=\boldsymbol{x}) = 2\epsilon(h)\,\mathbb{E}\left[K_h(T-t)(Y_{\text{obs}}-\mu_{t_0}(\boldsymbol{X})) \mid \boldsymbol{X}=\boldsymbol{x}\right] + \int_{t_{\min}}^{t-\epsilon(h)} \mathbb{E}\big[K_h(T-t')(\mu_t(\boldsymbol{X})-Y_{\text{obs}}) \\ \mid \boldsymbol{X}=\boldsymbol{x}\big]\,\mathrm{d}t' + \int_{t+\epsilon(h)}^{t_{\max}} \mathbb{E}\left[K_h(T-t')(\mu_t(\boldsymbol{X})-Y_{\text{obs}}) \mid \boldsymbol{X}=\boldsymbol{x}\right]\,\mathrm{d}t' + \int_{t_{\min}}^{t-\epsilon(h)} \mathbb{E}\Big[K_h(T-t') \\ (\mu_{t'}(\boldsymbol{X})-\mu_{t_0}(\boldsymbol{X})) \mid \boldsymbol{X}=\boldsymbol{x}\Big]\,\mathrm{d}t' + \int_{t+\epsilon(h)}^{t_{\max}} \mathbb{E}\left[K_h(T-t')(\mu_{t'}(\boldsymbol{X})-\mu_{t_0}(\boldsymbol{X})) \mid \boldsymbol{X}=\boldsymbol{x}\right]\,\mathrm{d}t'. \end{aligned} \tag{6.17}$$

By the dominated convergence theorem, we get

$$\begin{aligned}\mathbb{E}(Z^X_{t,h} \mid \boldsymbol{X}=\boldsymbol{x}) &\xrightarrow{h\to 0} \int_{t_{\min}}^{t} f_{T|\boldsymbol{X}}(t' \mid \boldsymbol{x})\,(\mu_t(\boldsymbol{x})-\mu_{t'}(\boldsymbol{x}))\,\mathrm{d}t' + \int_{t}^{t_{\max}} f_{T|\boldsymbol{X}}(t' \mid \boldsymbol{x})(\mu_t(\boldsymbol{x})-\mu_{t'}(\boldsymbol{x}))\,\mathrm{d}t' \\ &+ \int_{t_{\min}}^{t} f_{T|\boldsymbol{X}}(t' \mid \boldsymbol{x})(\mu_{t'}(\boldsymbol{x})-\mu_{t_0}(\boldsymbol{x}))\,\mathrm{d}t + \int_{t}^{t_{\max}} f_{T|\boldsymbol{X}}(t' \mid \boldsymbol{x})(\mu_{t'}(\boldsymbol{x})-\mu_{t_0}(\boldsymbol{x}))\,\mathrm{d}t'. \\ &= \left(\int_{t_{\min}}^{t} f_{T|\boldsymbol{X}}(t' \mid \boldsymbol{x})\,\mathrm{d}t' + \int_{t}^{t_{\max}} f_{T|\boldsymbol{X}}(t' \mid \boldsymbol{x})\,\mathrm{d}t'\right)(\mu_t(\boldsymbol{x})-\mu_{t_0}(\boldsymbol{x})) \\ &= (\mu_t(\boldsymbol{x})-\mu_{t_0}(\boldsymbol{x})) = \tau_t(\boldsymbol{x}).\end{aligned} \tag{6.18}$$

Therefore, we prove the consistency of the X-learner i.e.

$$\mathbb{E}(Z^X_{t,h} \mid \boldsymbol{X}=\boldsymbol{x}) \xrightarrow{h\to 0} \mu_t(\boldsymbol{x})-\mu_{t_0}(\boldsymbol{x}) = \tau_t(\boldsymbol{x}). \tag{6.19}$$

■

As a conclusion to this section, we can say that the generalization of pseudo-outcome meta-learners is feasible for continuous treatments, and introducing kernel methods allows for obtaining consistent estimators of the CATE.

## 6.4 Bias-Variance analysis of pseudo-outcome meta-learners

In this subsection, we propose to conduct the bias-variance analysis of pseudo-outcome meta-learners. We need to make the following assumptions (some of them are similar to the assumptions of Section 5.3 in Chapter 5) to control the behavior of meta-learners:

**Assumption 6.4.1.** *We assume that the outcomes $Y(t)$ are generated from a uniformly bounded function $f : \mathbb{R} \times \mathbb{R}^d \to \mathbb{R}$ respecting the causal assumptions (4.5.3-4.5.4) such that*

$$Y(t) = f(t, \boldsymbol{X}) + \epsilon, \tag{6.20}$$

*where $\epsilon \sim \mathcal{N}(0, \sigma^2)$ is an additive noise.*

**Assumption 6.4.2.** *For $t \in \mathcal{T}$, we assume the existence of $\boldsymbol{\beta}(t) \in \mathbb{R}^p$ such that, for all $t \in \mathcal{T}$ and $\boldsymbol{x} \in \mathcal{D}$:*

$$f(t, \boldsymbol{x}) = \sum_{j=1}^{p} \beta_j(t) f_j(\boldsymbol{x}) = \boldsymbol{f}(\boldsymbol{x})^\top \boldsymbol{\beta}(t), \tag{6.21}$$

*where $f_j$ are some predefined basis functions (e.g. polynomial functions $f_j(\boldsymbol{x}) = (\boldsymbol{x}_k^{j-1})_{1\le k\le d}$).*

The two previous assumptions, combined, are similar to the assumption made in the paper of Kaddour et al. (2021). The authors assume a product form on the outcome $Y$, that is $Y(t) = \boldsymbol{f}(\boldsymbol{X})^\top \boldsymbol{\beta}(t) + \epsilon$ for arbitrary functions $\boldsymbol{\beta}$ and $\boldsymbol{f}$ in $\mathbb{R}^p$, and $\epsilon$ is a random noise satisfying $\mathbb{E}(\epsilon \mid \boldsymbol{X}, T) = 0$.

The assumption of a product effect is reasonable. Indeed, one can show the universality of this representation in the Reproducing Kernel Hilbert Space (RKHS) (see Proposition 1 of Kaddour et al. (2021)) if we allow the dimension $p$ to grow enough.

Under the previous assumptions, the CATE in (6.1) can be written as

$$\begin{aligned} \tau_t(\boldsymbol{x}) &= f(t, \boldsymbol{x}) - f(t_0, \boldsymbol{x}) \\ &= \boldsymbol{f}(\boldsymbol{x})^\top (\boldsymbol{\beta}(t) - \boldsymbol{\beta}(t_0)) \\ &= \boldsymbol{f}(\boldsymbol{x})^\top \boldsymbol{\beta}^*(t), \end{aligned} \tag{6.22}$$

where $\boldsymbol{\beta}^*(t) = \boldsymbol{\beta}(t) - \boldsymbol{\beta}(t_0)$. In other words, the CATE can be learned by estimating $\boldsymbol{\beta}^*(t)$ or, equivalently, both $\boldsymbol{\beta}(t)$ and $\boldsymbol{\beta}(t_0)$. We call a *super S-learner*, a model able to learn "somehow" $\tau_t$ efficiently, that is, the super S-learner is unbiased $\mathbb{E}(\widehat{\boldsymbol{\beta}}_h^{SS}(t)) = \boldsymbol{\beta}^*(t)$ and has a minimal variance $\mathbb{V}(\widehat{\boldsymbol{\beta}}_h^{SS}(t)) = \mathbb{E}\big[(\boldsymbol{f}(\boldsymbol{x})^\top \boldsymbol{\beta}^*(t) - \boldsymbol{f}(\boldsymbol{x})^\top \widehat{\boldsymbol{\beta}}_h^{SS}(t))^2\big]$.

Finally, the following assumption is also necessary for the bias-variance analysis:

**Assumption 6.4.3.** *The outcome function $f$ and the conditional treatment density $f_{T|\boldsymbol{X}}$ are twice differentiable with respect to $t$. For technical reasons, we also assume that the third derivative exists and is uniformly bounded in $(t, \boldsymbol{x})$.*

**Assumption 6.4.4.** *Let $\widehat{\mu}$ and $\widehat{f}_{T|\boldsymbol{X}}$ denote arbitrary estimators of the outcome function $f$ and the conditional treatment density $f_{T|\boldsymbol{X}}$. We assume that $\widehat{\mu}$ and $\widehat{f}_{T|\boldsymbol{X}}$ have the same properties as $f$ and $f_{T|\boldsymbol{X}}$ (i.e. continuity, boundedness and differentiability).*

### Bias-Variance tradeoff of the M-Learner

For $t \in \mathcal{T}$, we consider the IPW pseudo-outcome with an arbitrary model (estimator) $\widehat{f}_{T|\boldsymbol{X}}$ of the conditional density $f_{T|\boldsymbol{X}}$:

$$Y_{t,h,i}^M = \frac{K_h(t_i - t)}{\widehat{f}_{T|\boldsymbol{X}}(t \mid \boldsymbol{x}^{(i)})} y_i, \quad i = 1, \dots, n, \tag{6.23}$$

and we denote in the following $\boldsymbol{y}_{t,h}^M = (Y_{t,h,i}^M)_{1 \le i \le n}$.

**Remark 6.4.5.** *We assume that the conditional density estimator $\widehat{f}_{T|\boldsymbol{X}}$ is estimated separately using a different unlabeled large sample $\mathbf{D}' = (t_i', \boldsymbol{x}'^{(i)})_{i=1}^{n'}$.*

The regression coefficient $\widehat{\boldsymbol{\beta}}_h^M(t)$ are given by the Ordinary Least Squares (OLS) method

$$\widehat{\boldsymbol{\beta}}_h^M(t) = (\mathbf{H}^\top \mathbf{H})^{-1} \mathbf{H}^\top \boldsymbol{y}_{t,h}^M, \tag{6.24}$$

where $\mathbf{H} = (\mathbf{H}_{ij}) = (f_j(\boldsymbol{x}^{(i)})) \in \mathbb{R}^{n \times p}$ is the regression matrix.

Some of the following calculations are similar to what has been done in Appendix B.2.

$$\begin{aligned}\widehat{\boldsymbol{\beta}}_h^M(t) &= (\mathbf{H}^\top\mathbf{H})^{-1}\mathbf{H}^\top \boldsymbol{y}_{t,h}^M = (\mathbf{H}^\top\mathbf{H})^{-1}\mathbf{H}^\top\Big(\frac{K_h(t_i-t)}{\widehat{f}_{T|\boldsymbol{X}}(t \mid \boldsymbol{x}^{(i)})}y_i\Big)_{i=1}^n \\ &= (\mathbf{H}^\top\mathbf{H})^{-1}\mathbf{H}^\top\Big(\frac{K_h(t_i-t)}{\widehat{f}_{T|\boldsymbol{X}}(t \mid \boldsymbol{x}^{(i)})}f(t+hu_i,\boldsymbol{x}^{(i)}) + \frac{K_h(t_i-t)}{\widehat{f}_{T|\boldsymbol{X}}(t \mid \boldsymbol{x}^{(i)})}\epsilon_i\Big)_{i=1}^n,\end{aligned} \tag{6.25}$$

where $u_i = (t_i - t)/h$ for $i = 1, \dots, n$.

With the differentiability assumption 6.4.3, we consider a second order Taylor expansion of $f$ and $f_{T|\boldsymbol{X}}$:

$$\begin{aligned}f(t+uh,\boldsymbol{x}) &= f(t,\boldsymbol{x}) + hu\frac{\partial f}{\partial t}(t,\boldsymbol{x}) + \frac{(hu)^2}{2}\frac{\partial^2 f}{\partial t^2}(t,\boldsymbol{x}) + (hu)^2\varepsilon_{1,\boldsymbol{x}}(uh), \\ f_{T|\boldsymbol{X}}(t+hu \mid \boldsymbol{x}) &= f_{T|\boldsymbol{X}}(t \mid \boldsymbol{x}) + hu\frac{\partial f_{T|\boldsymbol{X}}}{\partial t}(t \mid \boldsymbol{x}) + \frac{(hu)^2}{2}\frac{\partial^2 f_{T|\boldsymbol{X}}}{\partial t^2}(t \mid \boldsymbol{x}) + (hu)^2\varepsilon_{2,\boldsymbol{x}}(uh),\end{aligned} \tag{6.26}$$

where $\varepsilon_{j,\boldsymbol{x}}(t)$ are functions that are continuous in $t$, bounded uniformly in $(t, \boldsymbol{x})$, and such that $\varepsilon_{j,\boldsymbol{x}}(t) \to 0$ as $t \to 0$. In the following, $\tilde{\varepsilon}_{\boldsymbol{x}}, \tilde{\varepsilon}'_{\boldsymbol{x}}, \tilde{\varepsilon}^{(2)}_{\boldsymbol{x}}$ etc. refer to functions with similar properties.

**Lemma 6.4.6.** *Let $K$ be a kernel with the defined properties in Section 6.2. Let $\varepsilon : \mathbb{R} \to \mathbb{R}$ be a bounded function such that $\varepsilon(x) \to 0$ when $x \to 0$. The integrals $\int_{\mathbb{R}} K(u)\varepsilon(uh)\,\mathrm{d}u$, $\int_{\mathbb{R}} uK^2(u)\varepsilon(uh)\,\mathrm{d}u$ and $\int_{\mathbb{R}} u^2K(u)\varepsilon(uh)\,\mathrm{d}u$ converge to 0 as $h \to 0$.*

*Proof.* By Assumption 6.4.3, there exists $\varepsilon_\infty$ such that $\varepsilon(hu) \le \varepsilon_\infty$ for all $h, u > 0$.

For $n \in \{0, 1, 2\}$ and for all $h, u > 0$:

$$\big|u^n k(u)\varepsilon(hu)\big| \le |u|^n K(u)\varepsilon_\infty, \tag{6.27}$$

where $u \mapsto |u|^n K(u)\varepsilon_\infty$ is integrable by the properties given in Section 6.2. The result also holds for $K^2$.

Therefore, since $\varepsilon(hu) \xrightarrow{h\to 0} 0$ and by the dominated convergence theorem, we get the desired proof of the lemma. ■

Given these expansions, we can write:

$$\begin{aligned}\widehat{\boldsymbol{\beta}}_h^M(t) &= (\mathbf{H}^\top\mathbf{H})^{-1}\mathbf{H}^\top\Big(\frac{K_h(t_i-t)}{\widehat{f}_{T|\boldsymbol{X}}(t \mid \boldsymbol{x}^{(i)})}\big(f(t,\boldsymbol{x}^{(i)}) + hu_i\frac{\partial f}{\partial t}(t,\boldsymbol{x}^{(i)}) + \frac{(hu_i)^2}{2}\frac{\partial^2 f}{\partial t^2}(t,\boldsymbol{x}^{(i)}) + (hu_i)^2\varepsilon_{1,i}(hu_i)\big) \\ &\qquad + \frac{K_h(t_i-t)}{\widehat{f}_{T|\boldsymbol{X}}(t \mid \boldsymbol{x}^{(i)})}\epsilon_i\Big)_{i=1}^n \\ &= (\mathbf{H}^\top\mathbf{H})^{-1}\mathbf{H}^\top\big(f(t,\boldsymbol{x}^{(i)})\big)_{i=1}^n + (\mathbf{H}^\top\mathbf{H})^{-1}\mathbf{H}^\top\Big[\Big(\Big(\frac{K_h(t_i-t)}{\widehat{f}_{T|\boldsymbol{X}}(t \mid \boldsymbol{x}^{(i)})} - 1\Big)f(t,\boldsymbol{x}^{(i)})\Big)_{i=1}^n + \Big(\frac{K_h(t_i-t)}{\widehat{f}_{T|\boldsymbol{X}}(t \mid \boldsymbol{x}^{(i)})} \\ &\qquad \times \Big(hu_i\frac{\partial f}{\partial t}(t,\boldsymbol{x}^{(i)}) + \frac{h^2(u_i)^2}{2}\frac{\partial^2 f}{\partial t^2}(t,\boldsymbol{x}^{(i)})\Big)\Big)_{i=1}^n + \Big(\frac{K_h(t_i-t)}{\widehat{f}_{T|\boldsymbol{X}}(t \mid \boldsymbol{x}^{(i)})}\epsilon_i\Big)_{i=1}^n + \Big((hu_i)^2\varepsilon_{1,i}(hu_i)\Big)_{i=1}^n\Big] \\ &= \boldsymbol{\beta}(t) + (\mathbf{H}^\top\mathbf{H})^{-1}\mathbf{H}^\top(\boldsymbol{b}_{t,spec} + \boldsymbol{b}_{t,K,h} + \boldsymbol{b}_{t,\epsilon}) + h^2\boldsymbol{u}^2 \odot \varepsilon_{1,\mathbf{X}}(h\boldsymbol{u}),\end{aligned} \tag{6.28}$$

where

$$\boldsymbol{\beta}(t) = (\mathbf{H}^\top \mathbf{H})^{-1} \mathbf{H}^\top (f(t, \boldsymbol{x}^{(i)}))_{i=1}^n \tag{6.29}$$

is the true regression coefficients, and

$$\boldsymbol{b}_{t,spec}^M = \Big( \Big( \frac{K_h(t_i - t)}{\widehat{f}_{T|\boldsymbol{X}}(t \mid \boldsymbol{x}^{(i)})} - 1 \Big) f(t, \boldsymbol{x}^{(i)}) \Big)_{i=1}^n \tag{6.30}$$

is the bias term related to the misspecification of the conditional density estimator, and,

$$\boldsymbol{b}_{t,K,h} = \Big( \frac{K_h(t_i - t)}{\widehat{f}_{T|\boldsymbol{X}}(t \mid \boldsymbol{x}^{(i)})} \Big( h u_i \frac{\partial f}{\partial t}(t, \boldsymbol{x}^{(i)}) + \frac{h^2 u_i^2}{2} \frac{\partial^2 f}{\partial t^2}(t, \boldsymbol{x}^{(i)}) \Big) \Big)_{i=1}^n \tag{6.31}$$

is the bias term related to kernel methods estimation with bandwidth $h$, and,

$$\boldsymbol{b}_{t,\epsilon} = \Big( \frac{K_h(t_i - t)}{\widehat{f}_{T|\boldsymbol{X}}(t \mid \boldsymbol{x}^{(i)})} \epsilon_i \Big)_{i=1}^n \tag{6.32}$$

is an unbiased term due to the measurement errors, and,

$$\boldsymbol{u}^2 \odot \varepsilon_{1,\mathbf{X}}(h\boldsymbol{u}) = \Big( u_i^2 \varepsilon_{1,i}(h u_i) \Big)_{i=1}^n \tag{6.33}$$

with $\mathbb{E}\big[U^2 \varepsilon_{1,\boldsymbol{X}}(hU)\big] \to 0$ when $h \to 0$ by the dominated convergence theorem.

In the following, we denote $B_{t,spec}^M$ (respectively, $B_{t,K,h}$ and $B_{t,\epsilon}$) the random variable whose realizations correspond to $\boldsymbol{b}_{t,spec}^M$ (respectively, $\boldsymbol{b}_{t,K,h}$ and $\boldsymbol{b}_{t,\epsilon}$).

Let us consider the vector $\boldsymbol{Z}_t^{(n)}$

$$\begin{aligned} \boldsymbol{Z}_t^{(n)} = \Big( \frac{1}{n} \mathbf{H}^\top (\boldsymbol{b}_{t,spec} + \boldsymbol{b}_{t,K,h} + \boldsymbol{b}_{t,\epsilon}) \big)_1, \ldots, \frac{1}{n} \big( \mathbf{H}^\top (\boldsymbol{b}_{t,spec} + \boldsymbol{b}_{t,K,h} + \boldsymbol{b}_{t,\epsilon}) \big)_p, \\ \frac{1}{n} (\mathbf{H}^\top \mathbf{H})_{11}, \ldots, \frac{1}{n} (\mathbf{H}^\top \mathbf{H})_{pp} \Big)^\top \in \mathbb{R}^{p+p^2}, \end{aligned} \tag{6.34}$$

that allows us to write $\widehat{\boldsymbol{\beta}}_h^M(t) = \boldsymbol{\beta}(t) + \phi(\boldsymbol{Z}^{(n)}) + o(h^2)$ where $\phi : \mathbb{R}^{p+p^2} \longrightarrow \mathbb{R}^p$ is a $\mathcal{C}^1$-function.

The vector $\boldsymbol{Z}_t^{(n)}$ has mean $\boldsymbol{m}(h)$ such that:

$$\boldsymbol{m}(h) = \Big( h_{t,1}, \ldots, h_{t,p}, F_{11}, \ldots, F_{pp} \Big)^\top, \tag{6.35}$$

where, for $j = 1, \ldots, p$.

$$\begin{aligned} h_{t,j} &= \mathbb{E}\big[ f_j(\boldsymbol{X}) (B_{t,spec}^M + B_{t,K,h} + B_{t,\epsilon}) \big] \\ &= h_{t,spec,j}^M + h_{t,K,j} + h_{t,\epsilon,j}, \end{aligned} \tag{6.36}$$

and,

$$F_{jj'} = \mathbb{E}(f_j(\boldsymbol{X}) f_{j'}(\boldsymbol{X})). \tag{6.37}$$

In some cases, the polynomials $f_j$ are chosen to be orthonormal with respect to the distribution of $\boldsymbol{X}$ (e.g. Polynomials Chaos). A consequence of this choice would imply that $\mathbf{F}$ is the identity matrix.

**Lemma 6.4.7.** *For $j = 1, \dots, p$*

$$h^M_{t,j} = h^M_{t,spec,j} + h^2 \kappa_2(K) h^M_{t,Kern,j} + o(h^2), \tag{6.38}$$

*where*

$$h^M_{t,spec,j} = \mathbb{E}\big[f_j(\boldsymbol{X})\Big(\frac{f_{T|\boldsymbol{X}}(t \mid \boldsymbol{X})}{\widehat{f}_{T|\boldsymbol{X}}(t \mid \boldsymbol{X})} - 1\Big) f(t, \boldsymbol{X})\big] \tag{6.39}$$

*is the misspecification bias such that $h^M_{t,spec,j} = 0$ if the conditional density estimator is well-specified, and,*

$$h^M_{t,Kern,j} = \mathbb{E}\Big[\frac{f_j(\boldsymbol{X})}{\widehat{f}_{T|\boldsymbol{X}}(t \mid \boldsymbol{X})} C'_t(\boldsymbol{X})\Big], \tag{6.40}$$

*with*

$$C'_t(\boldsymbol{X}) = \frac{\partial f}{\partial t}(t, \boldsymbol{X}) \frac{\partial f_{T|\boldsymbol{X}}}{\partial t}(t \mid \boldsymbol{X}) + \frac{1}{2} f_{T|\boldsymbol{X}}(t \mid \boldsymbol{X}) \frac{\partial^2 f}{\partial t^2}(t, \boldsymbol{X}) + \frac{1}{2} \frac{\partial^2 f_{T|\boldsymbol{X}}}{\partial t^2}(t \mid \boldsymbol{X}), \tag{6.41}$$

*is the bias induced by the use of kernel methods.*

*Proof.* In this proof, we compute the terms $h^M_{t,spec,j}$, $h_{t,K,j}$ and $h_{t,\epsilon,j}$ separately.

For the specification term $h^M_{t,spec,j}$, by the continuity of $\widehat{f}_{T|\boldsymbol{X}}$ and the dominated convergence theorem, we have:

$$\begin{aligned} h^M_{t,spec,j} &= \mathbb{E}\Big[f_j(\boldsymbol{X})\Big(\frac{K_h(T-t)}{\widehat{f}_{T|\boldsymbol{X}}(t \mid \boldsymbol{X})} - 1\Big) f(t, \boldsymbol{X})\big] \\ &\overset{h \approx 0}{=} \mathbb{E}\big[\frac{f_j(\boldsymbol{X})}{\widehat{f}_{T|\boldsymbol{X}}(t \mid \boldsymbol{X})}\Big( \int_{\mathbb{R}} K(u)\big(f_{T|\boldsymbol{X}}(t \mid \boldsymbol{X}) + hu \frac{\partial f_{T|\boldsymbol{X}}}{\partial t}(t \mid \boldsymbol{X}) + \frac{(hu)^2}{2} \frac{\partial^2 f_{T|\boldsymbol{X}}}{\partial t^2}(t \mid \boldsymbol{X}) \\ &\qquad + (hu)^2 \varepsilon_{2,\boldsymbol{X}}(uh)\big)\, \mathrm{d}u - 1\Big) f(t, \boldsymbol{X})\Big]. \end{aligned} \tag{6.42}$$

The first-order moment integral vanishes by the symmetry property of the kernel $K$. Moreover, by Lemma 6.4.6, we have

$$\begin{aligned} h^M_{t,spec,j} &= \mathbb{E}\Big[f_j(\boldsymbol{X})\Big(\frac{f_{T|\boldsymbol{X}}(t \mid \boldsymbol{X}) + h^2 \frac{\partial^2 f_{T|\boldsymbol{X}}}{\partial t^2}(t \mid \boldsymbol{X})}{\widehat{f}_{T|\boldsymbol{X}}(t \mid \boldsymbol{X})} - 1\Big) f(t, \boldsymbol{X})\big] + o(h^2) \\ &= \mathbb{E}\Big[f_j(\boldsymbol{X})\Big(\frac{f_{T|\boldsymbol{X}}(t \mid \boldsymbol{X})}{\widehat{f}_{T|\boldsymbol{X}}(t \mid \boldsymbol{X})} - 1\Big) f(t, \boldsymbol{X})\Big] + h^2 \frac{\kappa_2(K)}{2} \mathbb{E}\Big[\frac{f_j(\boldsymbol{X})}{\widehat{f}_{T|\boldsymbol{X}}(t \mid \boldsymbol{X})} \frac{\partial^2 f_{T|\boldsymbol{X}}}{\partial t^2}(t \mid \boldsymbol{X})\Big] + o(h^2). \end{aligned} \tag{6.43}$$

For the Kernel estimations term, we have

$$
\begin{aligned}
h_{t,K,h,j} &= \mathbb{E}\Big[f_j(\boldsymbol{X})\frac{K_h(T-t)}{\widehat{f}_{T|\boldsymbol{X}}(t \mid \boldsymbol{X})}\Big(h(T-t)/h\frac{\partial f}{\partial t}(t,\boldsymbol{X}) + \frac{h^2((T-t)/h)^2}{2}\frac{\partial^2 f}{\partial t^2}(t,\boldsymbol{X})\Big)\Big] \\
&\overset{h\approx 0}{=} \mathbb{E}\Big[\frac{f_j(\boldsymbol{X})}{\widehat{f}_{T|\boldsymbol{X}}(t \mid \boldsymbol{X})}\Big(\int_{\mathbb{R}} K(u)\Big(hu\frac{\partial f}{\partial t}(t,\boldsymbol{X}) + \frac{(hu)^2}{2}\frac{\partial^2 f}{\partial t^2}(t,\boldsymbol{X}) + (hu)^2\varepsilon_{1,\boldsymbol{X}}(uh)\Big) \\
&\qquad\qquad \times \big(f_{T|\boldsymbol{X}}(t \mid \boldsymbol{X}) + hu\frac{\partial f_{T|\boldsymbol{X}}}{\partial t}(t \mid \boldsymbol{X}) + \frac{(hu)^2}{2}\frac{\partial^2 f_{T|\boldsymbol{X}}}{\partial t^2}(t \mid \boldsymbol{X}) + (hu)^2\varepsilon_{2,\boldsymbol{X}}(uh)\big)\,\mathrm{d}u\Big)\Big] \\
&= \mathbb{E}\Big[\frac{f_j(\boldsymbol{X})}{\widehat{f}_{T|\boldsymbol{X}}(t \mid \boldsymbol{X})}\Big(\int_{\mathbb{R}} K(u)\Big(huf_{T|\boldsymbol{X}}(t \mid \boldsymbol{X})\frac{\partial f}{\partial t}(t,\boldsymbol{X}) + (hu)^2\frac{\partial f}{\partial t}(t,\boldsymbol{X})\frac{\partial f_{T|\boldsymbol{X}}}{\partial t}(t \mid \boldsymbol{X}) \\
&\qquad\qquad + \frac{(hu)^2}{2}f_{T|\boldsymbol{X}}(t \mid \boldsymbol{X})\frac{\partial^2 f}{\partial t^2}(t,\boldsymbol{X}) + (hu)^2\tilde{\varepsilon}_{\boldsymbol{X}}(uh)\Big)\,\mathrm{d}u\Big)\Big] \\
&= h^2\mathbb{E}\Big[\frac{f_j(\boldsymbol{X})}{\widehat{f}_{T|\boldsymbol{X}}(t \mid \boldsymbol{X})}\Big(\int_{\mathbb{R}} u^2K(u)\,\mathrm{d}u\Big)\Big(\frac{\partial f}{\partial t}(t,\boldsymbol{X})\frac{\partial f_{T|\boldsymbol{X}}}{\partial t}(t \mid \boldsymbol{X}) + \frac{1}{2}f_{T|\boldsymbol{X}}(t \mid \boldsymbol{X})\frac{\partial^2 f}{\partial t^2}(t,\boldsymbol{X})\Big)\Big] \\
&\qquad + o(h^2) \quad \text{(The first moment order integral vanishes + Lemma 6.4.6 on } \tilde{\varepsilon}_{\boldsymbol{X}}) \\
&= h^2\kappa_2(K)\mathbb{E}\Big[\frac{f_j(\boldsymbol{X})}{\widehat{f}_{T|\boldsymbol{X}}(t \mid \boldsymbol{X})}C_t(\boldsymbol{X})\Big] + o(h^2),
\end{aligned}
\tag{6.44}
$$

where

$$
C_t(\boldsymbol{X}) = \frac{\partial f}{\partial t}(t,\boldsymbol{X})\frac{\partial f_{T|\boldsymbol{X}}}{\partial t}(t \mid \boldsymbol{X}) + \frac{1}{2}f_{T|\boldsymbol{X}}(t \mid \boldsymbol{X})\frac{\partial^2 f}{\partial t^2}(t,\boldsymbol{X}), \tag{6.45}
$$

and $\kappa_2(K) = \int_{\mathbb{R}} u^2K(u)\,\mathrm{d}u$ is the second moment of the Kernel $K$.

Finally, for the error measurement term, we have

$$
h_{t,\epsilon,j} = \mathbb{E}[f_j(\boldsymbol{X})\frac{K_h(T-t)}{\widehat{f}_{T|\boldsymbol{X}}(t \mid \boldsymbol{X})}\epsilon] = \mathbb{E}\Big[f_j(\boldsymbol{X})\frac{K_h(T-t)}{\widehat{f}_{T|\boldsymbol{X}}(t \mid \boldsymbol{X})}\Big]\mathbb{E}[\epsilon] = 0. \tag{6.46}
$$

By gathering the three previous terms, we obtain the desired result ■

The covariance matrix $\mathbf{C}^M$ of $\boldsymbol{Z}_t$ has entries

$$
\begin{aligned}
(\mathbf{C}^M)_{jj'} &= \mathrm{Cov}\left(\boldsymbol{Z}_{t,j}, \boldsymbol{Z}_{t,j'}\right) = \mathbb{E}(\boldsymbol{Z}_{t,j}, \boldsymbol{Z}_{t,j'}) - \mathbb{E}(\boldsymbol{Z}_{t,j})\mathbb{E}(\boldsymbol{Z}_{t,j'}) \\
&= \begin{cases}
\mathbb{E}[f_j(\boldsymbol{X})f_{j'}(\boldsymbol{X})(B_{t,spec} + B_{t,K,h} + B_{t,\epsilon})^2] - h^M_{t,j}h^M_{t,j'} & \text{if } j,j' \in \{1,\dots,p\} \\
\mathbb{E}[f_k(\boldsymbol{X})f_{k'}(\boldsymbol{X})f_l(\boldsymbol{X})f_{l'}(\boldsymbol{X})] - F_{kk'}F_{ll'} & \text{if } j,j' \in \{p+1,\dots,p^2\} \\
\mathbb{E}[f_k(\boldsymbol{X})f_{k'}(\boldsymbol{X})(B^M_{t,spec} + B_{t,K,h})] - (h^M_{t,spec,j'} + h_{t,K,j'})F_{kk'} & \text{otherwise.}
\end{cases}
\end{aligned}
\tag{6.47}
$$

where $k,k' = \eta^{-1}(j)$ (respectively, $l,k' = \eta^{-1}(j')$) such that $\eta$ is the correspondence indexes map between $\boldsymbol{m}(h)$ and $\mathbf{F}$ in $\boldsymbol{m}(h)_j = F_{kk'}$ when $j \geq p+1$ (respectively, $\boldsymbol{m}(h)_{j'} = F_{ll'}$ when $j' \geq p+1$). The last line holds because of the independence of $\epsilon$ and thus $B_{t,\epsilon}$ to the other variables $B_{t,spec}$ and $B_{t,K,h}$.

If $j, j' \in \{1, \dots, p\}$ then,

$$\begin{aligned}
(\mathbf{C}^M)_{jj'} &= \mathbb{E}\big[f_j(\boldsymbol{X}) f_{j'}(\boldsymbol{X})(B^M_{t,spec} + B_{t,K,h} + B_{t,\epsilon})^2\big] - h_{t,j} h_{t,j'} \\
&= \mathbb{E}\Big[f_j(\boldsymbol{X}) f_{j'}(\boldsymbol{X})(B^2_{t,spec} + B^2_{t,K,h} + B^2_{t,\epsilon} + 2B^M_{t,spec} B_{t,K,h} + 2B_{t,\epsilon}(B^M_{t,spec} + B_{t,K,h})\Big] - h_{t,j} h_{t,j'} \\
&= \mathbb{E}\Big[f_j(\boldsymbol{X}) f_{j'}(\boldsymbol{X}) B^2_{t,spec}\Big] + \mathbb{E}\Big[f_j(\boldsymbol{X}) f_{j'}(\boldsymbol{X}) B^2_{t,K,h}\Big] + \mathbb{E}\Big[f_j(\boldsymbol{X}) f_{j'}(\boldsymbol{X}) B^2_{t,\epsilon}\Big] \\
&\quad + 2\mathbb{E}\Big[f_j(\boldsymbol{X}) f_{j'}(\boldsymbol{X}) B^M_{t,spec} B_{t,K,h}\Big] - h_{t,j} h_{t,j'} \quad (b^{(n)}_{t,\epsilon} \text{ is independent of } b^{(n)}_{t,spec} \text{ and } b^{(n)}_{t,K,h}) \\
&= (\mathbf{C}^M_{t,spec})_{jj'} + (\mathbf{C}^M_{t,K,h})_{jj'} + (\mathbf{C}^M_{t,\epsilon})_{jj'} + 2(\mathbf{C}^M_{t,K,spec})_{jj'} + h_{t,j} h_{t,j'}.
\end{aligned} \tag{6.48}$$

The product $h_{t,j} h_{t,j'}$ can be computed easily using Lemma 6.4.7. In our case, we can write $h_{t,j} h_{t,j'} = h^M_{t,spec,j} h^M_{t,spec,j'} + o(1)$.

**Lemma 6.4.8.** *The entries of the covariance matrix $\mathbf{C}^M$ satisfy:*

$$(\mathbf{C}^M)_{jj'} = \begin{cases} \frac{1}{h} C^M_1 + C^M_0 + o(1) & \text{if } j, j' \in \{1, \dots, p\}, \\ C_2 & \text{if } j, j' \in \{p+1, \dots, p^2\}, \\ C^M_{spec} + o(1) & \text{otherwise.} \end{cases} \tag{6.49}$$

*where $C^M_1, C^M_0, C_2$ and $C^M_{spec}$ are some given terms such that $C^M_1 \neq 0$.*

*Proof.* Similarly to the proof of the previous lemma, we compute each term of (6.48) separately.

On the one hand, if $j, j' \in \{1, \dots, p\}$, the first term of $(\mathbf{C}^M_{t,spec})_{jj'}$ is equal to

$$
\begin{aligned}
(\mathbf{C}^M_{t,spec})_{jj'} &= \mathbb{E}\Big[f_j(\boldsymbol{X}) f_{j'}(\boldsymbol{X}) B^2_{t,spec}\Big] \\
&= \mathbb{E}\Big[f_j(\boldsymbol{X}) f_{j'}(\boldsymbol{X}) \Big(\frac{K_h(T-t)}{\widehat{f}_{T|\boldsymbol{X}}(t \mid \boldsymbol{X})} - 1\Big)^2 f^2(t, \boldsymbol{X})\Big] \\
&= \mathbb{E}\Big[f_j(\boldsymbol{X}) f_{j'}(\boldsymbol{X}) \Big(\frac{\mathbb{E}[K_h^2(T-t) \mid \boldsymbol{X}]}{\widehat{f}^2_{T|\boldsymbol{X}}(t \mid \boldsymbol{X})} - 2\, \frac{\mathbb{E}[K_h(T-t) \mid \boldsymbol{X}]}{\widehat{f}_{T|\boldsymbol{X}}(t \mid \boldsymbol{X})} + 1\Big) f^2(t, \boldsymbol{X})\Big] \\
&\overset{h \approx 0}{=} \mathbb{E}\Big[f_j(\boldsymbol{X}) f_{j'}(\boldsymbol{X}) \Big( \frac{1}{\widehat{f}^2_{T|\boldsymbol{X}}(t \mid \boldsymbol{X})} \Big( \int_{\mathbb{R}} \frac{K^2(u)}{h} \Big[ f_{T|\boldsymbol{X}}(t \mid \boldsymbol{X}) + hu \frac{\partial f_{T|\boldsymbol{X}}}{\partial t}(t \mid \boldsymbol{X}) + \frac{(hu)^2}{2} \frac{\partial^2 f_{T|\boldsymbol{X}}}{\partial t^2}(t \mid \boldsymbol{X}) \\
&\quad + (hu)^2 \varepsilon_{2,\boldsymbol{X}}(uh) \Big] \mathrm{d}u \Big) - \frac{2}{\widehat{f}_{T|\boldsymbol{X}}(t \mid \boldsymbol{X})} \Big( \int_{\mathbb{R}} K(u) \Big[ f_{T|\boldsymbol{X}}(t \mid \boldsymbol{X}) + hu \frac{\partial f_{T|\boldsymbol{X}}}{\partial t}(t \mid \boldsymbol{X}) + \frac{(hu)^2}{2} \frac{\partial^2 f_{T|\boldsymbol{X}}}{\partial t^2}(t \mid \boldsymbol{X}) \\
&\quad + (hu)^2 \varepsilon_{2,\boldsymbol{X}}(uh) \Big] \mathrm{d}u \Big) + 1 \Big) f^2(t, \boldsymbol{X})\Big] \\
&= \mathbb{E}\Big[f_j(\boldsymbol{X}) f_{j'}(\boldsymbol{X}) \Big( \frac{1}{\widehat{f}^2_{T|\boldsymbol{X}}(t \mid \boldsymbol{X})} \Big( \frac{1}{h} \Big( \int_{\mathbb{R}} K^2(u)\, \mathrm{d}u \Big) f_{T|\boldsymbol{X}}(t \mid \boldsymbol{X}) + \Big( \int_{\mathbb{R}} uK^2(u)\, \mathrm{d}u \Big) \frac{\partial f_{T|\boldsymbol{X}}}{\partial t}(t \mid \boldsymbol{X}) \\
&\quad + \Big( \int_{\mathbb{R}} uK^2(u) \tilde{\varepsilon}_{\boldsymbol{X}}(uh)\, \mathrm{d}u \Big) \Big) - 2\, \frac{f_{T|\boldsymbol{X}}(t \mid \boldsymbol{X})}{\widehat{f}_{T|\boldsymbol{X}}(t \mid \boldsymbol{X})} + \int_{\mathbb{R}} K(u) \tilde{\varepsilon}'_{\boldsymbol{X}}(uh)\, \mathrm{d}u + 1 \Big) f^2(t, \boldsymbol{X})\Big] \\
&= \mathbb{E}\Big[f_j(\boldsymbol{X}) f_{j'}(\boldsymbol{X}) \Big( \frac{1}{h} R(K) \frac{f_{T|\boldsymbol{X}}(t \mid \boldsymbol{X})}{\widehat{f}^2_{T|\boldsymbol{X}}(t \mid \boldsymbol{X})} + \kappa_1(K^2) \frac{\frac{\partial f_{T|\boldsymbol{X}}}{\partial t}(t \mid \boldsymbol{X})}{\widehat{f}^2_{T|\boldsymbol{X}}(t \mid \boldsymbol{X})} - 2\, \frac{f_{T|\boldsymbol{X}}(t \mid \boldsymbol{X})}{\widehat{f}_{T|\boldsymbol{X}}(t \mid \boldsymbol{X})} + 1 \\
&\quad + o(1) \Big) f^2(t, \boldsymbol{X})\Big] \quad \text{(By Lemma 6.4.6)} \\
&= \frac{1}{h} R(K) \mathbb{E}\Big[f_j(\boldsymbol{X}) f_{j'}(\boldsymbol{X}) \frac{f_{T|\boldsymbol{X}}(t \mid \boldsymbol{X})}{\widehat{f}^2_{T|\boldsymbol{X}}(t \mid \boldsymbol{X})} f^2(t, \boldsymbol{X})\Big] + \mathbb{E}\Big[f_j(\boldsymbol{X}) f_{j'}(\boldsymbol{X}) \Big(1 - 2\, \frac{f_{T|\boldsymbol{X}}(t \mid \boldsymbol{X})}{\widehat{f}_{T|\boldsymbol{X}}(t \mid \boldsymbol{X})}\Big) f^2(t, \boldsymbol{X})\Big] \\
&\quad + o(1),
\end{aligned}
\tag{6.50}
$$

where $R(K) = \int_{\mathbb{R}} K^2(u)\, \mathrm{d}u$ is the roughness of the kernel $K$.

For the second term $(\mathbf{C}^M_{t,K,h})_{jj'}$ and by similar argument, we have:

$$
\begin{aligned}
(\mathbf{C}^M_{t,K,h})_{jj'} &= \mathbb{E}\Big[f_j(\boldsymbol{X}) f_{j'}(\boldsymbol{X}) \frac{K_h^2(T-t)}{\widehat{f}^2_{T|\boldsymbol{X}}(t \mid \boldsymbol{X})} \Big(h(T-t)/h \frac{\partial f}{\partial t}(t, \boldsymbol{X}) + \frac{h^2((T-t)/h)^2}{2} \frac{\partial^2 f}{\partial t^2}(t, \boldsymbol{X})\Big)^2\Big] \\
&= \mathbb{E}\Big[\frac{f_j(\boldsymbol{X}) f_{j'}(\boldsymbol{X})}{\widehat{f}^2_{T|\boldsymbol{X}}(t \mid \boldsymbol{X})} \mathbb{E}\Big[K_h^2(T-t) \Big(h(T-t)/h \frac{\partial f}{\partial t}(t, \boldsymbol{X}) + \frac{h^2((T-t)/h)^2}{2} \frac{\partial^2 f}{\partial t^2}(t, \boldsymbol{X})\Big)^2 \mid \boldsymbol{X}\Big]\Big] \\
&\overset{h \approx 0}{=} \mathbb{E}\Big[\frac{f_j(\boldsymbol{X}) f_{j'}(\boldsymbol{X})}{\widehat{f}^2_{T|\boldsymbol{X}}(t \mid \boldsymbol{X})} \Big( \frac{1}{h} \int_{\mathbb{R}} K^2(u) \Big( f_{T|\boldsymbol{X}}(t \mid \boldsymbol{X}) + hu \frac{\partial f_{T|\boldsymbol{X}}}{\partial t}(t \mid \boldsymbol{X}) + \frac{(hu)^2}{2} \frac{\partial^2 f_{T|\boldsymbol{X}}}{\partial t^2}(t \mid \boldsymbol{X}) \\
&\quad + (hu)^2 \varepsilon_{2,\boldsymbol{X}}(uh) \Big) \Big(hu \frac{\partial f}{\partial t}(t, \boldsymbol{X}) + \frac{(hu)^2}{2} \frac{\partial^2 f}{\partial t^2}(t, \boldsymbol{X})\Big)^2 \mathrm{d}u \Big)\Big].
\end{aligned}
\tag{6.51}
$$

We consider only the first order term in the expansion of $f_{T|\boldsymbol{X}}$ and the second order term in

the second expression. Therefore,

$$
\begin{aligned}
(\mathbf{C}^M_{t,K,h})_{jj'} &= \mathbb{E}\Big[\frac{f_j(\boldsymbol{X})f_{j'}(\boldsymbol{X})}{\widehat{f}^2_{T|\boldsymbol{X}}(t \mid \boldsymbol{X})}\Big(\frac{1}{h}\int_{\mathbb{R}} K^2(u)\Big(f_{T|\boldsymbol{X}}(t \mid \boldsymbol{X}) + (hu)\tilde{\varepsilon}_{\boldsymbol{X}}(uh)\Big)\Big((hu)^2\Big(\frac{\partial f}{\partial t}(t, \boldsymbol{X})\Big)^2 \\
&\qquad + (hu)^2\tilde{\varepsilon}'_{\boldsymbol{X}}(uh)\Big)\,\mathrm{d}u\Big)\Big] \\
&= \frac{1}{h}\mathbb{E}\Big[\frac{f_j(\boldsymbol{X})f_{j'}(\boldsymbol{X})}{\widehat{f}^2_{T|\boldsymbol{X}}(t \mid \boldsymbol{X})}\Big(h^2\Big(\frac{\partial f}{\partial t}(t, \boldsymbol{X})\Big)^2 f_{T|\boldsymbol{X}}(t \mid \boldsymbol{X})\int_{\mathbb{R}} u^2K^2(u)\,\mathrm{d}u + h^2\int_{\mathbb{R}} u^2K^2(u)\tilde{\varepsilon}^{(2)}_{\boldsymbol{X}}(uh)\,\mathrm{d}u\Big)\Big] \\
&= h\Big(\int_{\mathbb{R}} u^2K^2(u)\,\mathrm{d}u\Big)\mathbb{E}\Big[\frac{f_j(\boldsymbol{X})f_{j'}(\boldsymbol{X})}{\widehat{f}^2_{T|\boldsymbol{X}}(t \mid \boldsymbol{X})}\Big(\frac{\partial f}{\partial t}(t, \boldsymbol{X})\Big)^2 f_{T|\boldsymbol{X}}(t \mid \boldsymbol{X})\Big] + o(h) \\
&= o(1).
\end{aligned}
\tag{6.52}
$$

For the third term $(\mathbf{C}^M_{t,\epsilon})_{jj'}$:

$$
\begin{aligned}
(\mathbf{C}^M_{t,\epsilon})_{jj'} &= \mathbb{E}\Big[f_j(\boldsymbol{X})f_{j'}(\boldsymbol{X})\frac{K^2_h(T-t)}{\widehat{f}^2_{T|\boldsymbol{X}}(t \mid \boldsymbol{X})}\epsilon^2\Big] \\
&= \mathbb{E}\Big[f_j(\boldsymbol{X})f_{j'}(\boldsymbol{X})\frac{\mathbb{E}[K^2_h(T-t) \mid \boldsymbol{X}]}{\widehat{f}^2_{T|\boldsymbol{X}}(t \mid \boldsymbol{X})}\Big]\mathbb{E}[\epsilon^2] \\
&\overset{h\approx 0}{=} \frac{\sigma^2}{h}\mathbb{E}\Big[f_j(\boldsymbol{X})f_{j'}(\boldsymbol{X})\frac{\int_{\mathbb{R}} K^2(u)\Big(f_{T|\boldsymbol{X}}(t \mid \boldsymbol{X}) + hu\frac{\partial f_{T|\boldsymbol{X}}}{\partial t}(t \mid \boldsymbol{X}) + \frac{(hu)^2}{2}\frac{\partial^2 f_{T|\boldsymbol{X}}}{\partial t^2}(t \mid \boldsymbol{X})}{\widehat{f}^2_{T|\boldsymbol{X}}(t \mid \boldsymbol{X})} \\
&\qquad + \frac{(hu)^2\varepsilon_{2,\boldsymbol{X}}(uh)\Big)\,\mathrm{d}u}{\widehat{f}^2_{T|\boldsymbol{X}}(t \mid \boldsymbol{X})}\Big] \\
&= \frac{\sigma^2}{h}\mathbb{E}\Big[f_j(\boldsymbol{X})f_{j'}(\boldsymbol{X})\frac{\Big(\int_{\mathbb{R}} K^2(u)\,\mathrm{d}u\Big)f_{T|\boldsymbol{X}}(t \mid \boldsymbol{X}) + h\Big(\int_{\mathbb{R}} uK^2(u)\,\mathrm{d}u\Big)\frac{\partial f_{T|\boldsymbol{X}}}{\partial t}(t \mid \boldsymbol{X})}{\widehat{f}^2_{T|\boldsymbol{X}}(t \mid \boldsymbol{X})} \\
&\qquad + h\frac{\int_{\mathbb{R}} uK^2(u)\tilde{\varepsilon}_{\boldsymbol{X}}(uh)\,\mathrm{d}u}{\widehat{f}^2_{T|\boldsymbol{X}}(t \mid \boldsymbol{X})}\Big] \\
&= \frac{\sigma^2}{h}R(K)\mathbb{E}\Big[f_j(\boldsymbol{X})f_{j'}(\boldsymbol{X})\frac{f_{T|\boldsymbol{X}}(t \mid \boldsymbol{X})}{\widehat{f}^2_{T|\boldsymbol{X}}(t \mid \boldsymbol{X})}\Big] + o(1),
\end{aligned}
\tag{6.53}
$$

where the last line holds using the fact that $K$ is even and Lemma 6.4.6 .

Finally, for the last term $(\mathbf{C}^M_{t,K,spec})_{jj'}$:

$$
\begin{aligned}
(\mathbf{C}^M{}_{t,K,spec})_{jj'} &= \mathbb{E}\Big[f_j(\boldsymbol{X})f_{j'}(\boldsymbol{X})B^M_{t,spec}B_{t,K,h}\Big] \\
&= \mathbb{E}\Big[f_j(\boldsymbol{X})f_{j'}(\boldsymbol{X})\Big(\frac{K_h(T-t)}{\widehat{f}_{T|\boldsymbol{X}}(t \mid \boldsymbol{X})} - 1\Big)f(t,\boldsymbol{X})\frac{K_h(T-t)}{\widehat{f}_{T|\boldsymbol{X}}(t \mid \boldsymbol{X})}\Big(h(T-t)/h\frac{\partial f}{\partial t}(t,\boldsymbol{X}) \\
&\qquad + \frac{h^2((T-t)/h)^2}{2}\frac{\partial^2 f}{\partial t^2}(t,\boldsymbol{X})\Big)\Big] \\
&= \mathbb{E}\Big[f_j(\boldsymbol{X})f_{j'}(\boldsymbol{X})f(t,\boldsymbol{X})\frac{K_h^2(T-t)}{\widehat{f}^2_{T|\boldsymbol{X}}(t \mid \boldsymbol{X})}\Big(h(T-t)/h\frac{\partial f}{\partial T}(t,\boldsymbol{X}) + \frac{h^2((T-t)/h)^2}{2}\frac{\partial^2 f}{\partial T^2}(t,\boldsymbol{X})\Big)\Big] \\
&\quad - \mathbb{E}\Big[f_j(\boldsymbol{X})f_{j'}(\boldsymbol{X})f(t,\boldsymbol{X})\frac{K_h(T-t)}{\widehat{f}_{T|\boldsymbol{X}}(t \mid \boldsymbol{X})}\Big(h(T-t)/h\frac{\partial f}{\partial T}(t,\boldsymbol{X}) + \frac{h^2((T-t)/h)^2}{2}\frac{\partial^2 f}{\partial T^2}(t,\boldsymbol{X})\Big)\Big] \\
&\overset{h\approx 0}{=} \mathbb{E}\Big[\frac{f_j(\boldsymbol{X})f_{j'}(\boldsymbol{X})f(t,\boldsymbol{X})}{\widehat{f}^2_{T|\boldsymbol{X}}(t \mid \boldsymbol{X})}\Big(\frac{1}{h}\int_{\mathbb{R}} K^2(u)\Big(hu\frac{\partial f}{\partial t}(t,\boldsymbol{X}) + \frac{(hu)^2}{2}\frac{\partial^2 f}{\partial t^2}(t,\boldsymbol{X})\Big)\Big(f_{T|\boldsymbol{X}}(t \mid \boldsymbol{X}) \\
&\qquad + hu\frac{\partial f_{T|\boldsymbol{X}}}{\partial t}(t \mid \boldsymbol{X}) + \frac{(hu)^2}{2}\frac{\partial^2 f_{T|\boldsymbol{X}}}{\partial t^2}(t \mid \boldsymbol{x}) + (hu)^2\varepsilon_{2,\boldsymbol{X}}(uh)\Big)\,\mathrm{d}u\Big)\Big] - \mathbb{E}\Big[\frac{f_j(\boldsymbol{X})f_{j'}(\boldsymbol{X})}{\widehat{f}_{T|\boldsymbol{X}}(t \mid \boldsymbol{X})} \\
&\qquad \times f(t,\boldsymbol{X})\Big(\int_{\mathbb{R}} K(u)\Big(hu\frac{\partial f}{\partial t}(t,\boldsymbol{X}) + \frac{(hu)^2}{2}\frac{\partial^2 f}{\partial t^2}(t,\boldsymbol{X})\Big)\Big(f_{T|\boldsymbol{X}}(t \mid \boldsymbol{X}) + hu\frac{\partial f_{T|\boldsymbol{X}}}{\partial t}(t \mid \boldsymbol{X}) \\
&\qquad\qquad + \frac{(hu)^2}{2}\frac{\partial^2 f_{T|\boldsymbol{X}}}{\partial t^2}(t \mid \boldsymbol{X}) + (hu)^2\varepsilon_{2,\boldsymbol{X}}(uh)\Big)\,\mathrm{d}u\Big)\Big].
\end{aligned}
\tag{6.54}
$$

As we want to have an expression in $o(1)$ or $\mathcal{O}(1)$, we keep only the first order term $hu$ in the expansion for the first expectation, and we neglect it for the second expectation. Therefore,

$$
\begin{aligned}
(\mathbf{C}^M{}_{t,K,spec})_{jj'} &= \mathbb{E}\Big[\frac{f_j(\boldsymbol{X})f_{j'}(\boldsymbol{X})f(t,\boldsymbol{X})}{\widehat{f}^2_{T|\boldsymbol{X}}(t \mid \boldsymbol{X})}\Big(\frac{1}{h}\int_{\mathbb{R}} K^2(u)\Big(hu\frac{\partial f}{\partial t}(t,\boldsymbol{X})f_{T|\boldsymbol{X}}(t \mid \boldsymbol{X}) + (hu)\tilde{\varepsilon}_{\boldsymbol{X}}(uh)\Big)\,\mathrm{d}u\Big)\Big] \\
&\qquad - \mathbb{E}\Big[\frac{f_j(\boldsymbol{X})f_{j'}(\boldsymbol{X})f(t,\boldsymbol{X})}{\widehat{f}_{T|\boldsymbol{X}}(t \mid \boldsymbol{X})}\Big(\int_{\mathbb{R}} K(u)\tilde{\varepsilon}'_{\boldsymbol{X}}(uh)\,\mathrm{d}u\Big)\Big] \\
&= \mathbb{E}\Big[\frac{f_j(\boldsymbol{X})f_{j'}(\boldsymbol{X})f(t,\boldsymbol{X})}{\widehat{f}^2_{T|\boldsymbol{X}}(t \mid \boldsymbol{X})}f_{T|\boldsymbol{X}}(t \mid \boldsymbol{X})\frac{\partial f}{\partial t}(t,\boldsymbol{X})\Big(\int_{\mathbb{R}} uK^2(u)\,\mathrm{d}u\Big)\Big] + o(1) \quad \text{(Lemma 6.4.6)} \\
&= o(1).
\end{aligned}
\tag{6.55}
$$

By gathering the four previous terms, we can write

$$
\mathbf{C}_{j,j'} = \frac{1}{h}C_1^M + C_0^M + o(1), \tag{6.56}
$$

where

$$
C_1^M = R(K)\mathbb{E}\Big[\frac{f_j(\boldsymbol{X})f_{j'}(\boldsymbol{X})}{\widehat{f}^2_{T|\boldsymbol{X}}(t \mid \boldsymbol{X})}f_{T|\boldsymbol{X}}(t \mid \boldsymbol{X})(\sigma^2 + f^2(t,\boldsymbol{X}))\Big], \tag{6.57}
$$

and

$$\begin{aligned} C_0^M = \mathbb{E}\Big[f_j(\boldsymbol{X})\Big(\frac{f_{T|\boldsymbol{X}}(t \mid \boldsymbol{X})}{\widehat{f}_{T|\boldsymbol{X}}(t \mid \boldsymbol{X})} - 1\Big) f(t, \boldsymbol{X})\Big] \mathbb{E}\Big[f_{j'}(\boldsymbol{X})\Big(\frac{f_{T|\boldsymbol{X}}(t \mid \boldsymbol{X})}{\widehat{f}_{T|\boldsymbol{X}}(t \mid \boldsymbol{X})} - 1\Big) f(t, \boldsymbol{X})\Big] \\ + \mathbb{E}\Big[f_j(\boldsymbol{X}) f_{j'}(\boldsymbol{X})\Big(1 - 2\,\frac{f_{T|\boldsymbol{X}}(t \mid \boldsymbol{X})}{\widehat{f}_{T|\boldsymbol{X}}(t \mid \boldsymbol{X})}\Big) f^2(t, \boldsymbol{X})\Big]. \end{aligned} \tag{6.58}$$

On the other hand, $j \in \{p+1, \ldots, p^2\}$ and $j' \in \{1, \ldots, p\}$ (or inversely by symmetry), then.

$$\begin{aligned} \mathbf{C}^M_{j,j'} &= \mathbb{E}\big[f_k(\boldsymbol{X}) f_{k'}(\boldsymbol{X})(B^M_{t,spec} + B_{t,K,h})\big] \\ &= \mathbb{E}\big[f_k(\boldsymbol{X}) f_{k'}(\boldsymbol{X}) B^M_{t,spec}\big] + \mathbb{E}\big[f_k(\boldsymbol{X}) f_{k'}(\boldsymbol{X}) B^M_{t,K,h}\big] \\ &\qquad \text{(By similar calculus to } h_{t,spec,j} \text{ and } h_{t,K,h,j}) \\ &= \mathbb{E}\Big[f_k(\boldsymbol{X}) f_{k'}(\boldsymbol{X})\Big(\frac{f_{T|\boldsymbol{X}}(t \mid \boldsymbol{X})}{\widehat{f}_{T|\boldsymbol{X}}(t \mid \boldsymbol{X})} - 1\Big) f(t, \boldsymbol{X})\Big] + o(1) \\ &= C^M_{spec} + o(1), \end{aligned} \tag{6.59}$$

where $C^M_{spec} = \mathbb{E}\Big[f_k(\boldsymbol{X}) f_{k'}(\boldsymbol{X})\Big(\frac{f_{T|\boldsymbol{X}}(t|\boldsymbol{X})}{\widehat{f}_{T|\boldsymbol{X}}(t|\boldsymbol{X})} - 1\Big) f(t, \boldsymbol{X})\Big]$ is the misspecification covariance term with $C^M_{spec} = 0$ if the conditional density estimator $\widehat{f}_{T|\boldsymbol{X}}$ is well-specified.

Thus, by gathering all the previous terms of the matrix $\mathbf{C}^M$, we get the desired result of the lemma. ■

**Proposition 6.4.9.** *If the conditional density estimator $\widehat{f}_{T|\boldsymbol{X}}$ is well specified, then the estimator $\widehat{\boldsymbol{\beta}}^M_h(t)$ has bias and variance such that*

$$\begin{aligned} \mathbb{E}(\widehat{\boldsymbol{\beta}}^M_h(t)) &\approx \boldsymbol{\beta}(t) + h^2 \mathbf{F}^{-1} \boldsymbol{h}^M_t, \\ \mathbb{V}(\widehat{\boldsymbol{\beta}}^M_h(t)) &\approx \frac{1}{nh}\, \mathbf{F}^{-1} \overline{\boldsymbol{\Sigma}}^M \mathbf{F}^{-1}, \end{aligned} \tag{6.60}$$

*where, for $j \in \{1, \ldots, p\}$,*

$$h^M_{t,j} = h^2 \kappa_2(K) \mathbb{E}\Big[\frac{f_j(\boldsymbol{X})}{\widehat{f}_{T|\boldsymbol{X}}(t \mid \boldsymbol{X})} C'_t(\boldsymbol{X})\Big] + o(h^2), \tag{6.61}$$

*with*

$$C'_t(\boldsymbol{X}) = \frac{\partial f}{\partial t}(t, \boldsymbol{X}) \frac{\partial f_{T|\boldsymbol{X}}}{\partial t}(t \mid \boldsymbol{X}) + \frac{1}{2} f_{T|\boldsymbol{X}}(t \mid \boldsymbol{X}) \frac{\partial^2 f}{\partial t^2}(t, \boldsymbol{X}) + \frac{1}{2} \frac{\partial^2 f_{T|\boldsymbol{X}}}{\partial t^2}(t \mid \boldsymbol{X}), \tag{6.62}$$

*and, for $j, j' \in \{1, \ldots, p\}$,*

$$\overline{\boldsymbol{\Sigma}}^M_{jj'} = (\overline{\mathbf{C}}_M)_{jj'} = R(K) \mathbb{E}\Big[\frac{f_j(\boldsymbol{X}) f_{j'}(\boldsymbol{X})}{f_{T|\boldsymbol{X}}(t \mid \boldsymbol{X})} (\sigma^2 + f^2(t, \boldsymbol{X}))\Big] + o(1). \tag{6.63}$$

*Proof.* The proof of this proposition is not too different from B.2. Indeed, with Lemmas 6.4.7 and 6.4.8, we have all the ingredients of the vector $\boldsymbol{Z}^{(n)}$ (i.e. all components of $\boldsymbol{m}(h)$ and $\mathbf{C}$).

We consider now the vector $\boldsymbol{S}^{(n)} = \sqrt{n}(\boldsymbol{Z}^{(n)} - \boldsymbol{m}(h))$ in a manner that $\widehat{\boldsymbol{\beta}}_h^M(t) = \boldsymbol{\beta}(t) + \Phi(\boldsymbol{S}^{(n)}, \boldsymbol{m}(h)) + o(h^2)$ where $\Phi : \mathbb{R}^{p+p^2} \times \mathbb{R}^{p+p^2} \to \mathbb{R}^p$ is also a $\mathcal{C}^1$-function.

Even without assuming the well-specification of $\widehat{f}_{T|\boldsymbol{X}}$, we can prove in the general case by the multivariate Central Limit Theorem and the Delta method:

$$\sqrt{n}\Big[\Phi(\boldsymbol{S}^{(n)}, \boldsymbol{m}(h)) - \Phi(\mathbf{0}, \boldsymbol{m}(h))\Big] \xrightarrow{\mathcal{L}} \mathcal{N}\left(\mathbf{0}, J_\Phi^{(1)}(\mathbf{0}, \boldsymbol{m}(h))^\top \mathbf{C}^M J_\Phi^{(1)}(\mathbf{0}, \boldsymbol{m}(h))\right), \tag{6.64}$$

where $J_\Phi^{(1)}(\mathbf{0}, \boldsymbol{m}(h))$ is the Jacobian matrix at the first $p + p^2$ coordinates of $\Phi$ at $(\mathbf{0}, \boldsymbol{m}(h))$.

$$\begin{aligned} J_\Phi^{(1)}(\mathbf{0}, \boldsymbol{m}(h)) &= J_\Phi^{(1)}(\mathbf{0}, \boldsymbol{m} + h^2(\boldsymbol{b}_{t,spec}^M, \mathbf{0})^\top) \\ &= J_\Phi^{(1)}(\mathbf{0}, \boldsymbol{m}) + h^2\Big(J^{(2)}_{J_\Phi^{(1)}}(\mathbf{0}, \boldsymbol{m})\Big)(\boldsymbol{b}_{t,spec}^M, \mathbf{0})^\top + o(h^2), \end{aligned} \tag{6.65}$$

where $J^{(2)}_{J_\Phi^{(1)}}(\mathbf{0}, \boldsymbol{m})$ is the Jacobian matrix at the second $p + p^2$ coordinates of $J_\Phi^{(1)}$ at $(\mathbf{0}, \boldsymbol{m}(h))$ and $\boldsymbol{b}_{t,spec}^M$ is the misspecification bias as defined in (6.30).

For $n$ big enough, the expansions of the first two moments are of the form:

$$\begin{aligned} \mathbb{E}(\widehat{\boldsymbol{\beta}}_h^M(t)) &\approx \boldsymbol{\beta}(t) + \Phi(\mathbf{0}, \boldsymbol{m}(h)) + o(h^2) \\ &= \boldsymbol{\beta}(t) + \Phi(\mathbf{0}, \boldsymbol{m}) + h^2 J_\Phi^{(1)}(\mathbf{0}, \boldsymbol{m})(\boldsymbol{b}_{t,spec}^M, \mathbf{0})^\top + o(h^2), \end{aligned} \tag{6.66}$$

and,

$$\begin{aligned} \mathbb{V}(\widehat{\boldsymbol{\beta}}_h^M(t)) &\approx \frac{1}{n}\ J_\Phi^{(1)}(\mathbf{0}, \boldsymbol{m}(h))^\top \mathbf{C}^M J_\Phi^{(1)}(\mathbf{0}, \boldsymbol{m}(h)) \\ &= \frac{1}{nh}\ J_\Phi^{(1)}(\mathbf{0}, \boldsymbol{m}(h))^\top \mathbf{C}^M J_\Phi^{(1)}(\mathbf{0}, \boldsymbol{m}(h)) \\ &= \frac{1}{nh}\Big[\Big(J_\Phi^{(1)}(\mathbf{0}, \boldsymbol{m})\Big)^\top \overline{\mathbf{C}}^M \Big(J_\Phi^{(1)}(\mathbf{0}, \boldsymbol{m})\Big) + h^2\Big(\Big(J_\Phi^{(1)}(\mathbf{0}, \boldsymbol{m})\Big)^\top \overline{\mathbf{C}}^M J^{(2)}_{J_\Phi^{(1)}}(\mathbf{0}, \boldsymbol{m}) \\ &\qquad + \Big(J^{(2)}_{J_\Phi^{(1)}}(\mathbf{0}, \boldsymbol{m})\Big)^\top \overline{\mathbf{C}}^M J_\Phi^{(1)}(\mathbf{0}, \boldsymbol{m})\Big) + o(h^2)\Big] \\ &= \frac{1}{nh}\Big(J_\Phi^{(1)}(\mathbf{0}, \boldsymbol{m})\Big)^\top \overline{\mathbf{C}}^M \Big(J_\Phi^{(1)}(\mathbf{0}, \boldsymbol{m})\Big) + o(1), \end{aligned} \tag{6.67}$$

where $\overline{\mathbf{C}}^M = h\mathbf{C}^M$ is a normalization matrix.

In the particular case where the conditional density estimator $\widehat{f}_{T|\boldsymbol{X}}$ is well-specified. We show that $\boldsymbol{h}_{t,spec}^M = \mathbf{0}$ and, thus, $\boldsymbol{h}_t^M = h^2\kappa_2(K)\boldsymbol{h}_{t,Kern}^M + o(h^2)$ where $\boldsymbol{h}_{t,Kern}^M = (h_{t,Kern,j}^M)_{j=1}^p$ are given in (6.40).

In the following, we denote $h^2\boldsymbol{b}^M = \boldsymbol{b}_{t,spec}^M + \boldsymbol{b}_{t,K,h}$, we neglect the term $o(h^2)$ and we apply the multivariate Central Theorem Limit (CTL)

$$\frac{1}{\sqrt{n}}\Big[\mathbf{H}^\top(\boldsymbol{b}_{t,spec}^M + \boldsymbol{b}_{t,K,h} + \boldsymbol{b}_{t,\epsilon}) - nh^2\boldsymbol{b}^M\Big] \xrightarrow{\mathcal{L}} \mathcal{N}(\mathbf{0}, \frac{1}{h}\overline{\boldsymbol{\Sigma}}^M), \tag{6.68}$$

where $\overline{\boldsymbol{\Sigma}}^M$ is a covariance matrix with the same entries as the first block matrix of $\overline{\mathbf{C}}_M$, i.e. for $j, j' \in 1, \ldots, p$:

$$\overline{\boldsymbol{\Sigma}}_{jj'}^M = (\overline{\mathbf{C}}_M)_{jj'} = C_1^M + hC_0^M + o(h), \tag{6.69}$$

where $C_1^M$ and $C_0^M$ are given in (6.58-6.57).

By Slutsky's theorem,

$$\begin{aligned}\sqrt{n}(\widehat{\boldsymbol{\beta}}_h^M(t) - \boldsymbol{\beta}(t) - nh^2(\mathbf{H}^\top\mathbf{H})^{-1}\boldsymbol{b}^M) &= n(\mathbf{H}^\top\mathbf{H})^{-1}\frac{1}{\sqrt{n}}\Big[\mathbf{H}^\top(\boldsymbol{b}_{t,spec} + \boldsymbol{b}_{t,K,h} + \boldsymbol{b}_{t,\epsilon}) - nh^2\boldsymbol{b}^M\Big] \\ &\xrightarrow{\mathcal{L}} \mathcal{N}(\mathbf{0}, \frac{1}{h}\mathbf{F}^{-1}\overline{\boldsymbol{\Sigma}}^M\mathbf{F}^{-1}),\end{aligned} \tag{6.70}$$

which leads finally to the important result of

$$\begin{aligned}\mathbb{E}(\widehat{\boldsymbol{\beta}}_h^M(t)) &\approx \boldsymbol{\beta}(t) + h^2\mathbf{F}^{-1}\boldsymbol{b}^M, \\ \mathbb{V}(\widehat{\boldsymbol{\beta}}_h^M(t)) &\approx \frac{1}{nh}\ \mathbf{F}^{-1}\overline{\boldsymbol{\Sigma}}^M\mathbf{F}^{-1}.\end{aligned} \tag{6.71}$$

■

**Discussion** Proposition 6.4.9 shows three main results: Firstly, the bias of the estimator $\widehat{\boldsymbol{\beta}}_h^M(t)$ is in $\mathcal{O}(h^2)$. It cannot provide a consistent estimation of $\boldsymbol{\beta}(t)$ unless if $h = 0$. Secondly, the variance of $\widehat{\boldsymbol{\beta}}_h^M(t)$ is in $1/(nh)$, which implies that choosing a small bandwidth $h$ would increase the variance of the estimator. Finally, the bias and variance of $\widehat{\boldsymbol{\beta}}_h^M(t)$ consider the conditional density estimator $\widehat{f}_{T|\boldsymbol{X}}$ in the denominator. Thus, both bias and variance of the M-learner are likely to be sensitive to the lower bound $r_{\min}$.

**Proposition 6.4.10.** *Under all previous assumptions, the asymptotic Mean Squared Error (MSE) of the M-Learner $\widehat{\boldsymbol{\beta}}_h^M(t)$ is given by*

$$MSE(\widehat{\boldsymbol{\beta}}_h^M(t)) = h^4\|\mathbf{F}^{-1}\boldsymbol{b}^M\|^2 + \frac{1}{nh}\operatorname{Tr}(\mathbf{F}^{-1}\overline{\boldsymbol{\Sigma}}^M\mathbf{F}^{-1}), \tag{6.72}$$

*and the optimal bandwidth $h_t^*$ that minimizes the asymptotic MSE satisfies:*

$$h_t^* = \left(\frac{\operatorname{Tr}(\mathbf{F}^{-1}\overline{\boldsymbol{\Sigma}}^M\mathbf{F}^{-1})}{4n\|\mathbf{F}^{-1}\boldsymbol{b}^M\|^2}\right)^{1/5}, \tag{6.73}$$

*Proof.* Under Assumptions (4.5.3-6.4.3) and if the conditional density estimator $\widehat{f}_{T|\boldsymbol{X}}$ is well-specified. The asymptotic Mean Squared Error (MSE) of the M-Learner is given by

$$\begin{aligned}\operatorname{MSE}(\widehat{\boldsymbol{\beta}}_h^M(t)) &= \mathbb{E}\big[\|\widehat{\boldsymbol{\beta}}_h^M(t) - \boldsymbol{\beta}(t)\|^2\big] \\ &= \sum_{j=1}^{p}\mathbb{E}\big[\big(\widehat{\beta}_j^M(t) - \beta_j(t)\big)^2\big] \\ &= \sum_{j=1}^{p}\Big(\operatorname{Bias}\big(\widehat{\beta}_j^M(t)\big)^2 + \mathbb{V}(\widehat{\beta}_j(t))\Big) \\ &= h^4\|\mathbf{F}^{-1}\boldsymbol{b}^M\|^2 + \frac{1}{nh}\operatorname{Tr}(\mathbf{F}^{-1}\overline{\boldsymbol{\Sigma}}^M\mathbf{F}^{-1}).\end{aligned} \tag{6.74}$$

Optimizing the bias-variance tradeoff of the asymptotic mean squared error, with respect to the bandwidth $h$, gives:

$$\frac{\partial}{\partial h}\operatorname{MSE}(\widehat{\boldsymbol{\beta}}_h^M(t)) = 4h^3\|\mathbf{F}^{-1}\boldsymbol{b}^M\|^2 - \frac{1}{nh^2}\operatorname{Tr}(\mathbf{F}^{-1}\overline{\boldsymbol{\Sigma}}^M\mathbf{F}^{-1}) = 0. \tag{6.75}$$

Therefore, if the bias term $\boldsymbol{b}^M$ is non-zero, the optimal bandwidth $h_t^*$ that minimizes the asymptotic MSE is

$$h_t^* = \left( \frac{\mathrm{Tr}(\mathbf{F}^{-1}\overline{\boldsymbol{\Sigma}}^M\mathbf{F}^{-1})}{4n\|\mathbf{F}^{-1}\boldsymbol{b}^M\|^2} \right)^{1/5}, \tag{6.76}$$

and its order is $\mathcal{O}(n^{-1/5})$. ■

We note that a similar result was also proven for the average treatment effects by Colangelo & Lee (2020).

**Bias-Variance trade-off of the DR-learner.**

For the DR-learner, the Bias-Variance analysis is quite similar to the M-learner. Indeed, for $t \in \mathcal{T}$, we consider the AIPW pseudo-outcome with arbitrary estimators of the outcome $\widehat{\mu}$ and the conditional density estimator $\widehat{f}_{T|\boldsymbol{X}}$

$$Y_{t,h,i}^{DR} = \frac{y_i - \widehat{\mu}_t(\boldsymbol{x}^{(i)})}{\widehat{f}_{T|\boldsymbol{X}}(t \mid \boldsymbol{x}^{(i)})} K_h(t_i - t) + \widehat{\mu}_t(\boldsymbol{x}^{(i)}), \quad i = 1, \dots, n. \tag{6.77}$$

The regression coefficient $\widehat{\boldsymbol{\beta}}_{t,h}$ are given by the Ordinary Least Squares (OLS) method

$$\widehat{\boldsymbol{\beta}}_h^{DR}(t) = (\mathbf{H}^\top\mathbf{H})^{-1}\mathbf{H}^\top\boldsymbol{y}_{t,h}^{DR}, \tag{6.78}$$

where $\boldsymbol{y}_{t,h}^{DR} = (Y_{t,h,i}^{DR})_{1\le i\le n}$ and $\mathbf{H} = (\mathbf{H}_{ij}) \in \mathbb{R}^{n\times p}$ is the regression matrix. Therefore,

$$\begin{aligned}
\widehat{\boldsymbol{\beta}}_h^{DR}(t) &= (\mathbf{H}^\top\mathbf{H})^{-1}\mathbf{H}^\top\boldsymbol{y}_{t,h}^{DR} \\
&= \widehat{\boldsymbol{\beta}}_h^{M}(t) - (\mathbf{H}^\top\mathbf{H})^{-1}\mathbf{H}^\top\left( \Big( \frac{K_h(t_i - t)}{\widehat{f}_{T|\boldsymbol{X}}(t \mid \boldsymbol{x}^{(i)})} - 1 \Big)\widehat{\mu}_t(\boldsymbol{x}^{(i)}) \right)_{i=1}^n \\
&= \boldsymbol{\beta}(t) + (\mathbf{H}^\top\mathbf{H})^{-1}\mathbf{H}^\top(\boldsymbol{b}_{t,spec}^{M} - \left( \Big( \frac{K_h(t_i - t)}{\widehat{f}_{T|\boldsymbol{X}}(t \mid \boldsymbol{x}^{(i)})} - 1 \Big)\widehat{\mu}_t(\boldsymbol{x}^{(i)}) \right)_{i=1}^n + \boldsymbol{b}_{t,K,h} + \boldsymbol{b}_{t,\epsilon}) + o(h^2) \\
&= \boldsymbol{\beta}(t) + (\mathbf{H}^\top\mathbf{H})^{-1}\mathbf{H}^\top(\boldsymbol{b}_{t,spec}^{DR} + \boldsymbol{b}_{t,K,h} + \boldsymbol{b}_{t,\epsilon}) + o(h^2).
\end{aligned} \tag{6.79}$$

Here,

$$\begin{aligned}
\boldsymbol{b}_{t,spec}^{DR} &= \left( \Big( \frac{K_h(t_i - t)}{\widehat{f}_{T|\boldsymbol{X}}(t \mid \boldsymbol{x}^{(i)})} - 1 \Big) f(t, \boldsymbol{x}^{(i)}) + \Big( \frac{K_h(t_i - t)}{\widehat{f}_{T|\boldsymbol{X}}(t \mid \boldsymbol{x}^{(i)})} - 1 \Big)\widehat{\mu}_t(\boldsymbol{x}^{(i)}) \right)_{i=1}^n \\
&= \left( \Big( \frac{K_h(t_i - t)}{\widehat{f}_{T|\boldsymbol{X}}(t \mid \boldsymbol{x}^{(i)})} - 1 \Big)\Big( f(t, \boldsymbol{x}^{(i)}) - \widehat{\mu}_t(\boldsymbol{x}^{(i)}) \Big) \right)_{i=1}^n
\end{aligned} \tag{6.80}$$

is the bias term related to the misspecification of the outcome model estimator or the conditional density estimator. The other terms $\boldsymbol{b}_{t,K,h}$ and $\boldsymbol{b}_{t,\epsilon}$ remain unchanged, and consequently the previous calculations are similar. The only changes are in the terms corresponding to $\boldsymbol{b}_{t,spec}^{DR}$.

**Lemma 6.4.11.** *For $j = 1, \dots, p$*

$$h_{t,j}^{DR} = h_{t,spec,j}^{DR} + h^2\kappa_2(K)h_{t,Kern,j}^{DR} + o(h^2), \tag{6.81}$$

*where*

$$h^{DR}_{t,spec,j} = \mathbb{E}\Big[f_j(\boldsymbol{X})\Big(\frac{K_h(T-t)}{\widehat{f}_{T|\boldsymbol{X}}(t \mid \boldsymbol{X})} - 1\Big)\Big(f(t,\boldsymbol{X}) - \widehat{\mu}_t(\boldsymbol{X})\Big)\Big] \tag{6.82}$$

*is the misspecification bias and is equal to zero under the Doubly-Robustness property of the DR-learner, and,*

$$h^{DR}_{t,Kern,j} = \mathbb{E}\Big[\frac{f_j(\boldsymbol{X})}{\widehat{f}_{T|\boldsymbol{X}}(t \mid \boldsymbol{X})} C'_t(\boldsymbol{X})\Big] = h^{M}_{t,Kern,j}, \tag{6.83}$$

*with*

$$C'_t(\boldsymbol{X}) = \frac{\partial f}{\partial t}(t,\boldsymbol{X})\frac{\partial f_{T|\boldsymbol{X}}}{\partial t}(t \mid \boldsymbol{X}) + \frac{1}{2} f_{T|\boldsymbol{X}}(t \mid \boldsymbol{X})\frac{\partial^2 f}{\partial t^2}(t,\boldsymbol{X}) + \frac{1}{2}\frac{\partial^2 f_{T|\boldsymbol{X}}}{\partial t^2}(t \mid \boldsymbol{X}) \tag{6.84}$$

*is the bias induced by the use of kernel methods.*

*Proof.* The proof holds immediately by adding the corresponding change in $h^{DR}_{t,spec,j}$. ■

**Lemma 6.4.12.** *The entries of the covariance matrix $\mathbf{C}^{DR}$ satisfy:*

$$(\mathbf{C}^{DR})_{jj'} = \begin{cases} \frac{1}{h} C_1^{DR} + C_0^{DR} + o(1) & \text{if } j,j' \in \{1,\dots,p\}, \\ C_2 & \text{if } j,j' \in \{p+1,\dots,p^2\}, \\ C_{spec}^{DR} + o(1) & \text{otherwise.} \end{cases} \tag{6.85}$$

*where $C_1^{DR}, C_0^{DR}, C_2$ and $C_{spec}^{DR}$ are some given terms such that $C_1^{DR} \neq 0$.*

*Proof.* This proof is similar to the proof of Lemma 6.4.8. The change in the expression of $\boldsymbol{b}^{DR}_{t,spec}$ implies only changes in $(\mathbf{C}^{DR}_{t,spec})_{jj'}$ for $j,j' \in \{1,\dots,p\}$ and $\mathbf{C}^{DR}_{j,j'}$ for $j \in \{p+1,\dots,p^2\}$ and $j' \in \{1,\dots,p\}$ (or inversely, by symmetry).

The first term $(\mathbf{C}^{DR}_{t,spec})_{jj'}$ can be computed using similar calculations for $(\mathbf{C}^{DR}_{t,spec})_{jj'}$. Indeed,

$$\begin{aligned}
(\mathbf{C}^{DR}_{t,spec})_{jj'} &= \mathbb{E}\Big[f_j(\boldsymbol{X}) f_{j'}(\boldsymbol{X}) B^2_{t,spec}\Big] \\
&= \mathbb{E}\Big[f_j(\boldsymbol{X}) f_{j'}(\boldsymbol{X})\Big(\frac{\mathbb{E}[K_h^2(T-t) \mid \boldsymbol{X}]}{\widehat{f}^2_{T|\boldsymbol{X}}(t \mid \boldsymbol{X})} - 2\,\frac{\mathbb{E}[K_h(T-t) \mid \boldsymbol{X}]}{\widehat{f}_{T|\boldsymbol{X}}(t \mid \boldsymbol{X})} + 1\Big)\Big(f(t,\boldsymbol{X}) - \widehat{\mu}_t(\boldsymbol{X})\Big)^2\Big] \\
&= \mathbb{E}\Big[f_j(\boldsymbol{X}) f_{j'}(\boldsymbol{X})\Big(\frac{1}{h} R(K)\frac{f_{T|\boldsymbol{X}}(t \mid \boldsymbol{X})}{\widehat{f}^2_{T|\boldsymbol{X}}(t \mid \boldsymbol{X})} + \Big(\kappa_1(K^2)\frac{\frac{\partial f_{T|\boldsymbol{X}}}{\partial t}(t \mid \boldsymbol{X})}{\widehat{f}^2_{T|\boldsymbol{X}}(t \mid \boldsymbol{X})} - 2\,\frac{f_{T|\boldsymbol{X}}(t \mid \boldsymbol{X})}{\widehat{f}_{T|\boldsymbol{X}}(t \mid \boldsymbol{X})} + 1 \\
&\qquad + o(1)\Big)\Big(f(t,\boldsymbol{X}) - \widehat{\mu}_t(\boldsymbol{X})\Big)^2\Big] \\
&= \frac{1}{h} R(K)\mathbb{E}\Big[f_j(\boldsymbol{X}) f_{j'}(\boldsymbol{X})\frac{f_{T|\boldsymbol{X}}(t \mid \boldsymbol{X})}{\widehat{f}^2_{T|\boldsymbol{X}}(t \mid \boldsymbol{X})}\big(f(t,\boldsymbol{X}) - \widehat{\mu}_t(\boldsymbol{X})\big)^2\Big] + \mathbb{E}\Big[f_j(\boldsymbol{X}) f_{j'}(\boldsymbol{X}) \\
&\qquad \times \Big(1 - 2\,\frac{f_{T|\boldsymbol{X}}(t \mid \boldsymbol{X})}{\widehat{f}_{T|\boldsymbol{X}}(t \mid \boldsymbol{X})}\Big)\big(f(t,\boldsymbol{X}) - \widehat{\mu}_t(\boldsymbol{X})\big)^2\Big] + o(1),
\end{aligned} \tag{6.86}$$

where $R(K) = \int_{\mathbb{R}} K^2(u)\,\mathrm{d}u$ is the roughness of the kernel $K$.

Thus, for $j, j' \in \{1, \dots, p\}$, we write:

$$\mathbf{C}_{j,j'} = \frac{1}{h} C_1^{DR} + C_0^{DR} + o(1), \tag{6.87}$$

where

$$C_1^{DR} = R(K)\mathbb{E}\Big[f_j(\boldsymbol{X}) f_{j'}(\boldsymbol{X}) \frac{f_{T|\boldsymbol{X}}(t \mid \boldsymbol{X})}{\widehat{f}^2_{T|\boldsymbol{X}}(t \mid \boldsymbol{X})} \Big(\sigma^2 + \big(f(t, \boldsymbol{X}) - \widehat{\mu}_t(\boldsymbol{X})\big)^2\Big)\Big], \tag{6.88}$$

and

$$\begin{aligned} C_0^{DR} &= h^{DR}_{t,spec,j} h^{DR}_{t,spec,j'} \\ &= \mathbb{E}\Big[f_j(\boldsymbol{X})\Big(\frac{K_h(T-t)}{\widehat{f}_{T|\boldsymbol{X}}(t \mid \boldsymbol{X})} - 1\Big)\Big(f(t, \boldsymbol{X}) - \widehat{\mu}_t(\boldsymbol{X})\Big)\Big]\mathbb{E}\Big[f_{j'}(\boldsymbol{X})\Big(\frac{K_h(T-t)}{\widehat{f}_{T|\boldsymbol{X}}(t \mid \boldsymbol{X})} - 1\Big) \\ &\quad \times \Big(f(t, \boldsymbol{X}) - \widehat{\mu}_t(\boldsymbol{X})\Big)\Big] + \mathbb{E}\Big[f_j(\boldsymbol{X}) f_{j'}(\boldsymbol{X})\Big(1 - 2\,\frac{f_{T|\boldsymbol{X}}(t \mid \boldsymbol{X})}{\widehat{f}_{T|\boldsymbol{X}}(t \mid \boldsymbol{X})}\Big)\big(f(t, \boldsymbol{X}) - \widehat{\mu}_t(\boldsymbol{X})\big)^2\Big]. \end{aligned} \tag{6.89}$$

For the second term, if $j \in \{p+1, \dots, p^2\}$ and $j' \in \{1, \dots, p\}$ (or inversely by symmetry), then

$$\begin{aligned} \mathbf{C}^{DR}_{j,j'} &= \mathbb{E}\big[f_k(\boldsymbol{X}) f_{k'}(\boldsymbol{X})(b_{t,spec} + b_{t,K,h})\big] \\ &= \mathbb{E}\big[f_k(\boldsymbol{X}) f_{k'}(\boldsymbol{X}) b_{t,spec}\big] + \mathbb{E}\big[f_k(\boldsymbol{X}) f_{k'}(\boldsymbol{X}) b_{t,K,h}\big] \\ &= \mathbb{E}\Big[f_k(\boldsymbol{X}) f_{k'}(\boldsymbol{X})\Big(\frac{f_{T|\boldsymbol{X}}(t \mid \boldsymbol{X})}{\widehat{f}_{T|\boldsymbol{X}}(t \mid \boldsymbol{X})} - 1\Big)\Big(f(t, \boldsymbol{X}) - \widehat{\mu}_t(\boldsymbol{X})\Big)\Big] + o(1) \\ &= C^{DR}_{spec} + o(1). \end{aligned} \tag{6.90}$$

Finally, by gathering all the previous terms,

$$(\mathbf{C}^{DR})_{jj'} = \begin{cases} \frac{1}{h} C_1^{DR} + C_0^{DR} + o(1) & \text{if } j, j' \in \{1, \dots, p\} \\ C_2 & \text{if } j, j' \in \{p+1, \dots, p^2\} \\ C^{DR}_{spec} + o(1) & \text{otherwise,} \end{cases} \tag{6.91}$$

with $C^{DR}_{spec} = \mathbb{E}\Big[f_k(\boldsymbol{X}) f_{k'}(\boldsymbol{X})\Big(\frac{f_{T|\boldsymbol{X}}(t|\boldsymbol{X})}{\widehat{f}_{T|\boldsymbol{X}}(t|\boldsymbol{X})} - 1\Big)\Big(f(t, \boldsymbol{X}) - \widehat{\mu}_t(\boldsymbol{X})\Big)\Big]$ is the misspecification covariance term with $C^{DR}_{spec} = 0$ if doubly robustness property is satisfied. ■

**Proposition 6.4.13.** *If the doubly-robustness property of the DR-learner holds, then the estimator* $\widehat{\boldsymbol{\beta}}_h^{DR}(t)$ *has bias and variance such that*

$$\begin{aligned} \mathbb{E}(\widehat{\boldsymbol{\beta}}_h^{DR}(t)) &\approx \boldsymbol{\beta}(t) + h^2 \mathbf{F}^{-1} \boldsymbol{h}_t^{DR}, \\ \mathbb{V}(\widehat{\boldsymbol{\beta}}_h^{DR}(t)) &\approx \frac{1}{nh}\, \mathbf{F}^{-1} \overline{\boldsymbol{\Sigma}}^{DR} \mathbf{F}^{-1}, \end{aligned} \tag{6.92}$$

*where, for* $j \in \{1, \dots, p\}$,

$$h^{DR}_{t,j} = h^2 \kappa_2(K) \mathbb{E}\Big[\frac{f_j(\boldsymbol{X})}{\widehat{f}_{T|\boldsymbol{X}}(t \mid \boldsymbol{X})} C'_t(\boldsymbol{X})\Big] + o(h^2), \tag{6.93}$$

*with*

$$C'_t(\boldsymbol{X}) = \frac{\partial f}{\partial t}(t, \boldsymbol{X})\frac{\partial f_{T|\boldsymbol{X}}}{\partial t}(t \mid \boldsymbol{X}) + \frac{1}{2} f_{T|\boldsymbol{X}}(t \mid \boldsymbol{X})\frac{\partial^2 f}{\partial t^2}(t, \boldsymbol{X}) + \frac{1}{2}\frac{\partial^2 f_{T|\boldsymbol{X}}}{\partial t^2}(t \mid \boldsymbol{X}), \tag{6.94}$$

*and, for* $j, j' \in \{1, \dots, p\}$,

$$\overline{\boldsymbol{\Sigma}}^{DR}_{jj'} = \begin{cases} \sigma^2 R(K)\mathbb{E}\Big[\frac{f_j(\boldsymbol{X})f_{j'}(\boldsymbol{X})}{f_{T|\boldsymbol{X}}(t|\boldsymbol{X})}\Big] + o(1) & \text{if both models are well-specified,} \\ \sigma^2 R(K)\mathbb{E}\Big[f_j(\boldsymbol{X})f_{j'}(\boldsymbol{X})\frac{f_{T|\boldsymbol{X}}(t|\boldsymbol{X})}{\widehat{f}^2_{T|\boldsymbol{X}}(t|\boldsymbol{X})}\Big] + o(1) & \text{if } \widehat{\mu} \text{ is well-specified,} \\ R(K)\mathbb{E}\Big[\frac{f_j(\boldsymbol{X})f_{j'}(\boldsymbol{X})}{f_{T|\boldsymbol{X}}(t|\boldsymbol{X})}(\sigma^2 + (f(t, \boldsymbol{X}) - \widehat{\mu}_t(\boldsymbol{X}))^2)\Big] + o(1), & \text{if } \widehat{f}_{T|\boldsymbol{X}} \text{ is well-specified.} \end{cases} \tag{6.95}$$

*Proof.* Based on Lemmas 6.4.11-6.4.12 and similar to the proof of Proposition 6.4.9. The expression of $\overline{\boldsymbol{\Sigma}}^{DR}$ under the well-specification of one or both models can be established easily. ■

**Discussion** For the DR-learner, the first two results of the M-learner are still valid: a bias in $\mathcal{O}(h^2)$ and variance in $1/(nh)$. The variance of the DR-learner is also sensitive to the lower bound $r_{\min}$ in the denominator. Regarding the numerator, it can be reduced by minimizing the Mean Squared Error of the outcome model $\widehat{\mu}$.

It is also interesting to see that, for both M- and DR-learners, the kernel $K$ impacts the induced bias and variance in a similar way: through the second moment $\kappa_2(K)$ for the bias and the roughness $R(K)$ for the variance.

**Proposition 6.4.14.** *Under all previous assumptions, the asymptotic Mean Squared Error (MSE) of the DR-learner* $\widehat{\boldsymbol{\beta}}^{DR}_h(t)$ *is given by*

$$MSE(\widehat{\boldsymbol{\beta}}^{DR}_h(t)) = h^4 \|\mathbf{F}^{-1}\boldsymbol{b}^{DR}\|^2 + \frac{1}{nh}\operatorname{Tr}(\mathbf{F}^{-1}\overline{\boldsymbol{\Sigma}}^{DR}\mathbf{F}^{-1}), \tag{6.96}$$

*and the optimal bandwidth* $h^*_t$ *that minimizes the asymptotic MSE satisfies:*

$$h^*_t = \left(\frac{\operatorname{Tr}(\mathbf{F}^{-1}\overline{\boldsymbol{\Sigma}}^{DR}\mathbf{F}^{-1})}{4n\|\mathbf{F}^{-1}\boldsymbol{b}^{DR}\|^2}\right)^{1/5}, \tag{6.97}$$

*Proof.* Similar to the proof of Proposition 6.4.10. ■

### Bias-Variance trade-off of the X-learner.

For the X-learner, the Bias-Variance analysis is different to M- and DR-learners and computationally heavy to establish. This time, we target $\boldsymbol{\beta}^*(t) = \boldsymbol{\beta}(t) - \boldsymbol{\beta}(t_0)$ instead of targeting $\boldsymbol{\beta}(t)$ or $\boldsymbol{\beta}(t)$ separately. For $t \in \mathcal{T}$, we recall the Regression-Adjustment pseudo-

outcome with an arbitrary estimator of the outcome $\widehat{\mu}$ such that

$$\begin{aligned} Z_{t,h}^X = 2\epsilon(h)\, K_h(T-t)(Y_{\text{obs}} - \widehat{\mu}_{t_0}(\boldsymbol{X})) + \int_{t_{\min}}^{t-\epsilon(h)} K_h(T-t')\big(\widehat{\mu}_t(\boldsymbol{X}) - Y_{\text{obs}}\big)\,\mathrm{d}t' + \\ \int_{t+\epsilon(h)}^{t_{\max}} K_h(T-t')\big(\widehat{\mu}_t(\boldsymbol{X}) - Y_{\text{obs}}\big)\,\mathrm{d}t' + \int_{t_{\min}}^{t-\epsilon(h)} K_h(T-t')\big(\widehat{\mu}_{t'}(\boldsymbol{X}) - \widehat{\mu}_{t_0}(\boldsymbol{X})\big)\,\mathrm{d}t' \\ + \int_{t+\epsilon(h)}^{t_{\max}} K_h(T-t')\big(\widehat{\mu}_{t'}(\boldsymbol{X}) - \widehat{\mu}_{t_0}(\boldsymbol{X})\big)\,\mathrm{d}t'. \end{aligned} \tag{6.98}$$

The regression coefficient $\widehat{\boldsymbol{\beta}}_h^X(t)$ are given by the Ordinary Least Squares (OLS) method

$$\widehat{\boldsymbol{\beta}}_h^X(t) = (\mathbf{H}^\top \mathbf{H})^{-1} \mathbf{H}^\top \boldsymbol{z}_{t,h}^X, \tag{6.99}$$

where $\boldsymbol{z}_t^X = (Z_{t,h,i}^X)_{1\le i\le n}$ and $\mathbf{H} = (\mathbf{H}_{ij}) \in \mathbb{R}^{n\times p}$ is the regression matrix.

Therefore,

$$
\begin{aligned}
&\widehat{\boldsymbol{\beta}}_h^X(t) = (\mathbf{H}^\top\mathbf{H})^{-1}\mathbf{H}^\top \boldsymbol{z}_{t,h}^X \\
&= (\mathbf{H}^\top\mathbf{H})^{-1}\mathbf{H}^\top\Bigg(2\epsilon(h)\,K_h(t_i-t)(y_i-\widehat{\mu}_{t_0}(\boldsymbol{x}^{(i)})) + \int_{t_{\min}}^{t-\epsilon(h)} K_h(t_i-t')(\widehat{\mu}_t(\boldsymbol{x}^{(i)})-y_i)\,\mathrm{d}t' + \\
&\quad \int_{t+\epsilon(h)}^{t_{\max}} K_h(t_i-t')(\widehat{\mu}_t(\boldsymbol{x}^{(i)})-y_i)\,\mathrm{d}t' + \int_{t_{\min}}^{t-\epsilon(h)} K_h(t_i-t')(\widehat{\mu}_{t'}(\boldsymbol{x}^{(i)})-\widehat{\mu}_{t_0}(\boldsymbol{x}^{(i)}))\,\mathrm{d}t' \\
&\quad + \int_{t+\epsilon(h)}^{t_{\max}} K_h(t_i-t')(\widehat{\mu}_{t'}(\boldsymbol{x}^{(i)})-\widehat{\mu}_{t_0}(\boldsymbol{x}^{(i)}))\,\mathrm{d}t'\Bigg)_{i=1}^n \\
&= (\mathbf{H}^\top\mathbf{H})^{-1}\mathbf{H}^\top\Bigg(2\epsilon(h)\,K_h(t_i-t)(f(t_i,\boldsymbol{x}^{(i)})+\epsilon_i-\widehat{\mu}_{t_0}(\boldsymbol{x}^{(i)})) + \int_{t_{\min}}^{t-\epsilon(h)} K_h(t_i-t')(\widehat{\mu}_t(\boldsymbol{x}^{(i)})-f(t_i,\boldsymbol{x}^{(i)}) \\
&- \epsilon_i)\,\mathrm{d}t' + \int_{t+\epsilon(h)}^{t_{\max}} K_h(t_i-t')(\widehat{\mu}_t(\boldsymbol{x}^{(i)})-f(t_i,\boldsymbol{x}^{(i)})-\epsilon_i)\,\mathrm{d}t' + \int_{t_{\min}}^{t-\epsilon(h)} K_h(t_i-t')(\widehat{\mu}_{t'}(\boldsymbol{x}^{(i)})-\widehat{\mu}_{t_0}(\boldsymbol{x}^{(i)}))\,\mathrm{d}t' \\
&\quad + \int_{t+\epsilon(h)}^{t_{\max}} K_h(t_i-t')(\widehat{\mu}_{t'}(\boldsymbol{x}^{(i)})-\widehat{\mu}_{t_0}(\boldsymbol{x}^{(i)}))\,\mathrm{d}t'\Bigg)_{i=1}^n \\
&= (\mathbf{H}^\top\mathbf{H})^{-1}\mathbf{H}^\top\Bigg((2\epsilon(h)\,K_h(t_i-t) - \int_{t_{\min}}^{t-\epsilon(h)} K_h(t_i-t')\,\mathrm{d}t' - \int_{t+\epsilon(h)}^{t_{\min}} K_h(t_i-t')\,\mathrm{d}t')\epsilon_i\Bigg)_{i=1}^n \\
&+ (\mathbf{H}^\top\mathbf{H})^{-1}\mathbf{H}^\top\Bigg(2\epsilon(h)\,K_h(t_i-t)(f(t_i,\boldsymbol{x}^{(i)})-\widehat{\mu}_{t_0}(\boldsymbol{x}^{(i)}))\Bigg)_{i=1}^n + (\mathbf{H}^\top\mathbf{H})^{-1}\mathbf{H}^\top\Bigg(\int_{t_{\min}}^{t-\epsilon(h)} K_h(t_i-t') \\
&\times (\widehat{\mu}_t(\boldsymbol{x}^{(i)})-f(t_i,\boldsymbol{x}^{(i)}))\,\mathrm{d}t' + \int_{t+\epsilon(h)}^{t_{\max}} K_h(t_i-t')(\widehat{\mu}_t(\boldsymbol{x}^{(i)})-f(t_i,\boldsymbol{x}^{(i)}))\,\mathrm{d}t'\Bigg)_{i=1}^n + (\mathbf{H}^\top\mathbf{H})^{-1}\mathbf{H}^\top \\
&\times \Bigg(\int_{t_{\min}}^{t-\epsilon(h)} K_h(t_i-t')(\widehat{\mu}_{t'}(\boldsymbol{x}^{(i)})-\widehat{\mu}_{t_0}(\boldsymbol{x}^{(i)}))\,\mathrm{d}t' + \int_{t+\epsilon(h)}^{t_{\max}} K_h(t_i-t')(\widehat{\mu}_{t'}(\boldsymbol{x}^{(i)})-\widehat{\mu}_{t_0}(\boldsymbol{x}^{(i)}))\,\mathrm{d}t'\Bigg)_{i=1}^n \\
&= (\mathbf{H}^\top\mathbf{H})^{-1}\mathbf{H}^\top\Bigg((2\epsilon(h)\,K_h(t_i-t) - \int_{t_{\min}}^{t-\epsilon(h)} K_h(t_i-t')\,\mathrm{d}t' - \int_{t+\epsilon(h)}^{t_{\min}} K_h(t_i-t')\,\mathrm{d}t')\epsilon_i\Bigg)_{i=1}^n \\
&+ 2\epsilon(h)(\mathbf{H}^\top\mathbf{H})^{-1}\mathbf{H}^\top\Bigg(K_h(t_i-t)(f(t_i,\boldsymbol{x}^{(i)})-\widehat{\mu}_{t_0}(\boldsymbol{x}^{(i)}))\Bigg)_{i=1}^n + (\mathbf{H}^\top\mathbf{H})^{-1}\mathbf{H}^\top\Bigg(\int_{t_{\min}}^{t-\epsilon(h)} K_h(t_i-t') \\
&\times (\widehat{\mu}_t(\boldsymbol{x}^{(i)})-f(t_i,\boldsymbol{x}^{(i)}))\,\mathrm{d}t' + \int_{t+\epsilon(h)}^{t_{\max}} K_h(t_i-t')(\widehat{\mu}_t(\boldsymbol{x}^{(i)})-f(t_i,\boldsymbol{x}^{(i)}))\,\mathrm{d}t'\Bigg)_{i=1}^n + (\mathbf{H}^\top\mathbf{H})^{-1} \\
&\times \mathbf{H}^\top\Bigg(\int_{t_{\min}}^{t-\epsilon(h)} K_h(t_i-t')(\widehat{\mu}_{t'}(\boldsymbol{x}^{(i)})-\widehat{\mu}_{t_0}(\boldsymbol{x}^{(i)}))\,\mathrm{d}t' + \int_{t+\epsilon(h)}^{t_{\max}} K_h(t_i-t')(\widehat{\mu}_{t'}(\boldsymbol{x}^{(i)})-\widehat{\mu}_{t_0}(\boldsymbol{x}^{(i)}))\,\mathrm{d}t'\Bigg)_{i=1}^n
\end{aligned}
\tag{6.100}
$$

To simplify calculus in the following, we denote

$$
\boldsymbol{b}_{t,\epsilon} = \Bigg((2\epsilon(h)\,K_h(t_i-t) - \int_{t_{\min}}^{t-\epsilon(h)} K_h(t_i-t')\,\mathrm{d}t' - \int_{t+\epsilon(h)}^{t_{\max}} K_h(t_i-t')\,\mathrm{d}t')\epsilon_i\Bigg)_{i=1}^n \tag{6.101}
$$

the term corresponding to the bias due to error measurement, and,

$$\begin{aligned}
\boldsymbol{b}_{f,t_0,K_h} &= \Big(K_h(t_i - t)\big(f(t_i, \boldsymbol{x}^{(i)}) - \widehat{\mu}_{t_0}(\boldsymbol{x}^{(i)})\big)\Big)_{i=1}^{n}, \\
\boldsymbol{b}_{t,f,K_h} &= \Big(K_h(t_i - t)\big(\widehat{\mu}_t(\boldsymbol{x}^{(i)}) - f(t_i, \boldsymbol{x}^{(i)})\big)\Big)_{i=1}^{n}, \\
\boldsymbol{b}_{t,t_0,K_h} &= \Big(K_h(t_i - t)\big(\widehat{\mu}_t(\boldsymbol{x}^{(i)}) - \widehat{\mu}_{t_0}(\boldsymbol{x}^{(i)})\big)\Big)_{i=1}^{n}.
\end{aligned} \tag{6.102}$$

The remaining terms where the integral appears can be rearranged as follows:

$$\begin{aligned}
I &= (\mathbf{H}^\top\mathbf{H})^{-1}\mathbf{H}^\top\Big(\int_{t_{\min}}^{t-\epsilon(h)} K_h(t_i - t')\big(\widehat{\mu}_t(\boldsymbol{x}^{(i)}) - f(t_i, \boldsymbol{x}^{(i)})\big)\,\mathrm{d}t' + \int_{t+\epsilon(h)}^{t_{\max}} K_h(t_i - t')\big(\widehat{\mu}_t(\boldsymbol{x}^{(i)}) \\
&\quad - f(t_i, \boldsymbol{x}^{(i)})\big)\,\mathrm{d}t'\Big)_{i=1}^{n} + (\mathbf{H}^\top\mathbf{H})^{-1}\mathbf{H}^\top\Big(\int_{t_{\min}}^{t-\epsilon(h)} K_h(t_i - t')\big(\widehat{\mu}_{t'}(\boldsymbol{x}^{(i)}) - \widehat{\mu}_{t_0}(\boldsymbol{x}^{(i)})\big)\,\mathrm{d}t' \\
&\quad + \int_{t+\epsilon(h)}^{t_{\max}} K_h(t_i - t')\big(\widehat{\mu}_{t'}(\boldsymbol{x}^{(i)}) - \widehat{\mu}_{t_0}(\boldsymbol{x}^{(i)})\big)\,\mathrm{d}t'\Big)_{i=1}^{n} \\
&= (\mathbf{H}^\top\mathbf{H})^{-1}\mathbf{H}^\top\Big(\Big(\int_{t_{\min}}^{t-\epsilon(h)} K_h(t_i - t')\,\mathrm{d}t' + \int_{t+\epsilon(h)}^{t_{\max}} K_h(t_i - t')\,\mathrm{d}t'\Big)\big(\widehat{\mu}_t(\boldsymbol{x}^{(i)}) - \widehat{\mu}_{t_0}(\boldsymbol{x}^{(i)})\big)\Big)_{i=1}^{n} + (\mathbf{H}^\top\mathbf{H})^{-1} \\
&\quad \times \mathbf{H}^\top\Big(\int_{t_{\min}}^{t-\epsilon(h)} K_h(t_i - t')\big(\widehat{\mu}_{t'}(\boldsymbol{x}^{(i)}) - f(t_i, \boldsymbol{x}^{(i)})\big)\,\mathrm{d}t' + \int_{t+\epsilon(h)}^{t_{\max}} K_h(t_i - t')\big(\widehat{\mu}_{t'}(\boldsymbol{x}^{(i)}) - f(t_i, \boldsymbol{x}^{(i)})\big)\,\mathrm{d}t'\Big)_{i=1}^{n} \\
&= I_1 + I_2,
\end{aligned} \tag{6.103}$$

where

$$\begin{aligned}
I_1 &= (\mathbf{H}^\top\mathbf{H})^{-1}\mathbf{H}^\top\Big(\Big(\int_{t_{\min}}^{t-\epsilon(h)} K_h(t_i - t')\,\mathrm{d}t' + \int_{t+\epsilon(h)}^{t_{\max}} K_h(t_i - t')\,\mathrm{d}t'\Big)\big(\widehat{\mu}_t(\boldsymbol{x}^{(i)}) - \widehat{\mu}_{t_0}(\boldsymbol{x}^{(i)})\big)\Big)_{i=1}^{n} \\
&= (\mathbf{H}^\top\mathbf{H})^{-1}\mathbf{H}^\top\Big(\Big(\int_{t_{\min}}^{t_{\max}} K_h(t_i - t')\,\mathrm{d}t' - \int_{t-\epsilon(h)}^{t-\epsilon(h)} K_h(t_i - t')\,\mathrm{d}t'\Big)\big(\widehat{\mu}_t(\boldsymbol{x}^{(i)}) - \widehat{\mu}_{t_0}(\boldsymbol{x}^{(i)})\big)\Big)_{i=1}^{n} \\
&= (\mathbf{H}^\top\mathbf{H})^{-1}\mathbf{H}^\top\Big(\Big(\int_{t_{\min}}^{t_{\max}} K_h(t_i - t')\,\mathrm{d}t' - 2\epsilon(h)\,K_h(t_i - t)\Big)\big(\widehat{\mu}_t(\boldsymbol{x}^{(i)}) - \widehat{\mu}_{t_0}(\boldsymbol{x}^{(i)})\big)\Big)_{i=1}^{n} \\
&= (\mathbf{H}^\top\mathbf{H})^{-1}\mathbf{H}^\top\Big(\Big(1 - 2\epsilon(h)\,K_h(t_i - t)\Big)\big(\widehat{\mu}_t(\boldsymbol{x}^{(i)}) - \widehat{\mu}_{t_0}(\boldsymbol{x}^{(i)})\big)\Big)_{i=1}^{n} \\
&= (\mathbf{H}^\top\mathbf{H})^{-1}\mathbf{H}^\top\Big(\big(\widehat{\mu}_t(\boldsymbol{x}^{(i)}) - \widehat{\mu}_{t_0}(\boldsymbol{x}^{(i)})\big)\Big)_{i=1}^{n} - 2\epsilon(h)(\mathbf{H}^\top\mathbf{H})^{-1}\mathbf{H}^\top\Big(K_h(t_i - t)\big(\widehat{\mu}_t(\boldsymbol{x}^{(i)}) - \widehat{\mu}_{t_0}(\boldsymbol{x}^{(i)})\big)\Big)_{i=1}^{n} \\
&= (\mathbf{H}^\top\mathbf{H})^{-1}\mathbf{H}^\top\Big(\big(\mu_t(\boldsymbol{x}^{(i)}) - \mu_{t_0}(\boldsymbol{x}^{(i)})\big)\Big)_{i=1}^{n} - 2\epsilon(h)(\mathbf{H}^\top\mathbf{H})^{-1}\mathbf{H}^\top\Big(K_h(t_i - t)\big(\widehat{\mu}_t(\boldsymbol{x}^{(i)}) - \widehat{\mu}_{t_0}(\boldsymbol{x}^{(i)})\big)\Big)_{i=1}^{n} \\
&= \boldsymbol{\beta}^*(t) + (\mathbf{H}^\top\mathbf{H})^{-1}\mathbf{H}^\top\Big(\big(\widehat{\mu}_t(\boldsymbol{x}^{(i)}) - f(t, \boldsymbol{x}^{(i)})\big) - \big(\widehat{\mu}_{t_0}(\boldsymbol{x}^{(i)}) - f(t_0, \boldsymbol{x}^{(i)})\big)\Big)_{i=1}^{n} - 2\epsilon(h)(\mathbf{H}^\top\mathbf{H})^{-1}\mathbf{H}^\top \\
&\quad \times \Big(K_h(t_i - t)\big(\widehat{\mu}_t(\boldsymbol{x}^{(i)}) - \widehat{\mu}_{t_0}(\boldsymbol{x}^{(i)})\big)\Big)_{i=1}^{n} \\
&= \boldsymbol{\beta}^*(t) + (\mathbf{H}^\top\mathbf{H})^{-1}\mathbf{H}^\top\Big(\big(\widehat{\mu}_t(\boldsymbol{x}^{(i)}) - f(t, \boldsymbol{x}^{(i)})\big) - \big(\widehat{\mu}_{t_0}(\boldsymbol{x}^{(i)}) - f(t_0, \boldsymbol{x}^{(i)})\big)\Big)_{i=1}^{n} - 2\epsilon(h)(\mathbf{H}^\top\mathbf{H})^{-1}\mathbf{H}^\top\boldsymbol{b}_{t,t_0,K_h}
\end{aligned} \tag{6.104}$$

and

$$
\begin{aligned}
I_2 &= (\mathbf{H}^\top\mathbf{H})^{-1}\mathbf{H}^\top\Big(\int_{t_{\min}}^{t-\epsilon(h)} K_h(t_i-t')(\widehat{\mu}_{t'}(\boldsymbol{x}^{(i)})-f(t_i,\boldsymbol{x}^{(i)}))\,\mathrm{d}t' \\
&\qquad + \int_{t+\epsilon(h)}^{t_{\max}} K_h(t_i-t')(\widehat{\mu}_{t'}(\boldsymbol{x}^{(i)})-f(t_i,\boldsymbol{x}^{(i)}))\,\mathrm{d}t'\Big)_{i=1}^n \\
&= (\mathbf{H}^\top\mathbf{H})^{-1}\mathbf{H}^\top\Big(\int_{t_{\min}}^{t_{\max}} K_h(t_i-t')(\widehat{\mu}_{t'}(\boldsymbol{x}^{(i)})-f(t_i,\boldsymbol{x}^{(i)}))\,\mathrm{d}t' - \int_{t-\epsilon(h)}^{t+\epsilon(h)} K_h(t_i-t')(\widehat{\mu}_{t}(\boldsymbol{x}^{(i)})-f(t_i,\boldsymbol{x}^{(i)}))\,\mathrm{d}t'\Big)_{i=1}^n \\
&= (\mathbf{H}^\top\mathbf{H})^{-1}\mathbf{H}^\top\Big(\Big(\int_{t_{\min}}^{t-\epsilon(h)} K_h(t_i-t')\,\mathrm{d}t' + \int_{t+\epsilon(h)}^{t_{\max}} K_h(t_i-t')\,\mathrm{d}t'\Big)(\widehat{\mu}_t(\boldsymbol{x}^{(i)})-\widehat{\mu}_{t_0}(\boldsymbol{x}^{(i)}))\Big)_{i=1}^n + \\
&= (\mathbf{H}^\top\mathbf{H})^{-1}\mathbf{H}^\top\Big(\int_{t_{\min}}^{t_{\max}} K_h(t_i-t')(\widehat{\mu}_{t'}(\boldsymbol{x}^{(i)})-f(t_i,\boldsymbol{x}^{(i)}))\,\mathrm{d}t' - 2\epsilon(h)K_h(t_i-t)(\widehat{\mu}_t(\boldsymbol{x}^{(i)})-f(t_i,\boldsymbol{x}^{(i)}))\Big)_{i=1}^n \\
&= (\mathbf{H}^\top\mathbf{H})^{-1}\mathbf{H}^\top\Big(\int_{t_{\min}}^{t_{\max}} K_h(t_i-t')(\widehat{\mu}_{t'}(\boldsymbol{x}^{(i)})-f(t_i,\boldsymbol{x}^{(i)}))\,\mathrm{d}t')\Big)_{i=1}^n - 2\epsilon(h)(\mathbf{H}^\top\mathbf{H})^{-1}\mathbf{H}^\top\boldsymbol{b}_{t,f,K_h}.
\end{aligned}
\tag{6.105}
$$

By gathering all previous terms:

$$
\begin{aligned}
\widehat{\boldsymbol{\beta}}_h^X(t) &= \boldsymbol{\beta}^*(t) + (\mathbf{H}^\top\mathbf{H})^{-1}\mathbf{H}^\top\Big((\widehat{\mu}_t(\boldsymbol{x}^{(i)})-f(t,\boldsymbol{x}^{(i)})) - (\widehat{\mu}_{t_0}(\boldsymbol{x}^{(i)})-f(t_0,\boldsymbol{x}^{(i)}))\Big)_{i=1}^n \\
&\quad + (\mathbf{H}^\top\mathbf{H})^{-1}\mathbf{H}^\top\boldsymbol{b}_{t,\epsilon} + 2\epsilon(h)(\mathbf{H}^\top\mathbf{H})^{-1}\mathbf{H}^\top(\boldsymbol{b}_{f,t_0,K_h} - \boldsymbol{b}^X_{t,t_0,K_h} - \boldsymbol{b}_{t,f,K_h}) \\
&\qquad + (\mathbf{H}^\top\mathbf{H})^{-1}\mathbf{H}^\top\Big(\int_{t_{\min}}^{t_{\max}} K_h(t_i-t')(\widehat{\mu}_{t'}(\boldsymbol{x}^{(i)})-f(t_i,\boldsymbol{x}^{(i)}))\,\mathrm{d}t')\Big)_{i=1}^n \\
&= \boldsymbol{\beta}^*(t) + (\mathbf{H}^\top\mathbf{H})^{-1}\mathbf{H}^\top\boldsymbol{b}_{t,\epsilon} - 4\epsilon(h)(\mathbf{H}^\top\mathbf{H})^{-1}\mathbf{H}^\top\boldsymbol{b}_{f,t,K_h} + (\mathbf{H}^\top\mathbf{H})^{-1}\mathbf{H}^\top\Big((\widehat{\mu}_t(\boldsymbol{x}^{(i)})-f(t,\boldsymbol{x}^{(i)})) \\
&\qquad - (\widehat{\mu}_{t_0}(\boldsymbol{x}^{(i)})-f(t_0,\boldsymbol{x}^{(i)})) + \int_{t_{\min}}^{t_{\max}} K_h(t_i-t')(\widehat{\mu}_{t'}(\boldsymbol{x}^{(i)})-f(t_i,\boldsymbol{x}^{(i)}))\,\mathrm{d}t')\Big)_{i=1}^n.
\end{aligned}
\tag{6.106}
$$

As seen previously with the M- and DR-learners, the bias terms are either in $\mathcal{O}(1)$ or $\mathcal{O}(h^2)$. It is sufficient to choose $\epsilon(h) = o(h^2)$ to neglect the bias term $\epsilon(h)(\mathbf{H}^\top\mathbf{H})^{-1}\mathbf{H}^\top\boldsymbol{b}_{f,t,K_h}$ that involves $\epsilon(h)$. This choice does not exclude other possible choices, typically in $o(h)$ or $o(1)$. However, our purpose is to eliminate as much as possible all bias with order term below $h^2$. The choice of $\epsilon(h) = o(h^2)$ is reasonable for that.

Under this condition, we can write:

$$
\boldsymbol{b}_{t,\epsilon} \approx -\Big(\int_{t_{\min}}^{t_{\max}} K_h(t_i-t')\,\mathrm{d}t'\,\epsilon_i\Big)_{i=1}^n = -\boldsymbol{\epsilon}. \tag{6.107}
$$

Therefore,

$$
\begin{aligned}
\widehat{\boldsymbol{\beta}}_h^X(t) &= \boldsymbol{\beta}^*(t) - (\mathbf{H}^\top\mathbf{H})^{-1}\mathbf{H}^\top\boldsymbol{\epsilon} + (\mathbf{H}^\top\mathbf{H})^{-1}\mathbf{H}^\top\Big((\widehat{\mu}_t(\boldsymbol{x}^{(i)})-f(t,\boldsymbol{x}^{(i)})) - (\widehat{\mu}_{t_0}(\boldsymbol{x}^{(i)})-f(t_0,\boldsymbol{x}^{(i)})) \\
&\qquad + \int_{t_{\min}}^{t_{\max}} K_h(t_i-t')(\widehat{\mu}_{t'}(\boldsymbol{x}^{(i)})-f(t_i,\boldsymbol{x}^{(i)}))\,\mathrm{d}t'\Big)_{i=1}^n \\
&= \boldsymbol{\beta}^*(t) + (\mathbf{H}^\top\mathbf{H})^{-1}\mathbf{H}^\top\boldsymbol{b}^X_{t,spec} + (\mathbf{H}^\top\mathbf{H})^{-1}\mathbf{H}^\top\boldsymbol{b}^X_{t,spec,K,h} - (\mathbf{H}^\top\mathbf{H})^{-1}\mathbf{H}^\top\boldsymbol{\epsilon},
\end{aligned}
\tag{6.108}
$$

where

$$\boldsymbol{b}^X_{t,spec} = \Big( (\widehat{\mu}_t(\boldsymbol{x}^{(i)}) - f(t, \boldsymbol{x}^{(i)})) - (\widehat{\mu}_{t_0}(\boldsymbol{x}^{(i)}) - f(t_0, \boldsymbol{x}^{(i)})) \Big)_{i=1}^{n}, \tag{6.109}$$

and

$$\boldsymbol{b}^X_{t,spec,K,h} = \Big( \int_{t_{\min}}^{t_{\max}} K_h(t_i - t')(\widehat{\mu}_{t'}(\boldsymbol{x}^{(i)}) - f(t_i, \boldsymbol{x}^{(i)}))\, \mathrm{d}t' \Big)_{i=1}^{n}, \tag{6.110}$$

are the bias terms due to the misspecification of the outcome model estimator $\widehat{\mu}$.

Since $\mathbb{E}(\epsilon) = 0$, the previous equation allows us to write, for $j = 1, \ldots, p$.

$$h_{t,j} = h^M_{t,spec,j} + h_{t,spec,K,j}. \tag{6.111}$$

.

**Lemma 6.4.15.** *For $j = 1, \ldots, p$*

$$h^X_{t,j} = h^X_{t,spec,j} + h^X_{t,spec,K,j}, \tag{6.112}$$

*where $h^X_{t,spec,j}$ and $h^X_{t,spec,K,h}$ are the model's misspecification bias such that*

$$h^X_{t,spec,j} = \mathbb{E}\Big[f_j(\boldsymbol{X})(\widehat{\mu}_t(\boldsymbol{X}) - f(t, \boldsymbol{X}))\Big] + \mathbb{E}\Big[f_j(\boldsymbol{X})(\widehat{\mu}_{t_0}(\boldsymbol{X}) - f(t, \boldsymbol{X}))\Big], \tag{6.113}$$

*and*

$$\begin{aligned} h^X_{t,spec,K,h,j} = \mathbb{E}\Big[f_j(\boldsymbol{X}) \int \Big((\widehat{\mu}_{t'}(\boldsymbol{X}) - f(t', \boldsymbol{X})\Big) f_{T|\boldsymbol{X}}(t' \mid \boldsymbol{X})\, \mathrm{d}t'\Big] - h^2 \kappa_2(K) \\ \times \mathbb{E}\Big[f_j(\boldsymbol{X}) \int_{t_{\min}}^{t_{\max}} C^{(2)}_{t'}(\boldsymbol{X})\, \mathrm{d}t'\Big] + o(h^2), \end{aligned} \tag{6.114}$$

*with*

$$\begin{aligned} C^{(2)}_{t'}(\boldsymbol{X}) = \frac{\partial f}{\partial t}(t', \boldsymbol{X}) \frac{\partial f_{T|\boldsymbol{X}}}{\partial t}(t' \mid \boldsymbol{X}) + \frac{1}{2} f_{T|\boldsymbol{X}}(t' \mid \boldsymbol{X}) \frac{\partial^2 f}{\partial t^2}(t', \boldsymbol{X}) \\ + \frac{1}{2} \frac{\partial^2 f_{T|\boldsymbol{X}}}{\partial t^2}(t' \mid \boldsymbol{X}) (f(t', \boldsymbol{X}) - \widehat{\mu}_{t'}(\boldsymbol{X})). \end{aligned} \tag{6.115}$$

*Proof.* The specification term $h^X_{t,spec,j}$ can be computed easily as

$$h^X_{t,spec,j} = \mathbb{E}\Big[f_j(\boldsymbol{X})(\widehat{\mu}_t(\boldsymbol{X}) - f(t, \boldsymbol{X}))\Big] + \mathbb{E}\Big[f_j(\boldsymbol{X})(\widehat{\mu}_{t_0}(\boldsymbol{X}) - f(t, \boldsymbol{X}))\Big]. \tag{6.116}$$

For the other specification term, we show that

$$\begin{aligned} h^X{}_{t,spec,K,h,j} &= \mathbb{E}\Big[f_j(\boldsymbol{X})\Big( \int K_h(T - t')(\widehat{\mu}_{t'}(\boldsymbol{X}) - f(T, \boldsymbol{X}))\, \mathrm{d}t'\Big)\Big] \\ &= \mathbb{E}\Big[f_j(\boldsymbol{X}) \int \mathbb{E}\Big[K_h(T - t')(\widehat{\mu}_{t'}(\boldsymbol{X}) - f(T, \boldsymbol{X}) \mid \boldsymbol{X}\Big]\, \mathrm{d}t'\Big] \\ &= \mathbb{E}\Big[f_j(\boldsymbol{X}) \int \Big( \int K_h(s - t')(\widehat{\mu}_{t'}(\boldsymbol{X}) - f(s, \boldsymbol{X})\Big) f_{T|\boldsymbol{X}}(s \mid \boldsymbol{X})\, \mathrm{d}s\Big)\, \mathrm{d}t'\Big] \\ &\overset{u=(s-t')/h}{=} \mathbb{E}\Big[f_j(\boldsymbol{X}) \int \Big( \int K(u)(\widehat{\mu}_{t'}(\boldsymbol{X}) - f(t' + uh, \boldsymbol{X})\Big) f_{T|\boldsymbol{X}}(t' + uh \mid \boldsymbol{X})\, \mathrm{d}u\Big)\, \mathrm{d}t'\Big] \\ &\overset{h\approx 0}{=} \mathbb{E}\Big[f_j(\boldsymbol{X}) \int \Big( \int_{\mathbb{R}} K(u)(\widehat{\mu}_{t'}(\boldsymbol{X}) - f(t, \boldsymbol{X}) - hu \frac{\partial f}{\partial t}(t, \boldsymbol{X}) - \frac{(hu)^2}{2} \frac{\partial^2 f}{\partial t^2}(t, \boldsymbol{X}) + (hu)^2 \varepsilon_{1,\boldsymbol{X}}(uh)\Big) \\ &\qquad \times \Big( f_{T|\boldsymbol{X}}(t \mid \boldsymbol{X}) + hu \frac{\partial f_{T|\boldsymbol{X}}}{\partial t}(t \mid \boldsymbol{X}) + \frac{(hu)^2}{2} \frac{\partial^2 f_{T|\boldsymbol{X}}}{\partial t^2}(t \mid \boldsymbol{X}) + (hu)^2 \varepsilon_{2,\boldsymbol{X}}(uh)\Big)\, \mathrm{d}s\Big)\, \mathrm{d}t'\Big]. \end{aligned} \tag{6.117}$$

By similar calculus to what have been done for the M- and DR-learners, we show that

$$\begin{aligned} h^X{}_{t,spec,K,h,j} &= \mathbb{E}\Big[f_j(\boldsymbol{X}) \int \Big( (\widehat{\mu}_{t'}(\boldsymbol{X}) - f(t', \boldsymbol{X})) f_{T|\boldsymbol{X}}(t' \mid \boldsymbol{X}) - h^2 \kappa_2(K) \Big( \frac{\partial f}{\partial t}(t', \boldsymbol{X}) \frac{\partial f_{T|\boldsymbol{X}}}{\partial t}(t' \mid \boldsymbol{X}) \\ &\quad + \frac{1}{2} f_{T|\boldsymbol{X}}(t' \mid \boldsymbol{X}) \frac{\partial^2 f}{\partial t^2}(t', \boldsymbol{X}) + \frac{1}{2} \frac{\partial^2 f_{T|\boldsymbol{X}}}{\partial t^2}(t' \mid \boldsymbol{X}) (f(t', \boldsymbol{X}) - \widehat{\mu}_{t'}(\boldsymbol{X})) \Big) \, \mathrm{d}t' + o(h^2) \Big] \\ &= \mathbb{E}\Big[f_j(\boldsymbol{X}) \int \Big( \widehat{\mu}_{t'}(\boldsymbol{X}) - f(t', \boldsymbol{X}) \Big) f_{T|\boldsymbol{X}}(t' \mid \boldsymbol{X}) \, \mathrm{d}t' \Big] - h^2 \kappa_2(K) \mathbb{E}\Big[ f_j(\boldsymbol{X}) \int C^{(2)}_{t'}(\boldsymbol{X}) \, \mathrm{d}t' \Big] \\ &\quad + o(h^2), \end{aligned} \tag{6.118}$$

where

$$\begin{aligned} C^{(2)}_{t'}(\boldsymbol{X}) = \frac{\partial f}{\partial t}(t', \boldsymbol{X}) \frac{\partial f_{T|\boldsymbol{X}}}{\partial t}(t' \mid \boldsymbol{X}) + \frac{1}{2} f_{T|\boldsymbol{X}}(t' \mid \boldsymbol{X}) \frac{\partial^2 f}{\partial t^2}(t', \boldsymbol{X}) \\ + \frac{1}{2} \frac{\partial^2 f_{T|\boldsymbol{X}}}{\partial t^2}(t' \mid \boldsymbol{X}) (f(t', \boldsymbol{X}) - \widehat{\mu}_{t'}(\boldsymbol{X})). \end{aligned} \tag{6.119}$$

■

**Lemma 6.4.16.** *The entries of the covariance matrix $\mathbf{C}^X$ satisfy:*

$$(\mathbf{C}^X)_{jj'} = \begin{cases} C_1^X + o(1) & \text{if } j, j' \in \{1, \dots, p\}, \\ C_2 & \text{if } j, j' \in \{p+1, \dots, p^2\}, \\ C^X_{spec} + o(1) & \text{otherwise.} \end{cases} \tag{6.120}$$

*where $C_1^X, C_2$ and $C^X_{spec}$ are some given terms such that $C_1^X \neq 0$.*

*Proof.* For $j, j' \in \{p+1, \dots, p^2\}$, the term $(\mathbf{C}^X_{t,spec})_{jj'}$ can be written as

$$\begin{aligned} (\mathbf{C}^X_{t,spec})_{jj'} &= \mathbb{E}\Big[ f_j(\boldsymbol{X}) f_{j'}(\boldsymbol{X}) B^2_{t,spec} \Big] \\ &= \mathbb{E}\Big[ f_j(\boldsymbol{X}) f_{j'}(\boldsymbol{X}) (\widehat{\mu}_t(\boldsymbol{X}) - f(t, \boldsymbol{X}))^2 \Big] + \mathbb{E}\Big[ f_j(\boldsymbol{X}) f_{j'}(\boldsymbol{X}) (\widehat{\mu}_{t_0}(\boldsymbol{X}) - f(t_0, \boldsymbol{X}))^2 \Big] \\ &\quad - 2\mathbb{E}\Big[ f_j(\boldsymbol{X}) f_{j'}(\boldsymbol{X}) (\widehat{\mu}_t(\boldsymbol{X}) - f(t, \boldsymbol{X})) (\widehat{\mu}_{t_0}(\boldsymbol{X}) - f(t_0, \boldsymbol{X})) \Big]. \end{aligned} \tag{6.121}$$

For the second term $(\mathbf{C}^X_{t,spec,K,h})_{jj'}$ and considering that we want to collect only expansion terms with order lower than $o(1)$, we have

$$\begin{aligned} (\mathbf{C}^X_{t,spec,K,h})_{jj'} &= \mathbb{E}\Big[ f_j(\boldsymbol{X}) f_{j'}(\boldsymbol{X}) B^2_{t,spec,K,h} \Big] \\ &= \mathbb{E}\Big[ f_j(\boldsymbol{X}) f_{j'}(\boldsymbol{X}) \Big( \int K_h(T - t') (\widehat{\mu}_{t'}(\boldsymbol{X}) - f(T, \boldsymbol{X})) \, \mathrm{d}t' \Big)^2 \Big] \\ &\overset{u=(T-t')/h}{=} \mathbb{E}\Big[ f_j(\boldsymbol{X}) f_{j'}(\boldsymbol{X}) \Big( \int K(u) (\widehat{\mu}_{T-hu}(\boldsymbol{X}) - f(T, \boldsymbol{X})) \, \mathrm{d}u \Big)^2 \Big] \end{aligned} \tag{6.122}$$

The function $\widehat{\mu}$ is continuous on $t$ by Assumption 6.4.4 and uniformly bounded on $(t, \boldsymbol{x})$. Thus, by the dominated convergence theorem:

$$(\mathbf{C}^X_{t,spec,K,h})_{jj'} \xrightarrow{h \to 0} \mathbb{E}\Big[ f_j(\boldsymbol{X}) f_{j'}(\boldsymbol{X}) \Big( \int_{\mathbb{R}} K(u) (\widehat{\mu}_T(\boldsymbol{X}) - f(T, \boldsymbol{X})) \, \mathrm{d}u \Big)^2 \Big]. \tag{6.123}$$

Therefore, we can write:

$$(\mathbf{C}^X_{t,spec,K,h})_{jj'} = \mathbb{E}\Big[f_j(\boldsymbol{X})f_{j'}(\boldsymbol{X})\big(\widehat{\mu}_T(\boldsymbol{X}) - f(T,\boldsymbol{X})\big)^2\Big] + o(1). \tag{6.124}$$

For the covariance term $(\mathbf{C}^X_{t,K,spec})_{jj'}$ between $B^X_{t,spec}$ and $B^X_{t,spec,K,h}$:

$$\begin{aligned}(\mathbf{C}^X_{t,K,spec})_{jj'} &= \mathbb{E}\Big[f_j(\boldsymbol{X})f_{j'}(\boldsymbol{X})B^X_{t,spec}B^X_{t,spec,K,h}\Big] \\ &= \mathbb{E}\Big[f_j(\boldsymbol{X})f_{j'}(\boldsymbol{X})B^X_{t,spec}\mathbb{E}\Big[B^X_{t,spec,K,h} \mid \boldsymbol{X}\Big]\Big] \\ &= \mathbb{E}\Big[f_j(\boldsymbol{X})f_{j'}(\boldsymbol{X})\int\Big(\widehat{\mu}_{t'}(\boldsymbol{X}) - f(t',\boldsymbol{X})\Big)f_{T|\boldsymbol{X}}(t' \mid \boldsymbol{X})\,\mathrm{d}t'\Big] + o(1).\end{aligned} \tag{6.125}$$

Finally, the covariance term $(\mathbf{C}^X_{t,\epsilon})_{jj'}$ corresponds to

$$(\mathbf{C}^X_{t,\epsilon})_{jj'} = \sigma^2. \tag{6.126}$$

If $j \in \{p+1,\dots,p^2\}$ and $j' \in \{1,\dots,p\}$ (or inversely by symmetry), then

$$\begin{aligned}\mathbf{C}^X_{j,j'} &= \mathbb{E}[f_k(\boldsymbol{X})f_{k'}(\boldsymbol{X})(b^X_{t,spec} + b^X_{t,spec,K,h})] \\ &= \mathbb{E}[f_k(\boldsymbol{X})f_{k'}(\boldsymbol{X})b^X_{t,spec}] + \mathbb{E}[f_k(\boldsymbol{X})f_{k'}(\boldsymbol{X})b^X_{t,spec,K,h}] \\ &\qquad \text{(By similar calculus to } h_{t,spec,j} \text{ and } h_{t,spec,K,h,j}\text{)} \\ &= \mathbb{E}\Big[f_k(\boldsymbol{X})f_{k'}(\boldsymbol{X})\Big(\int\Big(\widehat{\mu}_{t'}(\boldsymbol{X}) - f(t',\boldsymbol{X})\Big)f_{T|\boldsymbol{X}}(t' \mid \boldsymbol{X})\,\mathrm{d}t'\Big) \\ &\qquad\qquad + \big(\widehat{\mu}_t(\boldsymbol{X}) - f(t,\boldsymbol{X})\big) - \big(\widehat{\mu}_{t_0}(\boldsymbol{X}) - f(t_0,\boldsymbol{X})\big)\Big] + o(1) \\ &= C^X_{spec} + o(1),\end{aligned} \tag{6.127}$$

where $C^X_{spec}$ is the misspecification covariance term with $C^X_{spec} = 0$ if the conditional density estimator is well-specified for all $t \in \mathcal{T}$.

Thus, by gathering all the previous terms,

$$(\mathbf{C}^X)_{jj'} = \begin{cases} C^X_1 + o(1) & \text{if } j,j' \in \{1,\dots,p\} \\ C_2 & \text{if } j,j' \in \{p+1,\dots,p^2\} \\ C^X_{spec} + o(1) & \text{otherwise.}\end{cases} \tag{6.128}$$

with

$$\begin{aligned}C^X_1 = \mathbb{E}\Big[f_j(\boldsymbol{X})f_{j'}(\boldsymbol{X})\big(\widehat{\mu}_t(\boldsymbol{X}) - f(t,\boldsymbol{X})\big)^2\Big] + \mathbb{E}\Big[f_j(\boldsymbol{X})f_{j'}(\boldsymbol{X})\big(\widehat{\mu}_{t_0}(\boldsymbol{X}) - f(t_0,\boldsymbol{X})\big)^2\Big] \\ - 2\,\mathbb{E}\Big[f_j(\boldsymbol{X})f_{j'}(\boldsymbol{X})\big(\widehat{\mu}_t(\boldsymbol{X}) - f(t,\boldsymbol{X})\big)\big(\widehat{\mu}_{t_0}(\boldsymbol{X}) - f(t_0,\boldsymbol{X})\big)\Big] \\ + \mathbb{E}\Big[f_j(\boldsymbol{X})f_{j'}(\boldsymbol{X})\big(\widehat{\mu}_T(\boldsymbol{X}) - f(T,\boldsymbol{X})\big)^2\Big].\end{aligned} \tag{6.129}$$

■

**Proposition 6.4.17.** *If the outcome model $\widehat{\mu}$ is well-specified, then the estimator $\widehat{\boldsymbol{\beta}}_h^X(t)$ has bias and variance such that*

$$\begin{aligned} \mathbb{E}(\widehat{\boldsymbol{\beta}}_h^X(t)) &\approx \boldsymbol{\beta}(t) + h^2 \mathbf{F}^{-1} \boldsymbol{h}_t^X, \\ \mathbb{V}(\widehat{\boldsymbol{\beta}}_h^X(t)) &\approx \frac{1}{n}\gamma(h)\mathbf{C}, \end{aligned} \tag{6.130}$$

*where, for $j \in \{1, \dots, p\}$,*

$$h_{t,j}^X = \kappa_2(K)\mathbb{E}\Big[f_j(\boldsymbol{X}) \int C_{t'}(\boldsymbol{X})\, \mathrm{d}t'\Big], \tag{6.131}$$

*with*

$$C_t(\boldsymbol{X}) = \frac{\partial f}{\partial t}(t, \boldsymbol{X}) \frac{\partial f_{T|\boldsymbol{X}}}{\partial t}(t \mid \boldsymbol{X}) + \frac{1}{2} f_{T|\boldsymbol{X}}(t \mid \boldsymbol{X}) \frac{\partial^2 f}{\partial t^2}(t, \boldsymbol{X}), \tag{6.132}$$

*$\mathbf{C}$ is a fixed matrix and $\gamma : \mathbb{R} \to \mathbb{R}$ is a function such that $\gamma(h) \to 0$ when $h \to 0$.*

*Proof.* Based on Lemmas 6.4.15 and 6.4.16, and similar to the proof of Proposition 6.4.9. More precisely, when the outcome $\widehat{\mu}_t(\boldsymbol{X})$ is well specified then $\widehat{\mu}_t(\boldsymbol{x}) = f(t, \boldsymbol{x})$ and most terms become equal to zero. In the bias term $\boldsymbol{h}_t^X$, we obtain $C_{t'}^{(2)}(\boldsymbol{X}) = C_{t'}(\boldsymbol{X})$ where $C_{t'}(\boldsymbol{X})$ is given in (6.45). For the covariance term, we have simply $C_1^X = 0$ and therefore only terms in $o(1)$ remain in $(\mathbf{C}^X)_{jj'}$ for $j, j' \in \{1, \dots, p\}$. ■

**Discussion** The first result of the M- and DR-learners is also valid for the X-learner: Kernels methods induce a bias that is in $\mathcal{O}(h^2)$. Furthermore, the X-learner is likely to have the lowest variance compared to the M- and DR-learners. However, the comparison of the bias between X-learner and M- and DR-learners seems to be more challenging.

**Proposition 6.4.18.** *Under all previous assumptions, the asymptotic Mean Squared Error (MSE) of the X-Learner $\widehat{\boldsymbol{\beta}}_h^X(t)$ is given by*

$$MSE(\widehat{\boldsymbol{\beta}}_h^X(t)) = h^4 \|\mathbf{F}^{-1}\boldsymbol{b}^X\|^2 + \frac{\gamma(h)}{n} \operatorname{Tr}(\mathbf{C}), \tag{6.133}$$

*and the optimal bandwidth $h_t^*$ that minimizes the asymptotic MSE $h_t^* = 0$.*

Proposition 6.4.18 proves that the X-learner with kernel density methods does not bring any additional value. In addition, when considering the pseudo-outcome $Z_{t,h}^X$ with $\epsilon(h) = 0$ (or $h = 0$ if $\epsilon(0) = 0$, with an abuse of the notation), one gets $\mathbb{E}(Z_{t,h}^X \mid \boldsymbol{X}) = \mu_t(\boldsymbol{X}) - \mu_{t_0}(\boldsymbol{X})$. In conclusion: the optimal X-learner in continuous treatments is no more than a simple S-learner.

## 6.5 Discussion of the R-learner in the continuous treatment setting.

In this section, we consider the extension of the R-learner to continuous treatments, and we discuss some ideas elaborated in Zhang et al. (2022) work.

The generalization of the Robinson (1988) decomposition from multi-treatments to a continuous treatment is natural. Indeed, if we replace $\sum_{k=1}^K$ by $\int_{t \in \mathcal{T}}$ in Equation (5.2.6), we get:

$$Y_{\text{obs}} - m(\boldsymbol{X}) = \tau_T(\boldsymbol{X}) - \int_{t \in \mathcal{T}} \tau_t(\boldsymbol{X}) f_{T|\boldsymbol{X}}(t \mid \boldsymbol{X}) dt + \epsilon, \tag{6.134}$$

where $m(\boldsymbol{x}) = \mathbb{E}(Y_{\text{obs}} \mid \boldsymbol{X} = \boldsymbol{x})$ is the observed outcome model, and $\epsilon$ is the error and satisfies $\mathbb{E}(\epsilon \mid \boldsymbol{X}) = 0$ (Neyman Orthogonality). We may refer to the same paper for this extension.

With the previous equation, we define the generalized R-loss function as:

$$\ell_R(\overline{\tau}) = \mathbb{E}\Big[\big(Y_{\text{obs}} - m(\boldsymbol{X}) - \overline{\tau}_T(\boldsymbol{X}) + \int_{t\in\mathcal{T}} \overline{\tau}_t(\boldsymbol{X}) f_{T|\boldsymbol{X}}(t \mid \boldsymbol{X}) dt\big)^2\Big]. =, \tag{6.135}$$

for some function $\overline{\tau}$.

The generalized R-loss function $\ell_R$ has two main issues: On the one hand, and in contrast to binary or multi-treatment settings, it is impossible to solve this problem separately for each level $t \in \mathcal{T}$ as it would require estimating an infinite number of models $\{\widehat{\tau}_t\}_{t\neq t_0}$ in one problem. Instead, one must consider bi-variate functions $\overline{\tau} : \mathcal{T} \times \mathcal{D} \to \mathbb{R}$ and minimize the generalized R-loss function with respect to this functional form of $\overline{\tau}$. On the other hand, the problem of the non-identifiability of $\tau(t, \boldsymbol{x})$ for a given $t \in \mathcal{T}$ and $\boldsymbol{x} \in \mathcal{D}$ shows up. It has been shown by Zhang et al. (2022) that, if $\tau(t, \boldsymbol{x})$ is the true CATE, then all functions of the form $\tau(t, \boldsymbol{x}) + s(\boldsymbol{x})$ for a function $s$ with a finite second moment (i.e. $\|s\|_{\mathcal{L}^2} = \mathbb{E}(s^2(\boldsymbol{X})) < +\infty$) are also solution to the R-loss minimization problem.

To overcome the problem of the non-identifiability, Zhang et al. (2022) propose Tikhonov et al. (1995) regularization to the generalized R-loss function $\ell_R$ and define the $\ell_2$-penalized loss as:

$$\ell(\tau \mid \rho) = \ell_R(\tau) + \rho\|\tau\|^2_{\mathcal{L}^2}, \tag{6.136}$$

where $\rho$ is a penalty term and $\|\cdot\|^2_{\mathcal{L}^2}$ is the $\mathcal{L}_2$ norm. The so-proposed $\ell_2$-penalized R-learner identifies efficiently and uniquely the true CATE $\tau(t, \boldsymbol{x})$ (Zhang et al., 2022).

Another alternative to R-learning to continuous treatments is proposed by Kaddour et al. (2021). The approach considers both Assumptions 6.4.1 and 5.3.2 on the outcome $Y(t) = \boldsymbol{f}(\boldsymbol{X})^\top \boldsymbol{\beta}(t) + \epsilon$, then established the binarized Robinson (1988) decomposition such that

$$Y_{\text{obs}} - m(\boldsymbol{X}) = \boldsymbol{f}(\boldsymbol{X})^\top (\boldsymbol{\beta}(T) - e^{\boldsymbol{\beta}}(\boldsymbol{X})) + \epsilon, \tag{6.137}$$

where $m(\boldsymbol{x}) = \mathbb{E}(Y_{\text{obs}} \mid \boldsymbol{X} = \boldsymbol{x})$ and $e^{\boldsymbol{\beta}}(\boldsymbol{x}) = \mathbb{E}(\boldsymbol{\beta}(T) \mid \boldsymbol{X} = \boldsymbol{x})$.

Considering the mean squared error of $\epsilon$ as loss function and minimizing it allows us to identify the optimal functions $\widehat{\boldsymbol{f}}$ and $\widehat{\boldsymbol{\beta}}$ and therefore the CATE $\tau_t(\boldsymbol{X}) = \boldsymbol{f}(\boldsymbol{X})^\top (\boldsymbol{\beta}(t) - \boldsymbol{\beta}(t_0))$. One needs to solve the following problem:

$$\widehat{\boldsymbol{f}}, \widehat{\boldsymbol{\beta}} = \text{argmin}_{\boldsymbol{f},\boldsymbol{\beta}\in\mathcal{F}} \ \frac{1}{n}\sum_{i=1}^{n} \Big[ \Big(y_i - \widehat{m}(\boldsymbol{x}^{(i)})\Big) - \boldsymbol{f}(\boldsymbol{x}^{(i)})^\top (\boldsymbol{\beta}(t_i) - e^{\boldsymbol{\beta}}(\boldsymbol{x}^{(i)}))\Big], \tag{6.138}$$

where $e^{\boldsymbol{\beta}}(\boldsymbol{x}) = \mathbb{E}(\boldsymbol{\beta}(T) \mid \boldsymbol{X} = \boldsymbol{x})$ and $\mathcal{F}$ is the space of candidate models $\boldsymbol{f}$ and $\boldsymbol{\beta}$. In the end, although this approach is simple, it is computationally heavy. It also requires specifying the family of models $\mathcal{F}$ and precise the dimension $p$ for the assumption 6.4.2.

## 6.6 Discussion

In this chapter, we extended heterogeneous treatment effects estimation to continuous treatment. Using the adapted framework of potential outcomes under continuous treatments and kernel density methods, we developed pseudo-outcome meta-learners (M-, DR- and X-learners) under

continuous treatments. We performed a bias-variance analysis of this class of meta-learners. We compared their behavior with a super S-learner that would have learned to estimate heterogeneous treatment effects without relying on pseudo-outcome representations. Our results are surprising: For all pseudo-outcome meta-learners, kernel density methods induce an estimation bias, and this bias cannot be avoided and is intrinsic to kernels. The variance of both M- and DR-learner may increase significantly because of the bandwidth of the kernel and the lower bound of the generalized propensity score. The X-learner would perform as the S-learner in terms of variance but still suffers from induced bias. This claim is proper only when the outcome model is well-specified. The following question remains unanswered: under which conditions of a misspecified $\widehat{\mu}$ would the X-learner have an advantage over the S-learner? For the moment and unless proven otherwise, we recommend focusing on efficiently estimating heterogeneous treatment effects for continuous treatments using S-learning instead of relying on pseudo-outcome representations. This conclusion opens new perspectives on new methods and approaches for building efficient S-learners. One of them could be, for example, the binarized R-loss function in (6.137).

# CHAPTER 7

# Conclusion and Perspectives

## Conclusion

The interpretability of Machine Learning models in Energy industry motivated the work in this thesis. Beyond prediction, it was necessary to estimate the uncertainty of predictions and answer causal questions.

The objective of this thesis was twofold: The first goal was to develop a new method for uncertainty quantification for a misspecified Gaussian Process model. We were able to construct reliable prediction intervals with respect to some coverage and confidence criteria. This method was applied successfully to two real cases: natural gas fields and battery charging capacities. It could also be used for more generic problems for industrial and energy systems where a decline in production capacity may be observed in time.

The second goal was in the context of Causal Inference with multi-valued and continuous treatments. We developed frameworks and estimators for inferring the heterogeneity of treatment/intervention effects. However, because of the fundamental problem of Causal Inference (i.e. the counterfactual outcomes are unobserved, and only the factual outcome corresponding to the intervention are observed), we could not validate our estimators on a real-world dataset. We created, therefore, a semi-synthetic dataset simulating an enhanced geothermal system for this purpose. The ground truth effect of causal effects is known and used to assess different estimators. We provided some statistical guarantees and elaborated a detailed discussion about the use of meta-learners when inferring heterogeneous treatment effects under various circumstances and conditions. The possible real-case applications of this work include evaluating the impact of different insulation materials on energy efficiency and the effect of solar cell type on the energy storage of a solar panel. More generally, it can cover any application where the causal impact of a variable is crucial to optimize the outcome of the system (and also for decision-making).

The next step could be to combine the two parts of the thesis to address the issue of risk management and reliability in heterogeneous treatment effect inference. Indeed, providing causal estimations with a significance level is crucial to guarantee sufficient evidence and confidence for reliable decision-making. The Gaussian Process model seems to be the reasonable choice to tackle this problem. Nevertheless, treatment effects estimation can be seen as a missing-data problem (i.e. an extrapolation problem for the treatment $T$), whereas the Gaussian Process model is basically meant for interpolation. This issue would necessitate significant considerations in modelling the covariance function to estimate causal treatment effects with

their uncertainty. Another possibility is to consider the transposition of the Gaussian Process model from the learning domain to the extrapolation domain.

Despite the contributions to Gaussian process modelling and causal inference made in this thesis, several issues are not addressed. They are discussed in the following sections:

## On uncertainty quantification with Gaussian Process

Firstly, in our work, we did not consider the asymptotic properties of the RPIE method. The asymptotic results on an expansion-domain or in a fixed-domain (Bachoc, 2013; Stein, 1999) of the method may provide additional information about the consistency of the Leave-One-Out Coverage Probability and its convergence rate. Secondly, as already discussed in the conclusion of Chapter 3, the approach at this stage only considers continuous inputs. An extension with categorical and non-continuous variables should be developed in the future. Thirdly, the evaluation of the RPIE method on a new validation set is based on the assumption of random sampling (i.e. the training and the validation sets have the same distribution). In the case of sequential experimental designs, this could be problematic and compromises the RPIE method. Finally, it may be worthwhile to investigate the influence of outliers on misspecified models and the misspecification of the errors' $\epsilon$ distribution on the RPIE method.

## On Causal inference and treatment effects estimation

Although its efficiency and popularity, the Rubin Causal model has some critical limitations, primarily due to the untestable nature of causal assumptions (i.e. the unconfoundedness, Stable Unit Treatment Value Assumption).

The violation of these assumptions compromises the estimation of causal effects. The lack of unconfoundedness would lead to biased causal effects. Quantifying the influence of unmeasured confounding on estimating treatment effects should be considered in the future. Some recent works, namely Marginal Sensitivity Model (Jin et al., 2021; Kallus et al., 2019; Yin et al., 2022) has gained popularity for unmeasured confounding effects. It presents an exciting tool for characterizing the strength of unmeasured confounding necessary to explain causal effects estimands.

Another challenge we have not considered in the thesis is to train and optimize the S-learner: In addition to the binarized R-loss, a perspective could be developing specific representations of the covariates that will lead to a deconfounded S-learner. One of them is the Invariant Risk Minimization (Arjovsky et al., 2019; Shi et al., 2021).

Finally, in a completely different framework to this thesis, the SUTVA assumption may be relaxed using Networked Interference (Ma & Tresp, 2021) or multiple causal inference (D'Amour, 2019). Inferring heterogeneous treatment effects under this framework should also be investigated.

# Appendices

# APPENDIX A

# Appendix for Part I

## A.1 Proofs of Propositions 3.2.3 - 3.3.6.

### Preliminary lemmas

**Lemma A.1.1.** *Let* $\mathbf{F}$ *be a full rank matrix (Assumption 2.2.32), let* $\mathbf{K}$ *be a positive definite matrix and let* $\overline{\mathbf{K}}$ *defined by* $\overline{\mathbf{K}} = \mathbf{K}^{-1}\left(\mathbf{I}_n - \mathbf{F}\left(\mathbf{F}^\top\mathbf{K}^{-1}\mathbf{F}\right)^{-1}\mathbf{F}^\top\mathbf{K}^{-1}\right)$ *then* $\operatorname{Ker}\overline{\mathbf{K}} = \operatorname{Im}\mathbf{F}$ *and* $\overline{\mathbf{K}}$ *is singular.*

*Proof.* Let $\overline{\mathbf{K}}$ be the matrix defined above. Suppose that $\boldsymbol{x} \in \operatorname{Im}\mathbf{F}$, then there exists $\boldsymbol{y}$ such that $\boldsymbol{x} = \mathbf{F}\boldsymbol{y}$, and $\overline{\mathbf{K}}\boldsymbol{x} = \mathbf{K}^{-1}\left(\mathbf{F}\boldsymbol{y} - \mathbf{F}\left(\mathbf{F}^\top\mathbf{K}^{-1}\mathbf{F}\right)^{-1}\mathbf{F}^\top\mathbf{K}^{-1}\mathbf{F}\boldsymbol{y}\right) = \mathbf{K}^{-1}\left(\mathbf{F}\boldsymbol{y} - \mathbf{F}\boldsymbol{y}\right) = \mathbf{0}$. Thus $\boldsymbol{x} \in \operatorname{Ker}\overline{\mathbf{K}}$.

If $\boldsymbol{x} \in \operatorname{Ker}\overline{\mathbf{K}}$, then $\mathbf{K}\ \overline{\mathbf{K}}\boldsymbol{x} = \mathbf{0}$, and $\boldsymbol{x} = \mathbf{F}(\mathbf{F}^\top\mathbf{K}^{-1}\mathbf{F})^{-1}\mathbf{F}^\top\mathbf{K}^{-1}\boldsymbol{x} = \mathbf{F}\boldsymbol{x}' \in \operatorname{Im}\mathbf{F}$.

In case of Ordinary or Universal kriging, $p = \operatorname{rank}(\mathbf{F}) = \dim(\operatorname{Ker}\overline{\mathbf{K}}) \geq 1$ which means that $\overline{\mathbf{K}}$ is not invertible. ■

**Lemma A.1.2** ( de Oliveira (2007))**.** *Under the hypotheses of Lemma A.1.1 and given the full rank regression matrix* $\mathbf{F}$*, there exists a matrix* $\mathbf{W} \in \mathbb{R}^{n\times(n-p)}$ *satisfying :*

$$\mathbf{W}^\top\mathbf{W} = \mathbf{I}_{n-p}, \tag{A.1}$$

$$\mathbf{F}^\top\mathbf{W} = \mathbf{O}_{p\times(n-p)}, \tag{A.2}$$

*and*

$$\overline{\mathbf{K}} = \mathbf{W}\left(\mathbf{W}^\top\mathbf{K}\mathbf{W}\right)^{-1}\mathbf{W}^\top. \tag{A.3}$$

**Lemma A.1.3.** *Under the hypotheses of Lemma A.1.1, if additionally assumption 2.3.10 holds true, then* $\overline{\mathbf{K}}_{ii} > 0$ *for all* $i \in \{1, \ldots, n\}$.

*Proof.* $\overline{\mathbf{K}}$ is a positive semi-definite matrix by Lemma A.1.2 and we can write

$$\overline{\mathbf{K}} = \sum_{j=1}^{n} \lambda_j \boldsymbol{u}_j\boldsymbol{u}_j^\top, \tag{A.4}$$

with $\lambda_j \geq 0$ the eigenvalues of $\overline{\mathbf{K}}$ and $(\boldsymbol{u}_j)_{j=1}^n$ the orthonormal basis of the eigenvectors. We have

$$\overline{\mathbf{K}}_{ii} = \boldsymbol{e}_i^\top \overline{\mathbf{K}} \boldsymbol{e}_i = \sum_{j=1}^{n} \lambda_j (\boldsymbol{u}_j^\top \boldsymbol{e}_i)^2. \tag{A.5}$$

If $\overline{\mathbf{K}}_{ii} = 0$, then $\boldsymbol{u}_j^\top \boldsymbol{e}_i = 0$ for all $j$ such that $\lambda_j > 0$. Therefore

$$\overline{\mathbf{K}} \boldsymbol{e}_i = \sum_{j=1}^{n} \lambda_j (\boldsymbol{u}_j^\top \boldsymbol{e}_i) \boldsymbol{u}_j = \mathbf{0}, \tag{A.6}$$

which shows that $\boldsymbol{e}_i \in \operatorname{Ker} \overline{\mathbf{K}}$, that is, $\boldsymbol{e}_i \in \operatorname{Im} \mathbf{F}$ by Lemma A.1.1. ■

**Lemma A.1.4.** *Let* $\mathbf{\Pi} = \mathbf{W}\mathbf{W}^\top = \mathbf{I}_n - \mathbf{F}\left(\mathbf{F}^\top \mathbf{F}\right)^{-1} \mathbf{F}^\top$ *the orthogonal projection matrix on* $\mathrm{Im}\mathbf{F})^\perp$ *then, with the assumption 2.3.10,* $(\mathbf{\Pi})_{i,i} \neq 0$ *for all* $i \in \{1, \ldots, n\}$.

*Proof.* This lemma is a direct application of Lemma A.1.3 by choosing $\mathbf{K} = \mathbf{I}_n$. ■

## Proof of Proposition 3.2.3

From preliminary lemmas, we show now the stronger result (stronger than Proposition 3.2.3):

**Lemma A.1.5.** *Under the assumptions 2.2.32-3.2.1, for any* $\boldsymbol{\theta} \in (0, +\infty)^d$, *there exists* $\sigma^2 \in (0, +\infty)$ *such that* $(\sigma^2, \boldsymbol{\theta}) \in \mathcal{A}_{a,\delta}$.

*Proof.* Here $\sigma_\epsilon^2 > 0$. Let us assume that $a > 1/2$ (i.e. $q_a > 0$), then for $\boldsymbol{\theta}$ fixed in $(0, +\infty)^d$, the limit of $\overline{\mathbf{K}}$ when $\sigma^2 \to 0$ is well defined and is equal to

$$\lim_{\sigma^2 \to 0} \overline{\mathbf{K}} = \sigma_\epsilon^{-2}\, \mathbf{W}\mathbf{W}^\top = \sigma_\epsilon^{-2}\, \mathbf{\Pi}. \tag{A.7}$$

By Assumption 2.3.10 and from Lemma A.1.4, we can write for all $i \in \{1, \ldots, n\}$

$$\frac{\left(\overline{\mathbf{K}}\boldsymbol{y}\right)_i}{\sqrt{\left(\overline{\mathbf{K}}\right)_{i,i}}} \overset{\sigma^2 \to 0}{\longrightarrow} \frac{1}{\sigma_\epsilon} \frac{(\mathbf{\Pi}\boldsymbol{y})_i}{\sqrt{(\mathbf{\Pi})_{i,i}}}. \tag{A.8}$$

Since $h_\delta^+ \leq h$ for all $\delta > 0$, then

$$\lim_{\sigma^2 \to 0} \psi_a^{(\delta)}(\sigma^2, \boldsymbol{\theta}) \leq \lim_{\sigma^2 \to 0} \psi_a(\sigma^2, \boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} h\left(q_a - \frac{1}{\sigma_\epsilon} \frac{(\mathbf{\Pi}\boldsymbol{y})_i}{\sqrt{(\mathbf{\Pi})_{i,i}}}\right) = \frac{k_\epsilon}{n} \tag{A.9}$$

When $\sigma^2 \to +\infty$, we have

$$\overline{\mathbf{K}} \overset{\sigma^2 \to +\infty}{\sim} \sigma^{-2}\, \overline{\mathbf{R}}_{\boldsymbol{\theta}}, \tag{A.10}$$

where

$$\overline{\mathbf{R}}_{\boldsymbol{\theta}} = \mathbf{W}\left(\mathbf{W}^\top \mathbf{R}_{\boldsymbol{\theta}} \mathbf{W}\right)^{-1} \mathbf{W}^\top. \tag{A.11}$$

By lemma A.1.3, we have $\left(\overline{\mathbf{R}}_{\boldsymbol{\theta}}\right)_{i,i} > 0$ for all $i \in \{1, \ldots, n\}$ and we obtain that

$$\frac{1}{\sigma} \frac{\left(\overline{\mathbf{R}}_{\boldsymbol{\theta}} \boldsymbol{y}\right)_i}{\sqrt{\left(\overline{\mathbf{R}}_{\boldsymbol{\theta}}\right)_{i,i}}} \overset{\sigma^2 \to +\infty}{\longrightarrow} 0. \tag{A.12}$$

With $\delta$ small enough satisfying $\delta < q_a$, we obtain

$$\psi_a^{(\delta)}(\sigma^2, \boldsymbol{\theta}) \overset{\sigma^2 \to +\infty}{\longrightarrow} \frac{1}{n} \sum_{i=1}^{n} h_\delta^+ (q_a) = 1. \tag{A.13}$$

Since $k_\epsilon < an < n$ by Assumption 3.2.1 and since $\psi_a^{(\delta)}$ is continuous, the Intermediate Value Theorem gives the existence of $\sigma_\delta^2 \in (0, +\infty)$ such that

$$\psi_a^{(\delta)}(\sigma_\delta^2, \boldsymbol{\theta}) = a, \tag{A.14}$$

which gives the desired result.

Similarly, if $a < a/2$ then $q_a < 0$ and

$$\lim_{\sigma^2 \to 0} \psi_a^{(\delta)}(\sigma^2, \boldsymbol{\theta}) \geq \lim_{\sigma^2 \to 0} \psi_a(\sigma^2, \boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} h \left( q_a - \frac{1}{\sigma_\epsilon} \frac{(\boldsymbol{\Pi} \boldsymbol{y})_i}{\sqrt{(\boldsymbol{\Pi})_{i,i}}} \right) = \frac{k_\epsilon}{n} > a. \tag{A.15}$$

When $\delta < q_{1-a}$, one obtains

$$\psi_a^{(\delta)}(\sigma^2, \boldsymbol{\theta}) \overset{\sigma^2 \to +\infty}{\longrightarrow} \frac{1}{n} \sum_{i=1}^{n} h_\delta^- (q_a) = 0. \tag{A.16}$$

By the assumption 3.2.1, one has the existence of $\sigma_\delta^2 \in (0, +\infty)$ such that

$$\psi_a^{(\delta)}(\sigma_\delta^2, \boldsymbol{\theta}) = a, \tag{A.17}$$

which completes the proof of the lemma. ■

## Proof of Proposition 3.3.5

The existence of $\sigma_{\text{opt}}^2(\lambda)$ for all $\lambda \in (0, +\infty)$ results directly from the following lemma A.1.6 :

**Lemma A.1.6.** *For all $\lambda \in (0, +\infty)$, $H_\delta(\lambda)$ is a non-empty and compact subset of $\mathbb{R}^+$ i.e. $H_\delta$ is compact-valued.*

*Proof.* By Lemma A.1.5, $H_\delta(\lambda)$ is non-empty for all $\lambda \in (0, +\infty)$.

$H_\delta(\lambda)$ is closed since the functions $h_\delta^+, h_\delta^-$ are continuous and the map $(\sigma^2, \boldsymbol{\theta}) \mapsto \overline{\mathbf{K}}$ is also continuous for all $(\sigma^2, \boldsymbol{\theta})$ by the continuity of the kernel function $\boldsymbol{k}_{,,}^{\nu}(\boldsymbol{x}, \boldsymbol{x}')$ for any $\nu > 0$ and $\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{D}$.

We now prove that $H_\delta(\lambda)$ is bounded. Let us assume that $a \in (1/2, 1)$. If $H_\delta(\lambda)$ is not bounded then there exists a sequence $\left(\sigma_m^2\right)_{m\in\mathbb{N}}$ of $H_\delta(\lambda)$ such that $\lim_{m\to+\infty} \sigma_m^2 = +\infty$ and, by continuity of $\psi_a^{(\delta)}$

$$a = \lim_{m\to+\infty} \psi_a^{(\delta)}(\sigma_m^2, \lambda\boldsymbol{\theta}_0) = \frac{1}{n}\sum_{i=1}^{n} h_\delta^+\left(q_a\right) = 1, \tag{A.18}$$

which is a contradiction. Therefore, $H_\delta(\lambda)$ is closed and bounded, $H_\delta(\lambda)$ is compact. ■

$\sigma_{\text{opt}}^2(\lambda)$ can be seen the solution of a constrained maximization problem

$$\sigma_{\text{opt}}^2(\lambda) = -\max_{\sigma^2\in H_\delta(\lambda)} u(\sigma^2, \lambda), \quad \lambda \in (0, +\infty), \tag{A.19}$$

where $u(\sigma^2, \lambda) = -\sigma^2$ is a continuous function. $H_\delta$ is non-empty-valued and compact-valued by Lemma A.1.6, upper semi-continuous since $\psi_a^{(\delta)}$ is continuous on $[0, +\infty) \times (0, +\infty)^d$, and continuous if the assumption 3.3.4 is satisfied, the Maximum theorem (Berge (1963), p. 116) provides the continuity of $\sigma_{\text{opt}}^2$ on $(0, +\infty)$.

## Proof of Proposition 3.3.6

Let $\boldsymbol{\theta}_0$ be a solution of one of the problems described in (2.58) or (2.76). The continuity of $\mathcal{L}$ on $(0, +\infty)$ follows from the continuity of the trace function $\text{Tr}(\cdot)$, the continuity of the map $(\sigma^2, \boldsymbol{\theta}) \mapsto \overline{\mathbf{K}}$ and the continuity of $\sigma_{\text{opt}}^2$ by proposition 3.3.5.

Assume that $\lim_{\lambda\to+\infty} \sigma_{\text{opt}}^2(\lambda) < +\infty$, then there exists $M > 0$ such that for all $\lambda > 0$ there exists $\lambda' \geq \lambda$ and $\sigma_{\text{opt}}^2(\lambda') \leq M$. Hence, we can recursively build a sequence $(\lambda_m)_{m\in\mathbb{N}}$ of integers such that $\lambda_{m+1} \geq \lambda_m + 1$ and $\sigma_{\text{opt}}^2(\lambda_m) \leq M$ for all $m \in \mathbb{N}$.

By the Bolzano-Weierstrass theorem, we extract a convergent sub-sequence $\left(\lambda_{\phi(m)}\right)_{m\in\mathbb{N}}$ where $\phi : \mathbb{N} \to \mathbb{N}$ such that $\sigma_{\text{opt}}^2(\lambda_{\phi(m)}) \overset{m\to+\infty}{\longrightarrow} \sigma_\infty^2 < +\infty$ and

$$\mathbf{K}_{\sigma_{\text{opt}}^2(\lambda_{\phi(m)}),\lambda_{\phi(m)}\boldsymbol{\theta}_0} \overset{m\to+\infty}{\longrightarrow} \sigma_\infty^2\mathbf{J} + \sigma_\epsilon^2\mathbf{I}_n = \mathbf{K}_\infty. \tag{A.20}$$

When there is a nugget effect $\sigma_\epsilon^2 > 0$, the limit of $\overline{\mathbf{K}}_m := \overline{\mathbf{K}}_{\sigma_{\text{opt}}^2(\lambda_{\phi(m)}),\lambda_{\phi(m)}\boldsymbol{\theta}_0}$ when $m \to +\infty$ exists because the matrix $\mathbf{K}_\infty$ is nonsingular by the auxiliary fact 1 of Berger et al. (2001)

$$\det \mathbf{K}_\infty = \left(\frac{\sigma_\epsilon^2}{\sigma_\infty^2}\right)^n \left(1 + \frac{\sigma_\epsilon^2}{\sigma_\infty^2}\, \boldsymbol{e}^\top\mathbf{I}_n\boldsymbol{e}\right) = \left(\frac{\sigma_\epsilon^2}{\sigma_\infty^2}\right)^n \left(1 + n\frac{\sigma_\epsilon^2}{\sigma_\infty^2}\right) > 0. \tag{A.21}$$

From Assumption 2.2.32, $\boldsymbol{e}$ is a column of $\mathbf{F}$ and we can prove that

$$\begin{aligned} \overline{\mathbf{K}}_m \overset{m\to+\infty}{\longrightarrow} \overline{\mathbf{K}}_\infty :&= \mathbf{W}\left(\mathbf{W}^\top\left(\sigma_\infty^2\mathbf{J} + \sigma_\epsilon^2\mathbf{I}_n\right)\mathbf{W}\right)^{-1}\mathbf{W}^\top \\ &= \sigma_\epsilon^{-2}\,\mathbf{W}\left(\mathbf{W}^\top\mathbf{W}\right)^{-1}\mathbf{W}^\top = \sigma_\epsilon^{-2}\,\boldsymbol{\Pi}. \end{aligned} \tag{A.22}$$

By the assumption 2.3.10, the Leave-One-Out formulas (2.72-2.73) give for all $i \in \{1, \dots, n\}$

$$\frac{\left(\overline{\mathbf{K}}_m \boldsymbol{y}\right)_i}{\sqrt{\left(\overline{\mathbf{K}}_m\right)_{i,i}}} \overset{m\to+\infty}{\longrightarrow} \frac{1}{\sigma_\epsilon} \frac{(\boldsymbol{\Pi}\boldsymbol{y})_i}{\sqrt{(\boldsymbol{\Pi})_{i,i}}}. \tag{A.23}$$

If $a > 1/2$ for example and by definition of $\sigma^2_{\text{opt}}(\lambda_{\phi(m)})$, one obtains

$$\begin{aligned} a &= \frac{1}{n}\sum_{i=1}^{n} h_\delta^+ \left( q_a - \frac{\left(\overline{\mathbf{K}}_m \boldsymbol{y}\right)_i}{\sqrt{\left(\overline{\mathbf{K}}_m\right)_{i,i}}} \right) \\ &\overset{m\to+\infty}{\longrightarrow} \frac{1}{n}\sum_{i=1}^{n} h_\delta^+ \left( q_a - \frac{\left(\overline{\mathbf{K}}_\infty \boldsymbol{y}\right)_i}{\sqrt{\left(\overline{\mathbf{K}}_\infty\right)_{i,i}}} \right) \\ &= \frac{1}{n}\sum_{i=1}^{n} h_\delta^+ \left( q_a - \frac{1}{\sigma_\epsilon} \frac{(\boldsymbol{\Pi}\boldsymbol{y})_i}{\sqrt{(\boldsymbol{\Pi})_{i,i}}} \right) = \frac{k_\epsilon}{n} < a, \end{aligned} \tag{A.24}$$

which is contradictory. Therefore, $\lim_{\lambda\to+\infty} \sigma^2_{\text{opt}}(\lambda) = +\infty$ and $\mathcal{L}$ is coercive. The case $a < 1/2$ can be addressed in the same way.

## A.2 The no-nugget case.

### Proof of the existence of a solution to Problem (3.13)

In the absence of $\sigma^2_\epsilon = 0$, it follows from the Leave-One-Out formulas that, for all $i \in \{1, \dots, n\}$

$$\frac{\left(\overline{\mathbf{K}}\boldsymbol{y}\right)_i}{\sqrt{\left(\overline{\mathbf{K}}\right)_{i,i}}} = \frac{1}{\sigma} \frac{\left(\overline{\mathbf{R}}_{\boldsymbol{\theta}}\boldsymbol{y}\right)_i}{\sqrt{\left(\overline{\mathbf{R}}_{\boldsymbol{\theta}}\right)_{i,i}}}, \tag{A.25}$$

which is a monotonic function in $\sigma^2$ when $\boldsymbol{\theta}$ is fixed in $(0, \infty)^d$.

Let $\boldsymbol{\theta}$ be fixed in $(0, +\infty)^d$ and let $a > 1/2$. The proportion $\psi_a^{(\delta)}(\sigma^2, \boldsymbol{\theta})$ has the limit

$$\lim_{\sigma^2\to+\infty} \psi_a^{(\delta)}(\sigma^2, \boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} h_\delta^+(q_a) = 1, \tag{A.26}$$

and, if $\sigma^2 \to 0$, it has the limit

$$\lim_{\sigma^2\to 0} \psi_a^{(\delta)}(\sigma^2, \boldsymbol{\theta}) = \frac{1}{n} \text{Card}\left\{ i \in \{1, \dots, n\}, \left(\overline{\mathbf{R}}_{\boldsymbol{\theta}}\boldsymbol{y}\right)_i \leq 0 \right\} = \frac{k_{\boldsymbol{\theta}}}{n}. \tag{A.27}$$

Let $\boldsymbol{\theta}$ denote the norm of $\boldsymbol{\theta}$ (i.e. $\boldsymbol{\theta} = \|\boldsymbol{\theta}\|$) and consider the set $\mathcal{J} = \{i \in \{1, \dots, n\}, (\boldsymbol{\Pi}\boldsymbol{y})_i \leq 0\}$. For $i \in \mathcal{J}^c$, one has $(\boldsymbol{\Pi}\boldsymbol{y})_i > 0$, and, since $\overline{\mathbf{R}}_{\boldsymbol{\theta}}$ converges to $\boldsymbol{\Pi}$ when $\boldsymbol{\theta} \to 0$

$$\forall i \in \mathcal{J}^c \ : \left(\overline{\mathbf{R}}_{\boldsymbol{\theta}}\boldsymbol{y}\right)_i \overset{\boldsymbol{\theta}\to 0}{\longrightarrow} (\boldsymbol{\Pi}\boldsymbol{y})_i > 0. \tag{A.28}$$

It results that, there exists $\boldsymbol{\theta}_c > 0$ such that if $\boldsymbol{\theta} \in \mathcal{B}_r(\mathbf{0}, \boldsymbol{\theta}_c)$ (the open ball of radius $\boldsymbol{\theta}_c$ centered at $\mathbf{0}$) then $\left(\overline{\mathbf{R}}_{\boldsymbol{\theta}}\boldsymbol{y}\right)_i > 0$ for any $i \in \mathcal{J}^c$. Consequently, one gets for any $\boldsymbol{\theta} \in \mathcal{B}_r(\mathbf{0}, \boldsymbol{\theta}_c)$

$$\text{Card}\left\{i \in \{1, \ldots, n\}, \left(\overline{\mathbf{R}}_{\boldsymbol{\theta}}\boldsymbol{y}\right)_i > 0\right\} \geq \text{Card}(\mathcal{J}^c) = n - k_\epsilon. \tag{A.29}$$

Hence

$$k_{\boldsymbol{\theta}} = \text{Card}\left\{i \in \{1, \ldots, n\}, \left(\overline{\mathbf{R}}_{\boldsymbol{\theta}}\boldsymbol{y}\right)_i \leq 0\right\} \leq k_\epsilon. \tag{A.30}$$

Therefore, if $\boldsymbol{\theta}$ belongs to a neighborhood of $\mathbf{0}$, the condition $k_{\boldsymbol{\theta}} \leq k_\epsilon$ is satisfied and, under the assumption 3.2.1, the set of solutions $\mathcal{A}_{a,\delta}$ is also non-empty.

## Proof of the Coercivity

Let us assume that, under some conditions on $\boldsymbol{y}$, $\lambda \mapsto \sigma^2_{\text{opt}}(\lambda)$ is well-defined for all $\lambda \in (0, +\infty)$. In the absence of nugget effect $\sigma^2_\epsilon = 0$, the limit of $\overline{\mathbf{R}}_{\lambda\boldsymbol{\theta}_0}$ does not exist when $\lambda \to +\infty$. Still, we can assume that the correlation matrix $\mathbf{R}_{\lambda\boldsymbol{\theta}_0}$ satisfies (Berger et al., 2001)

$$\mathbf{R}_{\lambda\boldsymbol{\theta}_0} = \mathbf{F} + g_\lambda \left(\mathbf{D}_0 + o(1)\right), \tag{A.31}$$

where

- $\lambda \mapsto g_\lambda$ is a continuous function such that $\lim_{\lambda \to +\infty} g_\lambda = 0$.
- $\mathbf{D}_0$ and $\mathbf{J} = \boldsymbol{e}\boldsymbol{e}^\top$ are fixed symmetric matrices.

$\mathbf{D}_0$ can be singular or nonsingular depending on the chosen kernel $\boldsymbol{k}$. A review of Yagloom's book (Rosenblatt, 1989) shows that $\mathbf{D}_0$ is nonsingular only for Power-Exponential ($q < 2$) and Matérn kernels with smoothness parameter $\nu < 1$ like the Exponential kernel (See subsection 2.2). For the rest of Matérn kernels with smoothness parameter $\nu \geq 1$ $\mathbf{D}_0$ becomes singular.

### Case 1: $\mathbf{D_0}$ is nonsingular.

In this case, let $\mathbf{D}_\lambda = g_\lambda\ \mathbf{D}_0\,(1 + o(1))$ such that

$$\mathbf{R}_{\lambda\boldsymbol{\theta}_0} = \mathbf{J} + \mathbf{D}_\lambda. \tag{A.32}$$

We consider the matrix $\overline{\mathbf{R}}_{\lambda\boldsymbol{\theta}_0}$ in $\overline{\mathbf{K}} = \sigma^{-2}\ \overline{\mathbf{R}}_{\lambda\boldsymbol{\theta}_0}$, we have

$$\overline{\mathbf{R}}_{\lambda\boldsymbol{\theta}_0} = \mathbf{R}^{-1}_{\lambda\boldsymbol{\theta}_0}\left[\mathbf{I}_n - \mathbf{F}\left(\mathbf{F}^\top\mathbf{R}^{-1}_{\lambda\boldsymbol{\theta}_0}\mathbf{F}\right)^{-1}\mathbf{F}^\top\mathbf{R}^{-1}_{\lambda\boldsymbol{\theta}_0}\right]. \tag{A.33}$$

By using Lemma 4, Appendix B3 in Berger et al. (2001) and under assumption that $\boldsymbol{e} \in \text{Im}\mathbf{F}$ (hypothesis 2.2.32), we have

$$\overline{\mathbf{R}}_{\lambda\boldsymbol{\theta}_0} = \mathbf{D}^{-1}_\lambda\left[\mathbf{I}_n - \mathbf{F}\left(\mathbf{F}^\top\mathbf{D}^{-1}_\lambda\mathbf{F}\right)^{-1}\mathbf{F}^\top\mathbf{D}^{-1}_\lambda\right]. \tag{A.34}$$

Then we get

$$\overline{\mathbf{R}}_{\lambda\boldsymbol{\theta}_0} = g^{-1}_\lambda\left[\mathbf{D}^{-1}_0\left(\mathbf{I}_n - \mathbf{F}\left(\mathbf{F}^\top\mathbf{D}^{-1}_0\mathbf{F}\right)^{-1}\mathbf{F}^\top\mathbf{D}^{-1}_0\right) + o(1)\right]. \tag{A.35}$$

Finally

$$\overline{\mathbf{R}}_{\lambda\boldsymbol{\theta}_0} \overset{\lambda\to+\infty}{\sim} g_\lambda^{-1}\mathbf{A}, \tag{A.36}$$

where

$$\mathbf{A} = \mathbf{D}_0^{-1}\left(\mathbf{I}_n - \mathbf{F}\left(\mathbf{F}^\top\mathbf{D}_0^{-1}\mathbf{F}\right)^{-1}\mathbf{F}^\top\mathbf{D}_0^{-1}\right). \tag{A.37}$$

**Assumption A.2.1.** *Let* $\mathbf{A}$ *be the matrix defined in (A.37). We assume that* $\boldsymbol{y}$ *does not belong to a family of vectors such that* $(\mathbf{A}\boldsymbol{y})_i = 0$ *for all* $i \in \{1,\dots,n\}$ *and that* $\mathrm{Card}\{i \in \{1,\dots,n\}, (\mathbf{A}\boldsymbol{y})_i \leq 0\} \neq na$.

By applying Lemmas A.1.1 and A.1.2 on $\mathbf{D}_0$, we show that $(\mathbf{A})_{ii} \neq 0$ and we can write for all $i$ in $\{1,...,n\}$

$$\frac{\left(\overline{\mathbf{R}}_{\lambda\boldsymbol{\theta}_0}\boldsymbol{y}\right)_i}{\sqrt{\left(\overline{\mathbf{R}}_{\lambda\boldsymbol{\theta}_0}\right)_{ii}}} \overset{\lambda\to+\infty}{\sim} g_\lambda^{-1/2}\frac{(\mathbf{A}\boldsymbol{y})_i}{\sqrt{(\mathbf{A})_{ii}}}. \tag{A.38}$$

Analogously to the proof of Proposition 3.3.6, if we assume that $\lim_{\lambda\to+\infty}\sigma^2_{\mathrm{opt}}(\lambda) \neq +\infty$ and by taking a sub-sequence $\left(\sigma^2_{\mathrm{opt}}(\lambda_{\psi(m)})\right)_{m\in\mathbb{N}}$ converging to $\sigma^2_\infty$

$$\frac{1}{\sigma_\infty}\, g_{\lambda_{\psi(m)}}^{-1/2}\frac{(\mathbf{A}\boldsymbol{y})_i}{\sqrt{(\mathbf{A})_{ii}}} \overset{m\to+\infty}{\longrightarrow} \begin{cases} +\infty & \text{if } (\mathbf{A}\boldsymbol{y})_i > 0 \\ -\infty & \text{otherwise} \end{cases}. \tag{A.39}$$

The limit $\psi_a^{(\delta)}(\sigma^2_{\mathrm{opt}}(\lambda_{\psi(m)}), \lambda_{\psi(m)}\boldsymbol{\theta}_0)$ when $m \to +\infty$ exists and is equal to

$$a = \lim_{m\to+\infty}\psi_a^{(\delta)}(\sigma^2_{\mathrm{opt}}(\lambda_{\psi(m)}), \lambda_{\psi(m)}\boldsymbol{\theta}_0) = \frac{1}{n}\,\mathrm{Card}\{i \in \{1,\dots,n\}, (\mathbf{A}\boldsymbol{y})_i \leq 0\}, \tag{A.40}$$

which is contradictory and completes the proof.

### Case 2: $\mathrm{D}_0$ is singular.

In this case, one needs to go further in the Taylor expansion of $\overline{\mathbf{R}}_{\lambda\boldsymbol{\theta}_0}$. We consider the matrix $\mathbf{W}$ in Lemma A.1.3, by Lemma 6 of Ren et al. (2012)

$$\overline{\mathbf{R}}_{\lambda\boldsymbol{\theta}_0} = \mathbf{W}\left(\mathbf{W}^\top\mathbf{R}_{\lambda\boldsymbol{\theta}_0}\mathbf{W}\right)^{-1}\mathbf{W}^\top. \tag{A.41}$$

By setting $\boldsymbol{\Sigma}_\lambda = \mathbf{W}^\top\mathbf{R}_{\lambda\boldsymbol{\theta}_0}\mathbf{W}$, the asymptotic study of $\overline{\mathbf{R}}_{\lambda\boldsymbol{\theta}_0}$ is equivalent to the asymptotic study of $\boldsymbol{\Sigma}_\lambda$. In case of Matérn kernel with noninteger smoothness $\nu \geq 1$, the matrix $\boldsymbol{\Sigma}_\lambda$ can be written as (Muré, 2021)

$$\boldsymbol{\Sigma}_\lambda = g_\lambda\left(\mathbf{W}^\top\mathbf{D}_1\mathbf{W} + g_\lambda^*\mathbf{W}^\top\mathbf{D}_1^*\mathbf{W} + \mathbf{R}_g(\lambda)\right), \tag{A.42}$$

where

- Either $g_\lambda = c\lambda^{-2k_1}$ with $k_1$ a nonnegative integer, or $g_\lambda = c\lambda^{-2\nu}$.
- $g_\lambda^* = c^*\lambda^{-l}$ with $l \in (0,+\infty)$ .

- $\mathbf{R}_g$ is a differentiable mapping from $[0, +\infty)$ to $\mathcal{M}_n$ such that $\|\mathbf{R}_g(\lambda)\| = o(\lambda^{-2l})$.
- $\mathbf{D}_1$ and $\mathbf{D}_1^*$ are both fixed symmetric matrices with elements $\|x_i - x_j\|^{2k}$ where $k \in k_1 \cup \nu$ for $\mathbf{D}_1$ and $k = l$ for $\mathbf{D}_1^*$.

The matrix $\mathbf{W}^\top \mathbf{D}_1 \mathbf{W} + g_\lambda^* \mathbf{W}^\top \mathbf{D}_1^* \mathbf{W}$ is nonsingular when $\lambda \to +\infty$, whether if $\mathbf{W}^\top \mathbf{D}_1 \mathbf{W}$ is nonsingular or if it is singular.

The case where $\mathbf{W}^\top \mathbf{D}_1 \mathbf{W}$ is nonsingular happens for Matérn kernels with smoothness $1 \leq \nu < 2$ (Muré, 2021), whereas the other case occurs for regular and smooth Matérn kernels with $\nu \geq 2$.

### Case 2.a) $\mathbf{W}^\top \mathbf{D}_1 \mathbf{W}$ is non-singular.

In this case, we write $\mathbf{\Sigma}_\lambda$ in (A.42) as

$$\mathbf{\Sigma}_\lambda = g_\lambda \mathbf{W}^\top \mathbf{D}_1 \mathbf{W} \left( \mathbf{I}_n + g_\lambda^* \left( \mathbf{W}^\top \mathbf{D}_1 \mathbf{W} \right)^{-1} \left( \mathbf{W}^\top \mathbf{D}_1^* \mathbf{W} + \mathbf{R}_g(\lambda) \right) \right). \tag{A.43}$$

As $\mathbf{W}$ is full rank matrix, $\mathbf{\Sigma}_\lambda$ is non-singular and

$$\mathbf{\Sigma}_\lambda^{-1} = g_\lambda^{-1} \left( \mathbf{I}_n + g_\lambda^* \left( \mathbf{W}^\top \mathbf{D}_1 \mathbf{W} \right)^{-1} \left( \mathbf{W}^\top \mathbf{D}_1^* \mathbf{W} + \mathbf{R}_g(\lambda) \right) \right)^{-1} \left( \mathbf{W}^\top \mathbf{D}_1 \mathbf{W} \right)^{-1}. \tag{A.44}$$

Let $\mathbf{M}_\lambda = g_\lambda^* \left( \mathbf{W}^\top \mathbf{D}_1 \mathbf{W} \right)^{-1} \left( \mathbf{W}^\top \mathbf{D}_1^* \mathbf{W} + \mathbf{R}_g(\lambda) \right)$, since $\|\mathbf{M}_\lambda\| \overset{\lambda \to +\infty}{\longrightarrow} 0$, we can assume that $\|\mathbf{M}_\lambda\| < 1$ when $\lambda$ is large enough and apply the Taylor series expansion at order 1

$$\begin{aligned} \left[ \mathbf{I}_n + g_\lambda^* \left( \mathbf{W}^\top \mathbf{D}_1 \mathbf{W} \right)^{-1} \left( \mathbf{W}^\top \mathbf{D}_1^* \mathbf{W} + \mathbf{R}_g(\lambda) \right) \right]^{-1} &= \mathbf{I}_n - g_\lambda^* \left( \mathbf{W}^\top \mathbf{D}_1 \mathbf{W} \right)^{-1} \\ &\quad \times \left( \mathbf{W}^\top \mathbf{D}_1^* \mathbf{W} + \mathbf{R}_g(\lambda) + o(g_\lambda^*) \right). \end{aligned} \tag{A.45}$$

Then, we plug this quantity into the equation (A.44)

$$\begin{aligned} \mathbf{\Sigma}_\lambda^{-1} &= g_\lambda^{-1} \left( \mathbf{I}_n - g_\lambda^* \left( \mathbf{W}^\top \mathbf{D}_1 \mathbf{W} \right)^{-1} \left( \mathbf{W}^\top \mathbf{D}_1^* \mathbf{W} + \mathbf{R}_g(\lambda) \right) \right) \left( \mathbf{W}^\top \mathbf{D}_1 \mathbf{W} \right)^{-1} \\ &= g_\lambda^{-1} \left[ \left( \mathbf{W}^\top \mathbf{D}_1 \mathbf{W} \right)^{-1} - g_\lambda^* \left( \mathbf{W}^\top \mathbf{D}_1 \mathbf{W} \right)^{-1} \left( \mathbf{W}^\top \mathbf{D}_1^* \mathbf{W} + \mathbf{R}_g(\lambda) \right) \left( \mathbf{W}^\top \mathbf{D}_1 \mathbf{W} \right)^{-1} \right]. \end{aligned} \tag{A.46}$$

Finally, we can write the matrix $\overline{\mathbf{R}}_{\lambda \boldsymbol{\theta}_0}$ as

$$\overline{\mathbf{R}}_{\lambda \boldsymbol{\theta}_0} = g_\lambda^{-1} \mathbf{W} \left[ \left( \mathbf{W}^\top \mathbf{D}_1 \mathbf{W} \right)^{-1} - g_\lambda^* \left( \mathbf{W}^\top \mathbf{D}_1 \mathbf{W} \right)^{-1} \left( \mathbf{W}^\top \mathbf{D}_1^* \mathbf{W} + \mathbf{R}_g(\lambda) \right) \left( \mathbf{W}^\top \mathbf{D}_1 \mathbf{W} \right)^{-1} \right] \mathbf{W}^\top. \tag{A.47}$$

We can also simply the previous expression into

$$\overline{\mathbf{R}}_{\lambda \boldsymbol{\theta}_0} = g_\lambda^{-1} \left( \mathbf{A} - \mathbf{B}_\lambda \right), \tag{A.48}$$

where $\mathbf{A}$ is a fixed matrix and $\mathbf{B}_\lambda \overset{\lambda \to +\infty}{=} o(1)$ such that

$$\mathbf{A} = \mathbf{W} \left( \mathbf{W}^\top \mathbf{D}_1 \mathbf{W} \right)^{-1} \mathbf{W}^\top \tag{A.49}$$

$$\mathbf{B}_\lambda = g_\lambda^* \, \mathbf{W} \left(\mathbf{W}^\top \mathbf{D}_1 \mathbf{W}\right)^{-1} \left(\mathbf{W}^\top \mathbf{D}_1^* \mathbf{W} + \mathbf{R}_g(\lambda)\right) \left(\mathbf{W}^\top \mathbf{D}_1 \mathbf{W}\right)^{-1} \mathbf{W}^\top. \tag{A.50}$$

Or, equivalently,

$$\overline{\mathbf{R}}_{\lambda\boldsymbol{\theta}_0} \overset{\lambda\to+\infty}{\sim} g_\lambda^{-1} \mathbf{A}. \tag{A.51}$$

**Lemma A.2.2.** *Let* $\mathbf{A}$ *be the matrix defined in (A.49) , then* $\mathbf{A}_{ii} \neq 0$ *for all* $i \in \{1, \ldots, n\}$.

*Proof.* $\mathbf{A}$ is non-singular because

$$\det \mathbf{A} = \det \mathbf{W} \left(\mathbf{W}^\top \mathbf{D}_1 \mathbf{W}\right)^{-1} \mathbf{W}^\top = \det \left(\mathbf{W}^\top \mathbf{D}_1 \mathbf{W}\right)^{-1} \neq 0. \tag{A.52}$$

$\mathbf{A}$ is then a positive definite matrix

$$\mathbf{A}_{ii} = \boldsymbol{e}_i^\top \mathbf{A} \boldsymbol{e}_i > 0. \tag{A.53}$$

■

**Assumption A.2.3.** *Let* $\mathbf{A}$ *be the matrix defined in (A.49). We assume that* $\boldsymbol{y}$ *does not belong to a family of vectors such that* $(\mathbf{A}\boldsymbol{y})_i = 0$ *for all* $i \in \{1, \ldots, n\}$ *and that* $\operatorname{Card}\{i \in \{1, \ldots, n\}, (\mathbf{A}\boldsymbol{y})_i \leq 0\} \neq na$.

With Lemma A.1.6 and Assumption A.2.3, the proof of the divergence of $\sigma^2_{\text{opt}}(\lambda)$ when $\lambda \to +\infty$ is similar to the previous case when $\mathbf{D}_0$ is nonsingular.

**Remark A.2.4.** *The assumptions A.2.1 and A.2.3 are not restrictive, one can verify numerically, that each component of* $\mathbf{A}\boldsymbol{y}$ *is not null where* $\mathbf{A}$ *is one of the matrices defined in (A.37) or (A.49).*

### Case 2.b) $\mathbf{W}^\top \mathbf{D}_1 \mathbf{W}$ is singular.

Let us denote $\mathbf{A} = \mathbf{W}^\top \mathbf{D}_1 \mathbf{W}$ and $\mathbf{B} = \mathbf{W}^\top \mathbf{D}_1^* \mathbf{W}$ the two non-null symmetric matrices defined in (A.42). Let $\tilde{\boldsymbol{\Sigma}}_\lambda$ such that $\boldsymbol{\Sigma}_\lambda = g_\lambda \tilde{\boldsymbol{\Sigma}}_\lambda$, we consider $\boldsymbol{\Sigma}_\lambda$ as a Maclaurin serie:

$$\tilde{\boldsymbol{\Sigma}}_\lambda = \mathbf{A} + a_1(\lambda)\mathbf{B} + \mathbf{R}_g(\lambda), \tag{A.54}$$

where $a_1(\lambda) = g_\lambda^*$ with $a_1(\lambda) = o(1)$.

This case is complex because, due to the singularity of $\mathbf{A}$ , some eigenvalues tend to have unstable behaviour and compromise the convergence of some limits.

Indeed,

$$\begin{aligned} \frac{\left(\overline{\mathbf{R}}_{\lambda\boldsymbol{\theta}_0}\boldsymbol{y}\right)_i}{\sqrt{\left(\overline{\mathbf{R}}_{\lambda\boldsymbol{\theta}_0}\right)_{ii}}} &= \frac{g_\lambda^{-1}\left(\mathbf{W}^\top \tilde{\boldsymbol{\Sigma}}_\lambda^{-1} \mathbf{W}\boldsymbol{y}\right)_i}{\sqrt{g_\lambda \left(\mathbf{W}^\top \tilde{\boldsymbol{\Sigma}}_\lambda^{-1} \mathbf{W}\right)_{ii}}} = g_\lambda^{-1/2} \frac{(\mathbf{A}_\lambda \boldsymbol{y})_i}{\sqrt{(\mathbf{A}_\lambda)_{ii}}} \\ &= g_\lambda^{-1/2} \lambda_{n-p}^{-1/2}(\tilde{\boldsymbol{\Sigma}}_\lambda) \times \left(\lambda_{n-p}(\tilde{\boldsymbol{\Sigma}}_\lambda)\,(\mathbf{A}_\lambda)_{ii}\right)^{-1/2} \times \lambda_{n-p}(\tilde{\boldsymbol{\Sigma}}_\lambda)\,(\mathbf{A}_\lambda \boldsymbol{y})_i\,, \end{aligned} \tag{A.55}$$

where $\lambda_{n-p}(\tilde{\boldsymbol{\Sigma}}_\lambda)$ is the smallest eigenvalue of $\tilde{\boldsymbol{\Sigma}}_\lambda$.

We can summarize the key points of the proof in the following points:

- With Lemma A.2.5, we prove that $\lambda_{n-p}(\tilde{\mathbf{\Sigma}}_\lambda)$ has the same convergence rate as $g_\lambda^*$.
- With Lemma A.2.9, we prove that the limits of the upper and lower bounds of $\lambda_{n-p}(\tilde{\mathbf{\Sigma}}_\lambda)\,(\mathbf{A}_\lambda)_{ii}$ exist and are non null.
- We cannot prove, given the actual assumptions, that the limit $\lim_{\lambda\to\infty} \lambda_{n-p}(\tilde{\mathbf{\Sigma}}_\lambda)(\mathbf{A}_\lambda \boldsymbol{y})_i$, exist or examine if it is non null.

Indeed, $\mathbf{A}$, $\mathbf{B}$ and $\tilde{\mathbf{\Sigma}}_\lambda$ are real symmetric so they are orthogonally diagonalizables by the Spectral theorem, we denote $(\lambda_j(\mathbf{A}))_{j=1}^{n-p}$, $(\lambda_j(\mathbf{B}))_{j=1}^{n-p}$ and $\left(\lambda_j(\tilde{\mathbf{\Sigma}}_\lambda)\right)_{j=1}^{n-p}$ the sequences of ordered eigenvalues of each matrix.

$\mathbf{R}_\lambda$ is symmetric positive definite, and the kernel of $\mathbf{W}$ is trivial, this implies that $\mathbf{\Sigma}_\lambda$ and $\tilde{\mathbf{\Sigma}}_\lambda$ are both positive definite. The sequence $\left(\lambda_j(\tilde{\mathbf{\Sigma}}_\lambda)\right)_{j=1}^{n-p}$ satisfies

$$\lambda_1(\tilde{\mathbf{\Sigma}}_\lambda) \geq \lambda_2(\tilde{\mathbf{\Sigma}}_\lambda) \geq \ldots \geq \lambda_{n-p}(\tilde{\mathbf{\Sigma}}_\lambda) > 0. \tag{A.56}$$

By the singularity of $\mathbf{A}$, there exist $r$ positive eigenvalues ($r \geq 1$) such that

$$\lambda_1(\mathbf{A}) \geq \lambda_2(\mathbf{A}) \geq ... \geq \lambda_r(\mathbf{A}) > \lambda_{r+1}(\mathbf{A}) = \ldots = \lambda_{n-p}(\mathbf{A}) = 0 \tag{A.57}$$

The following inequalities hold when $\lambda$ is large enough:

$$\begin{gathered}\forall j \in \{1, ..., r\} : \lambda_j(\mathbf{A}) + g_\lambda^*\ \lambda_{n-p}(\mathbf{B}) \leq \lambda_j(\tilde{\mathbf{\Sigma}}_\lambda) \leq \lambda_j(\mathbf{A}) + g_\lambda^*\ \lambda_1(\mathbf{B}). \\ \forall j \in \{r+1, ..., n-p\} : 0 \leq \lambda_j(\tilde{\mathbf{\Sigma}}_\lambda) \leq g_\lambda^*\ \lambda_1(\mathbf{B}).\end{gathered} \tag{A.58}$$

These inequalities give in particular the convergence of the eigenvalues $\lambda_j(\tilde{\mathbf{\Sigma}}_\lambda) \overset{\lambda\to+\infty}{\longrightarrow} \lambda_j(\mathbf{A})$ for all $j \in \{1, \ldots, n-p\}$.

Let $\left(\boldsymbol{u}_j^{(1)}\right)_{j=1}^{r}$ be the orthonormal basis of the eigenvectors corresponding to the first eigenvalues $(\lambda_j(\mathbf{A}))_{j=1}^{r}$ and $\left(\boldsymbol{u}_j^{(2)}\right)_{j=r+1}^{n-p}$ be the orthonormal basis of the eigenvectors corresponding to the last eigenvalues $(\lambda_j(\mathbf{A}))_{j=r+1}^{n-p}$.

We denote $E_1$ the eigenspace spanned by the first $r$ eigenvalues of $\mathbf{A}$ and $E_2 = E_1^\perp$ the eigenspace

$$\mathbb{R}^{n-p} = E_1 \oplus E_2 \tag{A.59}$$

Note that $E_2 = \mathrm{Ker}(\mathbf{A})$ because $\mathbf{A}$ is diagonalizable.

Similarly, we denote $\left(\boldsymbol{u}_{\lambda,j}^{(1)}\right)_{j=1}^{r}$ the orthonormal basis of the eigenvectors corresponding to the first eigenvalues $\left(\lambda_j(\tilde{\mathbf{\Sigma}}_\lambda)\right)_{j=1}^{r}$ and $\left(\boldsymbol{u}_{\lambda,j}^{(2)}\right)_{j=r+1}^{n-p}$ be the orthonormal basis of the eigenvectors corresponding to the last eigenvalues $\left(\lambda_j(\tilde{\mathbf{\Sigma}}_\lambda)\right)_{j=r+1}^{n-p}$.

For a given $\lambda$, we denote $E_{\lambda,1}$ the eigenspace spanned by the first $r$ eigenvalue and $E_{\lambda,2}$ the eigenspace spanned by the last eigenvalues such that

$$\mathbb{R}^{n-p} = E_{\lambda,1} \oplus E_{\lambda,2} \tag{A.60}$$

We define the limit $\lim_{\lambda\to\infty} E_{\lambda,1} = E_1$ as the eigenspace spanned by the first $r$ eigenvectors of $\mathbf{A}$ and we define the limit $\lim_{\lambda\to\infty} E_{\lambda,2} = E_2$ by its orthogonal $E_2 = E_1^{\perp}$ because the eigenspaces are of $\mathbf{A}$ mutually orthogonal.

**Lemma A.2.5.** *there exist two positive constants $c_1, c_2 > 0$ such that :*

$$c_1 g_\lambda^* \leq \lambda_{n-p}(\tilde{\mathbf{\Sigma}}_\lambda) \leq c_2 g_\lambda^* \tag{A.61}$$

*Proof.* $\mathrm{Ker}(\mathbf{A})$ is non trivial, $\mathrm{Ker}(\mathbf{A}) \cap \mathrm{Ker}(\mathbf{B})$ is trivial and $\mathbf{\Sigma}_\lambda$ is positive definite for $\lambda$ large enough by Lemma 3.14 (Muré, 2018). The result holds directly from inequalities (A.58). ■

**Lemma A.2.6.** *For all $j \in \{r+1, \ldots, n-p\}$, the eigenvalue $\lambda_j(\tilde{\mathbf{\Sigma}}_\lambda)$ satisfies*

$$\lim_{\lambda\to\infty} \inf \lambda_{n-p}(\tilde{\mathbf{\Sigma}}_\lambda)\lambda_j^{-1}(\tilde{\mathbf{\Sigma}}_\lambda) > 0 \tag{A.62}$$

$$\lim_{\lambda\to\infty} \sup \lambda_{n-p}(\tilde{\mathbf{\Sigma}}_\lambda)\lambda_j^{-1}(\tilde{\mathbf{\Sigma}}_\lambda) \leq 1 \tag{A.63}$$

*Proof.* The lemma is a direct application of result of the inequality A.58 and A.61 which show the existence of two positive constants $c_1, c_2$ such that

$$\frac{c_1}{c_2} \leq \lambda_{n-p}(\tilde{\mathbf{\Sigma}}_\lambda)\lambda_j^{-1}(\tilde{\mathbf{\Sigma}}_\lambda) \leq \lambda_{n-p}(\tilde{\mathbf{\Sigma}}_\lambda)\lambda_{n-p}^{-1}(\tilde{\mathbf{\Sigma}}_\lambda) = 1 \tag{A.64}$$

■

**Assumption A.2.7.** *Let $(\boldsymbol{e}_i)_{i=1}^n$ be the canonical basis. We assume that $\mathbf{W}^\top \boldsymbol{e}_i \notin E_1$ for all $i \in \{1, \ldots, n\}$.*

**Lemma A.2.8.** *With the assumption A.2.7, if $p_{\lambda,1}(\boldsymbol{x})$ (resp. $p_1(\boldsymbol{x})$) designs the orthogonal projector of $\boldsymbol{x}$ on $E_{\lambda,1}$ (resp. $E_1$) then*

$$\lim_{\lambda\to\infty} \|p_{\lambda,1}(\boldsymbol{x})\|^2 = \|p_1(\boldsymbol{x})\|^2. \tag{A.65}$$

*Similarly, if $p_{\lambda,2}(\boldsymbol{x})$ (resp. $p_\lambda(\boldsymbol{x})$) designs the orthogonal projector of $\boldsymbol{x}$ on $E_{\lambda,2}$ (resp. $E_2$) then*

$$\lim_{\lambda\to\infty} \|p_{\lambda,2}(\boldsymbol{x})\|^2 = \|p_2(\boldsymbol{x})\|^2. \tag{A.66}$$

*Proof.* The lemma results directly from the convergence of the spaces $E_{\lambda,1}, E_{\lambda,2}$ to $E_1$ and $E_2$. ■

**Lemma A.2.9.** *With Assumption A.2.7, $\lim_{\lambda\to\infty} \inf \lambda_{n-p}(\tilde{\mathbf{\Sigma}}_\lambda)(\mathbf{A}_\lambda)_{ii}, \sup \lambda_{n-p}(\tilde{\mathbf{\Sigma}}_\lambda)(\mathbf{A}_\lambda)_{ii} > 0$ for all $i \in \{1, \ldots, n\}$.*

*Proof.* $\mathbf{A}_\lambda$ is a positive semi-definite matrix

$$\mathbf{A}_\lambda = \sum_{j=1}^{n-p} \lambda_j^{-1}(\tilde{\mathbf{\Sigma}}_\lambda)\ \mathbf{W}\boldsymbol{u}_{\lambda,j}\left(\mathbf{W}\boldsymbol{u}_{\lambda,j}\right)^\top \tag{A.67}$$

We have for all $i \in \{1, \dots, n\}$

$$\begin{aligned}
\lambda_{n-p}(\tilde{\mathbf{\Sigma}}_\lambda)(\mathbf{A}_\lambda)_{ii} &= \lambda_{n-p}(\tilde{\mathbf{\Sigma}}_\lambda)\ \boldsymbol{e}_i^\top \mathbf{A}_\lambda\ \boldsymbol{e}_i \\
&= \sum_{j=1}^{n-p} \lambda_{n-p}(\tilde{\mathbf{\Sigma}}_\lambda)\lambda_j^{-1}(\tilde{\mathbf{\Sigma}}_\lambda)\ \left[\left(\boldsymbol{e}_i^\top \mathbf{W}\right)\boldsymbol{u}_{\lambda,j}\right]^2 \\
&= \sum_{j=1}^{r} \lambda_{n-p}(\tilde{\mathbf{\Sigma}}_\lambda)\lambda_j^{-1}(\tilde{\mathbf{\Sigma}}_\lambda)\ \left[\left(\boldsymbol{e}_i^\top \mathbf{W}\right)\boldsymbol{u}_{\lambda,j}^{(1)}\right]^2 + \\
&\quad \sum_{j=r+1}^{n-p} \lambda_{n-p}(\tilde{\mathbf{\Sigma}}_\lambda)\lambda_j^{-1}(\tilde{\mathbf{\Sigma}}_\lambda)\ \left[\left(\boldsymbol{e}_i^\top \mathbf{W}\right)\boldsymbol{u}_{\lambda,j}^{(2)}\right]^2
\end{aligned} \tag{A.68}$$

On the one hand, from the first inequality of (A.58), the first term of the sum converges to 0. Thus $\liminf_{\lambda\to\infty} \lambda_{n-p}(\tilde{\mathbf{\Sigma}}_\lambda)\lambda_j^{-1}(\tilde{\mathbf{\Sigma}}_\lambda)\ \left[\left(\boldsymbol{e}_i^\top \mathbf{W}\right)\boldsymbol{u}_{\lambda,j}\right]^2 = 0$ for all $j \in \{1, \dots, r\}$.
On the other hand,

$$\begin{aligned}
\sum_{j=r+1}^{n-p} \lambda_{n-p}(\tilde{\mathbf{\Sigma}}_\lambda)\lambda_j^{-1}(\tilde{\mathbf{\Sigma}}_\lambda)\ \left[\left(\boldsymbol{e}_i^\top \mathbf{W}\right)\boldsymbol{u}_{\lambda,j}^{(2)}\right]^2 &\geq \left(\min_{j\in\mathcal{J}} \lambda_{n-p}(\tilde{\mathbf{\Sigma}}_\lambda)\lambda_j^{-1}(\tilde{\mathbf{\Sigma}}_\lambda)\right) \sum_{j=r+1}^{n-p} \left[\left(\boldsymbol{e}_i^\top \mathbf{W}\right)\boldsymbol{u}_{\lambda,j}^{(2)}\right]^2 \\
&\geq \left(\min_{j\in\mathcal{J}} \inf \lambda_{n-p}(\tilde{\mathbf{\Sigma}}_\lambda)\lambda_j^{-1}(\tilde{\mathbf{\Sigma}}_\lambda)\right) \sum_{j=r+1}^{n-p} \left[\left(\boldsymbol{e}_i^\top \mathbf{W}\right)\boldsymbol{u}_{\lambda,j}^{(2)}\right]^2 \\
&\geq \left(\min_{j\in\mathcal{J}} \inf \lambda_{n-p}(\tilde{\mathbf{\Sigma}}_\lambda)\lambda_j^{-1}(\tilde{\mathbf{\Sigma}}_\lambda)\right) \|p_{\lambda,2}(\mathbf{W}^\top \boldsymbol{e}_i)\|^2
\end{aligned} \tag{A.69}$$

Hence, for all $\lambda \in (0, \infty)$

$$\begin{aligned}
\lambda_{n-p}(\tilde{\mathbf{\Sigma}}_\lambda)(\mathbf{A}_\lambda)_{ii} \geq &\sum_{j=1}^{r} \inf \lambda_{n-p}(\tilde{\mathbf{\Sigma}}_\lambda)\lambda_j^{-1}(\tilde{\mathbf{\Sigma}}_\lambda)\ \left[\left(\boldsymbol{e}_i^\top \mathbf{W}\right)\boldsymbol{u}_{\lambda,j}\right]^2 \\
&+ \left(\min_{j\in\mathcal{J}} \inf \lambda_{n-p}(\tilde{\mathbf{\Sigma}}_\lambda)\lambda_j^{-1}(\tilde{\mathbf{\Sigma}}_\lambda)\right) \|p_{\lambda,2}(\mathbf{W}^\top \boldsymbol{e}_i)\|^2
\end{aligned} \tag{A.70}$$

Hence, considering that $\|p_{\lambda,2}(\mathbf{W}^\top \boldsymbol{e}_i)\|^2 \overset{\lambda\to+\infty}{\longrightarrow} \|p_2(\mathbf{W}^\top \boldsymbol{e}_i)\|^2 > 0$ by Lemma A.2.8 and $\liminf_{\lambda\to\infty} \lambda_{n-p}(\tilde{\mathbf{\Sigma}}_\lambda)\lambda_j^{-1}(\tilde{\mathbf{\Sigma}}_\lambda) > 0$ by Lemma A.2.6, we get the desired result on the limit of the lower bound $\liminf_{\lambda\to\infty} \lambda_{n-p}(\tilde{\mathbf{\Sigma}}_\lambda)(\mathbf{A}_\lambda)_{ii}$.

Similarly, we show that, for all $i \in \{1, \dots, n\}$

$$\begin{aligned}
\sum_{j=r+1}^{n-p} \lambda_{n-p}(\tilde{\mathbf{\Sigma}}_\lambda)\lambda_j^{-1}(\tilde{\mathbf{\Sigma}}_\lambda)\ \left[\left(\boldsymbol{e}_i^\top \mathbf{W}\right)\boldsymbol{u}_{\lambda,j}^{(2)}\right]^2 &\leq \left(\max_{j\in\mathcal{J}} \lambda_{n-p}(\tilde{\mathbf{\Sigma}}_\lambda)\lambda_j^{-1}(\tilde{\mathbf{\Sigma}}_\lambda)\right) \sum_{j=r+1}^{n-p} \left[\left(\boldsymbol{e}_i^\top \mathbf{W}\right)\boldsymbol{u}_{\lambda,j}^{(2)}\right]^2 \\
&\leq \left(\max_{j\in\mathcal{J}} \sup \lambda_{n-p}(\tilde{\mathbf{\Sigma}}_\lambda)\lambda_j^{-1}(\tilde{\mathbf{\Sigma}}_\lambda)\right) \sum_{j=r+1}^{n-p} \left[\left(\boldsymbol{e}_i^\top \mathbf{W}\right)\boldsymbol{u}_{\lambda,j}^{(2)}\right]^2 \\
&\leq \left(\max_{j\in\mathcal{J}} \sup \lambda_{n-p}(\tilde{\mathbf{\Sigma}}_\lambda)\lambda_j^{-1}(\tilde{\mathbf{\Sigma}}_\lambda)\right) \|p_{\lambda,2}\left(\mathbf{W}^\top \boldsymbol{e}_i\right)\|^2 \\
&\leq \|p_{\lambda,2}(\mathbf{W}^\top \boldsymbol{e}_i)\|^2. \qquad \text{(by Lemma A.2.6)}
\end{aligned} \tag{A.71}$$

and, for all $\lambda \in (0, \infty)$

$$\begin{aligned}\lambda_{n-p}(\tilde{\boldsymbol{\Sigma}}_\lambda)(\mathbf{A}_\lambda)_{ii} &\leq \sum_{j=1}^{r} \sup \lambda_{n-p}(\tilde{\boldsymbol{\Sigma}}_\lambda)\lambda_j^{-1}(\tilde{\boldsymbol{\Sigma}}_\lambda)\ \left[\left(\boldsymbol{e}_i^\top \mathbf{W}\right)\boldsymbol{u}_{\lambda,j}\right]^2 + \|p_{\lambda,2}(\mathbf{W}^\top \boldsymbol{e}_i)\|^2 \\ &\overset{\lambda\to+\infty}{\longrightarrow} \|p_2(\mathbf{W}^\top \boldsymbol{e}_i)\|^2 > 0,\end{aligned} \tag{A.72}$$

which gives finally the limit of the upper bound $\lim_{\lambda\to\infty} \sup \lambda_{n-p}(\tilde{\boldsymbol{\Sigma}}_\lambda)(\mathbf{A}_\lambda)_{ii}$ and ends the proof.

■

Now we consider the last term of (A.55). Let $i \in \{1, \ldots, n\}$ and let $\boldsymbol{y} \in \mathbb{R}^n$

$$\begin{aligned}\lambda_{n-p}(\tilde{\boldsymbol{\Sigma}}_\lambda)(\mathbf{A}_\lambda \boldsymbol{y})_i &= \sum_{j=1}^{n-p} \lambda_{n-p}(\tilde{\boldsymbol{\Sigma}}_\lambda)\lambda_j^{-1}(\tilde{\boldsymbol{\Sigma}}_\lambda)\ \boldsymbol{e}_i^\top \mathbf{W}\boldsymbol{u}_{\lambda,j}\left(\mathbf{W}\boldsymbol{u}_{\lambda,j}\right)^\top \boldsymbol{y} \\ &= \sum_{j=1}^{r} \lambda_{n-p}(\tilde{\boldsymbol{\Sigma}}_\lambda)\lambda_j^{-1}(\tilde{\boldsymbol{\Sigma}}_\lambda)\ \left(\boldsymbol{e}_i^\top \mathbf{W}\boldsymbol{u}_{\lambda,j}^{(1)}\right)\left(\mathbf{W}\boldsymbol{u}_{\lambda,j}^{(1)}\right)^\top \boldsymbol{y} \\ &\quad + \sum_{j=r+1}^{n-p} \lambda_{n-p}(\tilde{\boldsymbol{\Sigma}}_\lambda)\lambda_j^{-1}(\tilde{\boldsymbol{\Sigma}}_\lambda)\ \left(\boldsymbol{e}_i^\top \mathbf{W}\boldsymbol{u}_{\lambda,j}^{(2)}\right)\left(\mathbf{W}\boldsymbol{u}_{\lambda,j}^{(2)}\right)^\top \boldsymbol{y}.\end{aligned} \tag{A.73}$$

It is clear that the first term of the sum converges to zero by the first inequality of (A.58). Now by Lemma A.2.6, we consider the lower bound of the second term:

$$\begin{aligned}\frac{c_1}{c_2}\sum_{j=r+1}^{n-p}\left(\boldsymbol{e}_i^\top \mathbf{W}\boldsymbol{u}_{\lambda,j}^{(2)}\right)\left(\mathbf{W}\boldsymbol{u}_{\lambda,j}^{(2)}\right)^\top \boldsymbol{y} &= \frac{c_1}{c_2}\sum_{j=r+1}^{n-p}\left(\boldsymbol{e}_i^\top \mathbf{W}\boldsymbol{u}_{\lambda,j}^{(2)}\right)\left(\boldsymbol{u}_{\lambda,j}^{(2)}\right)^\top \mathbf{W}^\top \boldsymbol{y} \\ &= \frac{c_1}{c_2}\left(\sum_{j=r+1}^{n-p}\left(\boldsymbol{e}_i^\top \mathbf{W}\boldsymbol{u}_{\lambda,j}^{(2)}\right)\left(\boldsymbol{u}_{\lambda,j}^{(2)}\right)^\top\right)\mathbf{W}^\top \boldsymbol{y} \\ &= \frac{c_1}{c_2}\left[p_{\lambda,2}\left(\mathbf{W}^\top \boldsymbol{e}_i\right)\right]^\top \mathbf{W}^\top \boldsymbol{y} \\ &\overset{\lambda\to+\infty}{\longrightarrow} \frac{c_1}{c_2}\left[p_2\left(\mathbf{W}^\top \boldsymbol{e}_i\right)\right]^\top \mathbf{W}^\top \boldsymbol{y}.\end{aligned} \tag{A.74}$$

Similarly, the upper bound of the second term satisfies

$$\sum_{j=r+1}^{n-p}\left(\boldsymbol{e}_i^\top \mathbf{W}\boldsymbol{u}_{\lambda,j}^{(2)}\right)\left(\mathbf{W}\boldsymbol{u}_{\lambda,j}^{(2)}\right)^\top \boldsymbol{y} \overset{\lambda\to+\infty}{\longrightarrow} \left[p_2\left(\mathbf{W}^\top \boldsymbol{e}_i\right)\right]^\top \mathbf{W}^\top \boldsymbol{y}. \tag{A.75}$$

Unfortunately, we do not have any proof if the lower bound would be positive nor if the upper bound is negative. In addition, even though $p_2\left(\mathbf{W}^\top \boldsymbol{e}_i\right) \neq \mathbf{0}$ by Assumption A.2.7, it might happens that $\left[p_2\left(\mathbf{W}^\top \boldsymbol{e}_i\right)\right]^\top \mathbf{W}^\top \boldsymbol{y}$ gives zero, leading therefore to an indeterminate form. Requiring the vector product to be non zero would imply additional assumptions that may restrict heavily the set of observations $\boldsymbol{y}$. Finally, we have noticed numerically that the

asymptotic behaviour is unstable as $\left(\overline{\mathbf{R}}_{\lambda\boldsymbol{\theta}_0}\boldsymbol{y}\right)_i$ for $i \in \{1, \ldots, n\}$ oscillates randomly through zero.

For these considerations, we conclude that the coercivity cannot be guaranteed theoretically for Matèrn kernels with smoothness parameters $\nu > 2$.

# APPENDIX B

# Appendix for Part II

## B.1 Proofs of some propositions of Chapters 4 and 5

**Proof of Proposition 4.2.5: Identification of the counterfactual response using observed quantities.**

*Proof.* Let $(\boldsymbol{x}, t, y_{\text{obs}}, y_{\text{cf}}) \in \mathcal{D} \times \{0,1\} \times \mathbb{R} \times \mathbb{R}$ and let $p_{(\cdot)}$ refer to the joint/conditional distribution of the corresponding random variable. We have

$$
\begin{aligned}
p_{Y_{\text{cf}}|\boldsymbol{X},T,Y_{\text{obs}}}(y_{\text{cf}} \mid \boldsymbol{x}, t, y_{\text{obs}}) &= (1-t)\, p_{Y(1)|\boldsymbol{X},T,Y(0)}(y_{\text{cf}} \mid \boldsymbol{x}, t, y_{\text{obs}}) + t\, p_{Y(0)|\boldsymbol{X},T,Y(1)}(y_{\text{cf}} \mid \boldsymbol{x}, t, y_{\text{obs}}) \\
&= (1-t)\, \frac{p_{\boldsymbol{X},T,Y(0),Y(1)}(\boldsymbol{x}, 0, y_{\text{obs}}, y_{\text{cf}})}{p_{\boldsymbol{X},T,Y(0)}(\boldsymbol{x}, 0, y_{\text{obs}})} + t\, \frac{p_{\boldsymbol{X},T,Y(0),Y(1)}(\boldsymbol{x}, 1, y_{\text{cf}}, y_{\text{obs}})}{p_{\boldsymbol{X},T,Y(1)}(\boldsymbol{x}, 1, y_{\text{obs}})} \\
&= (1-t)\, \frac{p_{\boldsymbol{X},T,Y(0),Y(1)}(\boldsymbol{x}, 0, y_{\text{obs}}, y_{\text{cf}})}{\int p_{\boldsymbol{X},T,Y(0),Y(1)}(\boldsymbol{x}, 0, y_{\text{obs}}, y')\, \mathrm{d}y'} + t\, \frac{p_{\boldsymbol{X},T,Y(0),Y(1)}(\boldsymbol{x}, 1, y_{\text{cf}}, y_{\text{obs}})}{\int p_{\boldsymbol{X},T,Y(0),Y(1)}(\boldsymbol{x}, 1, y', y_{\text{obs}})\, \mathrm{d}y'} \\
&= (1-t)\, \frac{p_{\boldsymbol{X},T,Y(0),Y(1)}(\boldsymbol{x}, t, (1-t)y_{\text{obs}} + ty_{\text{cf}}, ty_{\text{obs}} + (1-t)y_{\text{cf}})}{\int p_{\boldsymbol{X},T,Y(0),Y(1)}(\boldsymbol{x}, t, (1-t)y_{\text{obs}} + ty', ty_{\text{obs}} + (1-t)y')\, \mathrm{d}y'} \\
&\qquad + t\, \frac{p_{\boldsymbol{X},T,Y(0),Y(1)}(\boldsymbol{x}, t, (1-t)y_{\text{obs}} + ty_{\text{cf}}, ty_{\text{obs}} + (1-t)y_{\text{cf}})}{\int p_{\boldsymbol{X},T,Y(0),Y(1)}(\boldsymbol{x}, t, (1-t)y_{\text{obs}} + ty', ty_{\text{obs}} + (1-t)y')\, \mathrm{d}y'}. \\
&= (1-t)\, \frac{p_{\boldsymbol{X},T,Y(0),Y(1)}(\boldsymbol{x}, t, y_0, y_1)}{\int p_{\boldsymbol{X},T,Y(0),Y(1)}(\boldsymbol{x}, t, (1-t)y_{\text{obs}} + ty', ty_{\text{obs}} + (1-t)y')\, \mathrm{d}y'} \\
&\qquad + t\, \frac{p_{\boldsymbol{X},T,Y(0),Y(1)}(\boldsymbol{x}, t, y_0, y_1)}{\int p_{\boldsymbol{X},T,Y(0),Y(1)}(\boldsymbol{x}, t, (1-t)y_{\text{obs}} + ty', ty_{\text{obs}} + (1-t)y')\, \mathrm{d}y'} \\
&= \frac{p_{\boldsymbol{X},T,Y(0),Y(1)}(\boldsymbol{x}, t, y_0, y_1)}{\int p_{\boldsymbol{X},T,Y(0),Y(1)}(\boldsymbol{x}, t, (1-t)y_{\text{obs}} + ty', ty_{\text{obs}} + (1-t)y')\, \mathrm{d}y'},
\end{aligned}
\tag{B.1}
$$

where $y_0 = (1-t)y_{\text{obs}} + ty_{\text{cf}}$, $y_1 = ty_{\text{obs}} + (1-t)y_{\text{cf}}$. The last line shows that, as function of $y_{\text{cf}}$, $p_{Y_{\text{cf}}|\boldsymbol{X},T,Y_{\text{obs}}}$ is proportional to $p_{\boldsymbol{X},T,Y(0),Y(1)}$.

Finally, given that $p_{\boldsymbol{X},T,Y(0),Y(1)}(\boldsymbol{x}, t, y_0, y_1) = p_{\boldsymbol{X},Y(0),Y(1)}(\boldsymbol{x}, y_0, y_1) p_{T|X,Y(0),Y(1)}(t \mid \boldsymbol{x}, y_0, y_1)$, we get the desired result of the proposition. ■

**Proof of Balancing properties (4.122 - 4.124) of the P-Function.**

*Proof.* Let $\phi : \mathbb{R} \to \mathbb{R}$ denote an arbitrary real function.

$$\mathbb{E}[\phi(T) \mid \Theta(\boldsymbol{X})] = \mathbb{E}[\mathbb{E}[\phi(T) \mid \boldsymbol{X}] \mid \Theta(\boldsymbol{X})] = \mathbb{E}\big[\int \phi(t)\pi(t \mid \boldsymbol{X})\,\mathrm{d}t \mid \Theta(\boldsymbol{X})\big]. \tag{B.2}$$

With Assumption (6.3.5), $\int \phi(t)\pi(t \mid \boldsymbol{X})\,\mathrm{d}t$ is $\Theta(\boldsymbol{X})$-measurable, thus

$$\mathbb{E}[\phi(T) \mid \Theta(\boldsymbol{X})] = \int \phi(t)\pi(t \mid \boldsymbol{X})\,\mathrm{d}t = \mathbb{E}[\phi(T) \mid \boldsymbol{X}], \tag{B.3}$$

which proves Equation (4.122): $f_{T|\boldsymbol{X}} = f_{T|\Theta(\boldsymbol{X})}$. ■

*Proof.* Let $\phi_1, \phi_2 : \mathbb{R} \to \mathbb{R}$ denote two arbitrary real bounded functions and let $t \in \mathcal{T}$ be an arbitrary treatment value:

$$\mathbb{E}[\phi_1(\boldsymbol{X})\phi_2(T) \mid \Theta(X)] = \mathbb{E}[\phi_1(\boldsymbol{X})\mathbb{E}[\phi_2(T) \mid \boldsymbol{X}] \mid \Theta(X)] \tag{B.4}$$

$$= \mathbb{E}[\phi_1(\boldsymbol{X}) \int \phi_2(t)\pi(t \mid \boldsymbol{X})dt \mid \Theta(\boldsymbol{X})]. \tag{B.5}$$

With Assumption (6.3.5):

$$\mathbb{E}[\phi_1(\boldsymbol{X})\phi_2(T) \mid \Theta(\boldsymbol{X})] = \mathbb{E}[\phi_1(\boldsymbol{X}) \mid \Theta(X)] \int \phi_2(t)\pi(t \mid \boldsymbol{X})dt \tag{B.6}$$

$$= \mathbb{E}[\phi_1(\boldsymbol{X}) \mid \Theta(\boldsymbol{X})]\mathbb{E}[\phi_2(T) \mid \Theta(\boldsymbol{X})], \tag{B.7}$$

which proves Equation (4.123): $\boldsymbol{X} \perp\!\!\!\perp T \mid (\pi(t \mid \boldsymbol{X}))_{t\in\mathcal{T}}$. ■

*Proof.* Let $\phi_1, \phi_2 : \mathbb{R} \to \mathbb{R}$ denote two arbitrary real bounded functions and let $t \in \mathcal{T}$ be an arbitrary treatment value:

$$\mathbb{E}[\phi_1(Y(t))\phi_2(T) \mid \Theta(\boldsymbol{X})] = \mathbb{E}[\mathbb{E}[\phi_1(Y(t))\phi_2(T) \mid \boldsymbol{X}] \mid \Theta(\boldsymbol{X})]. \tag{B.8}$$

By the uncounfoundedness assumption 4.5.3, $Y(t)$ and $T$ are conditionally independent to $\boldsymbol{X}$

$$\mathbb{E}[\phi_1(Y(t))\phi_2(T) \mid \Theta(\boldsymbol{X})] = \mathbb{E}[\mathbb{E}[\phi_1(Y(t)) \mid \boldsymbol{X}]\mathbb{E}[\phi_2(T) \mid \boldsymbol{X}] \mid \Theta(\boldsymbol{X})] \tag{B.9}$$

$$= \mathbb{E}[\mathbb{E}[\phi_1(Y(t)) \mid \boldsymbol{X}] \int \phi_2(t)\pi(t \mid \boldsymbol{X})\,\mathrm{d}t \mid \Theta(\boldsymbol{X})] \tag{B.10}$$

With Assumption (6.3.5), $\int \phi_2(t)\pi(t \mid \boldsymbol{X})\,\mathrm{d}t$ is $\Theta(\boldsymbol{X})$-measurable, thus

$$\mathbb{E}[\phi_1(Y(t))\phi_2(T) \mid \Theta(\boldsymbol{X})] = \mathbb{E}\big[\mathbb{E}[\phi_1(Y(t)) \mid \boldsymbol{X}] \mid \Theta(\boldsymbol{X})\big] \int \phi_2(t)\pi(t \mid \boldsymbol{X})\,\mathrm{d}t \tag{B.11}$$

$$= \mathbb{E}\big[\phi_1(Y(t)) \mid \Theta(\boldsymbol{X})\big] \int \phi_2(t)\pi(t \mid \boldsymbol{X})\,\mathrm{d}t \tag{B.12}$$

$$= \mathbb{E}\big[\phi_1(Y(t)) \mid \Theta(\boldsymbol{X})\big]\mathbb{E}\big[\phi_2(T) \mid \Theta(\boldsymbol{X})\big], \tag{B.13}$$

which proves Equation (4.124): $\forall t \in \mathcal{T} :\ Y(t) \perp\!\!\!\perp T \mid \Theta(\boldsymbol{X})$. ■

**Proof of Proposition 5.2.1: Regularizing the T-learner to selection bias.**

*Proof.* This proof is similar to the proof of equation (5) in supplementary of Curth & van der Schaar (2021a). Let $p_{\boldsymbol{X}}(\boldsymbol{x})$ denote the probability distribution function of $\boldsymbol{X}$, let $p(\boldsymbol{x} \mid T = t)$ denote the probability distribution function of $\boldsymbol{X}$ given $T = t$ and let $R_t = \int (\widehat{\mu}_t(\boldsymbol{x}) - \mu_t(\boldsymbol{x}))^2 p(\boldsymbol{x} \mid T = t) d\boldsymbol{x}$

$$
\begin{aligned}
&\mathbb{E}_{\boldsymbol{X}\sim p(\cdot)}\left[(\widehat{\mu}_t(\boldsymbol{X}) - \mu_t(\boldsymbol{X}))^2\right] = \int (\widehat{\mu}_t(\boldsymbol{x}) - \mu_t(\boldsymbol{x}))^2 p(\boldsymbol{x}) d\boldsymbol{x} \\
&= \mathbb{P}(T = t) \int (\widehat{\mu}_t(\boldsymbol{x}) - \mu_t(\boldsymbol{x}))^2 p(\boldsymbol{x} \mid T = t) d\boldsymbol{x} + \sum_{t' \neq t} \mathbb{P}(T = t') \int (\widehat{\mu}_t(\boldsymbol{x}) - \mu_t(\boldsymbol{x}))^2 p(\boldsymbol{x} \mid T = t') d\boldsymbol{x} \\
&= \mathbb{P}(T = t) R_t + \sum_{t' \neq t} \mathbb{P}(T = t') \int (\widehat{\mu}_t(\boldsymbol{x}) - \mu_t(\boldsymbol{x}))^2 \frac{p(\boldsymbol{x} \mid T = t')}{p(\boldsymbol{x} \mid T = t)} p(\boldsymbol{x} \mid T = t) d\boldsymbol{x} \\
&= \mathbb{P}(T = t) R_t + \sum_{t' \neq t} \mathbb{P}(T = t') \int (\widehat{\mu}_t(\boldsymbol{x}) - \mu_t(\boldsymbol{x}))^2 \frac{\frac{\mathbb{P}(T=t' \mid \boldsymbol{x}) p(\boldsymbol{x})}{\mathbb{P}(T=t')}}{\frac{\mathbb{P}(T=t \mid \boldsymbol{x}) p(\boldsymbol{x})}{\mathbb{P}(T=t)}} p(\boldsymbol{x} \mid T = t) d\boldsymbol{x} \qquad \text{(Bayes rule)} \\
&= \mathbb{P}(T = t) R_t + \mathbb{P}(T = t) \sum_{t' \neq t} \int (\widehat{\mu}_t(\boldsymbol{x}) - \mu_t(\boldsymbol{x}))^2 \frac{\mathbb{P}(T = t' \mid \boldsymbol{x})}{\mathbb{P}(T = t \mid \boldsymbol{x})} p(\boldsymbol{x} \mid T = t) d\boldsymbol{x} \\
&= \mathbb{P}(T = t) R_t + \mathbb{P}(T = t) \int (\widehat{\mu}_t(\boldsymbol{x}) - \mu_t(\boldsymbol{x}))^2 \frac{\sum_{t' \neq t} \mathbb{P}(T = t' \mid \boldsymbol{x})}{\mathbb{P}(T = t \mid \boldsymbol{x})} p(\boldsymbol{x} \mid T = t) d\boldsymbol{x} \\
&= \mathbb{P}(T = t) R_t + \mathbb{P}(T = t) \int \frac{1 - r(t, \boldsymbol{x})}{r(t, \boldsymbol{x})} (\widehat{\mu}_t(\boldsymbol{x}) - \mu_t(\boldsymbol{x}))^2 p(\boldsymbol{x} \mid T = t) d\boldsymbol{x} \\
&= \mathbb{P}(T = t) \int \left(1 + \frac{1 - r(t, \boldsymbol{x})}{r(t, \boldsymbol{x})}\right) (\widehat{\mu}_t(\boldsymbol{x}) - \mu_t(\boldsymbol{x}))^2 p(\boldsymbol{x} \mid T = t) d\boldsymbol{x} \\
&= \mathbb{E}_{\boldsymbol{X}\sim p_{(\cdot \mid T=t)}} \left[\frac{\mathbb{P}(T = t)}{r(t, \boldsymbol{X})} (\widehat{\mu}_t(\boldsymbol{X}) - \mu_t(\boldsymbol{X}))^2\right].
\end{aligned}
\tag{B.14}
$$

■

## B.2 Error estimation of two-step meta-learners.

In the following subsection, we will analyze the error estimation of each two-step meta-learner. Given the assumption (5.3.1) stating that the observations are generated from a function $f$ respecting the two causal assumptions (4.5.1-4.5.2), each unit $i$ has the following observed and potential outcomes

$$
\begin{aligned}
y_i = Y_{\text{obs},i} &= f(t_i, \boldsymbol{x}^{(i)}) + \epsilon_i, \\
Y_i(t) &= f(t, \boldsymbol{x}^{(i)}) + \epsilon_i(t), \\
Y_i(t^{(0)}) &= f(t^{(0)}, \boldsymbol{x}^{(i)}) + \epsilon_i(t^{(0)}).
\end{aligned}
\tag{B.15}
$$

where $\epsilon_i(t^{(0)})$ and $\epsilon_i(t^{(0)})$ are some Gaussian noise like $\epsilon$.

**Remark B.2.1.** *We recall that $(Y_i(t))_{1\leq i\leq n}$ and $(Y_i(t^{(0)}))_{1\leq i\leq n}$ are virtual vectors and cannot be observed.*

The CATE model $\tau_k$ for each $k = 1, \dots, K$ can be written as:

$$\begin{aligned}\tau_k(\boldsymbol{x}) &= \mathbb{E}(Y(t^{(k)}) - Y(t^{(0)}) \mid \boldsymbol{X} = \boldsymbol{x}) \\ &= \mathbb{E}(f(t^{(k)}, \boldsymbol{X}) - f(t^{(0)}, \boldsymbol{X}) + \epsilon^* \mid \boldsymbol{X} = \boldsymbol{x}) \\ &= f(t^{(k)}, \boldsymbol{x}) - f(t^{(0)}, \boldsymbol{x})\end{aligned} \tag{B.16}$$

with $\epsilon^*$ is a noise independent of $\boldsymbol{X}$ and satisfying $\mathbb{E}(\epsilon^*) = 0$.

Under the assumption 5.3.2, we write $\tau_k(\mathbf{X}) = f(t^{(k)}, \mathbf{X}) - f(t^{(0)}, \mathbf{X}) = \mathbf{H}\boldsymbol{\beta}_k^*$ where $\boldsymbol{\beta}_k^* = \boldsymbol{\beta}_{t^{(k)}} - \boldsymbol{\beta}_{t^{(0)}}$ and $\mathbf{H} = (\mathbf{H}_{ij}) \in \mathbb{R}^{n \times p}$ is the regression matrix, assumed to be full rank matrix, such that $\mathbf{H}_{ij} = f_j(\boldsymbol{x}^{(i)})$ for $i = 1, \dots, n$ and $j = 0, \dots, p-1$. With pseudo-outcome meta-learners, we consider a random variable $Z_k$ for a fixed $t^{(k)}$ such that

$$Z_{k,i} = A_{t^{(k)}}(t_i, \boldsymbol{x}^{(i)}) y_i + B_{t^{(k)}}(t_i, \boldsymbol{x}^{(i)}), \quad i = 1, \dots, n,$$

where the functions $A_{t^{(k)}}(T, \boldsymbol{X})$ and $B_{t^{(k)}}(T, \boldsymbol{X})$ are given for each pseudo-outcome meta-learners.

The regression coefficients $\widehat{\boldsymbol{\beta}}_k$ are given by the Ordinary Least Squares (OLS) method

$$\widehat{\boldsymbol{\beta}}_k = (\mathbf{H}^\top \mathbf{H})^{-1} \mathbf{H}^\top \boldsymbol{z}_k, \tag{B.17}$$

where $\boldsymbol{z}_k = (Z_{k,i})_{1 \leq i \leq n}$. Thus,

$$\begin{aligned}
&\widehat{\boldsymbol{\beta}}_k = (\mathbf{H}^\top \mathbf{H})^{-1} \mathbf{H}^\top \boldsymbol{z}_k \\
&= (\mathbf{H}^\top \mathbf{H})^{-1} \mathbf{H}^\top (A_{t^{(k)}}(t_i, \boldsymbol{x}^{(i)}) Y_{\text{obs},i} + B_{t^{(k)}}(t_i, \boldsymbol{x}^{(i)}))_{i=1}^n \\
&= (\mathbf{H}^\top \mathbf{H})^{-1} \mathbf{H}^\top (A_{t^{(k)}}(t_i, \boldsymbol{x}^{(i)}) f(t_i, \boldsymbol{x}^{(i)}) + B_{t^{(k)}}(t_i, \boldsymbol{x}^{(i)}) + A_{t^{(k)}}(t_i, \boldsymbol{x}^{(i)}) \epsilon_i)_{i=1}^n \\
&= (\mathbf{H}^\top \mathbf{H})^{-1} \mathbf{H}^\top (\tau_k(\boldsymbol{x}) + A_{t^{(k)}}(t_i, \boldsymbol{x}^{(i)}) f(t_i, \boldsymbol{x}^{(i)}) - \tau_k(\boldsymbol{x}) + B_{t^{(k)}}(t_i, \boldsymbol{x}^{(i)}) + A_{t^{(k)}}(t_i, \boldsymbol{x}^{(i)}) \epsilon_i)_{i=1}^n \\
&= (\mathbf{H}^\top \mathbf{H})^{-1} \mathbf{H}^\top (\mathbf{H}\boldsymbol{\beta}_k^* + A_{t^{(k)}}(t_i, \boldsymbol{x}^{(i)}) f(t_i, \boldsymbol{x}^{(i)}) - \tau_k(\boldsymbol{x}) + B_{t^{(k)}}(t_i, \boldsymbol{x}^{(i)}) + A_{t^{(k)}}(t_i, \boldsymbol{x}^{(i)}) \epsilon_i)_{i=1}^n \\
&= \boldsymbol{\beta}_k^* + (\mathbf{H}^\top \mathbf{H})^{-1} \mathbf{H}^\top (A_{t^{(k)}}(t_i, \boldsymbol{x}^{(i)}) f(t_i, \boldsymbol{x}^{(i)}) - \tau_k(\boldsymbol{x}) + B_{t^{(k)}}(t_i, \boldsymbol{x}^{(i)}) + A_{t^{(k)}}(t_i, \boldsymbol{x}^{(i)}) \epsilon_i)_{i=1}^n \\
&= \boldsymbol{\beta}_k^* + (\mathbf{H}^\top \mathbf{H})^{-1} \mathbf{H}^\top \tilde{\boldsymbol{\epsilon}}_k
\end{aligned}$$

where $\tilde{\epsilon}_{k,i} = \psi_k(t_i, \boldsymbol{x}^{(i)}) + A_{t^{(k)}}(t_i, \boldsymbol{x}^{(i)}) \epsilon_i$ and $\psi_k(t_i, \boldsymbol{x}^{(i)}) = A_{t^{(k)}}(t_i, \boldsymbol{x}^{(i)}) f(t_i, \boldsymbol{x}^{(i)}) - \tau_k(\boldsymbol{x}^{(i)}) + B_{t^{(k)}}(t_i, \boldsymbol{x}^{(i)})$ to simplify notations.

Let us consider the random vector $\boldsymbol{Z}_k^{(n)}$ such that

$$\boldsymbol{Z}_k^{(n)} = \left(\frac{1}{n}(\mathbf{H}^\top \tilde{\boldsymbol{\epsilon}}_k)_1, \dots, \frac{1}{n}(\mathbf{H}^\top \tilde{\boldsymbol{\epsilon}}_k)_p, \frac{1}{n}(\mathbf{H}^\top \mathbf{H})_{11}, \dots, \frac{1}{n}(\mathbf{H}^\top \mathbf{H})_{pp}\right)^\top \in \mathbb{R}^{p+p^2}, \tag{B.18}$$

that allows us to write $\widehat{\boldsymbol{\beta}}_k$ as

$$\begin{aligned}\widehat{\boldsymbol{\beta}}_k &= \boldsymbol{\beta}_k^* + (\mathbf{H}^\top \mathbf{H})^{-1} \mathbf{H}^\top \tilde{\boldsymbol{\epsilon}}_k \\ &= \boldsymbol{\beta}_k^* + \left(\frac{1}{n}\mathbf{H}^\top \mathbf{H}\right)^{-1} \left(\frac{1}{n}\mathbf{H}^\top \tilde{\boldsymbol{\epsilon}}_k\right) \\ &= \boldsymbol{\beta}_k^* + \phi(\boldsymbol{Z}_k^{(n)}),\end{aligned} \tag{B.19}$$

where $\phi : \mathbb{R}^{p+p^2} \longrightarrow \mathbb{R}^p$ is a $\mathcal{C}^1$-function.

In order to apply the Central Limit Theorem (CLT) later, we show that the vector $\boldsymbol{Z}_k^{(n)}$ can be written as sum of *i.i.d* random vectors $\boldsymbol{Z}_i^{(k)}$.

$$\begin{aligned}\boldsymbol{Z}_k^{(n)} &= (\frac{1}{n}(\mathbf{H}^\top\tilde{\boldsymbol{\epsilon}}_k)_1,\ldots,\frac{1}{n}(\mathbf{H}^\top\tilde{\boldsymbol{\epsilon}}_k)_p,\frac{1}{n}(\mathbf{H}^\top\mathbf{H})_{11},\ldots,\frac{1}{n}(\mathbf{H}^\top\mathbf{H})_{pp})^\top \in \mathbb{R}^{p+p^2}\\ &= (\frac{1}{n}\sum_{i=1}^n \mathbf{H}_{i1}\tilde{\epsilon}_i,\ldots,\mathbf{H}_{ip}\tilde{\epsilon}_i,\frac{1}{n}\sum_{i=1}^n \mathbf{H}_{i1}\mathbf{H}_{i1},\ldots,\frac{1}{n}\sum_{i=1}^n \mathbf{H}_{ip}\mathbf{H}_{ip})^\top\\ &= \frac{1}{n}\sum_{i=1}^n (\mathbf{H}_{i1}\tilde{\epsilon}_i,\ldots,\mathbf{H}_{ip}\tilde{\epsilon}_i,\mathbf{H}_{i1}\mathbf{H}_{i1},\ldots,\mathbf{H}_{ip}\mathbf{H}_{ip})^\top = \frac{1}{n}\sum_{i=1}^n \boldsymbol{Z}_i^{(k)}.\end{aligned} \tag{B.20}$$

The mean $\boldsymbol{m}$ of the vector $\boldsymbol{Z}_k^{(n)}$ satisfies

$$\begin{aligned}\boldsymbol{m} = \mathbb{E}(\boldsymbol{Z}_k^{(n)}) &= \frac{1}{n}\sum_{i=1}^n \mathbb{E}(\boldsymbol{Z}_i^{(k)}) = \mathbb{E}(\boldsymbol{Z}_i^{(k)})\\ &= \Big(h_1,\ldots,h_p,F_{11},\ldots,F_{pp}\Big)^\top,\end{aligned} \tag{B.21}$$

where

$$\begin{aligned}h_j = \mathbb{E}\big[f_j(\boldsymbol{X})(\psi_k(T,\boldsymbol{X}) + A_{t^{(k)}}(T,\boldsymbol{X})\epsilon)\big] = \mathbb{E}(f_j(\boldsymbol{X})\psi_k(T,\boldsymbol{X}))\\ F_{jj'} = \mathbb{E}(f_j(\boldsymbol{X})f_{j'}(\boldsymbol{X})),\end{aligned} \tag{B.22}$$

and a covariance matrix $\mathbf{C}$ with entries

$$\begin{aligned}\mathbf{C}_{jj'} &= \mathrm{Cov}\left(\boldsymbol{Z}_j^{(k)},\boldsymbol{Z}_{j'}^{(k)}\right) = \mathbb{E}(\boldsymbol{Z}_j^{(k)},\boldsymbol{Z}_{j'}^{(k)}) - \mathbb{E}(\boldsymbol{Z}_j^{(k)})\mathbb{E}(\boldsymbol{Z}_{j'}^{(k)})\\ &= \begin{cases}\mathbb{E}(f_j(\boldsymbol{X})f_{j'}(\boldsymbol{X})(\psi_k(T,\boldsymbol{X}) + A_{t^{(k)}}(T,\boldsymbol{X})\epsilon)^2) - h_jh_{j'} & \text{if } j,j'\in\{1,\ldots,p\}\\ \mathbb{E}(f_{\tilde{k}}(\boldsymbol{X})f_{\tilde{k}'}(\boldsymbol{X})f_l(\boldsymbol{X})f_{l'}(\boldsymbol{X})) - F_{kk'}F_{ll'} & \text{if } j,j'\in\{p+1,\ldots,p^2\}\\ \mathbb{E}(f_{\tilde{k}}(\boldsymbol{X})f_{\tilde{k}'}(\boldsymbol{X})f_j(\boldsymbol{X})(\psi_k(T,\boldsymbol{X}) + A_{t^{(k)}}(T,\boldsymbol{X})\epsilon)) - h_jF_{kk'} & \text{otherwise.}\end{cases}\\ &= \begin{cases}\mathbb{E}(f_j(\boldsymbol{X})f_{j'}(\boldsymbol{X})\psi_k^2(T,\boldsymbol{X})) + \sigma^2\mathbb{E}(f_j(\boldsymbol{X})f_{j'}(\boldsymbol{X})A_{t^{(k)}}^2(T,\boldsymbol{X})) - h_jh_{j'} & \text{if } j,j'\in\{1,\ldots,p\}\\ \mathbb{E}(f_{\tilde{k}}(\boldsymbol{X})f_{\tilde{k}'}(\boldsymbol{X})f_l(\boldsymbol{X})f_{l'}(\boldsymbol{X})) - F_{kk'}F_{ll'} & \text{if } j,j'\in\{p+1,\ldots,p^2\}\\ \mathbb{E}(f_{\tilde{k}}(\boldsymbol{X})f_{\tilde{k}'}(\boldsymbol{X})f_j(\boldsymbol{X})\psi_k(T,\boldsymbol{X})) - h_jF_{kk'} & \text{otherwise,}\end{cases}\end{aligned} \tag{B.23}$$

where $\tilde{k},\tilde{k}' = \eta^{-1}(j)$ (respectively, $l,l' = \eta^{-1}(j')$) such that $\eta$ is the correspondence indexes map between $\boldsymbol{m}$ and $F$ in $\boldsymbol{m}_j = F_{\tilde{k}\tilde{k}'}$ (respectively, $\boldsymbol{m}_{j'} = F_{ll'}$) when $j \geq p+1$ (respectively, $j' \geq p+1$).

By considering now the vector

$$\boldsymbol{S}^{(n)} = \sqrt{n}(\boldsymbol{Z}_k^{(n)} - \boldsymbol{m}) = \frac{1}{\sqrt{n}}\sum_{i=1}^n (\boldsymbol{Z}_i^{(k)} - \boldsymbol{m}), \tag{B.24}$$

one can show by the multivariate Central Limit Theorem (CLT) that

$$\boldsymbol{S}^{(n)} = \sqrt{n}(\boldsymbol{Z}_k^{(n)} - \boldsymbol{m}) \xrightarrow{\mathcal{L}} \mathcal{N}(\mathbf{0},\mathbf{C}). \tag{B.25}$$

This allows us to write $\widehat{\boldsymbol{\beta}}_k$ as function of $\boldsymbol{S}^{(n)}$ and $\boldsymbol{m}$. Indeed,

$$\begin{aligned}\widehat{\boldsymbol{\beta}}_k &= \boldsymbol{\beta}_k^* + (\mathbf{H}^\top \mathbf{H})^{-1}\mathbf{H}^\top \tilde{\boldsymbol{\epsilon}} \\ &= \boldsymbol{\beta}_k^* + \phi(\boldsymbol{Z}^{(n)}) \\ &= \boldsymbol{\beta}_k^* + \phi(\boldsymbol{m} + \boldsymbol{S}^{(n)}/\sqrt{n}) \\ &= \boldsymbol{\beta}_k^* + \Phi(\boldsymbol{S}^{(n)}, \boldsymbol{m}),\end{aligned} \tag{B.26}$$

where $\Phi : \mathbb{R}^{p+p^2} \times \mathbb{R}^{p+p^2} \longrightarrow \mathbb{R}^p$ is also $\mathcal{C}^1$-function.

Since $\sqrt{n}(\boldsymbol{S}^{(n)} - \mathbf{0}) \xrightarrow{\mathcal{L}} \mathcal{N}(\mathbf{0}, \mathbf{C})$, one obtains by the Delta method

$$\sqrt{n}\Big[\Phi(S^{(n)}, \boldsymbol{m}) - \Phi(\mathbf{0}, \boldsymbol{m})\Big] \xrightarrow{\mathcal{L}} \mathcal{N}\left(\mathbf{0}, J_\Phi^{(1)}(\mathbf{0}, \boldsymbol{m})^\top \mathbf{C} J_\Phi^{(1)}(\mathbf{0}, \boldsymbol{m})\right), \tag{B.27}$$

where $J_\Phi^{(1)}(\mathbf{0}, \boldsymbol{m})$ is the Jacobian matrix at the first $p + p^2$ coordinates of $\Phi$ at $(\mathbf{0}, \boldsymbol{m})$.

By denoting $\boldsymbol{g}_n$, a Gaussian noise with zero-mean and covariance matrix $\mathbf{C}' = J_\Phi^{(1)}(\mathbf{0}, \boldsymbol{m})^\top \mathbf{C} J_\Phi^{(1)}(\mathbf{0}, \boldsymbol{m})$, the previous equation is equivalent to

$$\widehat{\boldsymbol{\beta}}_k = \boldsymbol{\beta}_k^* + \Phi(\boldsymbol{S}_n, \boldsymbol{m}) \approx \boldsymbol{\beta}_k^* + \Phi(\mathbf{0}, \boldsymbol{m}) + \boldsymbol{g}_n/\sqrt{n}. \tag{B.28}$$

For $n$ large, the expansions of the first two moments are of the form:

$$\mathbb{E}(\widehat{\boldsymbol{\beta}}_k) \approx \boldsymbol{\beta}_k^* + \Phi(\mathbf{0}, \boldsymbol{m}). \tag{B.29}$$

and,

$$\mathbb{V}(\widehat{\boldsymbol{\beta}}_k) \approx \frac{1}{n}\, J_\Phi^{(1)}(\mathbf{0}, \boldsymbol{m})^\top \mathbf{C} J_\Phi^{(1)}(\mathbf{0}, \boldsymbol{m}). \tag{B.30}$$

This result holds whether the nuisance parameters in $A_t$ and $B_t$ are well-specified or not, so there is no guarantee that $\Phi(\mathbf{0}, \boldsymbol{m}) = 0$ and the estimator $\widehat{\boldsymbol{\beta}}_k$ may be biased.

In the following, we assume that the nuisance parameters in $A_t$ and $B_t$ are well-specified i.e. $\mathbb{E}(\psi_k(T, \boldsymbol{X})) \mid \boldsymbol{X} = \boldsymbol{x}) = 0$ in such way that $\mathbb{E}(Z_k \mid \boldsymbol{X} = \boldsymbol{x}) = \tau_k(\boldsymbol{x})$, or equivalently, $\mathbb{E}(\mathbf{H}^\top \tilde{\epsilon}_k) = \mathbf{0}$. Consequently, the estimator of $\widehat{\boldsymbol{\beta}}_k$ is unbiased. In this case, computing the variance $\mathbb{V}(\widehat{\boldsymbol{\beta}}_k)$ becomes much easier and explicit.

On the one hand, by the multivariate Central Theorem Limit (CTL)

$$\frac{1}{\sqrt{n}}\mathbf{H}^\top \tilde{\boldsymbol{\epsilon}}_k \xrightarrow{\mathcal{L}} \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}) \tag{B.31}$$

which is equivalent to

$$\frac{1}{\sqrt{n}}\, \mathbf{H}^\top \tilde{\boldsymbol{\epsilon}}_k \;\approx \boldsymbol{g}_n, \tag{B.32}$$

where $\boldsymbol{g}_n$ is a Gaussian noise with zero-mean and covariance matrix of $\boldsymbol{\Sigma}$ with entries

$$\begin{aligned}\boldsymbol{\Sigma}_{jj'} &= \mathbb{E}\big[f_j(\boldsymbol{X}) f_{j'}(\boldsymbol{X})\big(\psi_k(T, \boldsymbol{X}) + A_{t^{(k)}}(T, \boldsymbol{X})\epsilon\big)^2\big] \\ &= \mathbb{E}\big(f_j(\boldsymbol{X}) f_{j'}(\boldsymbol{X})\psi_k^2(T, \boldsymbol{X})\big) + \sigma^2 \mathbb{E}\big(f_j(\boldsymbol{X}) f_{j'}(\boldsymbol{X}) A_{t^{(k)}}^2(T, \boldsymbol{X})\big).\end{aligned} \tag{B.33}$$

On the other hand, by the law of large numbers, we have $1/n(\mathbf{H}^\top\mathbf{H}) \xrightarrow{a.s} \mathbf{F}$, thus $1/n(\mathbf{H}^\top\mathbf{H}) \xrightarrow{P} \mathbf{F}$. Since $\mathbf{F}$ is invertible, then

$$n(\mathbf{H}^\top\mathbf{H})^{-1} \xrightarrow{P} \mathbf{F}^{-1}, \tag{B.34}$$

where $\mathbf{F} = (F_{jj'})_{1\leq j,j'\leq p}$ and $F_{jj'} = \mathbb{E}(f_j(\boldsymbol{X})f_{j'}(\boldsymbol{X}))$.
By Slutsky's theorem,

$$\begin{aligned}\sqrt{n}(\widehat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k^*) &= n(\mathbf{H}^\top\mathbf{H})^{-1} \cdot 1/\sqrt{n}\ \mathbf{H}^\top\tilde{\boldsymbol{\epsilon}} \\ &\xrightarrow{\mathcal{L}} \mathcal{N}(\mathbf{0}, \mathbf{F}^{-1}\boldsymbol{\Sigma}\mathbf{F}^{-1}),\end{aligned} \tag{B.35}$$

which leads to

$$\begin{aligned}&\mathbb{E}(\widehat{\boldsymbol{\beta}}_k) = \boldsymbol{\beta}_k^*, \\ &\mathbb{V}(\widehat{\boldsymbol{\beta}}_k) \approx \frac{1}{n}\ \mathbf{F}^{-1}\boldsymbol{\Sigma}\mathbf{F}^{-1}.\end{aligned} \tag{B.36}$$

The determinant of the variance matrix, also known as the generalized variance by Wilks (1967, 1932) is usually used as a scalar measure of overall multidimensional scatter and can be useful to compare the variance of each meta-learner.

In our case, comparing the generalized variance is equivalent to comparing $\det\left(\frac{1}{n}\boldsymbol{\Sigma}\right)$ of each pseudo-outcome meta-learner since

$$\det\left(\mathbb{V}(\widehat{\boldsymbol{\beta}}_k)\right) = \left(\det\mathbf{F}^{-1}\right)^2 \det\left(\frac{1}{n}\boldsymbol{\Sigma}\right) = \frac{1}{\left(\det\mathbf{F}\right)^2}\det\left(\frac{1}{n}\boldsymbol{\Sigma}\right), \tag{B.37}$$

with, obviously, $\det(\boldsymbol{\Sigma}) > 0$ because $\boldsymbol{\Sigma}$ is symmetric positive definite.

The assumptions (4.5.2-5.3.3) will be used in the following calculations.

## Error estimation of the M-learner

**Lemma B.2.2.** *If $X_1, \ldots, X_m$ is a sequence of random variables and $b > 1$, then*

$$\begin{aligned}&\left|\mathbb{E}\Big[\Big(\sum_{i=1}^m X_i\Big)^2\Big]\right| \leq m\sum_{i=1}^m \mathbb{E}\Big[|X_i^2|\Big], \\ &\left|\mathbb{E}\Big[\Big(\sum_{i=1}^m X_i\Big)^b\Big]\right| \leq m^{(b-1)}\sum_{i=1}^m \mathbb{E}\Big[|X_i^b|\Big].\end{aligned} \tag{B.38}$$

*Proof.* The first inequality is obtained by Cauchy-Schwartz, whereas the second inequality can be proved by Jensen inequality. Indeed, for $b > 1$, the function $x \mapsto x^b$ is convex for $x > 0$ and

$$\left|\frac{\sum_{i=1}^m X_i}{m}\right|^b \leq \frac{\sum_{i=1}^m |X_i|^b}{m}. \tag{B.39}$$

Therefore,

$$\left|\mathbb{E}\Big[\Big(\sum_{i=1}^m X_i\Big)^b\Big]\right| \leq \mathbb{E}\Big[\Big|\sum_{i=1}^m X_i\Big|^b\Big] \leq m^{(b-1)}\sum_{i=1}^m \mathbb{E}\Big[|X_i^b|\Big]. \tag{B.40}$$

■

Let $a, b > 1$ such that $1/a + 1/b = 1$. We assume that $f_j(\boldsymbol{X}) \in L^a$ (i.e. $f_j(\boldsymbol{X})$ has all possible finite moments) for all $j \in \{1, \ldots, p\}$ and we denote $\delta_{jj'}^{(a)} = \Big|\mathbb{E}\big(f_j^a(\boldsymbol{X}) f_{j'}^a(\boldsymbol{X})\big)\Big|^{1/a}$. By Hölder inequality we show that for the M-learner:

$$
\begin{aligned}
\big|\mathbb{E}\big(f_j(\boldsymbol{X}) f_{j'}(\boldsymbol{X}) \psi_k^2(T, \boldsymbol{X})\big)\big| &\leq \Big|\mathbb{E}\big(f_j^a(\boldsymbol{X}) f_{j'}^a(\boldsymbol{X})\big)\Big|^{1/a} \cdot \Big|\mathbb{E}\big(\psi_k^{2b}(T, \boldsymbol{X})\big)\Big|^{1/b} \qquad \text{(Hölder)} \\
&\leq \delta_{jj'}^{(a)} \Big(2^{2b-1}\ \mathbb{E}\Big[\Big(\frac{\mathbf{1}\{T = t^{(k)}\}}{r(t^{(k)}, \boldsymbol{X})} - 1\Big)^{2b} f^{2b}(t^{(k)}, \boldsymbol{X}) \\
&\qquad + \Big(\frac{\mathbf{1}\{T = t^{(k)}\}}{r(t^{(k)}, \boldsymbol{X})} - 1\Big)^{2b} f^{2b}(t^{(0)}, \boldsymbol{X})\Big]\Big)^{1/b} \\
&\qquad \text{(Lemma B.2.2 with } m = 2) \\
&\leq 2^{(2b-1)/b}\ \delta_{jj'}^{(a)} \Big(\mathbb{E}\Big[2^{2b-1}\Big(\frac{\mathbf{1}\{T = t^{(k)}\}}{r^{2b}(t, \boldsymbol{X})} + 1\Big) f^{2b}(t^{(k)}, \boldsymbol{X})\Big] \\
&\qquad + \mathbb{E}\Big[2^{2b-1}\Big(\frac{\mathbf{1}\{T = t^{(0)}\}}{r^{2b}(t^{(0)}, \boldsymbol{X})} + 1\Big) f^{2b}(t^{(0)}, \boldsymbol{X})\Big]\Big)^{1/b} \quad \text{(Lemma B.2.2)} \\
&\leq 2^{2(2b-1)/b}\ \delta_{jj'}^{(a)} \Big(\mathbb{E}\Big[\mathbb{E}\Big(\frac{\mathbf{1}\{T = t^{(k)}\}}{r^{2b}(t, \boldsymbol{X})} + 1\Big) \mid \boldsymbol{X}\Big) f^{2b}(t^{(k)}, \boldsymbol{X})\Big] \\
&\qquad + \mathbb{E}\Big[\mathbb{E}\Big(\frac{\mathbf{1}\{T = t^{(0)}\}}{r^{2b}(t^{(0)}, \boldsymbol{X})} + 1\Big) \mid \boldsymbol{X}\Big) f^{2b}(t^{(0)}, \boldsymbol{X})\Big]\Big)^{1/b} \\
&\leq 2^{2(2b-1)/b}\ \delta_{jj'}^{(a)} \Big(\mathbb{E}\Big[\Big(\frac{1}{r^{2b-1}(t, \boldsymbol{X})} + 1\Big) f^{2b}(t^{(k)}, \boldsymbol{X})\Big] \\
&\qquad + \mathbb{E}\Big[\Big(\frac{1}{r^{2b-1}(t^{(0)}, \boldsymbol{X})} + 1\Big) f^{2b}(t^{(0)}, \boldsymbol{X})\Big]\Big)^{1/b} \\
&\leq 2^{2(2b-1)/b}\ \delta_{jj'}^{(a)} \Big(\frac{1}{r_{\min}^{2b-1}} + 1\Big)^{1/b} \Big(C^{2b} + C^{2b}\Big)^{1/b} \qquad \text{(Bounding } r \text{ and } f) \\
&\leq 2^{2(2b-1)/b}\ \delta_{jj'}^{(a)} \Big(\frac{1}{r_{\min}^{2b-1}} + \frac{1}{r_{\min}^{2b-1}}\Big)^{1/b} 2^{1/b} C^b \\
&\leq 2^{2(2b-1)/b}\ \delta_{jj'}^{(a)} \frac{2^{1/b}}{r_{\min}^{(2b-1)/b}} 2^{1/b} C^b \\
&\leq 2^4\ \delta_{jj'}^{(a)} \frac{1}{r_{\min}^{(2b-1)/b}} C^b = \frac{16}{r_{\min}^{(2b-1)/b}} \delta_{jj'}^{(a)} C^b.
\end{aligned}
\tag{B.41}
$$

On the other term, one obtains similarly:

$$\begin{aligned}
\left|\mathbb{E}(f_j(\boldsymbol{X})f_{j'}(\boldsymbol{X})A^2_{t^{(k)}}(T,\boldsymbol{X}))\right| &\leq \left|\mathbb{E}(f^a_j(\boldsymbol{X})f^a_{j'}(\boldsymbol{X})\right|^{1/a} \cdot \left|\mathbb{E}(A^{2b}_{t^{(k)}}(T,\boldsymbol{X}))\right|^{1/b} \qquad \text{(Hölder)}\\
&\leq \delta^{(a)}_{jj'} \left|\mathbb{E}(A^{2b}_{t^{(k)}}(T,\boldsymbol{X}))\right|^{1/b}\\
&\leq \delta^{(a)}_{jj'} \left(2^{2b-1}\ \mathbb{E}\Big(\frac{\mathbf{1}\{T=t^{(k)}\}}{r(t^{(k)},\boldsymbol{X})}\Big)^{2b} + \mathbb{E}\Big(\frac{\mathbf{1}\{T=t^{(0)}\}}{r(t^{(0)},\boldsymbol{X})}\Big)^{2b}\right)^{1/b} \qquad \text{(Lemma B.2.2)}\\
&\leq 2^{(2b-1)/b}\sigma^2\delta^{(a)}_{jj'} \left(\mathbb{E}\Big(\frac{\mathbf{1}\{T=t^{(k)}\}}{r^{2b}(t^{(k)},\boldsymbol{X})}\Big) + \mathbb{E}\Big(\frac{\mathbf{1}\{T=t^{(0)}\}}{r^{2b}(t^{(0)},\boldsymbol{X})}\Big)\right)^{1/b}\\
&\leq 2^{(2b-1)/b}\sigma^2\delta^{(a)}_{jj'}\Big(\frac{2}{r^{2b-1}_{\min}}\Big)^{1/b} = \frac{4}{r^{(2b-1)/b}_{\min}}\sigma^2\delta^{(a)}_{jj'}.
\end{aligned} \tag{B.42}$$

Thus, by combining the two terms, one gets:

$$\begin{aligned}
\left|\boldsymbol{\Sigma}^{(\mathrm{M})}_{jj'}\right| &\leq \left|\mathbb{E}(f_j(\boldsymbol{X})f_{j'}(\boldsymbol{X})\psi^2_k(T,\boldsymbol{X}))\right| + \sigma^2\left|\mathbb{E}(f_j(\boldsymbol{X})f_{j'}(\boldsymbol{X})A^2_{t^{(k)}}(T,\boldsymbol{X}))\right|\\
&\leq \frac{16}{r^{(2b-1)/b}_{\min}}\delta^{(a)}_{jj'}C^b + \frac{4}{r^{(2b-1)/b}_{\min}}\sigma^2\delta^{(a)}_{jj'}\\
&\leq \frac{1}{r^{(2b-1)/b}_{\min}}(16\ C^b + 4\sigma^2)\delta^{(b)}_*,
\end{aligned} \tag{B.43}$$

where $\delta^{(b)}_* = \max_{j,j'}\left|\mathbb{E}(f^{b/(b-1)}_j(\boldsymbol{X})f^{b/(b-1)}_{j'}(\boldsymbol{X}))\right|^{(b-1)/b} = \max_{j,j'}\delta^{(a)}_{jj'}$.

Therefore, for all $\epsilon = b-1 > 0$, there exists $C_M = 4C + \sigma^2$ such that

$$\left|\boldsymbol{\Sigma}^{(\mathrm{M})}_{jj'}\right| \leq 4r^{1/(1+\epsilon)-2}_{\min}\delta^{(1+\epsilon)}_* C_M. \tag{B.44}$$

In particular, if $\epsilon \ll 1$ then $1/(1+\epsilon) - 2 \approx -(1+\epsilon)$ and

$$\left|\boldsymbol{\Sigma}^{(\mathrm{M})}_{jj'}\right| \leq \frac{4}{r^{1+\epsilon}_{\min}}\delta^{(1+\epsilon)}_* C_M \tag{B.45}$$

### Error estimation of the DR-learner.

In this case, we have

$$A_{t^{(k)}}(T,\boldsymbol{X}) = \frac{\mathbf{1}\{T=t^{(k)}\}}{r(t^{(k)},\boldsymbol{X})} - \frac{\mathbf{1}\{T=t^{(0)}\}}{r(t^{(0)},\boldsymbol{X})}, \tag{B.46}$$

$$B_{t^{(k)}}(T,\boldsymbol{X}) = \mu_{t^{(k)}}(\boldsymbol{X}) - \mu_{t^{(0)}}(\boldsymbol{X}) - \left(\frac{\mathbf{1}\{T=t^{(k)}\}}{r(t^{(k)},\boldsymbol{X})} - \frac{\mathbf{1}\{T=t^{(0)}\}}{r(t^{(0)},\boldsymbol{X})}\right)\mu_T(\boldsymbol{X}). \tag{B.47}$$

We need just to compute the upper bound of $\mathbb{E}(f_j(\boldsymbol{X})f_{j'}(\boldsymbol{X})\psi_k^2(T,\boldsymbol{X}))$ such that

$$\begin{aligned}
&\psi_k(T,\boldsymbol{X}) = A_{t^{(k)}}(T,\boldsymbol{X})f(T,\boldsymbol{X}) - \tau_k(\boldsymbol{x}) + B_{t^{(k)}}(T,\boldsymbol{X}) \\
&= \Big(\frac{\mathbf{1}\{T=t^{(k)}\}}{r(t^{(k)},\boldsymbol{X})} - 1\Big)f(t^{(k)},\boldsymbol{X}) - \Big(\frac{\mathbf{1}\{T=t^{(0)}\}}{r(t^{(0)},\boldsymbol{X})} - 1\Big)f(t^{(0)},\boldsymbol{X}) + \mu_{t^{(k)}}(\boldsymbol{X})\left(1 - \frac{\mathbf{1}\{T=t^{(k)}\}}{r(t^{(k)},\boldsymbol{X})}\right) \\
&\qquad\qquad - \mu_{t^{(0)}}(\boldsymbol{X})\left(1 - \frac{\mathbf{1}\{T=t^{(0)}\}}{r(t^{(0)},\boldsymbol{X})}\right) \\
&= \Big(\frac{\mathbf{1}\{T=t^{(k)}\}}{r(t^{(k)},\boldsymbol{X})} - 1\Big)\big(f(t^{(k)},\boldsymbol{X}) - \mu_{t^{(k)}}(\boldsymbol{X})\big) - \Big(\frac{\mathbf{1}\{T=t^{(0)}\}}{r(t^{(0)},\boldsymbol{X})} - 1\Big)\big(f(t^{(0)},\boldsymbol{X}) - \mu_{t^{(0)}}(\boldsymbol{X})\big)
\end{aligned} \tag{B.48}$$

Similarly to the previous calculus, we show that for the DR-learner

$$\begin{aligned}
|\mathbb{E}(f_j(\boldsymbol{X})f_{j'}(\boldsymbol{X})\psi_k^2(T,\boldsymbol{X}))| &\leq \Big|\mathbb{E}(f_j^a(\boldsymbol{X})f_{j'}^a(\boldsymbol{X}))\Big|^{1/a} \cdot \Big|\mathbb{E}(\psi_k^{2b}(T,\boldsymbol{X}))\Big|^{1/b} \qquad \text{(Hölder)} \\
&\leq \delta_{jj'}^{(a)}\Big(2^{2b-1}\,\mathbb{E}\Big[\Big(\frac{\mathbf{1}\{T=t^{(k)}\}}{r(t^{(k)},\boldsymbol{X})} - 1\Big)^{2b}\big(f(t^{(k)},\boldsymbol{X}) - \mu_{t^{(k)}}(\boldsymbol{X})\big)^{2b} \\
&\qquad + \Big(\frac{\mathbf{1}\{T=t^{(k)}\}}{r(t^{(k)},\boldsymbol{X})} - 1\Big)^{2b}\big(f(t^{(0)},\boldsymbol{X}) - \mu_{t^{(0)}}(\boldsymbol{X})\big)^{2b}\Big]\Big)^{1/b} \qquad \text{(Lemma B.2.2)} \\
&\leq 2^{(2b-1)/b}\,\delta_{jj'}^{(a)}\Big(\mathbb{E}\Big[\Big(\frac{\mathbf{1}\{T=t^{(k)}\}}{r(t^{(k)},\boldsymbol{X})} - 1\Big)^{2b}\big(f(t^{(k)},\boldsymbol{X}) - \mu_{t^{(k)}}(\boldsymbol{X})\big)^{2b}\Big] \\
&\qquad + \mathbb{E}\Big[\Big(\frac{\mathbf{1}\{T=t^{(k)}\}}{r(t^{(k)},\boldsymbol{X})} - 1\Big)^{2b}\big(f(t^{(0)},\boldsymbol{X}) - \mu_{t^{(0)}}(\boldsymbol{X})\big)^{2b}\Big]\Big)^{1/b} \\
&\leq 2^{(2b-1)/b}\,\delta_{jj'}^{(a)}\Big(\mathbb{E}\Big[2^{2b-1}\,\Big(\frac{\mathbf{1}\{T=t^{(k)}\}}{r^{2b}(t,\boldsymbol{X})} + 1\Big)\big(f(t^{(k)},\boldsymbol{X}) - \mu_{t^{(k)}}(\boldsymbol{X})\big)^{2b}\Big] \\
&\qquad + \mathbb{E}\Big[2^{2b-1}\,\Big(\frac{\mathbf{1}\{T=t^{(0)}\}}{r^{2b}(t^{(0)},\boldsymbol{X})} + 1\Big)\big(f(t^{(0)},\boldsymbol{X}) - \mu_{t^{(0)}}(\boldsymbol{X})\big)^{2b}\Big]\Big)^{1/b} \qquad \text{(Lemma B.2.2)} \\
&\leq 2^{2(2b-1)/b}\,\delta_{jj'}^{(a)}\Big(\mathbb{E}\Big[\,\Big(\frac{1}{r^{2b-1}(t,\boldsymbol{X})} + 1\Big)\big(f(t^{(k)},\boldsymbol{X}) - \mu_{t^{(k)}}(\boldsymbol{X})\big)^{2b}\Big] \\
&\qquad + \mathbb{E}\Big[\Big(\frac{1}{r^{2b-1}(t^{(0)},\boldsymbol{X})} + 1\Big)\big(f(t^{(0)},\boldsymbol{X}) - \mu_{t^{(0)}}(\boldsymbol{X})\big)^{2b}\Big]\Big)^{1/b} \\
&\leq 2^{2(2b-1)/b}\,\delta_{jj'}^{(a)}\Big(\frac{1}{r_{\min}^{(2b-1)/b}} + 1\Big)\Big(\mathbb{E}\Big[\,\big(f(t^{(k)},\boldsymbol{X}) - \mu_{t^{(k)}}(\boldsymbol{X})\big)^{2b}\Big] \\
&\qquad + \mathbb{E}\Big[\big(f(t^{(0)},\boldsymbol{X}) - \mu_{t^{(0)}}(\boldsymbol{X})\big)^{2b}\Big]\Big)^{1/b} \\
&\leq 2^{2(2b-1)/b}\,\delta_{jj'}^{(a)}\Big(\frac{1}{r_{\min}^{(2b-1)/b}} + 1\Big)\Big[\Big(\mathbb{E}\big(f(t^{(k)},\boldsymbol{X}) - \mu_{t^{(k)}}(\boldsymbol{X})\big)^{2b}\Big)^{1/b} \\
&\qquad + \mathbb{E}\big(f(t^{(0)},\boldsymbol{X}) - \mu_{t^{(0)}}(\boldsymbol{X})\big)^{2b}\Big)^{1/b}\Big] \qquad \text{(Subadditivity of } |\,\boldsymbol{X}|^{1/b}\text{)}
\end{aligned} \tag{B.49}$$

Hence,

$$\begin{aligned}\left|\boldsymbol{\Sigma}_{jj'}^{(\mathrm{DR})}\right| &\leq 2^{2(2b-1)/b}\ \delta_{jj'}^{(a)}\Big(\frac{1}{r_{\min}^{(2b-1)/b}}+1\Big)\Big[\Big(\mathbb{E}\big(f(t^{(k)},\boldsymbol{X})-\mu_{t^{(k)}}(\boldsymbol{X})\big)^{2b}\Big)^{1/b} \\ &\qquad + \Big(\mathbb{E}\big(f(t^{(0)},\boldsymbol{X})-\mu_{t^{(0)}}(\boldsymbol{X})\big)^{2b}\Big)^{1/b}\Big] + \frac{4}{r_{\min}^{(2b-1)/b}}\sigma^2\delta_{jj'}^{(a)} \\ &\leq 2^{2(2b-1)/b}\ \delta_{*}^{(b)}\Big(\frac{1}{r_{\min}^{(2b-1)/b}}+1\Big)\Big[\Big(\mathbb{E}\big(f(t^{(k)},\boldsymbol{X})-\mu_{t^{(k)}}(\boldsymbol{X})\big)^{2b}\Big)^{1/b} \\ &\qquad + \Big(\mathbb{E}\big(f(t^{(0)},\boldsymbol{X})-\mu_{t^{(0)}}(\boldsymbol{X})\big)^{2b}\Big)^{1/b}\Big] + \frac{4}{r_{\min}^{(2b-1)/b}}\sigma^2\delta_{*}^{(b)}\end{aligned} \tag{B.50}$$

We consider now $\epsilon = b-1>0$, and we assume that $\epsilon \ll 1$, then

$$\begin{aligned}&2^{2(2b-1)/b}\ \delta_{*}^{(b)}\Big(\frac{1}{r_{\min}^{(2b-1)/b}}+1\Big)\Big[\Big(\mathbb{E}\big(f(t^{(k)},\boldsymbol{X})-\mu_{t^{(k)}}(\boldsymbol{X})\big)^{2b}\Big)^{1/b}+\Big(\mathbb{E}\big(f(t^{(0)},\boldsymbol{X})-\mu_{t^{(0)}}(\boldsymbol{X})\big)^{2b}\Big)^{1/b}\Big] \\ &+\frac{4}{r_{\min}^{(2b-1)/b}}\sigma^2\delta_{*}^{(b)} \approx 4\ \delta_{*}^{(1+\epsilon)}\Big(\frac{1}{r_{\min}^{1+\epsilon}}+1\Big)\Big(\mathbb{E}\big(f(t^{(k)},\boldsymbol{X})-\mu_{t^{(k)}}(\boldsymbol{X})\big)^2+\mathbb{E}\big(f(t^{(0)},\boldsymbol{X})-\mu_{t^{(0)}}(\boldsymbol{X})\big)^2\Big) \\ &\qquad + 4\ \sigma^2\delta_{*}^{(1+\epsilon)}\frac{1}{r_{\min}^{1+\epsilon}}.\end{aligned} \tag{B.51}$$

Consequently,

$$\left|\boldsymbol{\Sigma}_{jj'}^{(\mathrm{DR})}\right| \leq 4\Big(\frac{C_{DR}^{*}+\sigma^2}{r_{\min}^{1+\epsilon}}+C_{DR}^{*}\Big)\delta_{*}^{(1+\epsilon)}, \tag{B.52}$$

where $C_{DR}^{*} = \mathbb{E}\big(f(t^{(k)},\boldsymbol{X})-\mu_{t^{(k)}}(\boldsymbol{X})\big)^2+\mathbb{E}\big(f(t^{(0)},\boldsymbol{X})-\mu_{t^{(0)}}(\boldsymbol{X})\big)^2 = \mathrm{err}(\mu_{t^{(k)}})+\mathrm{err}(\mu_{t^{(0)}})$.

### Error estimation of the X-learner.

In this case, we have

$$A_{t^{(k)}}(T,\boldsymbol{X}) = 2\times\mathbf{1}\{T=t^{(k)}\}-1, \tag{B.53}$$

$$B_{t^{(k)}}(T,\boldsymbol{X}) = (1-\mathbf{1}\{T=t^{(k)}\})\mu_{t^{(k)}}(\boldsymbol{X})-\mu_{t^{(0)}}(\boldsymbol{X})+\sum_{l\neq k}\mathbf{1}\{T=t^{(l)}\}\mu_{t^{(l)}}(\boldsymbol{X}). \tag{B.54}$$

One can write $\psi_k$ as

$$\begin{aligned}\psi_k(T,\boldsymbol{X}) &= A_{t^{(k)}}(T,\boldsymbol{X})f(T,\boldsymbol{X})-\tau_k(\boldsymbol{x})+B_{t^{(k)}}(T,\boldsymbol{X}) \\ &= \big(2\ \mathbf{1}\{T=t^{(k)}\}-1\big)f(T,\boldsymbol{X})-\big(f(t^{(k)},\boldsymbol{X})-f(t^{(0)},\boldsymbol{X})\big)+\big(1-\mathbf{1}\{T=t^{(k)}\}\big) \\ &\qquad \mu_{t^{(k)}}(\boldsymbol{X})-\mu_{t^{(0)}}(\boldsymbol{X})+\sum_{l\neq k}\mathbf{1}\{T=t^{(l)}\}\mu_{t^{(l)}}(\boldsymbol{X}) \\ &= (1-\mathbf{1}\{T=t^{(k)}\})(\mu_{t^{(k)}}(\boldsymbol{X})-f(t^{(k)},\boldsymbol{X}))-(\mu_{t^{(0)}}(\boldsymbol{X})-f(t^{(0)},\boldsymbol{X})) \\ &\qquad +\sum_{l\neq k}\mathbf{1}\{T=t^{(l)}\}\big(\mu_{t^{(l)}}(\boldsymbol{X})-f(t^{(l)},\boldsymbol{X})\big) = a_k+\sum_{l\neq k}b_l.\end{aligned} \tag{B.55}$$

where

$$a_k = \big(1 - \mathbf{1}\{T = t^{(k)}\}\big)(\mu_{t^{(k)}}(\boldsymbol{X}) - f(t^{(k)}, \boldsymbol{X})) - (\mu_{t^{(0)}}(\boldsymbol{X}) - f(t^{(0)}, \boldsymbol{X})), \tag{B.56}$$

$$b_l = \mathbf{1}\{T = t^{(l)}\}(\mu_{t^{(l)}}(\boldsymbol{X}) - f(t^{(l)}, \boldsymbol{X})). \tag{B.57}$$

Similarly to the M- and DR-learners calculus, and using lemma B.2.2:

$$\begin{aligned}
\big|\mathbb{E}(f_j(\boldsymbol{X})f_{j'}(\boldsymbol{X})\psi_k^2(T,\boldsymbol{X}))\big| &\leq \Big|\mathbb{E}(f_j^a(\boldsymbol{X})f_{j'}^a(\boldsymbol{X}))\Big|^{1/a} \cdot \Big|\mathbb{E}(\psi_k^{2b}(T,\boldsymbol{X}))\Big|^{1/b} \\
&\leq \delta_{jj'}^{(a)}\Big|\mathbb{E}\big(a_t + \sum_{l\neq k} b_l\big)^{2b}\Big|^{1/b} \qquad \text{(Hölder)} \\
&\leq \delta_{jj'}^{(a)}\Big(2^{2b-1}\Big(\mathbb{E}(a_t^{2b}) + \mathbb{E}\big(\sum_{l\neq k} b_l\big)^{2b}\Big)\Big)^{1/b} \qquad \text{(Lemma B.2.2 with } m=2) \\
&\leq 2^{(2b-1)/b}\ \delta_{jj'}^{(a)}\Big(\mathbb{E}(a_t^{2b}) + \mathbb{E}\big(\sum_{l\neq k} b_l\big)^{2b}\Big)^{1/b} \\
&\leq 2^{(2b-1)/b}\ \delta_{jj'}^{(a)}\Big[2^{2b-1}\Big(\mathbb{E}\big((1-\mathbf{1}\{T=t^{(k)}\})^{2b}(\mu_{t^{(k)}}(\boldsymbol{X}) - f(t^{(k)},\boldsymbol{X}))^{2b}\big) \\
&\qquad + \mathbb{E}\big(\mu_{t^{(0)}}(\boldsymbol{X}) - f(t^{(0)},\boldsymbol{X})\big)^{2b}\Big) + (K-1)^{2b-1} \\
&\qquad \times \sum_{l\neq k}\mathbb{E}\big(\mathbf{1}\{T=t^{(l)}\}\big(\mu_{t^{(l)}}(\boldsymbol{X}) - f(t^{(l)},\boldsymbol{X})\big)^{2b}\Big]^{1/b} \\
&\quad \text{(Lemma B.2.2 with } m=2 \text{ on the } 1^{\text{st}} \text{ term, and } m=(K-1) \text{ on the } 2^{\text{nd}} \text{ term)} \\
&\leq 2^{(2b-1)/b}\ \delta_{jj'}^{(a)}\Big[2^{2b-1}\Big(\mathbb{E}\big(\mu_{t^{(k)}}(\boldsymbol{X}) - f(t^{(k)},\boldsymbol{X})\big)^{2b} + \mathbb{E}\big(\mu_{t^{(0)}}(\boldsymbol{X}) - f(t^{(0)},\boldsymbol{X})\big)^{2b}\Big) \\
&\qquad + (K-1)^{2b-1}\sum_{l\neq k}\mathbb{E}\big(\mu_{t^{(l)}}(\boldsymbol{X}) - f(t^{(l)},\boldsymbol{X})\big)^{2b}\Big]^{1/b} \\
&\leq 2^{(2b-1)/b}\ \delta_{jj'}^{(a)}\Big[2^{(2b-1)/b}\Big(\mathbb{E}\big(\mu_{t^{(k)}}(\boldsymbol{X}) - f(t^{(k)},\boldsymbol{X})\big)^{2b}\Big)^{1/b} + 2^{(2b-1)/b}\Big(\mathbb{E}\big(\mu_{t^{(0)}}(\boldsymbol{X}) \\
&\quad - f(t^{(0)},\boldsymbol{X})\big)^{2b}\Big)^{1/b} + (K-1)^{(2b-1)/b}\sum_{l\neq k}\Big(\mathbb{E}\big(\mu_{t^{(l)}}(\boldsymbol{X}) - f(t^{(l)},\boldsymbol{X})\big)^{2b}\Big)^{1/b}\Big] \\
&\leq 2^{2(2b-1)/b}\ \delta_{jj'}^{(a)}\Big[\Big(\mathbb{E}\big(\mu_{t^{(k)}}(\boldsymbol{X}) - f(t^{(k)},\boldsymbol{X})\big)^{2b}\Big)^{1/b} + \Big(\mathbb{E}\big(\mu_{t^{(0)}}(\boldsymbol{X}) - f(t^{(0)},\boldsymbol{X})\big)^{2b}\Big)^{1/b} \\
&\quad + \big(\frac{K-1}{2}\big)^{(2b-1)/b}\sum_{l\neq k}\Big(\mathbb{E}\big(\mu_{t^{(l)}}(\boldsymbol{X}) - f(t^{(l)},\boldsymbol{X})\big)^{2b}\Big)^{1/b}\Big].
\end{aligned} \tag{B.58}$$

Given that $\mathbb{E}(f_j(\boldsymbol{X})f_{j'}(\boldsymbol{X})A^2_{t^{(k)}}(T,\boldsymbol{X})) = \mathbb{E}(f_j(\boldsymbol{X})f_{j'}(\boldsymbol{X})) = \delta^{(1)}_{jj'}$, we deduce finally

$$\begin{aligned}
\left|\boldsymbol{\Sigma}^{(\mathrm{X})}_{jj'}\right| &\leq \left|\mathbb{E}(f_j(\boldsymbol{X})f_{j'}(\boldsymbol{X})\psi^2_k(T,\boldsymbol{X}))\right| + \sigma^2\left|\mathbb{E}(f_j(\boldsymbol{X})f_{j'}(\boldsymbol{X})A^2_{t^{(k)}}(T,\boldsymbol{X}))\right| \\
&\leq 2^{2(2b-1)/b}\ \delta^{(a)}_{jj'}\Big[\Big(\mathbb{E}\big(\mu_{t^{(k)}}(\boldsymbol{X}) - f(t^{(k)},\boldsymbol{X})\big)^{2b}\Big)^{1/b} + \Big(\mathbb{E}\big(\mu_{t^{(0)}}(\boldsymbol{X}) - f(t^{(0)},\boldsymbol{X})\big)^{2b}\Big)^{1/b} \\
&\qquad + \big(\frac{K-1}{2}\big)^{(2b-1)/b}\sum_{l\neq k}\Big(\mathbb{E}\big(\mu_{t^{(l)}}(\boldsymbol{X}) - f(t^{(l)},\boldsymbol{X})\big)^{2b}\Big)^{1/b}\Big] + \sigma^2\delta^{(1)}_{jj'} \\
&\leq 2^{2(2b-1)/b}\ \delta^{(b)}_{*}\Big[\Big(\mathbb{E}\big(\mu_{t^{(k)}}(\boldsymbol{X}) - f(t^{(k)},\boldsymbol{X})\big)^{2b}\Big)^{1/b} + \Big(\mathbb{E}\big(\mu_{t^{(0)}}(\boldsymbol{X}) - f(t^{(0)},\boldsymbol{X})\big)^{2b}\Big)^{1/b} \\
&\qquad + \big(\frac{K-1}{2}\big)^{(2b-1)/b}\sum_{l\neq k}\Big(\mathbb{E}\big(\mu_{t^{(l)}}(\boldsymbol{X}) - f(t^{(l)},\boldsymbol{X})\big)^{2b}\Big)^{1/b}\Big] + \sigma^2\delta^{(1)}_{*}
\end{aligned} \tag{B.59}$$

where $\delta^{(1)}_* = \max_{j,j'}\mathbb{E}(f_j(\boldsymbol{X})f_{j'}(\boldsymbol{X}))$.

As in the previous cases, we consider now $\epsilon = b-1 > 0$ with $\epsilon \ll 1$, then

$$\begin{aligned}
&2^{2(2b-1)/b}\ \delta^{(b)}_{*}\Big[\Big(\mathbb{E}\big(\mu_{t^{(k)}}(\boldsymbol{X}) - f(t^{(k)},\boldsymbol{X})\big)^{2b}\Big)^{1/b} + \Big(\mathbb{E}\big(\mu_{t^{(0)}}(\boldsymbol{X}) - f(t^{(0)},\boldsymbol{X})\big)^{2b}\Big)^{1/b} \\
&\qquad + \big(\frac{K-1}{2}\big)^{(2b-1)/b}\sum_{l\neq k}\Big(\mathbb{E}\big(\mu_{t^{(l)}}(\boldsymbol{X}) - f(t^{(l)},\boldsymbol{X})\big)^{2b}\Big)^{1/b}\Big] + \sigma^2\delta^{(1)}_{*} \\
&\approx 4\ \delta^{(1+\epsilon)}_{*}\Big(\mathbb{E}\big(f(t^{(k)},\boldsymbol{X}) - \mu_{t^{(k)}}(\boldsymbol{X})\big)^2 + \mathbb{E}\big(f(t^{(0)},\boldsymbol{X}) - \mu_{t^{(0)}}(\boldsymbol{X})\big)^2 \\
&\qquad + \frac{(K-1)^2}{4}\sum_{l\neq k}\mathbb{E}\big(\mu_{t^{(l)}}(\boldsymbol{X}) - f(t^{(l)},\boldsymbol{X})\big)^2 + \sigma^2\delta^{(1)}_{*}.
\end{aligned} \tag{B.60}$$

Therefore,

$$\left|\boldsymbol{\Sigma}^{(\mathrm{X})}_{jj'}\right| \leq 4\delta^{(1+\epsilon)}_{*}C_X + \sigma^2\delta^{(1)}_{*}. \tag{B.61}$$

where $C_X = \mathrm{err}(\mu_{t^{(k)}}) + \mathrm{err}(\mu_{t^{(0)}}) + \frac{(K-1)^2}{4}\sum_{l\neq k}\mathrm{err}(\mu_{t^{(l)}})$.

## Analysis and comparison:

From equation (B.44), (B.52) and (B.61), one can deduce that:

**M-learner.** The M-learner has the largest variance and its variance upper bound is constant.

**M- and DR-learners.** As the term $r_{\min}$ is present in the denominator of the upper bounds of both M-learners and DR-learners. The variance is likely to be high when there is a lack of overlap in the propensity score, i.e. when $r_{\min}$ is close to 0. In addition, having more treatments values $K$ makes the lower bound $r_{\min}$ smaller because $r_{\min} \leq 1/K$.

**X-learner.** Since the upper bounds of the X-learner and DR-learner depend on the expected squared error $\mathrm{err}(\mu_t) = \mathbb{E}\big[f(t,\boldsymbol{X}) - \mu_t(\boldsymbol{X}))^2\big]$. One can expect that, the more outcome models are precise, the lower the variance is.

**M-learner vs DR-learner.** If the potential outcome models are well-specified, then the expected squared error $\mu_t$ is minimal and the upper bound of $\mathbf{\Sigma}_{jj'}^{(DR)}$ is expected to be lower for the DR-learner. One can anticipate the estimator $\widehat{\boldsymbol{\beta}}_k$ of the DR-learner would have a variance smaller than the M-learner. Controversially, suppose the outcome models are misspecified (but the propensity score is well-specified). In that case, there is no guarantee that the DR-learner would perform better than M-learner, and it may perform even worse.

**X-learner vs M-learner.** The X-learner is likely to have low variance if the expected squared errors of all outcome models $\mu_{t^{(l)}}$ are not big enough. We do not establish the discussion here about conditions on $K$ and $r_{\min}$ under which the X-learner may perform less than the M-learner. The idea is to take both error upper bounds and obtain properly these conditions. Unfortunately, the general comparison of $r_{\min}$ and $K$ is very difficult, we would require to specify the form of $r_{\min}$ given $K$ to make it simpler.

**X-learner vs DR-learner.** It is difficult to anticipate which meta-learner would perform better in terms of variance. This will depends mainly on the expected squared error $\text{err}(\mu_{t^{(l)}})$ for $l \neq k \in \{1, \ldots, K\}$, $K$ and $r_{\min}$, whom, in some cases, will make the X-learner having less variance than the DR-learner, and the opposite in the other cases.

## B.3 Additional details about simulated analytical functions in section 5.5.

In this section, we consider a treatment $T$ with $K+1=10$ possible values in $\mathcal{T}=\{t^{(k)}:=\frac{k}{K}, k\in\{0,\dots,K\}\}$, drawn from an uniform distribution, and the following outcome functions. The linear model outcome for $X\in\mathbb{R}$:

$$Y(t)\mid X\sim\mathcal{N}((1+t)X,\sigma^2). \tag{B.62}$$

The multivariate hazard rate (Imbens, 2000) outcome satisfies for $\boldsymbol{X}\in\mathbb{R}^5$:

$$Y(t)\mid \boldsymbol{X}\sim\mathcal{N}\left(t+\|\boldsymbol{X}\|\exp\left(-t\|\boldsymbol{X}\|\right),\sigma^2\right). \tag{B.63}$$

We compute in the following subsections the exact components of each model: the GPS $r$, the potential outcome models $\mu_t$ and the observed outcome model $m$.

### The Generalized Propensity Score.

#### Randomized Controlled Trials (RCT) setting.

In the first design (RCT), we sample $n$ units such that $T$ and $\boldsymbol{X}$ are independent. The true propensity score is known

$$r(t,\boldsymbol{X})=\mathbb{P}(T=t)=1/(K+1)\text{ for }t\in\mathcal{T}. \tag{B.64}$$

#### Observational non-randomized setting.

In the second design (observational studies), we combine $K+2$ samples in a single sample of $n$ units. The first sample $\mathbf{D}_{K+1}$ contains $n_{K+1}=n/2$ units where the treatment is assigned randomly: $\boldsymbol{X}$ and $T$ are independent, $\mathbb{P}(T=t)=1/(K+1)$, $\boldsymbol{X}\sim\mathcal{N}(\mathbf{0},\mathbf{I}_5)$ when the hazard rate model is applied and $X\sim\mathcal{U}(0,1)$ when the linear model is applied. For $k=0,\dots,K$, the sample $\mathbf{D}_k$ contains $n_k=n/(2(K+1))$ units and the distribution of $(\boldsymbol{X},T)$ does not respect a RCT setting. For the linear model, the joint distribution of $(X,T)$ is given by:

$$T=\frac{k}{K}\text{ and }X\text{ follows a uniform distribution }\mathcal{U}(I_k)\text{ with }I_k=\Big[\frac{k}{K+1},\frac{k+1}{K+1}\Big). \tag{B.65}$$

For the hazard rate model, the joint distribution of $(\boldsymbol{X},T)$ is given by:

$$\begin{aligned}T=\frac{k}{K},\ X_1\text{ follows a truncated standardized normal distribution on }I_k=\big[q_{\frac{k}{K+1}},q_{\frac{k+1}{K+1}}\big)\\ \text{and }X_j\text{ follow a standardized normal distribution }\mathcal{N}(0,1)\text{ for }j\geq 2,\end{aligned} \tag{B.66}$$

where $q_\alpha$ is the $\alpha$-quantile of the standardized normal distribution. This strategy of selecting preferentially only observations with certain characteristics is called *preferential selection* sampling and creates thus a selection bias on observed data.

For all $k\in\{0,\dots,K\}$, the true propensity score satisfies for the linear model:

$$r(t^{(k)},x)=\begin{cases}\frac{K+2}{2(K+1)} & \text{if }x\in I_k,\\ \frac{1}{2(K+1)} & \text{otherwise.}\end{cases} \tag{B.67}$$

and, for the hazard rate model, it satisfies for $\boldsymbol{x} \in \mathbb{R}^5$:

$$r(t^{(k)}, \boldsymbol{x}) = \begin{cases} \frac{K+2}{2(K+1)} & \text{if } x_1 \in I_k, \\ \frac{1}{2(K+1)} & \text{otherwise.} \end{cases} \tag{B.68}$$

*Proof.* We show the proof for the hazard rate model with normal distribution. The proof remains the same for the linear model in a non-randomized setting.

Let $A$ be a random event, then

$$\mathbb{P}(A) = \sum_{k=0}^{K+1} \frac{n_k}{n} \mathbb{P}_k(A), \tag{B.69}$$

where $\mathbb{P}$ is the observed probability distribution of the combined sample and $\mathbb{P}_k$ denotes the probability measure induced by (B.64), (B.66) and the unconfoundedness assumption 4.5.1.

Given the treatment $T = t^{(j)}$ and covariate vector $\boldsymbol{x} = (x, x_2, \ldots, x_5)$, we have

$$\begin{aligned} r(T = t^{(j)}, \boldsymbol{x}) &= \mathbb{P}(T = t^{(j)} \mid X_1 = x) \\ &= \lim_{\delta \to 0} \mathbb{P}(T = t^{(j)} \mid X_1 \in [x, x+\delta]) \\ &= \lim_{\delta \to 0} \frac{\mathbb{P}\left(T = t^{(j)}, X_1 \in [x, x+\delta]\right)}{\mathbb{P}\left(X_1 \in [x, x+\delta]\right)}. \end{aligned} \tag{B.70}$$

On the one hand,

$$\begin{aligned} \mathbb{P}(T = t^{(j)}, X_1 \in [x, x+\delta]) &= \sum_{k=0}^{K+1} \frac{n_k}{n} \mathbb{P}_k(T = t^{(j)}, X_1 \in [x, x+\delta]) \\ &= \frac{n_j}{n}\, \mathbb{P}_j(T = t^{(j)}, X_1 \in [x, x+\delta]) + \frac{n_{K+1}}{n}\, \mathbb{P}_{K+1}(T = t^{(j)}, X_1 \in [x, x+\delta]) \\ &= \frac{n_j}{n}\, \mathbb{P}_j(X_1 \in [x, x+\delta]) + \frac{n_{K+1}}{n}\, \mathbb{P}_{K+1}(T = t^{(j)}) \mathbb{P}_{K+1}(X_1 \in [x, x+\delta]) \\ &= \frac{1}{2(K+1)}\, \mathbb{P}_j(X_1 \in [x, x+\delta]) + \frac{1}{2(K+1)}\, \mathbb{P}_{K+1}(X_1 \in [x, x+\delta]). \end{aligned} \tag{B.71}$$

For $x \in \mathbb{R}$, there exists a unique $j_0$ such that $x \in I_{j_0}$. For $\delta$ small enough, we have $[x, x+\delta] \subset I_{j_0}$ and, consequently, $[x, x+\delta] \cap I_j = \emptyset$ for all $j \neq j_0$. This implies:

$$\mathbb{P}_j(X_1 \in [x, x+\delta]) = \frac{\mathbb{P}_{K+1}\left(X_1 \in [x, x+\delta], X_1 \in I_j\right)}{\mathbb{P}_{K+1}\left(X_1 \in I_j\right)} = \frac{\mathbb{P}_{K+1}\left(X_1 \in [x, x+\delta]\right)}{\mathbb{P}_{K+1}\left(X_1 \in I_j\right)} \mathbf{1}\{j = j_0\}. \tag{B.72}$$

Therefore,

$$\begin{aligned} \mathbb{P}(T = t^{(j)}, X_1 \in [x, x+\delta]) &= \frac{1}{2(K+1)} \mathbb{P}_{K+1}(X_1 \in [x, x+\delta])\, \left(\frac{\mathbf{1}\{j = j_0\}}{\mathbb{P}_{K+1}(X_1 \in I_{j_0})} + 1\right) \\ &= \Big(\frac{1}{2}\mathbf{1}\{j = j_0\} + \frac{1}{2(K+1)}\Big) \mathbb{P}_{K+1}(X_1 \in [x, x+\delta]). \end{aligned} \tag{B.73}$$

On the other hand,

$$\begin{aligned}
\mathbb{P}(X_1 \in [x, x+\delta]) &= \sum_{k=0}^{K+1} \frac{n_k}{n} \mathbb{P}_k(X_1 \in [x, x+\delta]) \\
&= \frac{1}{2(K+1)} \sum_{k=0}^{K} \frac{\mathbb{P}_{K+1}\left(X_1 \in [x, x+\delta], X_1 \in I_k\right)}{\mathbb{P}_{K+1}\left(X_1 \in I_k\right)} + \frac{1}{2}\mathbb{P}_{K+1}(X_1 \in [x, x+\delta]) \\
&= \frac{1}{2(K+1)} \frac{\mathbb{P}_{K+1}(X_1 \in [x, x+\delta])}{\mathbb{P}_{K+1}(X_1 \in I_{j_0})} + \frac{1}{2}\mathbb{P}_{K+1}(X_1 \in [x, x+\delta]) \\
&= \frac{1}{2}\mathbb{P}_{K+1}(X_1 \in [x, x+\delta]) + \frac{1}{2}\mathbb{P}_{K+1}(X_1 \in [x, x+\delta]) \\
&= \mathbb{P}_{K+1}(X_1 \in [x, x+\delta])
\end{aligned} \tag{B.74}$$

Finally,

$$\begin{aligned}
r(t^{(j)}, \boldsymbol{x}) &= \lim_{\delta \to 0} \frac{\mathbb{P}\left(T = t^{(j)}, X_1 \in [x, x+\delta]\right)}{\mathbb{P}\left(X_1 \in [x, x+\delta]\right)} \\
&= \lim_{\delta \to 0} \frac{\left(\frac{1}{2}\mathbf{1}\{j = j_0\} + \frac{1}{2(K+1)}\right)\mathbb{P}_{K+1}(X_1 \in [x, x+\delta])}{\mathbb{P}_{K+1}(X_1 \in [x, x+\delta])} \\
&= \frac{1}{2}\mathbf{1}\{j = j_0\} + \frac{1}{2(K+1)} \\
&= \frac{(K+1)\mathbf{1}\{j = j_0\} + 1}{2(K+1)} \\
&= \begin{cases} \frac{K+2}{2(K+1)} & \text{if } x \in I_j, \\ \frac{1}{2(K+1)} & \text{otherwise.} \end{cases}
\end{aligned} \tag{B.75}$$

■

**Triple treatment toy example:**

In this section, we assess the performance of the three different GPS estimators in the case of three-level treatment $T \in \{0, 1, 2\}$ drawn from an uniform distribution. We consider 1-dimensional covariate ($\boldsymbol{X} = X$) where $X$ follows a discrete uniform distribution in $\{\frac{100}{1000}k, k \in \{1, \ldots, 10\}\}$.

In the first setting design, we sample $n = 10000$ units following Randomized Controlled Trials setting, the true propensity score is known

$$r(t, X) = 1/3 \text{ for } t \in \{0, 1, 2\}. \tag{B.76}$$

In the second setting design, we combine two samples in a single sample of $n = 10000$ units. The first sample $\mathbf{D}_3$ contains $n_3 = n/2$ units where the treatment is assigned randomly (RCT), and for $j = 0, 1, 2$ the sample $\mathbf{D}_j$ contains $n_j = n/6$ units satisfying, with $x_1 = 300$ and $x_2 = 600$,

$$\begin{aligned}
\text{In } \mathbf{D}_0,\ T_i = 0 \text{ and the } X_i \text{ are i.i.d uniformly distributed over } [100, x_1] = I_0. \\
\text{In } \mathbf{D}_1,\ T_i = 1 \text{ and the } X_i \text{ are i.i.d uniformly distributed over } (x_1, x_2] = I_1. \\
\text{In } \mathbf{D}_2,\ T_i = 2 \text{ and the } X_i \text{ are i.i.d uniformly distributed over } (x_2, 1000] = I_2.
\end{aligned}$$

This case corresponds closely to an observational study where the treatment $T$ is confounded with the covariate $X$ (e.g. the larger $X$ is, the more likely we have chance to receive the treatment $T = 2$).

The true propensity score (can be proved with similarly to B.68) to is a step-wise function such that:

$$
\begin{aligned}
r(0,x) &= \begin{cases} \frac{13}{19} & \text{if } x \leq x_1 \\ \frac{3}{19} & \text{if } x_1 < x \leq x_2 \\ \frac{4}{22} & \text{if } x > x_2 \end{cases} \\
r(1,x) &= \begin{cases} \frac{3}{19} & \text{if } x \leq x_1 \\ \frac{13}{19} & \text{if } x_1 < x \leq x_2 \\ \frac{4}{22} & \text{if } x > x_2 \end{cases} \\
r(2,x) &= \begin{cases} \frac{3}{19} & \text{if } x \leq x_1 \\ \frac{13}{19} & \text{if } x_1 < x \leq x_2 \\ \frac{14}{22} & \text{if } x > x_2 \end{cases}
\end{aligned}
\tag{B.77}
$$

The following figures show GPS's estimation for a given estimation method:

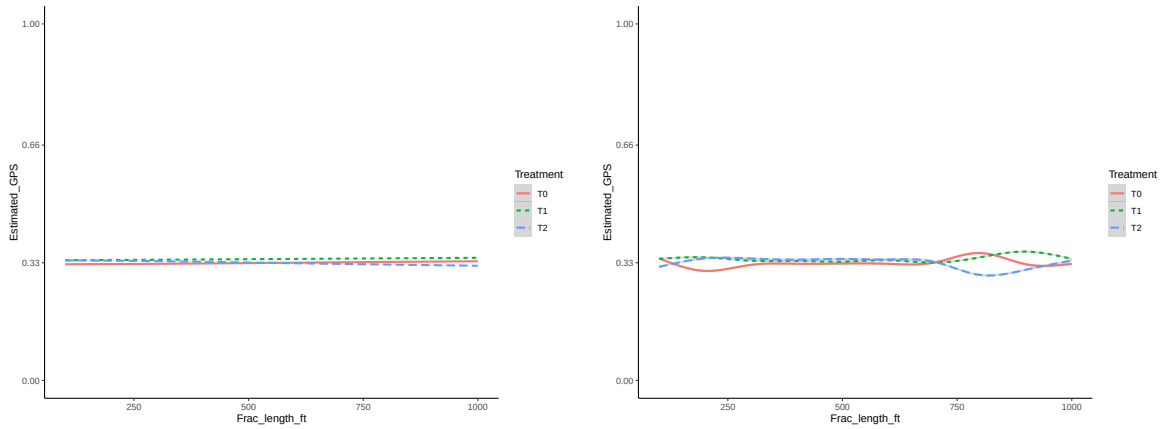



Figure B.1: Estimation of the GPS in the first setting design. a: Using the Generalized Linear Models; b: Using XGBoost model.

## The potential outcome models.

The potential outcome models are given directly by the conditional mean. For the linear model, $\mu_t$ satisfies for all $t \in \mathcal{T}$ and $x \in [0,1]$:

$$\mu_t(x) = (1+t)x, \tag{B.78}$$

and, for the hazard rate model, $\mu_t$ is given by:

$$\mu_t(\boldsymbol{x}) = t + \|\boldsymbol{x}\| \exp\left(-t\|\boldsymbol{x}\|\right). \tag{B.79}$$

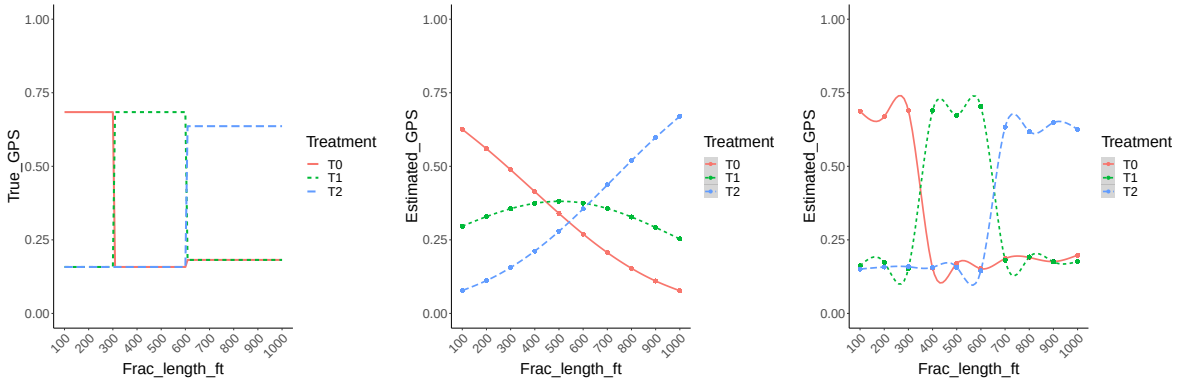



Figure B.2: a: The true GPS; b: Using the Generalized Linear Models; c: Using XGBoost model.

### The observed outcome models.

For the linear model, the observed outcome model $m$ can be computed as:

$$\begin{aligned} m(x) &= \mathbb{E}(Y_{\text{obs}} \mid X = x) \\ &= \mathbb{E}((1+T)\boldsymbol{X} \mid X = x) \\ &= (1 + \mathbb{E}(T \mid X = x))x \\ &= (1 + \sum_{k=1}^{K} r(t^{(k)}, x)t^{(k)})x, \end{aligned} \tag{B.80}$$

where $r$ is given by (B.67).

and, for the hazard rate model, $m$ can be computed as:

$$\begin{aligned} m(\boldsymbol{x}) &= \mathbb{E}(\mathbb{E}(Y_{\text{obs}} \mid \boldsymbol{X}, T) \mid \boldsymbol{X} = \boldsymbol{x}) \\ &= \mathbb{E}(T + \|\boldsymbol{X}\| \exp(-T\|\boldsymbol{X}\|) \mid \boldsymbol{X} = \boldsymbol{x}) \\ &= \mathbb{E}(T \mid \boldsymbol{X} = \boldsymbol{x}) + \|\boldsymbol{x}\| \ \mathbb{E}(\exp(-T\|\boldsymbol{X}\|) \mid \boldsymbol{X} = \boldsymbol{x}) \\ &= \sum_{k=1}^{K} r(t^{(k)}, \boldsymbol{x})t^{(k)} + \sum_{k=1}^{K} \|\boldsymbol{x}\| \ r(t^{(k)}, \boldsymbol{x}) \exp(-t^{(k)}\|\boldsymbol{x}\|), \end{aligned} \tag{B.81}$$

where $r$ is given by (B.68).

## B.4 Additional numerical results and plots.

In this section, we present the results of different simulations and scenarios for both linear (5.60) and hazard rate (5.61) models with $K+1=10$, $n=2000$ for the linear model, and $n=10000$ for the Hazard rate model. In randomized setting, the sample $\mathbf{D}_{\text{obs}}$ is sampled randomly and the propensity score is given by (B.64). In non-randomized setting, the sample $\mathbf{D}_{\text{obs}}$ is given by preferential selection as described in Section B.3 and the GPS is given by (B.67). When we say that the models nuisance components are exact, then we replace the expression of $\mu_t, m$ or $r$ by the expressions obtained in Section B.3.

### Linear model (5.60) in randomized setting.

Table B.1: **mPEHE** for three different Machine Learning base-learners; Case where nuisance components are exact.

| Meta-learner | XGBoost | RandomForest | Linear Model |
|---|---|---|---|
| M-Learner | 2.248 | 2.07 | 0.099 |
| DR-Learner | 0.159 | 0.134 | $7.04\ 10^{-3}$ |
| X-Learner | **0.022** | **0.028** | $\mathbf{1.53\ 10^{-3}}$ |
| RLin-Learner | | $7.33\ 10^{-3}$ | |

Table B.2: **mPEHE** for three different Machine Learning base-learners; Case when nuisance components are well-specified.

| Meta-learner | XGBoost | RandomForest | Linear Model |
|---|---|---|---|
| T-Learner | 0.061 | **0.037** | $7.37\ 10^{-3}$ |
| S-Learner | **0.029** | 0.040 | $3.65\ 10^{-3}$ |
| M-Learner | 1.23 | 1.15 | 0.210 |
| DR-Learner | 0.063 - 0.063 | 0.060 - 0.060 | 7.22 - $\mathbf{3.39\ 10^{-3}}$ |
| X-Learner | 0.059 - 0.030 | 0.041 - 0.079 | 7.36 - $3.59\ 10^{-3}$ |
| RLin-Learner | 0.122 | 0.112 | 0.046 |

For the DR and X-learners: $\mu_t$ are estimated by T-learning (left value) or S-learning (right value).

Table B.3: **mPEHE** for three different Machine Learning base-learners; Case when the propensity score is misspecified.

| Meta-learner | XGBoost | RandomForest | Linear Model |
|---|---|---|---|
| M-Learner | 3.54 | 3.31 | 1.31 |
| DR-Learner | **0.119** | **0.104** | **0.011** |
| X-Learner | 0.030 | 0.041 | $3.59\ 10^{-3}$ |
| RLin-Learner | 0.318 | 0.313 | 0.334 |

Table B.4: **mPEHE** for three different Machine Learning base-learners; Case when the outcome models are misspecified.

| Meta-learner | XGBoost | RandomForest | Linear Model |
|---|---|---|---|
| M-Learner | 1.23 | 1.15 | 0.210 |
| DR-Learner | 0.737 | 0.800 | 0.217 |
| X-Learner | 0.282 | 0.282 | 0.246 |
| RLin-Learner | | **0.045** | |

Table B.5: **mPEHE** for three different Machine Learning base-learners; Case when nuisance components are misspecified.

| Meta-learner | XGBoost | RandomForest | Linear Model |
|---|---|---|---|
| M-Learner | 3.54 | 3.31 | 1.31 |
| DR-Learner | 1.66 | 1.85 | 0.758 |
| X-Learner | 0.282 | 0.282 | **0.246** |
| RLin-Learner | | **0.280** | |

## Linear model (5.60) in non-randomized setting

Table B.6: **mPEHE** for three different Machine Learning base-learners; Case where nuisance components are exact.

| Meta-learner | XGBoost | RandomForest | Linear Model |
|---|---|---|---|
| M-Learner | 3.68 | 2.33 | 0.68 |
| DR-Learner | 0.287 | 0.147 | 0.014 |
| X-Learner | **0.023** | **0.030** | **1.57 $10^{-3}$** |
| RLin-Learner | | 9.44 $10^{-3}$ | |

Table B.7: **mPEHE** for three different Machine Learning base-learners; Case when nuisance components are well-specified.

| Meta-learner | XGBoost | RandomForest | Linear Model |
|---|---|---|---|
| T-Learner | 0.061 | **0.042** | 7.37 $10^{-3}$ |
| *Reg*T-Learner | 0.052 | **0.042** | 7.60 $10^{-3}$ |
| S-Learner | **0.029** | 0.050 | **3.65 $10^{-3}$** |
| M-Learner | 1.23 | 1.15 | 0.209 |
| DR-Learner | 0.060 - 0.055 | 0.068 - 0.095 | 7.60 - 3.95 $10^{-3}$ |
| X-Learner | 0.051 - 0.030 | 0.045 - 0.079 | 7.33 - 3.95 $10^{-3}$ |
| RLin-Learner | 0.122 | 0.127 | 0.046 |

For the DR and X-learners: $\mu_t$ are estimated by T-learning (left value) or S-learning (right value).

### Hazard rate model (5.61) in randomized setting

Table B.8: **mPEHE** for three different Machine Learning base-learners; Case where nuisance components are exact.

| Meta-learner | XGBoost | RandomForest | Linear Model |
|---|---|---|---|
| M-Learner | 4.25 | 4.22 | 0.52 |
| DR-Learner | 0.127 | 0.139 | 0.099 |
| X-Learner | **0.045** | **0.085** | **0.098** |
| RLin-Learner | | 0.100 | |

Table B.9: **mPEHE** for three different Machine Learning base-learners; Case when nuisance components are well-specified.

| Meta-learner | XGBoost | RandomForest | Linear Model |
|---|---|---|---|
| T-Learner | 0.171 | 0.267 | **0.105** |
| S-Learner | 0.154 | 0.267 | 0.649 |
| M-Learner | 1.52 | 1.76 | 0.792 |
| DR-Learner | 0.154 - 0.163 | 0.286 - 0.282 | 0.106 - 0.461 |
| X-Learner | (**0.149**) 0.161 | 0.284 - 0.285 | **0.105** - 0.637 |
| RLin-Learner | 0.227 | **0.241** | 0.691 |

For the DR and X-learners: $\mu_t$ are estimated by T-learning (left value) or S-learning (right value).

## Hazard rate model (5.61) in non-randomized setting

Table B.10: **mPEHE** for three different Machine Learning base-learners; Case where nuisance components are exact.

| Meta-learner | XGBoost | RandomForest | Linear Model |
|---|---|---|---|
| M-Learner | 6.33 | 5.81 | 3.52 |
| DR-Learner | 0.138 | 0.140 | 0.100 |
| X-Learner | **0.044** | **0.085** | **0.098** |
| RLin-Learner | | 0.290 | |

Table B.11: **mPEHE** for three different Machine Learning base-learners; Case when nuisance components are well-specified.

| Meta-learner | XGboost | RandomForest | Linear Model |
|---|---|---|---|
| T-Learner | 0.184 | 0.251 | 0.128 |
| *Reg*T-Learner | 0.158 | 0.253 | **0.111** |
| S-Learner | 0.166 | 0.269 | 0.642 |
| M-Learner | 1.56 | 1.55 | 0.866 |
| DR-Learner | 0.151 - 0.171 | 0.275 - 0.288 | **0.111**- 0.495 |
| X-Learner | **0.149** - 0.162 | 0.270 - 0.286 | 0.114 - 0.627 |
| RLin-Learner | 0.235 | **0.178** | 1.00 |

For the DR and X-learners: $\mu_t$ are estimated by T-learning (left value) or S-learning (right value).

**Asymptotic performances when $n$ and $K$ increase.**

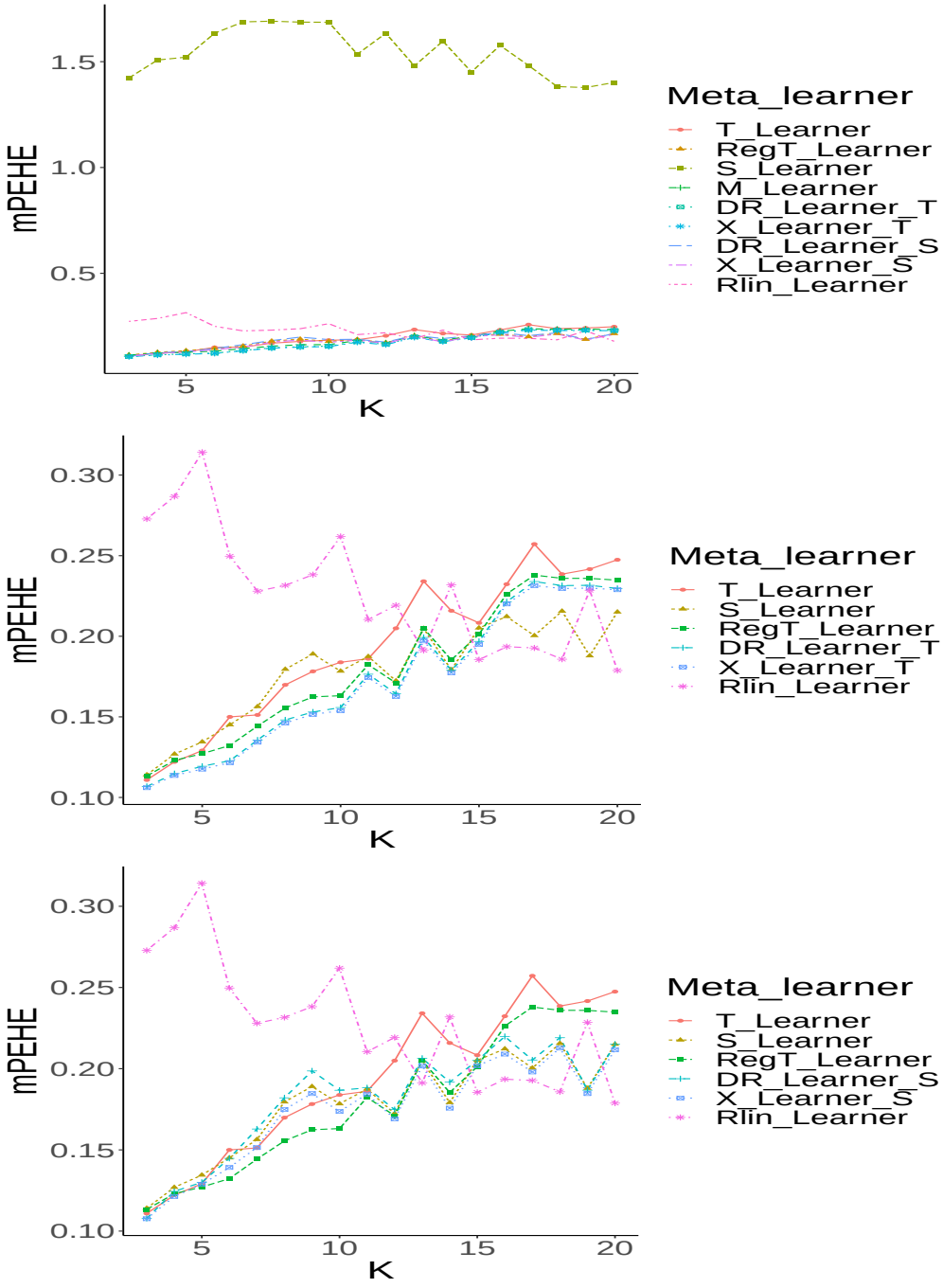



Figure B.3: Variation of meta-learner's performances when number of possible treatment values $K$ for the hazard rate function in observational design setting. a: All meta-learners; b: When the potential outcome models $\mu_\cdot$ are estimated by *reg*T-learning; c: When the potential outcome models $\mu_\cdot$ are estimated by S-learning.

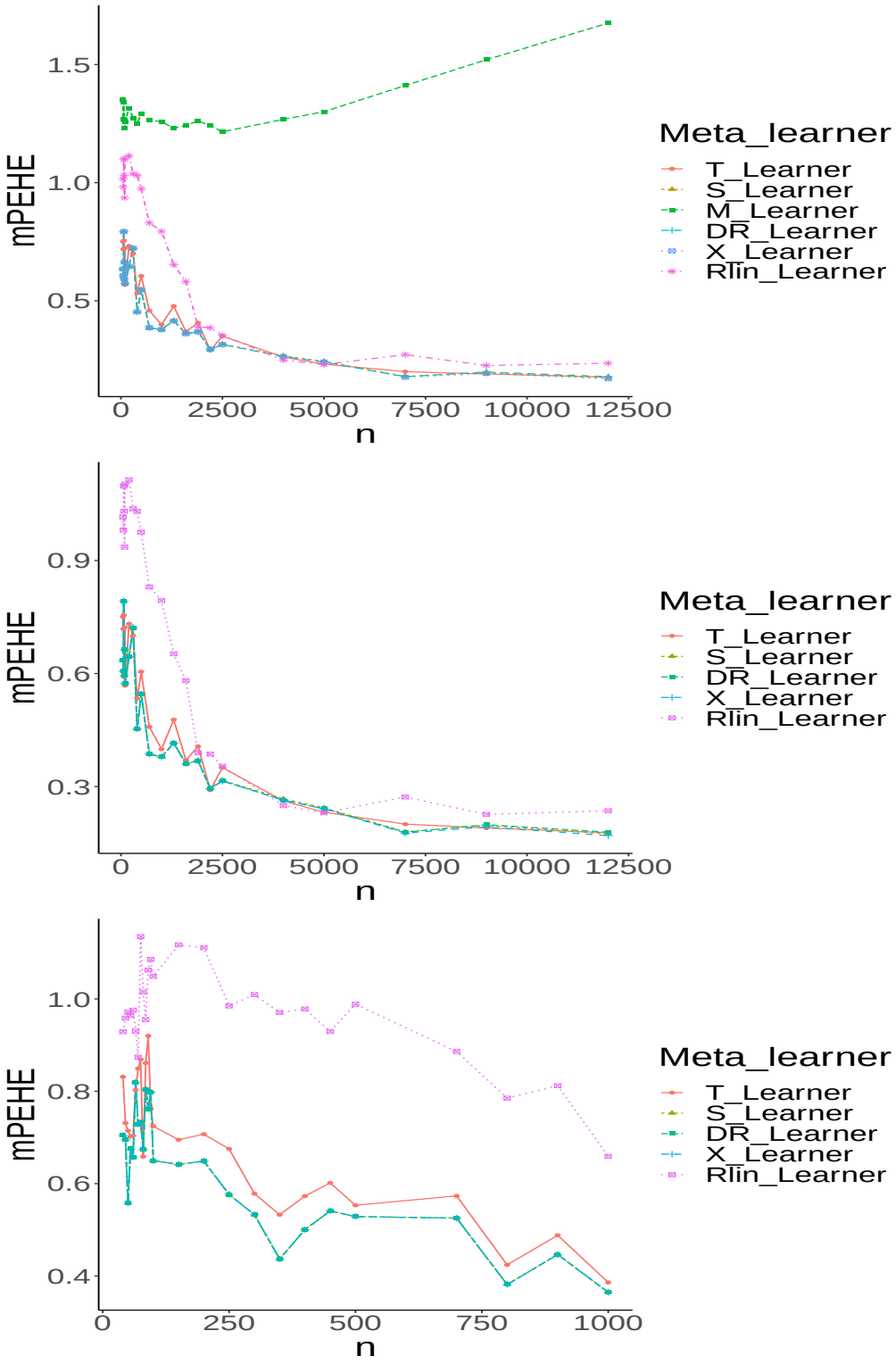



Figure B.4: Variation of meta-learner's performances with the observed sample size $n$ for the hazard rate function in observational design setting. a: All meta-learners; b: Without the M-learner; c: Without the M-learner with a focus on low sample regime.

# Bibliography


Abadie, A. and Imbens, G. W. Matching on the estimated propensity score. *Econometrica*, 84 (2):781–807, 2016.

Acharki, N., Garnier, J., Bertoncello, A., and Lugo, R. Heterogeneous treatment effects estimation: When machine learning meets multiple treatment regime. *arXiv preprint arXiv:2205.14714*, 2022.

Acharki, N., Bertoncello, A., and Garnier, J. Robust prediction interval estimation for gaussian processes by cross-validation method. *Computational Statistics & Data Analysis*, 178:107597, 2023. ISSN 0167-9473. doi: 10.1016/j.csda.2022.107597.

Ahmad, T. and Chen, H. Potential of three variant machine-learning models for forecasting district level medium-term and long-term energy demand in smart grid environment. *Energy*, 160:1008–1020, 2018. ISSN 0360-5442. doi: 10.1016/j.energy.2018.07.084.

Alaa, A. and van der Schaar, M. Limits of estimating heterogeneous treatment effects: Guidelines for practical algorithm design. In Dy, J. and Krause, A. (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 129–138. PMLR, 10–15 Jul 2018.

Alaa, A. M. and van der Schaar, M. Bayesian inference of individualized treatment effects using multi-task gaussian processes. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pp. 3427–3435, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

Alvarado, V., Ranson, A., Hernandez, K., Manrique, E., Matheus, J., Liscano, T., and Prosperi, N. Selection of EOR/IOR Opportunities Based on Machine Learning. In *SPE Europec featured at EAGE Conference and Exhibition*, volume All Days, 10 2002. doi: 10.2118/78332-MS.

Andrianakis, I. and Challenor, P. G. The effect of the nugget on gaussian process emulators of computer models. *Computational Statistics & Data Analysis*, 56(12):4215–4228, 12 2012. ISSN 0167-9473. doi: 10.1016/j.csda.2012.04.020.

Angelopoulos, A. N. and Bates, S. A gentle introduction to conformal prediction and distribution-free uncertainty quantification, 2021.

Angrist, J. D. Estimating the labor market impact of voluntary military service using social security data on military applicants. *Econometrica*, 66(2):249–288, 1998.

Angrist, J. D., Imbens, G. W., and Rubin, D. B. Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 91(434):444–455, 1996. ISSN 01621459.

Arceneaux, K., Gerber, A. S., and Green, D. P. Comparing experimental and matching methods using a large-scale voter mobilization experiment. *Political Analysis*, 14(1):37–62, 2006.

Arjovsky, M., Bottou, L., Gulrajani, I., and Lopez-Paz, D. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

Aronszajn, N. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68:337–404, 1950.

Arsigny, V., Fillard, P., Pennec, X., and Ayache, N. Geometric means in a novel vector space structure on symmetric positive-definite matrices. *SIAM Journal on Matrix Analysis and Applications*, 29(1):328–347, 2007. doi: 10.1137/050637996.

Arslan, O. and Yetik, O. Ann based optimization of supercritical orc-binary geothermal power plant: Simav case study. *Applied Thermal Engineering*, 31(17):3922–3928, 2011. ISSN 1359-4311. doi: 10.1016/j.applthermaleng.2011.07.041. SET 2010 Special Issue.

Arto, I., Capellán-Pérez, I., Lago, R., Bueno, G., and Bermejo, R. The energy requirements of a developed world. *Energy for Sustainable Development*, 33:1–13, 2016. ISSN 0973-0826. doi: 10.1016/j.esd.2016.04.001.

Assouline, D., Mohajeri, N., Gudmundsson, A., and Scartezzini, J.-L. A machine learning approach for mapping the very shallow theoretical geothermal potential. *Geothermal Energy*, 7, 12 2019. doi: 10.1186/s40517-019-0135-6.

Athey, S. and Imbens, G. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360, 2016. doi: 10.1073/pnas.1510489113.

Austin, P. C. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate behavioral research*, 46(3):399–424, 2011.

Bachoc, F. Cross validation and maximum likelihood estimations of hyper-parameters of Gaussian processes with model misspecification. *Computational Statistics & Data Analysis*, 66:55–69, 2013.

Bachoc, F. *Parametric estimation of covariance function in Gaussian-process based Kriging models. Application to uncertainty quantification for computer experiments.* PhD thesis, Université Paris-Diderot-Paris VII, 2013. URL http://www.theses.fr/2013PA077111. 2013PA077111.

Bachoc, F. Asymptotic analysis of covariance parameter estimation for Gaussian processes in the misspecified case. *Bernoulli*, 24(2):1531 – 1575, 2018. doi: 10.3150/16-BEJ906.

Bachoc, F., Gamboa, F., Loubes, J.-M., and Venet, N. A gaussian process regression model for distribution inputs. *IEEE Transactions on Information Theory*, 64(10):6620–6637, 2018. doi: 10.1109/TIT.2017.2762322.

Bahadori, T., Tchetgen, E. T., and Heckerman, D. End-to-end balancing for causal continuous treatment-effect estimation. In Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., and Sabato, S. (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 1313–1326. PMLR, 17–23 Jul 2022.

Bai, Y., Mei, S., Wang, H., and Xiong, C. Understanding the under-coverage bias in uncertainty estimation. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 18307–18319. Curran Associates, Inc., 2021.

Baker, G. A. The Probability That the Mean of a Second Sample Will Differ from the Mean of a First Sample By Less Than a Certain Multiple of the Standard Deviation of the First Sample. *The Annals of Mathematical Statistics*, 6(4):197 – 201, 1935. doi: 10.1214/aoms/1177732565.

Barber, R. F., Candès, E. J., Ramdas, A., and Tibshirani, R. J. Predictive inference with the jackknife+. *The Annals of Statistics*, 49(1):486 – 507, 2021. doi: 10.1214/20-AOS1965.

Barber, R. F., Candes, E. J., Ramdas, A., and Tibshirani, R. J. Conformal prediction beyond exchangeability, 2022.

Begoli, E., Bhattacharya, T., and Kusnezov, D. F. The need for uncertainty quantification in machine-assisted medical decision making. *Nature Machine Intelligence*, 1(1), 1 2019. doi: 10.1038/s42256-018-0004-1.

Belloni, A., Chernozhukov, V., and Kato, K. Valid post-selection inference in high-dimensional approximately sparse quantile regression models. *Journal of the American Statistical Association*, 114(526):749–758, 2019.

Beran, R. Controlling Conditional Coverage Probability in Prediction. *The Annals of Statistics*, 20(2):1110 – 1119, 1992. doi: 10.1214/aos/1176348673.

Berge, C. *Topological Spaces: Including a Treatment of Multi-valued Functions, Vector Spaces and Convexity*. Oliver & Boyd, 1963.

Berger, J. O., Oliveira, V. D., and Sansó, B. Objective bayesian analysis of spatially correlated data. *Journal of the American Statistical Association*, 96(456):1361–1374, 2001. doi: 10.1198/016214501753382282.

Berlinet, A. and Thomas-Agnan, C. *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Kluwer Academic Publishers, 2004.

Bhatia, S. *Advanced renewable energy systems,(Part 1 and 2)*. CRC Press, 2014.

Bica, I., Alaa, A. M., Lambert, C., and van der Schaar, M. From real-world patient data to individualized treatment effects using machine learning: Current and future methods to address underlying challenges. *Clinical Pharmacology & Therapeutics*, 109(1):87–100, 2021. doi: 10.1002/cpt.1907.

Billingsley, P. *Probability and Measure*. Wiley Series in Probability and Statistics. Wiley, 1995. ISBN 9780471007104.

Bogachev, V. I. *Gaussian measures*. American Mathematical Soc., 1998.

Bonvini, M. and Kennedy, E. H. Fast convergence rates for dose-response estimation. *arXiv preprint arXiv:2207.11825*, 2022.

Bostanabad, R., Kearney, T., Tao, S., Apley, D. W., and Chen, W. Leveraging the nugget parameter for efficient gaussian process modeling. *International Journal for Numerical Methods in Engineering*, 114(5):501–516, 2018.

Bouktif, S., Fiaz, A., Ouni, A., and Serhani, M. A. Optimal deep learning lstm model for electric load forecasting using feature selection and genetic algorithm: Comparison with machine learning approaches. *Energies*, 11(7), 2018. ISSN 1996-1073. doi: 10.3390/en11071636.

Breiman, L. Random forests. *Machine Learning*, 45(1):5–32, Oct 2001. ISSN 1573-0565. doi: 10.1023/A:1010933404324.

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. *Classification and Regression Trees*. Wadsworth and Brooks, 1984.

Briggs, L. L. and Division, N. E. Uncertainty quantification approaches for advanced reactor analyses. 3 2009. doi: 10.2172/956921.

British Petroleum. Statistical review of world energy, 2020. https://www.bp.com/content/dam/bp/business-sites/en/global/corporate/pdfs/energy-economics/statistical-review/bp-stats-review-2020-full-report.pdf.

Bruckner, T., Bashmakov, I., Mulugetta, Y., Chum, H., Navarro, A., Edmonds, J., Faaij, A., Fungtammasan, B., Garg, A., Hertwich, E., Honnery, D., Infield, D., Kainuma, M., Khennas, S., Kim, S., Nimir, H., Riahi, K., Strachan, N., Wiser, R., and Upadhyay, J. *Energy Systems. Climate Change 2014: Mitigation of Climate Change. Contribution of Working Group III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*, pp. 511–598. Cambridge University Press, 11 2014. ISBN 9781107415416. doi: 10.1017/CBO9781107415416.013.

Bughin, J., Hazan, E., Ramaswamy, S., Chui, M., Allas, T., Dahlstrom, P., Henke, N., and Trench, M. Artificial intelligence: the next digital frontier? *McKinsey Global Institute*, 2017.

Cao, Q., Banerjee, R., Gupta, S., Li, J., Zhou, W., and Jeyachandra, B. Data Driven Production Forecasting Using Machine Learning. In *SPE Argentina Exploration and Production of Unconventional Resources Symposium*, volume Day 2 Thu, June 02, 2016, 06 2016. doi: 10.2118/180984-MS.

Caron, A., Baio, G., and Manolopoulou, I. Estimating individual treatment effects using non-parametric regression models: A review. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 185(3):1115–1149, 2022a. doi: 10.1111/rssa.12824.

Caron, A., Baio, G., and Manolopoulou, I. Shrinkage bayesian causal forests for heterogeneous treatment effects estimation. *Journal of Computational and Graphical Statistics*, pp. 1–13, 2022b.

Cepeda, M. S., Boston, R., Farrar, J. T., and Strom, B. L. Comparison of logistic regression versus propensity score when the number of events is low and there are multiple confounders. *American journal of epidemiology*, 158(3):280–287, 2003.

Chen, H., Huang, Z., Lam, H., Qian, H., and Zhang, H. Learning prediction intervals for regression: Generalization and calibration. In Banerjee, A. and Fukumizu, K. (eds.), *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pp. 820–828. PMLR, 13–15 Apr 2021a.

Chen, T. and Guestrin, C. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785–794, 2016.

Chen, X. and White, H. Improved rates and asymptotic normality for nonparametric neural network estimators. *IEEE Transactions on Information Theory*, 45(2):682–691, 1999.

Chen, Y., Owhadi, H., and Stuart, A. Consistency of empirical bayes and kernel flow for hierarchical parameter estimation. *Mathematics of Computation*, 90(332):2527–2578, 2021b.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 01 2018. ISSN 1368-4221. doi: 10.1111/ectj.12097.

Chickering, D. M. Optimal structure identification with greedy search. *Journal of machine learning research*, 3(Nov):507–554, 2002.

Colangelo, K. and Lee, Y.-Y. Double debiased machine learning nonparametric inference with continuous treatments. *arXiv preprint arXiv:2004.03036*, 2020.

Cottrell, F. *Energy & society: The relation between energy, social change, and economic development*. AuthorHouse, 2009.

Cox, D. D. Best unbiased prediction for gaussian and log-gaussian processes. *Lecture Notes-Monograph Series*, 44:125–132, 2004. ISSN 07492170.

Cox, D. R. Prediction intervals and empirical bayes confidence intervals. *Journal of Applied Probability*, 12(S1):47–55, 1975. doi: 10.1017/S0021900200047550.

Cressie, N. *Statistics for spatial data*. Wiley series in probability and mathematical statistics: Applied probability and statistics. J. Wiley, 1993. ISBN 9780471002550.

Crump, R. K., Hotz, V. J., Imbens, G. W., and Mitnik, O. A. Nonparametric tests for treatment effect heterogeneity. *The Review of Economics and Statistics*, 90(3):389–405, 2008.

Currin, C., Mitchell, T. J., Morris, M. D., and Ylvisaker, D. Bayesian prediction of deterministic functions, with applications to the design and analysis of computer experiments. *Journal of the American Statistical Association*, 86(416):953–963, 1991. doi: 10.1080/01621459.1991.10475138.

Curth, A. and van der Schaar, M. Nonparametric estimation of heterogeneous treatment effects: From theory to learning algorithms. In Banerjee, A. and Fukumizu, K. (eds.), *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pp. 1810–1818. PMLR, 13–15 Apr 2021a.

Curth, A. and van der Schaar, M. On inductive biases for heterogeneous treatment effect estimation. In Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*. PMLR, 2021b.

Curth, A., Svensson, D., Weatherall, J., and van der Schaar, M. Really doing great at estimating CATE? a critical look at ML benchmarking practices in treatment effect estimation. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021.

D'Amour, A. On multi-cause causal inference with unobserved confounding: Counterexamples, impossibility, and alternatives. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pp. 3478–3486. PMLR, 04 2019.

Dawid, A. P. Causal inference without counterfactuals. *Journal of the American Statistical Association*, 95(450):407–424, 2000. doi: 10.1080/01621459.2000.10474210.

de Oliveira, V. Objective bayesian analysis of spatial data with measurement error. *The Canadian Journal of Statistics / La Revue Canadienne de Statistique*, 35(2):283–301, 2007. ISSN 03195724.

Deville, Y., Ginsbourger, D., and Durrande., O. R. C. N. *kergp: Gaussian Process Laboratory*, 2019. URL https://CRAN.R-project.org/package=kergp. R package version 0.5.0.

Dewolf, N. and Baets, Bernard Deand Waegeman, W. Valid prediction intervals for regression problems. *Artificial Intelligence Review*, Apr 2022. doi: 10.1007/s10462-022-10178-5.

Dhaou, A., Bertoncello, A., Gourvénec, S., Garnier, J., and Le Pennec, E. Causal and interpretable rules for time series analysis. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, KDD '21, pp. 2764–2772, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383325. doi: 10.1145/3447548.3467161.

Dietterich, T. G. Ensemble methods in machine learning. In *Multiple Classifier Systems*, pp. 1–15, Berlin, Heidelberg, 2000. Springer Berlin Heidelberg. ISBN 978-3-540-45014-6.

Dominici, F., Daniels, M., Zeger, S. L., and Samet, J. M. Air pollution and mortality: estimating regional and national dose-response relationships. *Journal of the American Statistical Association*, 97(457):100–111, 2002.

Dorie, V., Hill, J., Shalit, U., Scott, M., and Cervone, D. Automated versus do-it-yourself methods for causal inference: Lessons learned from a data analysis competition. *Statistical Science*, 34(1):43–68, 2019.

Dowson, D. and Landau, B. The fréchet distance between multivariate normal distributions. *Journal of Multivariate Analysis*, 12(3):450–455, 1982. ISSN 0047-259X. doi: 10.1016/0047-259X(82)90077-X.

Drgoňa, J., Picard, D., Kvasnica, M., and Helsen, L. Approximate model predictive building control via machine learning. *Applied Energy*, 218:199–216, 2018. ISSN 0306-2619. doi: 10.1016/j.apenergy.2018.02.156.

Dryden, I., Koloydenko, A., and Zhou, D. Non-euclidean statistics for covariance matrices, with applications to diffusion tensor imaging. *The Annals of Applied Statistics*, 3, 10 2009.

Dubrule, O. Cross validation of kriging in a unique neighborhood. *Journal of the International Association for Mathematical Geology*, 15(6):687–699, Dec 1983. ISSN 1573-8868. doi: 10.1007/BF01033232.

Dudley, R. M. *Real Analysis and Probability*. Cambridge Studies in Advanced Mathematics. Cambridge University Press, 2 edition, 2002. doi: 10.1017/CBO9780511755347.

Efron, B. Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics*, 7(1):1 – 26, 1979. doi: 10.1214/aos/1176344552.

Efron, B. Jackknife-after-bootstrap standard errors and influence functions. *Journal of the Royal Statistical Society. Series B (Methodological)*, 54(1):83–127, 1992. ISSN 00359246.

Efron, B. and Gong, G. A leisurely look at the bootstrap, the jackknife, and cross-validation. *The American Statistician*, 37(1):36–48, 1983. doi: 10.1080/00031305.1983.10483087.

Efron, B. and Tibshirani, R. J. *An Introduction to the Bootstrap*. Number 57 in Monographs on Statistics and Applied Probability. Chapman & Hall/CRC, Boca Raton, Florida, USA, 1993.

Fan, Q., Hsu, Y.-C., Lieli, R. P., and Zhang, Y. Estimation of conditional average treatment effects with high-dimensional data. *Journal of Business & Economic Statistics*, 40(1):313–327, 2022.

Feng, P., Zhou, X.-H., Zou, Q.-M., Fan, M.-Y., and Li, X.-S. Generalized propensity score for estimating the average treatment effect of multiple treatments. *Statistics in medicine*, 31(7): 681–697, 2012.

Fisher, R. *Statistical methods for research workers*. Edinburgh Oliver & Boyd, 1925.

Fisher, R. *The Design of Experiments*. Oliver and Boyd, 1935.

Flores, C. A. Estimation of Dose-Response Functions and Optimal Doses with a Continuous Treatment. Technical Report 0707, University of Miami, Department of Economics, November 2007.

Flores, C. A., Flores-Lagunes, A., Gonzalez, A., and Neumann, T. C. Estimating the effects of length of exposure to instruction in a training program: the case of job corps. *Review of Economics and Statistics*, 94(1):153–171, 2012.

Foley, A. M., Leahy, P. G., Marvuglia, A., and McKeogh, E. J. Current methods and advances in forecasting of wind power generation. *Renewable Energy*, 37(1):1–8, 2012. ISSN 0960-1481. doi: 10.1016/j.renene.2011.05.033.

Fong, C., Hazlett, C., and Imai, K. Covariate balancing propensity score for a continuous treatment: Application to the efficacy of political advertisements. *The Annals of Applied Statistics*, 12(1):156–177, 2018.

Forrester, A. I. J., Sóbester, A., and Keane, A. J. *Engineering Design Via Surrogate Modelling: A Practical Guide*. Progress in Astronautics and Aeronautics. American Institute of Aeronautics and Astronautics, 2008. ISBN 9781563479557.

Förstner, W. and Moonen, B. *A Metric for Covariance Matrices*. Springer Berlin Heidelberg, Berlin, Heidelberg, 2003.

Foster, J., Taylor, J. M., and Ruberg, S. Subgroup identification from randomized clinical trial data. *Statistics in medicine*, 30 24:2867–80, 2011.

Foygel Barber, R., Candès, E. J., Ramdas, A., and Tibshirani, R. J. The limits of distribution-free conditional predictive inference. *Information and Inference: A Journal of the IMA*, 10 (2):455–482, 08 2020. ISSN 2049-8772. doi: 10.1093/imaiai/iaaa017.

Friedman, J. H. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5):1189 – 1232, 2001.

Frölich, M. Programme evaluation with multiple treatments. *Wiley-Blackwell: Journal of Economic Surveys*, 2002.

Fulford, D. S., Bowie, B., Berry, M. E., Bowen, B., and Turk, D. W. Machine Learning as a Reliable Technology for Evaluating Time-Rate Performance of Unconventional Wells. In *SPE Annual Technical Conference and Exhibition*, volume Day 3 Wed, September 30, 2015, 09 2015. doi: 10.2118/174784-MS.

Galagate, D. *Causal inference with a continuous treatment and outcome: Alternative estimators for parametric dose-response functions with applications*. PhD thesis, University of Maryland, College Park, 2016.

Galilei, G. a. *On Motion, and On Mechanics: Comprising De Motu (ca. 1590) – Le meccaniche (ca. 1600)*. Publications in medieval science. University of Wisconsin Press, 1600. Translated with introduction by Drabkin, I.E. and Drake, S. in 1960.

Galton, F. Regression towards mediocrity in hereditary stature. *The Journal of the Anthropological Institute of Great Britain and Ireland*, 15:246–263, 1886.

Galvao, A. F. and Wang, L. Uniformly semiparametric efficient estimation of treatment effects with a continuous treatment. *Journal of the American Statistical Association*, 110(512): 1528–1542, 2015.

Gammerman, A., Vovk, V., and Vapnik, V. Learning by transduction. In *UAI '98: Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence*, UAI'98, pp. 148–155, San Francisco, CA, USA, 1998. Morgan Kaufmann Publishers Inc. ISBN 155860555X.

Gao, J. Machine Learning applications for data center optimization, 2014.

Gao, J. and Evans, R. DeepMind AI reduces google data centre cooling bill by 40%, 2014.

Gates, B. Dear class of 2017. *GatesNotes: The blog of Bill Gates*, 2017. https://www.gatesnotes.com/About-Bill-Gates/Dear-Class-of-2017.

Gensler, A., Henze, J., Sick, B., and Raabe, N. Deep learning for solar power forecasting — an approach using autoencoder and lstm neural networks. In *2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pp. 002858–002865, 2016. doi: 10.1109/SMC.2016.7844673.

Geothermal Technologies Office. How an enhanced geothermal system works, 2022. https://www.energy.gov/eere/geothermal/geothermal-technologies-office.

Ghahramani, Z. Probabilistic machine learning and artificial intelligence. *Nature*, 521(7553): 452–459, May 2015. ISSN 1476-4687. doi: 10.1038/nature14541.

Gibbs, I. and Candes, E. Adaptive conformal inference under distribution shift. In Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, 2021.

Ginsbourger, D. and Schärer, C. Fast calculation of gaussian process multiple-fold cross-validation residuals and their covariances, 2021.

Golub, G. and Van Loan, C. *Matrix Computations*. Johns Hopkins Studies in the Mathematical Sciences. Johns Hopkins University Press, 2013. ISBN 9781421408590.

Gretton, A., Fukumizu, K., Teo, C., Song, L., Schölkopf, B., and Smola, A. A kernel statistical test of independence. In Platt, J., Koller, D., Singer, Y., and Roweis, S. (eds.), *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc., 2007.

Guardabascio, B. and Ventura, M. Estimating the dose–response function through a generalized linear model approach. *The Stata Journal*, 14(1):141–158, 2014.

Guo, R., Cheng, L., Li, J., Hahn, P. R., and Liu, H. A survey of learning causality with data: Problems and methods. *ACM Computing Surveys (CSUR)*, 53(4):1–37, 2020.

Gupta, S., Kambli, R., Wagh, S., and Kazi, F. Support-vector-machine-based proactive cascade prediction in smart grid using probabilistic framework. *IEEE Transactions on Industrial Electronics*, 62(4):2478–2486, 2015. doi: 10.1109/TIE.2014.2361493.

Guttman, I. *Statistical Tolerance Regions: Classical and Bayesian*. Griffin's statistical monographs & courses. Hafner Publishing Company, 1970.

Hahn, P. R., Murray, J. S., and Carvalho, C. M. Bayesian regression tree models for causal inference: Regularization, confounding, and heterogeneous effects (with discussion). *Bayesian Analysis*, 15(3):965–1056, 2020.

Han, M., May, R., Zhang, X., Wang, X., Pan, S., Da, Y., and Jin, Y. A novel reinforcement learning method for improving occupant comfort via window opening and closing. *Sustainable Cities and Society*, 61:102247, 2020. ISSN 2210-6707. doi: 10.1016/j.scs.2020.102247.

Hansen, L. and Salamon, P. Neural network ensembles. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(10):993–1001, 1990. doi: 10.1109/34.58871.

Harada, S. and Kashima, H. Graphite: Estimating individual effects of graph-structured treatments. In *Proceedings of the 30th ACM International Conference on Information and Knowledge Management*, pp. 659–668, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450384469.

Hassanpour, N. and Greiner, R. Counterfactual regression with importance sampling weights. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pp. 5880–5887. International Joint Conferences on Artificial Intelligence Organization, 7 2019. doi: 10.24963/ijcai.2019/815.

Hastie, T., Tibshirani, R., and Friedman, J. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., New York, NY, USA, 2001.

Hastie, T., Tibshirani, R., and Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*. Springer Series in Statistics. Springer New York, 2009.

Heiler, P. and Knaus, M. C. Effect or treatment heterogeneity? policy evaluation with aggregated and disaggregated treatments. *arXiv preprint arXiv:2110.01427*, 2021.

Heinermann, J. and Kramer, O. Machine learning ensembles for wind power prediction. *Renewable Energy*, 89:671–679, 2016. ISSN 0960-1481. doi: 10.1016/j.renene.2015.11.073.

Hernan, M. and Robins, J. *Causal Inference*. Chapman & Hall/CRC Monographs on Statistics & Applied Probab. Taylor & Francis, 2020. ISBN 9781420076165.

Heskes, T. Practical confidence and prediction intervals. In Mozer, M., Jordan, M., and Petsche, T. (eds.), *Advances in Neural Information Processing Systems*, volume 9. MIT Press, 1996.

Hill, J. L. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240, 2011. doi: 10.1198/jcgs.2010.08162.

Holland, P. and Rubin, D. Causal inference in retrospective studies. *Evaluation Review - EVALUATION REV*, 12:203–231, 06 1988. doi: 10.1177/0193841X8801200301.

Holland, P. W. Statistics and causal inference. *Journal of the American Statistical Association*, 81(396):945–960, 1986.

Horvitz, D. G. and Thompson, D. J. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260):663–685, 1952. ISSN 01621459.

Hu, L., Gu, C., Lopez, M., Ji, J., and Wisnivesky, J. Estimation of causal effects of multiple treatments in observational studies with a binary outcome. *Statistical methods in medical research*, 29(11):3218–3234, 2020.

Hu, R., Huang, Q., Chang, S., Wang, H., and He, J. The mbpep: A deep ensemble pruning algorithm providing high quality uncertainty prediction. *Applied Intelligence*, 49(8):2942–2955, aug 2019. doi: 10.1007/s10489-019-01421-8.

Hüllermeier, E. and Waegeman, W. Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods. *Machine Learning*, 110(3):457–506, 2021. doi: 10.1007/s10994-021-05946-3.

Hume, D. *A Treatise of Human Nature (1739-40)*. Oxford University Press, 1978.

Idowu, S., Saguna, S., Åhlund, C., and Schelén, O. Applied machine learning: Forecasting heat load in district heating system. *Energy and Buildings*, 133:478–488, 2016. ISSN 0378-7788. doi: 10.1016/j.enbuild.2016.09.068.

Imai, K. and Dyk, D. A. V. Causal inference with general treatment regimes: Generalizing the propensity score. *Journal of the American Statistical Association*, 99(467):854–866, 2004.

Imai, K. and Ratkovic, M. Estimating treatment effect heterogeneity in randomized program evaluation. *The Annals of Applied Statistics*, 7(1):443–470, 2013.

Imai, K. and Ratkovic, M. Covariate balancing propensity score. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):243–263, 2014.

Imai, K. and Van Dyk, D. A. Causal inference with general treatment regimes: Generalizing the propensity score. *Journal of the American Statistical Association*, 99(467):854–866, 2004.

Imbens, G. W. The role of the propensity score in estimating dose-response functions. *Biometrika*, 87(3):706–710, 2000.

Imbens, G. W. Nonparametric Estimation of Average Treatment Effects Under Exogeneity: A Review. *The Review of Economics and Statistics*, 86(1):4–29, 02 2004. ISSN 0034-6535. doi: 10.1162/003465304323023651.

Imbens, G. W. Potential outcome and directed acyclic graph approaches to causality: Relevance for empirical practice in economics. *Journal of Economic Literature*, 58(4):1129–79, 2020.

Imbens, G. W. and Rubin, D. B. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction.* Cambridge University Press, 2015. doi: 10.1017/CBO9781139025751.

Imberman, S. A. How effective are financial incentives for teachers. *The IZA World of Labor*, pp. 158–158, 2015.

Intergovernmental Panel on Climate Change. Climate change 2014: Mitigation of climate change: Working group iii contribution to the ipcc fifth assessment report. *Cambridge University Press*, 1454:147, 2015. doi: 10.1017/CBO9781107415416.

International Energy Agency. *World Energy Outlook 2016.* International Energy Agency, 2016. doi: 10.1787/weo-2016-en.

International Energy Agency. *World Energy Outlook 2017.* International Energy Agency, 2017. doi: 10.1787/weo-2017-en.

International Energy Agency. *World Energy Outlook 2021.* International Energy Agency, 2021. doi: 10.1787/14fcb638-en.

International Energy Agency. The annual energy outlook, 2022. https://www.eia.gov/outlooks/aeo/pdf/AEO2022_ReleasePresentation.pdf.

Iooss, B. and Le Gratiet, L. Uncertainty and sensitivity analysis of functional risk curves based on gaussian processes. *Reliability Engineering & System Safety*, 187:58–66, 2019. ISSN 0951-8320. doi: 10.1016/j.ress.2017.11.022. Sensitivity Analysis of Model Output.

Iooss, B. and Marrel, A. An efficient methodology for the analysis and modeling of computer experiments with large number of inputs. In *UNCECOMP 2017 2nd ECCOMAS Thematic Conference on Uncertainty Quantification in Computational Sciences and Engineering*, pp. 187–197, Rhodes Island, Greece, June 2017.

James, G., Witten, D., Hastie, T., and Tibshirani, R. *An Introduction to Statistical Learning: With Applications in R*. Springer Publishing Company, Incorporated, 2014. ISBN 1461471370.

Jarvis, A. J., Leedal, D. T., and Hewitt, C. N. Climate–society feedbacks and the avoidance of dangerous climate change. *Nature Climate Change*, 2(9):668–671, Sep 2012. ISSN 1758-6798. doi: 10.1038/nclimate1586.

Jin, Y., Ren, Z., and Candès, E. J. Sensitivity analysis of individual treatment effects: A robust conformal inference approach. *arXiv preprint arXiv:2111.12161*, 2021.

Johansson, F., Shalit, U., and Sontag, D. Learning representations for counterfactual inference. In Balcan, M. F. and Weinberger, K. Q. (eds.), *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 3020–3029, New York, New York, USA, 20–22 Jun 2016. PMLR.

Jovanović, R. Ž., Sretenović, A. A., and Živković, B. D. Ensemble of various neural networks for prediction of heating energy consumption. *Energy and Buildings*, 94:189–199, 2015. ISSN 0378-7788. doi: 10.1016/j.enbuild.2015.02.052.

Jursa, R. and Rohrig, K. Short-term wind power forecasting using evolutionary algorithms for the automated specification of artificial intelligence models. *International Journal of Forecasting*, 24(4):694–709, 2008. ISSN 0169-2070. doi: 10.1016/j.ijforecast.2008.08.007. Energy Forecasting.

Kaddour, J., Zhu, Y., Liu, Q., Kusner, M. J., and Silva, R. Causal effect inference for structured treatments. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 24841–24854. Curran Associates, Inc., 2021.

Kallus, N., Mao, X., and Zhou, A. Interval estimation of individual-level causal effects under unobserved confounding. In Chaudhuri, K. and Sugiyama, M. (eds.), *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pp. 2281–2290. PMLR, 04 2019.

Kang, J. D. and Schafer, J. L. Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical science*, 22(4): 523–539, 2007.

Karvonen, T. Asymptotic bounds for smoothness parameter estimates in gaussian process interpolation, 2022.

Karvonen, T. and Oates, C. J. Maximum likelihood estimation in gaussian process regression is ill-posed, 2022.

Karvonen, T., Wynne, G., Tronarp, F., Oates, C., and Särkkä, S. Maximum likelihood estimation and uncertainty quantification for gaussian process approximation of deterministic functions. *SIAM/ASA Journal on Uncertainty Quantification*, 8(3):926–958, 2020. doi: 10.1137/20M1315968.

Kendall, A. and Gal, Y. What uncertainties do we need in bayesian deep learning for computer vision? In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

Kennedy, E. H. Optimal doubly robust estimation of heterogeneous causal effects. *arXiv preprint arXiv:2004.14497*, 2020.

Kennedy, E. H., Ma, Z., McHugh, M. D., and Small, D. S. Non-parametric methods for doubly robust estimation of continuous treatment effects. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(4):1229–1245, 2017.

Kennedy, M. C. and O'Hagan, A. Predicting the output from a complex computer code when fast approximations are available. *Biometrika*, 87(1):1–13, 2000. ISSN 00063444.

Kerleguer, B. Multi-fidelity surrogate modeling for time-series outputs, 2021.

Khosravi, A., Nahavandi, S., and Creighton, D. A prediction interval-based approach to determine optimal structures of neural network metamodels. *Expert Syst. Appl.*, 37:2377–2387, 03 2010. doi: 10.1016/j.eswa.2009.07.059.

Khosravi, A., Nahavandi, S., Creighton, D., and Atiya, A. F. Lower upper bound estimation method for construction of neural network-based prediction intervals. *IEEE Transactions on Neural Networks*, 22(3):337–346, 2011.

Kleijnen, J. P. C. and Sargent, R. G. A methodology for fitting and validating metamodels in simulation. *European Journal of Operational Research*, 120:14–29, 2000.

Knaus, M. C., Lechner, M., and Strittmatter, A. Machine learning estimation of heterogeneous causal effects: Empirical Monte Carlo evidence. *The Econometrics Journal*, 24(1):134–161, 06 2020a.

Knaus, M. C., Lechner, M., and Strittmatter, A. Heterogeneous employment effects of job search programmes: A machine learning approach. *Journal of Human Resources*, Mar 2020b. doi: 10.3368/jhr.57.2.0718-9615r1.

Koenker, R. *Quantile Regression*. Econometric Society Monographs. Cambridge University Press, 2005. doi: 10.1017/CBO9780511754098.

Koenker, R. and Bassett, G. Regression quantiles. *Econometrica*, 46(1):33–50, 1978.

Koenker, R. and Hallock, K. F. Quantile regression. *Journal of Economic Perspectives*, 15(4): 143–156, December 2001. doi: 10.1257/jep.15.4.143.

Kolmogorov, A. Interpolation and extrapolation of stationary sequences. *Izvestiya the Academy of Sciences of the USSR, Ser. Math.*, 52(5):3–14, 1941.

Krige, D. A statistical approach to some basic mine valuation problems on the witwatersrand. *Journal of the Southern African Institute of Mining and Metallurgy*, 52(6):119–139, 1951. doi: 10.10520/AJA0038223X_4792.

Kunsch, H. R. The jackknife and the bootstrap for general stationary observations. *The Annals of Statistics*, 17(3):1217–1241, 1989. ISSN 00905364.

Künzel, S. R., Sekhon, J. S., Bickel, P. J., and Yu, B. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences*, 116(10):4156–4165, 02 2019. ISSN 1091-6490. doi: 10.1073/pnas.1804597116.

Lakshminarayanan, B., Pritzel, A., and Blundell, C. Simple and scalable predictive uncertainty estimation using deep ensembles. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

LaLonde, R. J. Evaluating the econometric evaluations of training programs with experimental data. *The American Economic Review*, 76(4):604–620, 1986. ISSN 00028282.

Lawless, J. F. and Fredette, M. Frequentist prediction intervals and predictive distributions. *Biometrika*, 92(3):529–542, 2005.

Le Gratiet, L. Bayesian analysis of hierarchical multifidelity codes. *SIAM/ASA Journal on Uncertainty Quantification*, 1(1):244–269, 2013. doi: 10.1137/120884122.

Le Gratiet, L., Marelli, S., and Sudret, B. *Metamodel-Based Sensitivity Analysis: Polynomial Chaos Expansions and Gaussian Processes*, pp. 1289–1325. Springer International Publishing, Cham, 2017. ISBN 978-3-319-12385-1. doi: 10.1007/978-3-319-12385-1_38.

Lechner, M. Identification and estimation of causal effects of multiple treatments under the conditional independence assumption. In Lechner, M. and Pfeiffer, F. (eds.), *Econometric Evaluation of Labour Market Policies*, pp. 43–58, Heidelberg, 2001. Physica-Verlag HD.

Lechner, M. Modified causal forests for estimating heterogeneous causal effects. *arXiv preprint arXiv:1812.09487*, 2018.

Lee, B. K., Lessler, J., and Stuart, E. A. Improving propensity score weighting using machine learning. *Statistics in medicine*, 29(3):337–346, 2010.

Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R., and Wasserman, L. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113, 04 2016. doi: 10.1080/01621459.2017.1307116.

Lewis, D. Causation. *Journal of Philosophy*, 70(17):556–567, 1973. doi: 10.2307/2025310.

Lewis, D. *Philosophical Papers, Volume II*. Oxford University Press, 1986.

Lewis, D. Causation as influence. *The Journal of Philosophy*, 97(4):182–197, 2000.

Li, M. and Lior, N. Energy analysis for guiding the design of well systems of deep enhanced geothermal systems. *Energy*, 93:1173–1188, 2015. ISSN 0360-5442. doi: 10.1016/j.energy.2015.09.113.

Li, Y., Liu, K., Foley, A. M., Zülke, A., Berecibar, M., Nanini-Maury, E., Van Mierlo, J., and Hoster, H. E. Data-driven health estimation and lifetime prediction of lithium-ion batteries: A review. *Renewable and Sustainable Energy Reviews*, 113:109254, 2019. ISSN 1364-0321. doi: 10.1016/j.rser.2019.109254.

Li, Y., Kuang, K., Li, B., Cui, P., Tao, J., Yang, H., and Wu, F. Continuous treatment effect estimation via generative adversarial de-confounding. In *Proceedings of the 2020 KDD Workshop on Causal Discovery*, volume 127 of *Proceedings of Machine Learning Research*, pp. 4–22, San Diego, CA, USA, 08 2020. PMLR.

Li, Z., Rahman, S. M., Vega, R., and Dong, B. A hierarchical approach using machine learning methods in solar photovoltaic energy production forecasting. *Energies*, 9(1), 2016. ISSN 1996-1073. doi: 10.3390/en9010055.

Lin, L., Zhu, Y., and Chen, L. Causal inference for multi-level treatments with machine-learned propensity scores. *Health Services and Outcomes Research Methodology*, 19(2):106–126, 2019.

Liu, H., Ong, Y.-S., Shen, X., and Cai, J. When gaussian process meets big data: a review of scalable gps. *IEEE Transactions on Neural Networks and Learning Systems*, 31(11): 4405–4423, January 2020. ISSN 2162-237X. doi: 10.1109/TNNLS.2019.2957109.

Lopez, M. J. and Gutman, R. Estimation of causal effects with multiple treatments: a review and new ideas. *Statistical Science*, pp. 432–454, 2017.

López-Lopera, A. F., Bachoc, F., Durrande, N., and Roustant, O. Finite-dimensional gaussian approximation with linear inequality constraints. *SIAM/ASA Journal on Uncertainty Quantification*, 6(3):1224–1255, 2018. doi: 10.1137/17M1153157.

Louizos, C., Shalit, U., Mooij, J. M., Sontag, D., Zemel, R., and Welling, M. Causal effect inference with deep latent-variable models. In *Advances in neural information processing systems*, volume 30. PMLR, 2017.

Lukić, M. and Beder, J. Stochastic processes with sample paths in reproducing kernel hilbert spaces. *Transactions of the American Mathematical Society*, 353(10):3945–3969, 2001.

Luna, S. S. and Young, A. The Bootstrap and Kriging Prediction Intervals. *Scandinavian Journal of Statistics*, 30(1):175–192, 03 2003. doi: 10.1111/1467-9469.00325.

Ma, Y. and Tresp, V. Causal inference under networked interference and intervention policy enhancement. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pp. 3700–3708. PMLR, 04 2021.

MacKay, D. J. C. A Practical Bayesian Framework for Backpropagation Networks. *Neural Computation*, 4(3):448–472, 05 1992. ISSN 0899-7667. doi: 10.1162/neco.1992.4.3.448.

Mackie, J. L. Causes and conditions. *American Philosophical Quarterly*, 2(4):245–264, 1965.

Mackie, J. L. *The Cement of the Universe: A Study of Causation*. Oxford, England: Oxford, Clarendon Press, 1974.

Makala, B. and Bakovic, T. Artificial intelligence in the power sector. *EMCompass Notes*, 08 2020. doi: 10.13140/RG.2.2.34011.18729.

Mardia, K. V. and Marshall, R. J. Maximum likelihood estimation of models for residual covariance in spatial regression. *Biometrika*, 71(1):135–146, 1984.

Marrel, A., Iooss, B., Van Dorpe, F., and Volkova, E. An efficient methodology for modeling complex computer codes with gaussian processes. *Computational Statistics & Data Analysis*, 52(10):4731–4744, 2008. ISSN 0167-9473. doi: 10.1016/j.csda.2008.03.026.

Masarotto, V., Panaretos, V. M., and Zemel, Y. Procrustes metrics on covariance operators and optimal transportation of gaussian processes. *Sankhya A*, 81(1):172–213, Feb 2019. ISSN 0976-8378. doi: 10.1007/s13171-018-0130-1.

Matheron, G. *La Théorie des variables régionalisées, et ses applications*. Les Cahiers du Centre de morphologie mathématique de Fontainebleau. Ecole Nationale Supérieure des Mines de Paris, 1970.

Mccaffrey, D., Griffin, B. A., Almirall, D., Slaughter, M., Ramchand, R., and Burgette, L. A tutorial on propensity score estimation for multiple treatments using generalized boosted models. *Statistics in medicine*, 32, 08 2013. doi: 10.1002/sim.5753.

McCaffrey, D. F., Ridgeway, G., and Morral, A. R. Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological methods*, 9(4): 403, 2004.

McCullagh, P. and Nelder, J. A. *Generalized Linear Models*. Chapman & Hall / CRC, London, 1989.

Meinshausen, N. Quantile regression forests. *Journal of Machine Learning Research*, 7:983–999, 12 2006. ISSN 1532-4435.

Ministère de la transition écologique. Chiffres clés de l'énergie - Édition 2021, 2021. https://www.statistiques.developpement-durable.gouv.fr/chiffres-cles-de-lenergie-edition-2021.

Mohaghegh, S. D., Grujic, O., Zargari, S., and Kalantari, M. Modeling, History Matching, Forecasting and Analysis of Shale Reservoirs Performance Using Artificial Intelligence. In *SPE Digital Energy Conference and Exhibition*, volume All Days, 04 2011. doi: 10.2118/143875-MS.

Moon, H. *Design and Analysis of Computer Experiments for Screening Input Variables*. PhD thesis, The Ohio State University, 01 2010.

Morgan, M. G. and Henrion, M. *Uncertainty: A Guide to Dealing with Uncertainty in Quantitative Risk and Policy Analysis*. Cambridge University Press, 1990. doi: 10.1017/CBO9780511840609.

Morgan, S. L. and Winship, C. *Counterfactuals and Causal Inference: Methods and Principles for Social Research*. Analytical Methods for Social Research. Cambridge University Press, 2 edition, 2014. doi: 10.1017/CBO9781107587991.

Morokoff, W. J. and Caflisch, R. E. Quasi-monte carlo integration. *Journal of computational physics*, 122:218–230, 1995.

Muré, J. *Objective Bayesian analysis of Kriging models with anisotropic correlation kernel*. PhD thesis, Sorbonne Paris Cité, 2018. URL http://www.theses.fr/2018USPCC069.

Murnane, R. J. and Willett, J. B. *Methods matter: Improving causal inference in educational and social science research*. Oxford University Press, 2010.

Muré, J. Propriety of the reference posterior distribution in Gaussian process modeling. *The Annals of Statistics*, 49(4):2356 – 2377, 2021. doi: 10.1214/20-AOS2040.

Neal, R. M. Probabilistic inference using markov chain monte carlo methods. Technical report, Dept. of Computer Science, University of Toronto., 1993.

Neal, R. M. *Bayesian Learning for Neural Networks*. Springer New York, 1996. doi: 10.1007/978-1-4612-0745-0.

Nelder, J. A. and Wedderburn, R. W. M. Generalized linear models. *Journal of the Royal Statistical Society. Series A (General)*, 135(3):370–384, 1972. ISSN 00359238.

Nelson, W. B. Two sample prediction. Technical report, General Electric TIS Report 68-C-404, 1968.

Neyman, J. On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9. *Statistical Science*, 5:465 – 472, 1923. Translated and published in English in 1990.

Ng, M.-F., Zhao, J., Yan, Q., Conduit, G. J., and Seh, Z. W. Predicting the state of charge and health of batteries using data-driven machine learning. *Nature Machine Intelligence*, 2 (3):161–170, Mar 2020. ISSN 2522-5839. doi: 10.1038/s42256-020-0156-7.

Nie, X. and Wager, S. Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika*, 09 2020. doi: 10.1093/biomet/asaa076.

Nix, D. A. and Weigend, A. S. Estimating the mean and variance of the target probability distribution. In *Proceedings of 1994 IEEE International Conference on Neural Networks (ICNN'94)*, volume 1, pp. 55–60 vol.1, 1994.

Nwachukwu, A., Jeong, H., Sun, A., Pyrcz, M., and Lake, L. W. Machine Learning-Based Optimization of Well Locations and WAG Parameters under Geologic Uncertainty. In *SPE Improved Oil Recovery Conference*, volume Day 3 Mon, April 16, 2018, 04 2018. doi: 10.2118/190239-MS.

Oakley, J., Hagan, A., and O'Hagan, A. Probabilistic sensitivity analysis of complex models: A bayesian approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66:751 – 769, 08 2004. doi: 10.1111/j.1467-9868.2004.05304.x.

OCDE. *Energy: The Next Fifty Years*. Organization for Economic Cooperation and Development (OECD), 1999. doi: 10.1787/9789264173163-en.

Ohio Department of Natural Resources. The Ohio oil and gas well locator, 2022. https://ohiodnr.gov/wps/portal/gov/odnr/discover-and-learn/safety-conservation/about-odnr/oil-gas/oil-gas-resources/well-locator.

Okasa, G. Meta-learners for estimation of causal effects: Finite sample cross-fit performance. *arXiv preprint arXiv:2201.12692*, 2022.

Ornstein, L. S. and Uhlenbeck, G. E. On the theory of the brownian motion. *Phys. Rev.*, 36: 823–841, Sep 1930. doi: 10.1103/PhysRev.36.823.

Paananen, T., Piironen, J., Andersen, M. R., and Vehtari, A. Variable selection for gaussian processes via sensitivity analysis of the posterior predictive distribution. In Chaudhuri, K. and Sugiyama, M. (eds.), *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pp. 1743–1752. PMLR, 16–18 Apr 2019.

Pang, J., Liu, D., Peng, Y., and Peng, X. Optimize the coverage probability of prediction interval for anomaly detection of sensor-based monitoring series. *Sensors (Basel, Switzerland)*, 18, 03 2018. doi: 10.3390/s18040967.

Papadopoulos, H. Inductive conformal prediction: Theory and application to neural networks. In Fritzsche, P. (ed.), *Tools in Artificial Intelligence*, chapter 18. IntechOpen, Rijeka, 2008. doi: 10.5772/6078.

Patel, J. K. Prediction intervals - a review. *Communications in Statistics - Theory and Methods*, 18(7):2393–2465, 1989. doi: 10.1080/03610928908830043.

Pearce, T., Brintrup, A., Zaki, M., and Neely, A. High-quality prediction intervals for deep learning: A distribution-free, ensembled approach. In Dy, J. and Krause, A. (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 4075–4084. PMLR, 10–15 Jul 2018.

Pearl, J. Causal diagrams for empirical research. *Biometrika*, 82(4):669–688, 1995.

Pearl, J. *Causality*. Cambridge University Press, 2 edition, 2009. doi: 10.1017/CBO9780511803161.

Pearl, J. Aspects of graphical models connected with causality. 2011.

Pearl, J. Causes of effects and effects of causes. *Sociological Methods & Research*, 44(1):149–164, 2015. doi: 10.1177/0049124114562614.

Pearl, J. The seven tools of causal inference, with reflections on machine learning. *Commun. ACM*, 62(3):54–60, 02 2019. ISSN 0001-0782. doi: 10.1145/3241036.

Pearl, J. and Mackenzie, D. *The book of why: the new science of cause and effect*. Basic books, 2018.

Pearl, J. and Shafer, G. Probabilistic reasoning in intelligent systems: Networks of plausible inference. *Synthese-Dordrecht*, 104(1):161, 1995.

Pearson, K. *The Grammar of Science*. Adam and Charles Black, 1892.

Pearson, K. Mathematical contributions to the theory of evolution. iii. regression, heredity, and panmixia. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, pp. 253–318, 1896.

Pennec, X., Fillard, P., and Ayache, N. A riemannian framework for tensor computing. *International Journal of Computer Vision*, 66(1):41–66, Jan 2006. ISSN 1573-1405. doi: 10.1007/s11263-005-3222-z.

Pepelyshev, A. The role of the nugget term in the gaussian process method. In *mODa 9 – Advances in Model-Oriented Design and Analysis*, pp. 149–156, Heidelberg, 2010. Physica-Verlag HD. ISBN 978-3-7908-2410-0.

Perrin, G., Soize, C., Marque-Pucheu, S., and Garnier, J. Nested polynomial trends for the improvement of gaussian process-based predictors. *Journal of Computational Physics*, 346: 389–402, 2017. ISSN 0021-9991. doi: 10.1016/j.jcp.2017.05.051.

Peters, J., Janzing, D., and Schlkopf, B. *Elements of Causal Inference: Foundations and Learning Algorithms*. The MIT Press, 2017. ISBN 0262037319.

Pigoli, D., Aston, J. A. D., Dryden, I. L., and Secchi, P. Distances and inference for covariance operators. *Biometrika*, 101(2):409–422, 04 2014.

Podofillini, L., Sudret, B., Stojadinovic, B., Zio, E., and Kröger, W. *Safety and Reliability of Complex Engineered Systems: ESREL 2015*. CRC Press, 2015. ISBN 9781315648415.

Powers, S., Qian, J., Jung, K., Schuler, A., Shah, N. H., Hastie, T., and Tibshirani, R. Some methods for heterogeneous treatment effect estimation in high dimensions. *Statistics in medicine*, 37(11):1767–1787, 2018.

Probst, P., Wright, M. N., and Boulesteix, A. Hyperparameters and tuning strategies for random forest. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 9, 2019.

Proschan, F. Confidence and tolerance intervals for the normal distribution. *Journal of the American Statistical Association*, 48(263):550–564, 1953. ISSN 01621459.

Quenouille, M. H. Approximate tests of correlation in time-series. *Journal of the Royal Statistical Society. Series B (Methodological)*, 11(1):68–84, 1949. ISSN 00359246.

Rapp, M., Mencía, E. L., Fürnkranz, J., Nguyen, V.-L., and Hüllermeier, E. Learning gradient boosted multi-label classification rules. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 124–140. Springer, 2020.

Rasmussen, C. E. and Williams, C. K. I. *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press, 2005.

Raza, M. Q. and Khosravi, A. A review on artificial intelligence based load demand forecasting techniques for smart grid and buildings. *Renewable and Sustainable Energy Reviews*, 50: 1352–1372, 2015. ISSN 1364-0321. doi: 10.1016/j.rser.2015.04.065.

Ren, C., Sun, D., and He, C. Z. Objective bayesian analysis for a spatial model with nugget effects. *Journal of Statistical Planning and Inference*, 142:1933–1946, 2012.

Richardson, T. S. and Robins, J. M. Single world intervention graphs (swigs): A unification of the counterfactual and graphical approaches to causality. *Center for the Statistics and the Social Sciences, University of Washington Series. Working Paper*, 128(30):2013, 2013.

Ripley, B. *Spatial Statistics*. Wiley Series in Probability and Statistics. Wiley, 1981. ISBN 9780471083672.

Ritter, K. *Average-case analysis of numerical problems*. Springer Science & Business Media, 2000.

Robert, C. P. and Casella, G. *Monte Carlo Statistical Methods*. Springer New York, 2004. doi: 10.1007/978-1-4757-4145-2.

Roberts, S., Osborne, M., Ebden, M., Reece, S., Gibson, N., and Aigrain, S. Gaussian processes for time-series modelling. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 371(1984):20110550, 2013.

Robins, J., Hernán, M., and Brumback, B. Marginal structural models and causal inference in epidemiology. *Epidemiology*, 11:550–560, 10 2000. doi: 10.1097/00001648-200009000-00011.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89(427): 846–866, 1994.

Robinson, P. M. Root-n-consistent semiparametric regression. *Econometrica*, 56(4):931–954, 1988.

Romano, Y., Patterson, E., and Candes, E. Conformalized quantile regression. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Rosenbaum, P. R. Model-based direct adjustment. *Journal of the American Statistical Association*, 82(398):387–394, 1987. ISSN 01621459.

Rosenbaum, P. R. and Rubin, D. B. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 04 1983.

Rosenbaum, P. R. and Rubin, D. B. Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association*, 79(387):516–524, 1984.

Rosenbaum, P. R. and Rubin, D. B. The bias due to incomplete matching. *Biometrics*, pp. 103–116, 1985.

Rosenblatt, M. Review: A. M. Yaglom, correlation theory of stationary and random functions vol. i; basic results, vol. ii, supplementary notes and references. *Bulletin (New Series) of the American Mathematical Society*, 20(2):207–211, 04 1989.

Roustant, O., Ginsbourger, D., and Deville, Y. DiceKriging, DiceOptim: Two R packages for the analysis of computer experiments by kriging-based metamodeling and optimization. *Journal of Statistical Software*, 51(1):1–55, 2012. URL http://www.jstatsoft.org/v51/i01/.

Roustant, O., Padonou, E., Deville, Y., Clément, A., Perrin, G., Giorla, J., and Wynn, H. Group kernels for Gaussian process metamodels with categorical inputs. *SIAM/ASA Journal on Uncertainty Quantification*, 8(2):775–806, 2020. doi: 10.1137/18M1209386.

Rubin, D. Estimating causal effects if treatment in randomized and nonrandomized studies. *J. Educ. Psychol.*, 66, 1974.

Rubin, D. Bayesian inference for causal effects: The role of randomization. *The Annals of Statistics*, 6, 1978. doi: 10.1214/aos/1176344064.

Rubin, D. Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100:322–331, 02 2005. doi: 10.2307/27590541.

Rubin, D. B. Using multivariate matched sampling and regression adjustment to control bias in observational studies. *Journal of the American Statistical Association*, 74(366a):318–328, 1979.

Rubin, D. B. Comment: Neyman (1923) and causal inference in experiments and observational studies. *Statistical Science*, 5(4):472–480, 1990.

Rullière, D., Durrande, N., Bachoc, F., and Chevalier, C. Nested kriging predictions for datasets with large number of observations. *Statistics and Computing*, 28, 07 2018. doi: 10.1007/s11222-017-9766-2.

Russell, B. On the notion of cause. *Proceedings of the Aristotelian Society*, 13:1–26, 1912.

Sacks, J., Welch, W. J., Mitchell, T., and Wynn, H. Design and analysis of computer experiments. *Statist. Sci.*, 4(4):409–423, 11 1989. doi: 10.1214/ss/1177012413.

Saini, S. K., Dhamnani, S., Aakash, Ibrahim, A. A., and Chavan, P. Multiple treatment effect estimation using deep generative model with task embedding. In *The World Wide Web Conference*, WWW '19, pp. 1601–1611, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450366748. doi: 10.1145/3308558.3313744.

Santner, T. J., Williams, B. J., and Notz, W. I. *The Design and Analysis of Computer Experiments*. Springer New York, 2003. doi: 10.1007/978-1-4757-3799-8.

Santosh, T., Vinod, G., Saraf, R., Ghosh, A., and Kushwaha, H. Application of artificial neural networks to nuclear power plant transient diagnosis. *Reliability Engineering & System Safety*, 92(10):1468–1472, 2007. ISSN 0951-8320. doi: 10.1016/j.ress.2006.10.009.

Schlumberger. Petrel reservoir engineering, 2015. https://www.software.slb.com/products/petrel/petrel-reservoir-engineering.

Schlumberger. Eclipse simulators, 2017. https://www.software.slb.com/products/eclipse/simulators.

Schölkopf, B. and Smola, A. J. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. Adaptive computation and machine learning. MIT Press, 2002.

Schwab, P., Linhardt, L., Bauer, S., Buhmann, J., and Karlen, W. Learning counterfactual representations for estimating individual dose-response curves. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34:5612–5619, 04 2020. doi: 10.1609/aaai.v34i04.6014.

Schölkopf, B., Janzing, D., Peters, J., Sgouritsa, E., Zhang, K., and Mooij, J. M. On causal and anticausal learning. In *n Proceedings of the 29th International Conference on Machine Learning*, 2012.

Securities and Exchange Commission. Modernization of oil and gas reporting, revisions and additions to the definition section in rule 4-10 of regulation s-x, January 2010. https://www.sec.gov/rules/final/2008/33-8995.pdf.

Severson, K. A., Attia, P. M., Jin, N., Perkins, N., Jiang, B., Yang, Z., Chen, M. H., Aykol, M., Herring, P. K., Fraggedakis, D., Bazant, M. Z., Harris, S. J., Chueh, W. C., and Braatz, R. D. Data-driven prediction of battery cycle life before capacity degradation. *Nature Energy*, 4(5), 3 2019. doi: 10.1038/s41560-019-0356-8.

Shalit, U., Johansson, F. D., and Sontag, D. Estimating individual treatment effect: generalization bounds and algorithms. In *International Conference on Machine Learning*, pp. 3076–3085. PMLR, 2017.

Shapiro, S. S. and Wilk, M. B. An analysis of variance test for normality (complete samples). *Biometrika*, 52(3/4):591–611, 1965. ISSN 00063444.

Shi, C., Blei, D., and Veitch, V. Adapting neural networks for the estimation of treatment effects. In *Advances in neural information processing systems*, volume 32. PMLR, 2019.

Shi, C., Veitch, V., and Blei, D. M. Invariant representation learning for treatment effect estimation. In *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence*, volume 161 of *Proceedings of Machine Learning Research*, pp. 1546–1555. PMLR, 07 2021.

Shimizu, S. Lingam: Non-gaussian methods for estimating causal structures. *Behaviormetrika*, 41(1):65–98, 2014.

Shimizu, S., Hoyer, P. O., Hyvärinen, A., Kerminen, A., and Jordan, M. A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7(10), 2006.

Shrestha, D. L. and Solomatine, D. P. Machine learning approaches for estimation of prediction interval for the model output. *Neural Networks*, 19(2):225–235, 2006. ISSN 0893-6080. doi: 10.1016/j.neunet.2006.01.012. Earth Sciences and Environmental Applications of Computational Intelligence.

Sobel, M. E. Causal inference in the social and behavioral sciences. In *Handbook of statistical modeling for the social and behavioral sciences*, pp. 1–38. Springer, 1995.

Society of Petroleum Engineers. Petroleum reserves and resources definitions, 2022. https://www.spe.org/en/industry/reserves/.

Spirtes, P. and Glymour, C. An algorithm for fast recovery of sparse causal graphs. *Social science computer review*, 9(1):62–72, 1991.

Spirtes, P., Glymour, C. N., Scheines, R., and Heckerman, D. *Causation, prediction, and search*. MIT press, 2000.

Spreeuwenberg, M. D., Bartak, A., Croon, M. A., Hagenaars, J. A., Busschbach, J. J. V., Andrea, H., Twisk, J., and Stijnen, T. The multiple propensity score as control for bias in the comparison of more than two treatment arms: An introduction from a case study in mental health. *Medical Care*, 48(2):166–174, 2010.

Stein, M. L. *Interpolation of spatial data: some theory for kriging*. Springer New York, 1999. doi: 10.1007/978-1-4612-1494-6.

Steinberger, L. and Leeb, H. Conditional predictive inference for stable algorithms, 2018.

Stuart, E. A. Matching methods for causal inference: A review and a look forward. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 25(1):1, 2010.

Su, L., Ura, T., and Zhang, Y. Non-separable models with high-dimensional data. *Journal of Econometrics*, 212(2):646–677, 2019.

Sullivan, T. *Introduction to Uncertainty Quantification*. Texts in Applied Mathematics. Springer International Publishing, 2015. ISBN 9783319233956.

Suppes, P. *A Probabilistic Theory of Causality*. Amsterdam: North-Holland Pub. Co., 1968.

Swiler, L. P., Gulian, M., Frankel, A. L., Safta, C., and Jakeman, J. D. A survey of constrained gaussian process regression: Approaches and implementation challenges. *Journal of Machine Learning for Modeling and Computing*, 1(2):119–156, 2020. ISSN 2689-3967.

Teimouri, H., Milani, A. S., Loeppky, J., and Seethaler, R. A gaussian process–based approach to cope with uncertainty in structural health monitoring. *Structural Health Monitoring*, 16 (2):174–184, 2017. doi: 10.1177/1475921716669722.

Thatcher, A. R. Relationships between bayesian and confidence limits for predictions. *Journal of the Royal Statistical Society. Series B (Methodological)*, 26(2):176–210, 1964. ISSN 00359246.

The World Economic Forum. Harnessing artificial intelligence to accelerate the energy transition. 2021. URL https://www3.weforum.org/docs/WEF_Harnessing_AI_to_accelerate_the_Energy_Transition_2021.pdf.

Tian, Q., Nordman, D. J., and Meeker, W. Q. Methods to compute prediction intervals: A review and new results, 2020.

Tibshirani, R. J., Barber, R. F., Candès, E. J., and Ramdas, A. Conformal prediction under covariate shift. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, 2019. Curran Associates Inc.

Tikhonov, A. N., Goncharsky, A., Stepanov, V., and Yagola, A. G. *Numerical methods for the solution of ill-posed problems*, volume 328. Springer Science & Business Media, 1995.

Tolba, H., Dkhili, N., Nou, J., Eynard, J., Thil, S., and Grieu, S. GHI forecasting using Gaussian process regression. In *IFAC Workshop on Control of Smart Grid and Renewable Energy Systems*, Jeju, South Korea, June 2019.

TotalEnergies. From net zero ambition to total strategy, 2020. https://totalenergies.com/sites/g/files/nytnzq121/files/documents/2020-09/strategy-and-outlook-2020.pdf.

TotalEnergies. Energy outlook 2021, 2021. https://totalenergies.com/system/files/documents/2021-09/2021_TotalEnergies_Energy_Outlook.pdf.

TotalEnergies. Sustainability & climate 2022 progress report, 2022. https://totalenergies.com/system/files/documents/2022-03/Sustainability_Climate_2022_Progress_Report_EN_0.pdf.

Tübbicke, S. Entropy balancing for continuous treatments. *Journal of Econometric Methods*, 11(1):71–89, 2022.

Tukey, J. Bias and confidence in not-quite large sample. *Annals of Mathematical Statistics*, 29: 614, 1958.

UN General Assembly. Transforming our world : the 2030 agenda for sustainable development, 2015. https://www.un.org/ga/search/view_doc.asp?symbol=A/RES/70/1&Lang=E.

United Nations. Goal 7: Affordable and clean energy, 2021. https://www.statistiques.developpement-durable.gouv.fr/chiffres-cles-de-lenergie-edition-2021.

van Asselt, M. and Rotmans, J. Uncertainty in perspective. *Global Environmental Change*, 6 (2):121–157, 1996. ISSN 0959-3780.

Varshney, K. R. Engineering safety in machine learning. In *2016 Information Theory and Applications Workshop (ITA)*, pp. 1–5, 2016. doi: 10.1109/ITA.2016.7888195.

Vegetabile, B. G., Griffin, B. A., Coffman, D. L., Cefalu, M., Robbins, M. W., and McCaffrey, D. F. Nonparametric estimation of population average dose-response curves using entropy balancing weights for continuous exposures. *Health Services and Outcomes Research Methodology*, 21(1):69–110, 2021.

Veiga, S. D. and Marrel, A. Gaussian process modeling with inequality constraints. *Annales de la Faculté des Sciences de Toulouse*, 21:529–555, 2012.

Villani, C. *The Wasserstein distances*, pp. 93–111. Springer Berlin Heidelberg, Berlin, Heidelberg, 2009.

Villani, C., Bonnet, Y., Berthet, C., Levin, F., Schoenauer, M., Cornut, A., and Rondepierre, B. *Donner un sens à l'intelligence artificielle: pour une stratégie nationale et européenne*. éditeur inconnu, 2018. ISBN 9782111457003.

Von Mises, R. Chapter viii - more on distributions. In Von Mises, R. (ed.), *Mathematical Theory of Probability and Statistics*, pp. 368–430. Academic Press, 1964. ISBN 978-1-4832-3213-3.

Vovk, V. Conditional validity of inductive conformal predictors. In Hoi, S. C. H. and Buntine, W. (eds.), *Proceedings of the Asian Conference on Machine Learning*, volume 25 of *Proceedings of Machine Learning Research*, pp. 475–490, Singapore Management University, Singapore, 04–06 Nov 2012. PMLR.

Vovk, V., Gammerman, A., and Saunders, C. Machine-learning applications of algorithmic randomness. In *Proceedings of the Sixteenth International Conference on Machine Learning*, ICML '99, pp. 444–453, San Francisco, CA, USA, 1999. Morgan Kaufmann Publishers Inc. ISBN 1558606122.

Vovk, V., Gammerman, A., and Shafer, G. *Algorithmic Learning in a Random World*. Springer US, 01 2005. doi: 10.1007/b106715.

Vovk, V., Nouretdinov, I., Manokhin, V., and Gammerman, A. Cross-conformal predictive distributions. In Gammerman, A., Vovk, V., Luo, Z., Smirnov, E., and Peeters, R. (eds.), *Proceedings of the Seventh Workshop on Conformal and Probabilistic Prediction and Applications*, volume 91 of *Proceedings of Machine Learning Research*, pp. 37–51. PMLR, 11–13 Jun 2018.

Voyant, C., Notton, G., Kalogirou, S., Nivet, M.-L., Paoli, C., Motte, F., and Fouilloy, A. Machine learning methods for solar radiation forecasting: A review. *Renewable Energy*, 105: 569–582, 2017. ISSN 0960-1481. doi: 10.1016/j.renene.2016.12.095.

Wager, S. and Athey, S. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018.

Wager, S., Hastie, T., and Efron, B. Confidence intervals for random forests: The jackknife and the infinitesimal jackknife. *Journal of Machine Learning Research*, 15(48):1625–1651, 2014.

Wang, W. On the inference of applying gaussian process modeling to a deterministic function. *Electronic Journal of Statistics*, 15, 01 2021. doi: 10.1214/21-EJS1912.

Wang, Z., Hong, T., and Piette, M. A. Building thermal load prediction through shallow machine learning and deep learning. *Applied Energy*, 263:114683, 2020. ISSN 0306-2619. doi: 10.1016/j.apenergy.2020.114683.

Washington, W., Bader, D., Collins, B., Drake, J., Taylor, M., Kirtman, B., Williams, D., and Middleton, D. Scientific grand challenges: Challenges in climate change science and the role of computing at the extreme scale. *U.S. Department of Energy.*, pp. 135, 01 2008. doi: 10.2172/990597.

Wendland, H. *Scattered Data Approximation*. Cambridge Monographs on Applied and Computational Mathematics. Cambridge University Press, 2004. doi: 10.1017/CBO9780511617539.

Wiener, N. *Extrapolation, Interpolation, and Smoothing of Stationary Time Series: With Engineering Applications*. The MIT Press, 08 1949. ISBN 9780262257190. doi: 10.7551/mitpress/2946.001.0001.

Wiens, J., Saria, S., Sendak, M., Ghassemi, M., Liu, V. X., Doshi-Velez, F., Jung, K., Heller, K., Kale, D., Saeed, M., Ossorio, P. N., Thadaney-Israni, S., and Goldenberg, A. Do no harm: a roadmap for responsible machine learning for health care. *Nature Medicine*, 25(9): 1337–1340, Sep 2019. ISSN 1546-170X. doi: 10.1038/s41591-019-0548-6.

Wilks, S. Multidimensional statistical scatter. *Collected Papers, Contributions to Mathematical Statistics*, 01 1967.

Wilks, S. S. Certain generalizations in the analysis of variance. *Biometrika*, 24(3/4):471–494, 1932.

Williams, C. and Barber, D. Bayesian classification with Gaussian processes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(12):1342–1351, 1998. doi: 10.1109/34.735807.

Williams, C. and Rasmussen, C. Gaussian processes for regression. In Touretzky, D., Mozer, M., and Hasselmo, M. (eds.), *Advances in Neural Information Processing Systems*, volume 8. MIT Press, 1995.

Wu, X., Mealli, F., Kioumourtzoglou, M.-A., Dominici, F., and Braun, D. Matching on generalized propensity scores with continuous exposures. *arXiv preprint arXiv:1812.06575*, 2018.

Wynne, G., Briol, F.-X., and Girolami, M. Convergence guarantees for gaussian process means with misspecified likelihoods and smoothness. *Journal of Machine Learning Research*, 22 (123):1–40, 2021.

Wyss, R., Ellis, A. R., Brookhart, M. A., Girman, C. J., Jonsson Funk, M., LoCasale, R., and Stürmer, T. The role of prediction modeling in propensity score estimation: an evaluation of logistic regression, bcart, and the covariate-balancing propensity score. *American journal of epidemiology*, 180(6):645–655, 2014.

Xu, C. and Xie, Y. Conformal prediction interval for dynamic time-series. In Meila, M. and Zhang, T. (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 11559–11569. PMLR, 18–24 Jul 2021.

Xu, W. and Stein, M. L. Maximum likelihood estimation for a smooth gaussian random field model. *SIAM/ASA Journal on Uncertainty Quantification*, 5(1):138–175, 2017. doi: 10.1137/15M105358X.

Yan, X., Abdia, Y., Datta, S., Kulasekera, K., Ugiliweneza, B., Boakye, M., and Kong, M. Estimation of average treatment effects among multiple treatment groups by using an ensemble approach. *Statistics in Medicine*, 38, 04 2019. doi: 10.1002/sim.8146.

Yang, S., Imbens, G. W., Cui, Z., Faries, D. E., and Kadziola, Z. Propensity score matching and subclassification in observational studies with multi-level treatments. *Biometrics*, 72(4): 1055–1065, 2016.

Yang, S., Wan, M. P., Chen, W., Ng, B. F., and Dubey, S. Model predictive control with adaptive machine-learning-based model for building energy efficiency and comfort optimization. *Applied Energy*, 271:115147, 2020. ISSN 0306-2619. doi: 10.1016/j.apenergy.2020.115147.

Yao, L., Li, S., Li, Y., Huai, M., Gao, J., and Zhang, A. Representation learning for treatment effect estimation from observational data. In *Advances in Neural Information Processing Systems*, volume 31. PMLR, 2018.

Yao, L., Chu, Z., Li, S., Li, Y., Gao, J., and Zhang, A. A survey on causal inference. *ACM Trans. Knowl. Discov. Data*, 15(5):74:1–74:46, 2021.

Yin, M., Shi, C., Wang, Y., and Blei, D. M. Conformal sensitivity analysis for individual treatment effects. *Journal of the American Statistical Association*, 0:1–30, 2022.

Yiu, S. and Su, L. Covariate association eliminating weights: a unified weighting framework for causal effect estimation. *Biometrika*, 105(3):709–722, 2018.

Yoon, J., Jordon, J., and van der Schaar, M. GANITE: estimation of individualized treatment effects using generative adversarial nets. In *6th International Conference on Learning Representations, ICLR 2018*. OpenReview.net, 2018.

Yuce, B., Li, H., Rezgui, Y., Petri, I., Jayan, B., and Yang, C. Utilizing artificial neural network to predict energy consumption and thermal comfort level: An indoor swimming pool case study. *Energy and Buildings*, 80:45–56, 2014. ISSN 0378-7788. doi: 10.1016/j.enbuild.2014.04.052.

Zanutto, E., Lu, B., and Hornik, R. Using propensity score subclassification for multiple treatment doses to evaluate a national antidrug media campaign. *Journal of Educational and Behavioral Statistics*, 30(1):59–73, 2005. doi: 10.3102/10769986030001059.

Zhang, H. Inconsistent estimation and asymptotically equal interpolations in model-based geostatistics. *Journal of the American Statistical Association*, 99(465):250–261, 2004.

Zhang, H. and Wang, Y. Kriging and cross-validation for massive spatial data. *Environmetrics*, 21(3-4):290–304, 2010.

Zhang, H., Zimmerman, J., Nettleton, D., and Nordman, D. J. Random forest prediction intervals. *The American Statistician*, 74(4):392–406, 2020. doi: 10.1080/00031305.2019.1585288.

Zhang, Q., Chien, P., Liu, Q., Xu, L., and Hong, Y. Mixed-input gaussian process emulators for computer experiments with a large number of categorical levels. *Journal of Quality Technology*, 53(4):410–420, 2021. doi: 10.1080/00224065.2020.1778431.

Zhang, X., Tan, S., Koch, P., Lou, Y., Chajewska, U., and Caruana, R. Axiomatic interpretability for multiclass additive models. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 226–234, 2019.

Zhang, Y., Kong, D., and Yang, S. Towards R-learner of conditional average treatment effects with a continuous treatment: T-identification, estimation, and inference. *arXiv preprint arXiv:2208.00872*, 2022.

Zhang, Z., Zhou, J., Cao, W., and Zhang, J. Causal inference with a quantitative exposure. *Statistical methods in medical research*, 25(1):315–335, 2016.

Zhao, J. The lower semicontinuity of optimal solution sets. *Journal of Mathematical Analysis and Applications*, 207(1):240 – 254, 1997.

Zhao, Q., Small, D. S., Ertefaie, A., et al. Selective inference for effect modification via the lasso. *Journal of the Royal Statistical Society Series B*, 84(2):382–413, 2022.

Zhao, S., van Dyk, D. A., and Imai, K. Propensity score-based methods for causal inference in observational studies with non-binary treatments. *Statistical methods in medical research*, 29 (3):709–727, 2020.

Zhao, Y., Li, T., Zhang, X., and Zhang, C. Artificial intelligence-based fault detection and diagnosis methods for building energy systems: Advantages, challenges and the future. *Renewable and Sustainable Energy Reviews*, 109(C):85–101, 2019. doi: 10.1016/j.rser.2019.04.02.

Zhizhou Wang, Vemuri, B. C., Chen, Y., and Mareci, T. H. A constrained variational principle for direct estimation and smoothing of the diffusion tensor field from complex dwi. *IEEE Transactions on Medical Imaging*, 23(8):930–939, Aug 2004.

Zhou, Y. *Adaptive Importance Sampling for Integration*. PhD thesis, Stanford University, 1998.

Zhu, Y., Coffman, D. L., and Ghosh, D. A boosting algorithm for estimating generalized propensity scores with continuous treatments. *Journal of Causal Inference*, 3(1):25–40, 2015. doi: doi:10.1515/jci-2014-0022.

Zimmert, M. and Lechner, M. Nonparametric estimation of causal heterogeneity under high-dimensional confounding. *arXiv preprint arXiv:1908.08779*, 2019.

ECOLE
DOCTORALE
DE MATHEMATIQUES
HADAMARD

**Titre :** Apprentissage statistique et inférence causale pour la production de l'énergie

**Mots clés :** Apprentissage statistique, Inference causale, uncertitudes, effets hétérogènes, optimisation, production

**Résumé :** Les modèles d'apprentissage automatique offrent des solutions efficaces pour répondre aux besoins du domaine énergétique. Les résultats de ces modèles peuvent être contestables. Il est donc nécessaire de quantifier les incertitudes de prédictions et prédire l'effet causal d'un changement ou d'une intervention. Ce travail de recherche développe des approches data-driven pour l'optimisation de la production d'énergie : l'une est prédictive pour améliorer la quantification d'incertitudes du modèle. L'autre est causale pour évaluer l'impact des interventions sur le syst 'eme. Ces approches servent à l'identification des stratégies optimales pour augmenter la production et la prise de décision. L'approche prédictive est basée sur le modèle de processus gaussiens et la méthode de validation croisée pour calibrer les intervalles de prédiction. L'approche causale est basée sur des cadres statistiques et estime les effets hétérogènes de l'intervention pour des variables discrètes et continues.

**Title :** Statistical learning and causal inference for energy production

**Keywords :** Statistical learning, Causal Inference, uncertainty, heterogeneous effects, optimization, production

**Abstract :** The energy domain is growing rapidly to meet the needs of the economy. Machine learning models can support the field in facing challenges in an efficient manner. Sometimes, the results of these models are not always convincing. One needs to make reliable predictions whose uncertainties can be quantified and predicts the causal effect of a change or an intervention. This research work develops data-driven approaches for energy production optimization: one is predictive to improve the uncertainty quantification of the model. The other is causal to evaluate the impact of interventions in the system. Such approaches serve to find the optimal strategies to increase production and for decision-making. The predictive approach uses the Gaussian Process model and the cross-validation method to calibrate of prediction intervals. The causal approach is based on statistical frameworks to estimate the heterogeneous effects of intervention for discrete and continuous variables.

**Institut Polytechnique de Paris**
91120 Palaiseau, France