Draft: Please Do Not Quote or Cite

Machine Learning for Causal Inference

Proposal for Cambridge University Press

April 2026

Dr. Melvyn Weeks
Faculty of Economics and Clare College
University of Cambridge

Contents

1 Proposal Narrative: Machine Learning for Causal Inference

This proposal addresses the key criteria that Cambridge University Press considers when evaluating academic book projects. These include: why a comprehensive treatment of machine learning for causal inference is needed now, what unique conceptual framework and themes the book will develop, market positioning and intended audience, pedagogical value, and practical feasibility.

The proposal details the book's pedagogical foundations, drawing on course materials developed and tested at the Faculty of Economics, University of Cambridge. Critical to the evaluation is the book's competitive positioning - how it complements rather than competes with established texts while filling a genuine gap in the literature.

The following sections address each of these considerations systematically.

The Imperative for This Book

The modern researcher faces an unprecedented challenge in conducting empirical analysis: traditional econometric methods, while theoretically sound, often struggle with high-dimensional data and complex heterogeneity patterns that characterise contemporary datasets. Simultaneously, machine learning approaches, though powerful in prediction, lack the architecture to facilitate causal estimation and inference, essential for policy and decision-making. This fundamental tension - what Leo Breiman termed the divide between two statistical cultures - demands a synthesis that this book provides.

The book's central innovation lies in its presentation of modern causal machine learning methods within a coherent economic framework. Core themes include the reconciliation of prediction and causation, the treatment of high-dimensional nuisance parameter estimation, and the development of cross-fitting and sample-splitting procedures that enable valid statistical inference with flexible machine learning algorithms. The book is also centred on a number of key points of departure that span the conditional expectation function, Frisch-Waugh-Lovell Thereom, and machine learning for prediction.

... cross-fitting and sample-splitting procedures Need more of this in proposal, lst Chapter

The book's methodological contribution centers on double/debiased machine learning techniques that allow researchers to combine high-dimensional controls, nonparametric functional f

methodological contribution centers on double/debiased machine learning techniques ... Vague

The tension between parametric and nonparametric approaches reflects deeper disciplinary differences in priorities. Econometricians have traditionally relied on parametric models with explicit functional form assumptions because they yield interpretable parameters and structural understanding of economic relationships. Machine learning, by contrast, is fundamentally algorithmic and nonparametric, treating the data mechanism as unknown while prioritising external validity and robust generalization to unseen data. Modern causal machine learning faces the challenge of achieving both structural understanding and robust generalisation.

prioritizing external validity? See Athey slides

Causal inference — whether parametric or nonparametric - asks fundamentally different questions: what happens to an outcome Y when an element of X changes? A model that accurately predicts ice cream sales using historical trends tells us nothing about how sales would respond to a price increase. For causal inference, we need variation that isolates the causal mechanism of interest, ensuring internal validity where observed associations reflect genuine causal relationships. Causal machine learning methods also ensure external validity through techniques like sample splitting in Double ML and causal forests, which help causal estimates generalise beyond the specific sample used for estimation.

Causal machine learning methods also ensure external validity ... Also ... or external validity follows from ML focus ... ?

1.1 Market Position and Competition

This book targets three overlapping audiences:

Graduate Students in Economics: The book can serve as the primary text for a graduate course on machine learning for economists or as a supplementary text for courses in econometrics or applied microeconomics. For this audience, we emphasize the connection between ML methods and familiar

econometric concepts, using the Conditional Expectation Function and Frisch-Waugh-Lovell theorem as bridges. The material in the book forms the lecture material for a core course on *Machine Learning for Causal Inference*, part of the MPhil degree in *Economics and Data Science*.

See blurb for new Econ and Data Science course

Data Scientists: Practitioners trained in ML who work on economic or business problems will learn how to move beyond prediction to causal inference. For this audience, we emphasise the fundamental differences between prediction and causal estimation, building intuition for identification strategies.

Applied Researchers: Economists and social scientists conducting empirical research will find practical guidance on incorporating ML methods while maintaining causal rigor. The pedagogical approach emphasises when and how to apply specific methods, with clear guidance on interpretation and inference.

For practicing researchers, it offers a reference guide with practical implementation guidance. For policymakers and practitioners, it bridges the gap between theoretical developments and applied work. This multi-audience appeal, combined with the book's grounding in classroom-tested material, positions it to become a standard reference in this rapidly growing field.

Code examples in R and Python will be integrated throughout, with accompanying online resources providing datasets and replication materials. This multi-modal approach recognizes that modern empirical work requires both theoretical understanding and practical implementation skills.

Competition

Several excellent books address portions of this space:

- *The Elements of Statistical Learning* (Hastie, Tibshirani, Friedman)
- *Mostly Harmless Econometrics* (Angrist, Pischke)
- *Applied Causal Inference Powered by ML and AI* (Chernozhukov et al.)

While computer science texts like Hastie, Tibshirani, and Friedman's *Elements of Statistical Learning* provide comprehensive coverage of machine learning algorithms, they lack the causal perspective central to economic analysis. Conversely, established econometrics texts such as Angrist and Pischke's *Mostly Harmless Econometrics* excel in causal inference but predate the machine learning revolution. Recent developments like Athey and Imbens' influential papers provide excellent conceptual overviews but lack the pedagogical structure and comprehensive coverage needed for systematic learning.

The closest complement to this work is Chernozhukov et al.'s excellent online resource *Applied Causal Inference Powered by* ML *and* AI, which provides comprehensive technical coverage and cutting-edge methodological developments. This book differs by offering a structured textbook format with systematic chapter progression, extensive exercises, and pedagogical structure built around the conditional expectation function and Frisch-Waugh-Lovell theorem as unifying frameworks. Where Chernozhukov et al. excels in breadth and methodological innovation, this book emphasizes depth in foundational concepts and classroom-ready organisation for systematic learning.

classroom-ready organisation .. ??

The book fills this space by maintaining economists' emphasis on identification and structural interpretation while fully embracing machine learning's flexibility and power. Unlike purely technical treatments, each method is motivated by economic questions and illustrated through applications that demonstrate how these tools enhance our ability to answer substantive research questions. This approach ensures that readers understand not just how these methods work, but when and why to apply them in economic contexts.

1.2 Pedagogical Foundations

The pedagogical landscape for empirical economics confronts a distinctive challenge: how to teach students to harness the computational power and flexibility of machine learning while maintaining the causal rigor that defines economic inquiry. Traditional econometric pedagogy, with its emphasis on identification strategies and structural parameters, must now incorporate algorithmic methods originally

designed for prediction tasks - yet do so in ways that preserve and enhance our ability to draw causal conclusions. Students must learn not to choose between parametric interpretation and nonparametric flexibility, but rather to synthesise these approaches through partial linear model ... causal forests.

but rather to synthesise these approaches through partial linear model .. ??

The volume addresses this need by developing a unified framework that bridges algorithmic prediction with causal identification. Rather than treating machine learning as merely a set of new estimation techniques, our approach reconceptualises how we teach empirical methods when traditional assumptions about data generating processes no longer hold. This perspective shift is essential as we navigate an era where the complexity of available data often exceeds our traditional analytical capabilities.

Each chapter follows a consistent structure designed to facilitate learning:

- **Theory**: Core concepts are developed rigorously but accessibly, with technical details in starred sections
- **Implementation**: Practical guidance on applying methods, including code snippets and computational tips
- **Application**: Detailed empirical application using real data
- **Exercises**: Problems ranging from theoretical derivations to empirical exercises with provided datasets

Code examples in R and Python will be integrated throughout, with accompanying online resources providing datasets and replication materials.

with accompanying online resources??

The book provides multiple Points of Departure - entry points for readers with different backgrounds—alongside a smaller set of Unifying Principles that recur throughout. Points of Departure determine where a reader begins; Unifying Principles determine what they will repeatedly encounter as they progress. A reader entering through the CEF will find the same orthogonalisation principle surfacing in double ML, causal forests, and heterogeneous treatment effects. A reader entering through high-dimensional methods will encounter the same insight that prediction serves causation. The Principles provide coherence; the Points of Departure provide accessibility.

The points of departure are:

- ➢ Conditional Expectation Function
- ➢ Parametric versus Nonparametric Methods
- ➢ High Dimensional Methods in Statistics
- ➢ Causal Inference and Treatment Effects
- ➢ Machine Learning Methods for Prediction

Further details on these points of departure are given in Section 4.

Remark. ***Note to reviewers:*** *The material in Sections 2, 4 of this proposal, covering unifying principles and foundational departures, will be systematically developed in Chapter 2 of the book, providing the theoretical framework and key concepts that underpin the technical methods presented in subsequent chapters.*

2 Unifying Principles

The book develops a number of interconnected unifying principles that establish the conceptual foundation for causal machine learning. In Section 2.1 .... The Frisch-Waugh-Lovell (FWL) theorem and its nonparametric extensions unify classical econometrics with modern causal ML. As we will demonstrate complex moment conditions reduce to familiar formulas after residualisation. For readers trained in econometrics, FWL theorem provides familiar ground from which to approach new material.

But CEF also embodies a deeper insight—that partialling out confounding variation enables identification—which generalises far beyond the linear case. This orthogonalisation principle recurs throughout the book: in the partial linear model, in double machine learning, in instrumental variable extensions, and in heterogeneous treatment effect estimation.

In Section 5 we explore the *Scalar to Heterogeneous Effects* theme, examining the transition from estimating single average treatment effects to discovering patterns of effect variation across populations. This theme examines challenges that arise when moving from scalar to individual-level estimation, including the curse of dimensionality and the increasingly stringent overlap requirements needed for credible heterogeneous effect identification.

In Section 2.4 we explore the theme *Value Added: Where Machine Learning Enhances Causal Inference*, examining the specific domains where ML methods provide improvements to traditional econometric approaches - from high-dimensional control selection to discovering heterogeneous treatment effects. We also acknowledge the fundamental limitations that ML cannot overcome, particularly the inability to create quasi-experimental variation or resolve identification failures.

Finally, we address the central challenge of *Re-optimizing Machine Learning for Causal Inference* (Section 5.3), examining how methods originally designed for prediction—with their emphasis on regularisation and cross-validation, must be re-engineered through techniques such as orthogonalisation, double debiasing, and sample splitting to facilitate causal identification and valid statistical inference.

2.1 Regression and the Frisch-Waugh-Lovell Theorem

The tension between *parametric* and *nonparametric* approaches shapes our approach to statistical modeling, creating what Breiman (2001) characterised as two distinct cultures in statistical practice.

> *There are two cultures in the use of statistical modeling to reach conclusions from data. One assumes that the data are generated by a given stochastic data model. The other uses algorithmic models and treats the data mechanism as unknown.*
>
> Breiman [2001], p199.

In a world where the analyst cares about a specific causal parameter, the Frisch-Waugh-Lovell (FWL) theorem provides a fundamental insight: we can transform a multivariate regression problem into a bivariate one through the process of "partialing out."

Consider the causal model:

$$Y = \alpha + \tau d + \beta X + \varepsilon,$$

where d represents our treatment of interest, X represents confounding variables, and τ is our target causal parameter. Internal validity - establishing that observed associations reflect genuine causal relationships within our sample—requires correctly identifying and controlling for all relevant confounders. The FWL theorem reveals that $\widehat{\tau}$ can be obtained by first removing (partialing out) the linear influence of X from both Y and d, then regressing the residualised Y on the residualised d. This two-step procedure yields exactly the same estimate as the full multivariate regression, demonstrating that causal estimation fundamentally involves isolating variation in treatment that is orthogonal to confounders.

The key insight here is that residualisation creates orthogonality. When we partial out X from d, we obtain residuals $\tilde{d}$ that represent the component of treatment variation that cannot be explained by the confounders; mathematically, these residuals are uncorrelated with (orthogonal to) X. Similarly, the residualised outcome $\tilde{Y}$ captures variation in Y that is orthogonal to X.[1] By working with $\tilde{d}$ and $\tilde{Y}$, we isolate precisely the exogenous variation in treatment that identifies the causal effect, stripped of any confounding correlations. The two-step procedure yields exactly the same estimate as the full multivariate regression, demonstrating that causal estimation fundamentally involves isolating variation in treatment that is orthogonal to confounders.

The FWL theorem provides a systematic solution: by decomposing the problem into separate prediction tasks for estimating $E[Y|X]$ and $E[d|X]$, we can harness machine learning's flexibility to capture complex confounding relationships without sacrificing causal identification. The orthogonalisation ensures that

[1] In geometric terms, we are projecting both d and Y onto the space perpendicular to the span of X, ensuring that the resulting variables contain only the variation that is independent of confounders.

errors in estimating these nuisance functions—inevitable when using regularized methods like LASSO - do not contaminate our causal estimates. This addresses internal validity by maintaining unbiased estimation of τ even when our confounding controls are imperfectly specified.

In high-dimensional settings where X contains hundreds of potential confounders, traditional parametric methods risk either omitted variable bias (from excluding relevant controls) or overfitting bias (from including too many controls relative to sample size). Make distinction between automatic use of FWL through algebra od least squares, and when FWL needs to be manually applied as with Double Lasso, for example ..

2.2 The Prediction Versus Causation Overlay

The parametric/nonparametric distinction explored in Section **??**, does not fully capture the methodological complexity facing modern empirical analysts. Cutting across both statistical cultures lies an equally fundamental choice of purpose: Are we trying to predict what will happen, or understand what would happen if we intervened? This prediction versus causation divide creates a second methodological dimension that overlays the two cultures framework. Prediction asks: Given features X, what is our best estimate for outcome Y? This question drives much of machine learning, where the goal is to minimize expected prediction error, often measured by out-of-sample performance. A predictive model might use temperature, day of the week, and historical sales patterns to forecast ice cream sales—all legitimate if they improve predictive accuracy.

Causation asks fundamentally different questions: What happens to Y if we intervene to change an element of X? A model that accurately predicts ice cream sales using historical trends tells us nothing about how sales would respond to a price increase. For causal inference, we need variation that isolates the causal mechanism of interest, whether through experimental design or quasi-experimental identification strategies.

When we combine these two dimensions - parametric/nonparametric from the two cultures, and prediction/causation - we obtain four distinct methodological approaches that define the landscape of modern empirical practice. This 2×2 framework, presented in Table 1, reveals both the established foundations of existing methods and the emerging frontiers where synthesis is most needed.

This prediction versus causation distinction creates a crucial overlay on our methodological frameworks, as in Table 1

TABLE 1
Four methodological approaches: Parametric/Nonparametric and Prediction/Causation dimensions

	Prediction	**Causation**
Parametric Methods	Traditional econometric forecasting models with known functional forms	Instrumental variables, difference-in-differences, regression discontinuity
Nonparametric Methods	Standard machine learning optimizing for out-of-sample performance	Double LASSO, double/debiased ML, causal forests, and other methods adapting algorithmic flexibility for causal questions

2.3 Re-optimizing Machine Learning for Causal Inference

The four methodological approaches outlined in Table 1 - parametric/nonparametric crossed with prediction/causation - demonstrate the evolving landscape of modern empirical practice. While we have mature methods for parametric causal models (IV, DID, RDD) and algorithmic prediction (standard ML), the nonparametric causation quadrant represents the emerging synthesis. This quadrant embodies the central methodological challenge: how to harness nonparametric flexibility for causal identification without sacrificing the rigor of statistical inference.

The fundamental challenge lies in the architectural mismatch between prediction and causation objectives. Methods optimized for prediction—minimising expected loss through regularization and cross-validation—deliberately introduce bias to reduce variance and improve out-of-sample

performance. Causal inference, by contrast, requires orthogonal moment conditions that yield unbiased estimation of specific structural parameters with valid confidence intervals. This tension necessitates a complete re-engineering: developing estimators that leverage ML's flexibility for nuisance parameter estimation while maintaining the orthogonality conditions essential for causal identification.

This re-optimisation involves several key innovations that form the backbone of this book:

- **Orthogonalisation**: Making causal estimators locally insensitive to first-order errors in nuisance parameter estimation
- **Double debiasing**: Removing the bias introduced by regularisation while maintaining dimension reduction benefits
- **Sample splitting**: Separating model selection from parameter estimation to preserve valid inference
- **Flexible functional forms**: Moving from the linear specification $y = \alpha + \tau d + X'\beta + \varepsilon$ to partial linear models $y = g(X) + \tau d + \varepsilon$, where nuisance functions are estimated nonparametrically while preserving the interpretability of the scalar treatment parameter
- **Heterogeneous treatment effects**: Extending from constant treatment effects to covariate-specific effects $\tau(X)$, yielding the fully nonparametric specification $y = \alpha(X) + \tau(X)d + \varepsilon$, enabling discovery of treatment effect heterogeneity across subpopulations

2.4 Value Added: Where Machine Learning Enhances Causal Inference

A critical question facing empirical economists is not whether to use machine learning methods, but when and how these methods add substantive value to causal analysis. The answer depends critically on features of the data-generating process—the degree of sparsity, the extent of multicollinearity, sample size relative to dimensionality, and the presence of meaningful nonlinearities—many of which are difficult to verify in practice. Baiardi and Naghi (2024), in a systematic re-analysis of published causal studies, provide an empirical audit of where causal ML delivers genuine gains and where it does not. Their findings underscore that ML methods are most productively understood as a complement to well-specified parametric benchmarks, capable of validating, sharpening, or in specific circumstances challenging their conclusions, rather than as a wholesale replacement for traditional identification strategies.

Machine learning methods provide substantial improvements to causal inference in several well-defined areas. Perhaps the most reliable of these is their role in **specification sensitivity analysis**.
Post-double-selection and related methods are most consistently valuable not as primary identification strategies but as diagnostic tools for assessing whether conclusions from a well-specified parametric benchmark are robust to alternative approaches to control selection. The key diagnostic question is not whether the selected set of controls changes across implementation variants—it often does—but whether the causal parameter estimate moves with it. Stability of $\hat{\tau}$ across variants, including double selection versus outcome-only selection and alternative penalty representations such as the plug-in rigorous penalty (`rlasso`) and cross-validation minimising mean squared prediction error, provides positive evidence that the underlying causal claim is robust to specification choices. Co-movement of the control set and the point estimate, by contrast, signals residual confounding that the parametric benchmark may not have resolved. This diagnostic use of ML is more reliable than its use as a primary estimator precisely because it leverages the familiarity and transparency of the parametric benchmark while using ML to probe its boundaries.

The challenge of high-dimensional control selection represents another domain where machine learning delivers clear benefits. When the set of potential confounders is large relative to sample size, traditional methods struggle to balance the competing demands of controlling for confounding while maintaining statistical power. Double LASSO addresses this challenge by selecting controls through consideration of both the outcome and treatment equations, ensuring that important confounders are retained even when their direct predictive effects appear modest. This dual selection process addresses a fundamental weakness of single-equation methods that may inadvertently exclude variables crucial for identification, a risk that becomes particularly acute when confounders have strong effects on treatment but weak direct effects on outcomes.

The distinction between estimation and identification remains fundamental yet often blurred in applied work. Machine learning methods function primarily as tools for estimation—they help us better approximate unknown functions and manage high-dimensional data. However, they do not address identification, which concerns whether the variation we analyze has a causal interpretation. The absence of an identifiable causal effect in the data structure represents a fundamental problem that cannot be overcome through more sophisticated estimation techniques, regardless of their flexibility or predictive accuracy.

Moreover, while machine learning can enhance the efficiency of control selection, it cannot resolve selection bias in the absence of valid conditional independence assumptions. The flexibility of these methods may even exacerbate identification problems if researchers mistake predictive accuracy for causal validity, or if algorithms inadvertently learn and amplify the very confounding patterns that compromise identification. This risk is particularly acute with methods that excel at finding complex patterns in data, as they may discover spurious relationships that superficially appear to support causal interpretations.

Machine learning also substantially enhances instrumental variable estimation through two distinct channels. First, when researchers face many potential instruments, LASSO-based selection provides a principled approach that improves upon traditional 2SLS by mitigating the bias arising from overfitting in finite samples. The regularization inherent in these methods prevents the mechanical fitting of noise that plagues conventional many - instrument approaches. Second, and perhaps more fundamentally, the first stage of IV estimation is essentially a prediction problem - we seek to predict the endogenous variable using exogenous instruments. Here, flexible ML methods can capture nonlinear relationships between instruments and endogenous variables that linear projections would miss entirely, potentially transforming weak instruments into strong ones and enabling credible identification where linear methods fail.

Perhaps the most transformative contribution of machine learning to causal inference lies in its capacity to discover heterogeneous treatment effects. Traditional econometric methods typically estimate average effects, possibly with simple interactions to capture basic heterogeneity. In contrast, methods like causal forests and generalized random forests can reveal complex patterns of effect variation across the covariate space without requiring researchers to pre-specify interaction terms. This capability fundamentally transforms our understanding of policy interventions, moving from questions about whether something works on average to understanding for whom and under what conditions interventions are most effective—insights that are crucial for policy design and targeting.

Even in cases where machine learning methods do not alter fundamental causal conclusions, they often deliver meaningful efficiency improvements. By better approximating complex nuisance functions and deriving optimal weighting schemes, these methods extract more information from available data, yielding tighter confidence intervals and more precise estimates. This efficiency gain proves particularly valuable in settings with limited sample sizes or expensive data collection, where every observation must be leveraged to its fullest potential.

The Costs of Flexibility: Method-Specific Risks

The benefits of causal ML methods are real but conditional. A balanced assessment requires acknowledging three method-specific risks that can silently undermine causal validity in ways that are difficult to detect after the fact.

The Detectability Threshold. Post-double-selection LASSO can generate omitted variable bias when relevant controls have coefficients too small to reliably trigger selection yet too large for their omission to be innocuous. This detectability problem arises because the penalty parameter must be calibrated to control false discovery rates at the selection stage, but this calibration is agnostic to the magnitude of confounding induced by excluded controls. In finite samples, controls may fall below the selection threshold due to sampling variation rather than genuine irrelevance, and the severity of the resulting bias depends on sample size, correlation structure, and penalty choice in ways that theory alone cannot anticipate. The risk is greatest precisely in the settings—moderately sized samples, dense but weakly confounding covariate spaces—where LASSO is most commonly applied.

Multicollinearity and Selection Instability. LASSO's selection consistency relies on restricted eigenvalue conditions that break down under high multicollinearity, while its implicit sparsity assumption fails when many controls contribute small but non-negligible confounding effects. Under these

conditions—common in economic datasets with many correlated demographic and regional controls—the selected control set can vary substantially across bootstrap samples or minor specification changes, even when the underlying causal parameter is stable. Regularisation-based selection may then deliver unreliable causal estimates not because the method is wrong in principle but because the structural conditions on the covariate matrix required by its theoretical guarantees are not satisfied in the specific application.

Instrument Contamination in IV Settings. When ML methods are used to select instruments by maximising first-stage fit, they risk selecting on the first-stage error, inducing correlation between the selected instrument and the structural error that invalidates the exclusion restriction. Sample splitting provides a partial remedy but does not resolve the more fundamental problem that automated instrument selection based on predictive power conflates relevance with validity. In IV settings, the diagnostic use of ML—checking whether a pre-specified instrument remains relevant conditional on flexible nonlinear controls—is considerably safer than automated selection from a candidate pool.

The Boundaries of Machine Learning's Contribution

Understanding where machine learning cannot add value is equally crucial, as these boundaries reflect fundamental principles of causal identification that no amount of algorithmic sophistication can transcend. As Angrist et al (2018) emphasise, "ML methods do not create quasi-experimental variation." This simple statement encapsulates a profound limitation: no algorithm, however sophisticated, can transform correlation into causation. The causal interpretation of any ML estimate ultimately rests on maintained assumptions about independence, exclusion restrictions, or other identifying conditions. When these assumptions fail, machine learning offers no remedy—it cannot generate the exogenous variation necessary for credible causal inference.

This limitation becomes particularly clear when considering well-identified zero effects from careful parametric analyses. If traditional methods with appropriate controls for selection and confounding yield null results, should we expect the use of ML to overturn these findings. The Dale and Krueger analysis of elite college effects provides an illuminating example: their post-LASSO estimates closely mirror the original estimates, confirming rather than contradicting the absence of causal effects. Machine learning serves here as a valuable robustness check, demonstrating that the zero effect is not an artifact of functional form assumptions or variable selection, but it cannot manufacture effects where none exist in the identified variation.

This value-added framework implies a disciplined approach to applying machine learning in causal inference that begins with clear identification strategies grounded in economic theory and institutional knowledge. Machine learning methods then serve to implement these strategies more effectively rather than replacing the fundamental need for identification. The greatest returns to these methods arise in settings characterized by rich covariate information, complex functional forms, and potential effect heterogeneity, while in simpler settings with clear identification and limited covariates, traditional methods may suffice and even be preferable for their transparency.

3 Causal Machine Learning

??

Section needs Integrating /extending

Modern empirical analysis demands the flexibility of machine learning methods to estimate complex conditional expectation functions, yet these methods must be carefully re-engineered to preserve causal validity. LASSO excels at prediction through coefficient shrinkage and automatic variable selection, yet this same regularisation introduces bias that contaminates causal estimates. The solution embodies a deep principle: separating prediction tasks from causal estimation through systematic application of the Frisch-Waugh-Lovell theorem in high-dimensional settings.

The evolution from LASSO to Double LASSO and Double Machine Learning exemplifies this fundamental re-optimisation challenge.

Double Machine Learning implements this separation via a two-stage process. First, machine learning methods estimate nuisance functions $E[Y|X]$ and $E[d|X]$ - modeling how confounders relate to outcomes and treatments through flexible algorithms like random forests or neural networks. Second, the FWL

principle orthogonalises these relationships, partialling out confounding variation before estimating causal effects. By decomposing causal problems into prediction sub-problems, namely , we can deploy any machine learning method within this framework. This decomposition allows us to harness ML's predictive flexibility while preserving the orthogonality conditions essential for valid causal inference.

The methods we have developed for estimating scalar treatment effects - from the FWL theorem through double machine learning – provide the conceptual foundation for a more ambitious goal: understanding how treatment effects vary across the population. While the average treatment effect τ provides valuable summary information, it masks potentially rich heterogeneity that is crucial for both economic understanding and policy design. A job training program might substantially boost earnings for workers with certain skill profiles while having minimal impact on others. A monetary policy intervention might stimulate different sectors to varying degrees. The scalar effect tells us something important, but not everything we need to know.

4 Points of Departure

The book is based on a number of points of departure that serve as *scaffolding*, allowing readers with different backgrounds to find their entry point into Causal ML:

- *Economists* can start with familiar CEF and FWL concepts and build toward ML flexibility
- *Statisticians* can begin with high-dimensional methods and add causal structure utilising the Frisch Waugh Lovell Theorem ....
- *Machine Learning practitioners* can start with prediction methods and learn causal adaptations
- *Policy researchers* can focus on treatment effects and see how ML enhances traditional methods

This pedagogical approach is particularly valuable because it prevents the common mistake of treating machine learning as a "black box" solution to causal problems, instead emphasizing that successful causal machine learning requires deep understanding of both the prediction capabilities of ML and the identification requirements of causal inference.

The points of departure that follow - the CEF, high-dimensional methods, the FWL theorem, machine learning for prediction and treatment effects - are not mutually exclusive categories but rather overlapping perspectives on the same fundamental challenge: how to combine the flexibility of modern algorithmic methods with the rigor of causal identification. Each provides a different lens through which to view this synthesis, and readers may find value in exploring multiple entry points ..

4.1 The Conditional Expectation Function

At the heart of both prediction and causal inference lies a fundamental mathematical object: the Conditional Expectation Function (CEF). The CEF, denoted $E[Y|X]$, represents the expectation about the outcome Y given knowledge of covariates X. Under the assumption that $E[Y|X] = X'\beta$, the conditional expectation function coincides with the best linear predictor obtained through OLS estimation. This mathematical equivalence forms the foundation that connects statistical prediction and econometric inference in the linear framework.

While the linear specification provides the best linear approximation to $E[Y|X]$, "best linear" does not mean "best overall." The true CEF may be fundamentally nonlinear, such that constraining ourselves to linear approximations may miss important patterns in the data that are essential for both accurate prediction and causal inference.

Understanding this limitation requires distinguishing between two related issues. First, while the BLP minimises prediction error among all linear functions, the CEF itself minimizes prediction error among all possible functions. The BLP represents the best we can do under a linearity constraint, not the best we can do overall. Second, when the true CEF exhibits substantial nonlinearity, even the optimal linear approximation may provide a poor fit to the underlying relationship. In such cases, the approximation error from imposing linearity - the gap between the true CEF and its best linear approximation - may obscure economically important patterns, compromises predictive accuracy, and potentially biases our causal estimates.

Consider point of functional form issues impacting bias and std error. The latter follows, check the first - bias. Reduce length of above?

This recognition that linear approximations may be inadequate leads to a fundamental choice in how we estimate the CEF.

4.2 Parametric versus Nonparametric Methods

The choice between parametric and nonparametric methods represents a fundamental decision about how we approach estimating the CEF. Parametric methods assume a specific functional form - typically $E[Y|X] = X\beta$ - offering interpretability and established tools for statistical inference. However, this approach requires us to know, or at least correctly specify, the functional relationship *a priori*.

.. his approach requires us to know, or at least correctly specify, ... Unclear

Nonparametric methods take a different approach by allowing the data to reveal the shape of the CEF. Consider the covariate space $\mathcal{X}$ - the set of all possible values our features can take. Rather than imposing a global linear structure across $\mathcal{X}$, nonparametric methods partition this space into local regions and estimate conditional expectations within each region. This can be as simple as computing average outcomes within regions defined by covariate values, or as sophisticated as using kernels, trees, or neural networks to approximate $E[Y|X] = f(X)$ without pre-specifying the form of $f(\cdot)$.

Rewrite the above parag ...

This flexibility comes at a cost that manifests differently in predictive and causal settings. In pure prediction problems, in accommodating complex patterns we face the classic curse of dimensionality - the number of observations required to maintain a given level of estimation precision grows exponentially with the dimension of $\mathcal{X}$. In causal inference, we confront an additional and often more binding constraint: the overlap or common support condition. As the dimension of the covariate space $\mathcal{X}$ increases, it becomes increasingly difficult to find comparable units across treatment conditions. Even with abundant data, we may lack sufficient observations in specific regions of the covariate space to credibly estimate treatment effects.

This sparsity problem means that while nonparametric methods can flexibly capture the CEF's shape where data exists, they may leave us unable to make causal statements precisely where we need them most - in regions of $\mathcal{X}$ where treated and control units rarely overlap. The tension between flexibility and interpretability thus becomes especially acute in causal inference, where we need not just accurate predictions but also valid counterfactual comparisons across the entire support of the covariate distribution.

Make distinction between data sparsity, as above, and the sparsity assumption which underpins LASSO ....

4.3 High Dimensional Methods in Statistics

The question of why we need to move beyond simple linear models becomes even more pressing in our modern era of 'big data'. When the dimension of our covariate vector X is large - potentially larger than sample size n - traditional methods break down entirely. This high-dimensional setting arises naturally in two ways: either we observe many covariates directly (hundreds of demographic variables, thousands of product features), or we wish to consider many transformations and interactions of a smaller set of base covariates to capture nonlinearities.

High-dimensional methods like the LASSO address this challenge by adding penalty terms to the estimation criterion, trading off model complexity against fit. This form of regularisation performs automatic variable selection, identifying which features are most predictive while avoiding overfitting. However, the regularisation that makes these methods work for prediction introduces bias that can invalidate causal inference. This problem motivates the development of "double-debiased" machine learning methods that separate the tasks of model selection and parameter estimation.

4.4 Causal Inference and Treatment Effects

The high-dimensional challenges we have outlined become even more complex when we move from prediction to the central concern of economics: causal inference. Machine learning excels at prediction for

an outcome Y given features X. But economics typically asks causal questions: what happens to Y if an element of X, say d changes? This distinction is fundamental because while predictive models can exploit all available correlations in the high-dimensional space $\mathcal{X}$, causal inference requires us to separate spurious correlations from genuine causal relationships.

The potential outcomes framework formalizes this causal reasoning. A given individual i faced with a binary treatment d faces two potential outcomes: Y_{0i} if untreated and Y_{1i} if treated. The fundamental problem of causal inference is that we observe only one of these outcomes, $Y_i = Y_{0i} + d_i(Y_{1i} - Y_{0i})$, while the counterfactual remains forever unobserved. This missing data problem lies at the heart of all causal analysis: how can we learn about the causal effect $Y_{1i} - Y_{0i}$ when we never observe both potential outcomes for the same unit?

In the high-dimensional settings we have been considering, identification of causal effects can proceed through different strategies, each placing different demands on our ability to estimate functions over the covariate space $\mathcal{X}$. Inverse propensity weighting requires an estimate of propensity score $e(x_i) = \Pr(d_i = 1|X_i = x)$, a challenging prediction problem when $\mathcal{X}$ is high-dimensional. Alternatively, outcome regression approaches require modeling the conditional outcome functions $\mu_0(x) = E[Y_{0i}|X_i = x]$ and $\mu_1(x) = E[Y_{1i}|X_i = x]$, precisely the conditional expectation functions we have been discussing, but now separately for treated and control potential outcomes.

Doubly robust estimators, which have become increasingly important in high-dimensional settings, combine both approaches by using the propensity score and the outcome functions together. This dual approach provides insurance against misspecification: the estimator remains consistent if either the propensity score or the outcome models are correctly specified (though not necessarily both). Under unconfoundedness – the assumption that treatment is as good as randomly assigned conditional on X - any of these approaches can identify causal effects from observational data.

While machine learning methods offer powerful tools for estimating these high-dimensional nuisance functions, their implementation introduces new challenges for causal inference. The very flexibility that makes machine learning attractive for prediction -through regularisation, cross-validation, and model selection - can introduce bias that invalidates standard statistical inference procedures. This regularisation bias necessitates specialised double-debiased approaches that carefully separate the prediction problem (estimating nuisance functions) from the causal estimation problem, ensuring that the bias introduced by machine learning estimation does not contaminate our causal conclusions.

Some duplication around this point ..

Relocate parag below

Traditional methods often assume constant treatment effects, while reality reveals heterogeneous responses that vary smoothly as an unknown function of covariates. That is, we seek to estimate $\tau(x) = \mu_1(x) - \mu_0(x)$ as a function over $\mathcal{X}$. Does job training help all workers equally, or are gains concentrated among certain groups? As we search for this heterogeneity in the high-dimensional space, we encounter precisely the overlap problem discussed earlier: we must locate sufficiently rich sets of comparable treated and control units throughout $\mathcal{X}$ -a requirement that becomes increasingly difficult to satisfy as dimensionality grows.

4.5 Machine Learning Methods for Prediction

In a linear setting regularisation methods like LASSO, ridge regression, and Elastic Net, handle settings where $p > n$ by trading bias for reduced variance through automatic variable selection and coefficient shrinkage, prioritising out-of-sample predictive performance over unbiased estimation of specific parameters. Nonparametric methods like random forests and neural networks approximate complex functional relationships $E[Y|X] = f(X)$ without imposing linearity constraints, using ensemble averaging and hierarchical representations to manage the bias-variance tradeoff inherent in flexible function approximation.

The success of these methods rests on techniques that are fundamentally oriented toward prediction. Cross-validation selects tuning parameters by maximising held-out sample performance, ensuring that models generalise beyond their training data. Regularisation paths systematically vary penalty parameters from $\lambda = 0$ (yielding unpenalized OLS estimates) to $\lambda = \infty$ (forcing all coefficients to zero), tracing how coefficient estimates shrink as penalty strength increases. For LASSO, this path reveals the sequence in which variables enter or exit the active set, providing a complete characterization of the

model selection process across all possible penalty levels. Cross-validation then selects the penalty parameter λ^* that minimizes out-of-sample prediction error, typically using k-fold CV to estimate $E[(Y - \hat{Y}(\lambda))^2]$ for each λ along the path.

Ensemble methods combine multiple models to reduce prediction error, averaging away individual model uncertainties. These techniques have proven remarkably successful at estimating CEFs in settings ranging from image recognition to natural language processing to economic forecasting. However, the very features that make these methods powerful for prediction — aggressive regularisation, complex model averaging, and optimisation for predictive accuracy — create fundamental challenges when our goal shifts from predicting outcomes to estimating causal effects.

The Frisch-Waugh-Lovell Theorem: Nonparametric Case

This is where the FWL theorem reveals its modern relevance. When we apply regularisation methods like LASSO to select which variables from a high-dimensional X to include, we cannot simply run a penalised regression of Y on (d, X) and read off the coefficient on d; the regularisation bias contaminates our causal estimate. The automatic isolation of variation that makes FWL work in the linear case breaks down under regularisation because the penalty term does not respect orthogonality required for causal identification; it shrinks all coefficients toward zero based on predictive considerations rather than preserving the specific variation needed to identify τ. What we can do is manually implement the FWL logic: first use regularisation to select the relationship between X and Y, then between X and d, and finally estimate τ from the residualised regression.

Consider the partial linear model (PLM):

$$\begin{aligned} Y_i &= d_i\tau + g(X_i) + \varepsilon_i \\ d_i &= m(X_i) + v_i \end{aligned} \tag{1}$$

The partial linear model requires solving for τ in a way that respects the orthogonality conditions essential for causal identification. In this semi-parametric setting, the orthogonality condition is expressed through the moment equation

$$E[(Y_i - d_i\tau - g(X_i))(d_i - m(X_i))] = 0$$

This orthogonal estimating equation embodies the FWL logic: the residualised outcome $(Y_i - d_i\tau - g(X_i))$ must be uncorrelated with the residualised treatment $(d_i - m(X_i))$ for the estimator to be unbiased. Solving this moment condition by setting the sample analog to zero and rearranging for τ reveals:

$$\tau_0 = \frac{E[(Y - g_0(X))(d - m_0(X))]}{E[(d - m_0(X))^2]} \tag{2}$$

Although (2) appears complex - we have an unknown function $g(\cdot)$ capturing how confounders X affect the outcome Y, and another unknown function $m(\cdot)$ describing how X affects treatment d, through orthogonalisation - after partialing out the influence of confounders through the nonparametric functions $g_0(X)$ and $m_0(X)$, the causal parameter emerges from the familiar covariance realtionship. To see define residualised variables: $Y^* = Y - g_0(X)$, the part of Y not explained by confounders X; and $d^* = d - m_0(X)$, the part of treatment d not explained by confounders X. Substituting these into our equation and noting that residuals have zero mean, we can write:

$$\tau_0 = \frac{\text{Cov}(Y^*, d^*)}{\text{Var}(d^*)}$$

The apparent complexity of the partial linear model is recast as the familiar OLS formula we've known all along. The key difference being that after locating estimates of $g_0(X)$ and $m_0(X)$, estimation is based upon residualised variables that have been purged of confounding influences.

This principle extends naturally to instrumental variables settings. When treatment is endogenous, we must partial out confounders from the instrument as well, creating what amounts to a triple residualization problem: $Y^* = Y - g_0(X)$, $d^* = d - m_0(X)$, and $Z^* = Z - r_0(X)$. The classical IV

formula $\text{Cov}(Z, Y)/\text{Cov}(Z, d)$ then applies to these residualised variables, demonstrating how the FWL logic scales to increasingly complex identification strategies while maintaining its fundamental insight about isolating causal variation.

Need to anticipate bias from estimating nuisance functions - edit below

External validity - ensuring our estimates generalise beyond the specific sample - benefits from the same decomposition through different mechanisms. Sample splitting techniques, inherent to double machine learning approaches, protect against overfitting to idiosyncratic patterns in our data. By using separate samples for model selection (choosing which confounders to include) and parameter estimation (calculating treatment effects), we prevent our estimates from being contaminated by spurious correlations that exist only in our training data. This separation helps ensure that identified treatment effects reflect genuine structural relationships rather than sample-specific noise, enhancing the generalisability of causal conclusions to new populations and contexts.

5 From Scalar to Heterogeneous Effects

Traditional econometric approaches estimate a single parameter representing the average treatment effect (ATE), providing information about whether an intervention works on average. Consider the scalar treatment effect model

$$Y_i = \alpha + \tau d_i + \beta' X_i + \varepsilon_i$$

, where τ denotes the impact that may mask substantial variation across individuals or groups.

This scalar approach obscures potential heterogeneity that is crucial for policy targeting, resource allocation, and economic understanding. A job training program might boost earnings dramatically for some workers while having minimal impact on others. Educational interventions may benefit certain student populations while showing no effect for others. Monetary policy might stimulate some sectors while leaving others unchanged. The scalar approach treats these diverse responses as noise around a single mean effect, potentially leading to misguided policy conclusions.

The challenge becomes particularly acute when the distribution of treatment effects is bimodal or highly skewed. An intervention might help 30% of recipients substantially while harming 20% and having no effect on the remaining 50%. The average effect might appear close to zero, leading to the incorrect conclusion that the intervention is ineffective, when in reality it creates both winners and losers who could potentially be identified ex ante.

At the frontier lies fully heterogeneous estimation, where the conditional average treatment effect (CATE) is defined as $\tau(x) = E[Y(1) - Y(0)|X = x]$, transforming our estimating equation to the fully nonparametric specification $Y_i = \alpha(X_i) + \tau(X_i)d_i + \varepsilon_i$. This creates a fundamental tension: the curse of dimensionality makes local estimation increasingly difficult as X grows (??), yet heterogeneous effects precisely require us to estimate complex functions locally rather than globally.

The evolution from scalar to heterogeneous effects thus represents not merely a technical extension but a fundamental reconceptualization of what we seek to learn from causal analysis. By combining the identification insights of traditional econometrics with the functional approximation capabilities of machine learning, we can move beyond asking "what is the effect?" to understanding "for whom does the effect occur, and why?" – questions that are essential for both economic theory and optimal policy design.

Group Average Treatment Effects (GATEs) provide a practical compromise between the oversimplification of scalar effects and the data (??) requirements of fully individual-level estimation. Following the methodology developed by Chernozhukov et al., the GATE approach uses machine learning methods to estimate individual treatment effects $S(X_i) = E[Y_i(1) - Y_i(0)|X_i]$, then partitions the population into discrete groups based on these predicted effects.

The standard implementation divides the sample into quartiles, where G_1 represents the 25% of observations with the lowest (most negative) treatment effects, and G_4 represents the 25% with the highest treatment effects. Group-specific treatment effects are then estimated through:

$$Y_i = \alpha_1 B(X_i) + \sum_{k=1}^{K} \gamma_k I(G_k) + \varepsilon_i$$

where $B(X_i)$ represents the baseline outcome function, $I(G_k)$ indicates group membership, and γ_k captures the average treatment effect for group k. This approach concentrates statistical power while maintaining meaningful differentiation across the treatment effect distribution.

The move to heterogeneous effects follows naturally from our points of departure. The conditional average treatment effect (CATE) extends the CEF framework by defining treatment effects as functions over the covariate space:

$$\tau(x) = E[Y(1) - Y(0)|X = x]$$

where $Y(1)$ and $Y(0)$ denote potential outcomes under treatment and control.[2] This formulation transforms our estimating equation from the partially linear model $Y_i = \tau d_i + g(X_i) + \varepsilon_i$ to the fully nonparametric specification:

$$Y_i = \alpha(X_i) + \tau(X_i)d_i + \varepsilon_i \tag{3}$$

where both the baseline function $\alpha(\cdot)$ and the treatment effect function $\tau(\cdot)$ vary flexibly with covariates.

The challenges we identified for scalar effects become more severe for heterogeneous effects. The overlap problem is no longer global but local: we require adequate numbers of treated and control units not just overall, but in each region of $\mathcal{X}$ where we seek to estimate $\tau(x)$. The curse of dimensionality thus manifests doubly: we need sufficient data to estimate complex functions $\alpha(\cdot)$ and $\tau(\cdot)$, while simultaneously requiring strong local overlap to identify treatment effects at each point in the covariate space.

The distinction between prediction problems and causal inference becomes even sharper when targeting heterogeneous effects. Traditional machine learning optimizes for mean squared prediction (P) error: $\text{MSE}_\text{P} = E[(Y - \hat{Y})^2]$. For treatment effect estimation, the relevant loss becomes

$$\text{MSE}_\text{CATE} = E[(\hat{\tau}(X) - \tau(X))^2],$$

where the target function $\tau(X)$ is never directly observed.

Machine learning methods adapted for heterogeneous effects address these challenges through localisation strategies. Causal forests, for example, build on the random forest framework but modify the splitting criterion to maximize heterogeneity in treatment effects rather than predictive accuracy.

5.1 What Does "Local" Mean in Heterogeneous Treatment Effect Estimation?

Recall that in the partially linear model (PLM), we residualized both outcomes and treatment to isolate the variation identifying τ. The key insight was that after partialling out confounders through the functions $g_0(X)$ and $m_0(X)$, causal estimation reduced to the familiar covariance formula: $\tau_0 = \text{Cov}(Y^*, d^*)/\text{Var}(d^*)$, where $Y^* = Y - g_0(X)$ and $d^* = d - m_0(X)$ represent residualized variables purged of confounding influences.

The transition from scalar to heterogeneous treatment effects changes our conception of statistical estimation from global to local. But what does local mean here? If we utilise random forests for causal effects in the same manner was when we predict outcomes, local implies X_i resides in the same partition as x, and we then estimate the moment function for observations in that partition. Given this implies that the treatment effect function $\tau(x)$ is discontinuous, a different method is used based on the use of a weight function, say ωX_i

In modern causal forests (GRF), "local" means continuously varying weights $\omega(X_i)$ that determine each observation's influence on the estimate at point x. This creates a smooth, adaptive neighborhood where each observation i receives a weight $\omega(X_i)$ when estimating $\tau(x)$. These weights vary continuously with the distance/similarity between X_i and x, more "similar" to x receive higher weights. In this sense the locality is defined by the weight function, not by hard partitions

The R-Learner Framework: From Theory to Implementation

Having established how the orthogonalization principle applies locally through weighting, we now formalize this approach in the R-learner framework. The R-learner operationalizes the weighted

[2]The overall ATE can be recovered through the law of iterated expectations: $\tau = E[\tau(X)]$.

orthogonalization by minimizing:

$$\hat{\tau}(x) = \arg\min_{\tau} E_n \left[(Y_i - \hat{g}(X_i) - \tau(x)(d_i - \hat{m}(X_i)))^2 \cdot \omega(X_i) \right] \tag{4}$$

where $\alpha(X_i) = K_h(X_i - x)$ is the kernel weighting function with bandwidth h controlling the bias-variance tradeoff: small h focuses on very similar observations (low bias, high variance), while large h includes more distant observations (higher bias, lower variance).

The first-order condition yields the localized moment condition:

$$E\left[(Y_i - \hat{g}(X_i) - \tau(x)(d_i - \hat{m}(X_i)))(d_i - \hat{m}(X_i)) \cdot \omega(X_i)\right] = 0 \tag{5}$$

Solving this moment condition gives the weighted covariance formula:

$$\hat{\tau}(x) = \frac{\sum_i \alpha(X_i)[(Y_i - \hat{g}(X_i))(d_i - \hat{m}(X_i))]}{\sum_i \omega(X_i)[(d_i - \hat{m}(X_i))^2]} \tag{6}$$

This retains the familiar $\text{Cov}(Y^*, d^*)/\text{Var}(d^*)$ structure from the scalar case, now with localization weights $\alpha(X_i)$ that concentrate the estimation around the target point x.

This weighted approach preserves the continuity and smoothness of the treatment effect function $\tau(x)$, unlike partition-based methods that would create discontinuities at region boundaries. Such an approach provides for greater efficiency because it doesn't force discrete boundaries. Instead of asking "is X_i in the same partition as x?", it asks "how much should X_i influence our estimate at x?" This allows for smooth treatment effect surfaces and better utilizes information from observations at varying distances from the target point.

The key insight is that the accuracy of the nuisance function estimates matters most in regions where $\alpha(X_i)$ places substantial weight. If we estimate $g(X_i)$ poorly for observations where $\omega(X_i)$ is large, the residual $(Y_i - \hat{g}(X_i))$ will contain estimation error $(g(X_i) - \hat{g}(X_i))$ that directly contaminates our local treatment effect estimate $\hat{\tau}(x)$. This contamination occurs because the weighting function $\omega(X_i)$ amplifies these errors precisely where they matter most for estimating $\tau(x)$.

5.2 Practical Implementation Through Machine Learning Methods

Consider a concrete example: estimating the effect of a job training program on a 35-year-old high school graduate with two years of work experience. Global estimation would use all observations in the dataset, potentially including 22-year-old college graduates and 50-year-old workers with decades of experience. Local estimation recognizes that these distant observations may provide misleading counterfactuals—the treatment effect for our target individual should be estimated primarily using observations from similar individuals in their neighborhood of the covariate space.

The R-learner implements this localization through kernel weights $K_h(X_i - x)$ that assign higher importance to observations with covariates X_i similar to the target point x. The bandwidth parameter h controls the neighborhood size: when h is small, we focus on very similar observations (high precision, potentially high variance); when h is large, we include more distant observations (lower variance, potentially higher bias). This is precisely analogous to choosing window width in kernel regression or the number of neighbors in k-nearest neighbor estimation.

In practice, machine learning methods implement this localization through different adaptive mechanisms. Causal forests create data-driven neighborhoods through recursive partitioning, where each tree builds a different partition of the covariate space and treatment effects are estimated within each leaf. The final estimate $\hat{\tau}(x)$ emerges from averaging predictions across all trees, with each tree contributing based on which leaf contains the target point x. Neural network approaches can implement localization through attention mechanisms or local linear approximations around each prediction point.

How the R-Learner Generates Treatment Effects for Every X_i: A Step-by-Step Walkthrough

The R-learner's procedure for generating treatment effects at any point X_i combines the three core principles in a systematic sequence that maintains the fundamental simplicity of the OLS formula while adapting it to local neighborhoods:

Step 1: Global Nuisance Function Estimation. Using the full sample, estimate the outcome regression $\hat{g}(X) = E[Y|X]$ and propensity score $\hat{m}(X) = E[d|X]$ via machine learning methods (random forests, neural networks, gradient boosting). These functions capture confounding relationships globally across the entire covariate space.

Step 2: Residualization (Orthogonalization). For every observation in the dataset, compute residualized outcomes $Y_i - \hat{g}(X_i)$ and residualized treatments $d_i - \hat{m}(X_i)$. This step implements the Frisch-Waugh-Lovell logic by removing the predictable component of both outcomes and treatment assignment, leaving only the variation orthogonal to confounders.

Theoretically, it ensures that estimation errors in the nuisance functions $g(X)$ and $m(X)$ do not contaminate our treatment effect estimates—a property known as Neyman orthogonality. When we residualize both outcomes and treatments, we isolate precisely the causal variation needed to identify $\tau(x)$, stripped of confounding influences.

Practically, orthogonalization transforms a complex estimation problem into something familiar. Without residualization, estimating heterogeneous treatment effects requires simultaneously modeling baseline outcomes $\alpha(X)$, treatment effects $\tau(X)$, and their complex interactions in the specification $Y_i = \alpha(X_i) + \tau(X_i)d_i + \varepsilon_i$. After orthogonalization, the problem reduces to locally weighted regression of residualized outcomes on residualized treatments—a much simpler computational task.

Step 3: Adaptive Weighting and Localization. For the target point X_i, construct forest-based Kernel weights $\omega_i = K_h(X_j - X_i)$ for all observations $j \neq i$. Observations 00with covariates X_j similar to X_i receive high weights; distant observations receive low weights. This creates a data-adaptive neighborhood around each target point.

Step 4: Local Weighted Regression. Apply the weighted OLS formula using the residualized variables and adaptive weights:

$$\hat{\tau}(X_i) = \frac{\sum_i \omega_i[(Y_i - \hat{g}(X_i))(d_i - \hat{m}(X_i))]}{\sum_i \omega_i[(d_i - \hat{m}(X_i))^2]} \tag{7}$$

This final step reveals the simplicity underlying the apparent complexity. The treatment effect $\tau(X_i)$ emerges from exactly the same covariance structure as scalar OLS: $\text{Cov}(Y^*, d^*)/\text{Var}(d^*)$, where $Y^* = Y - \hat{g}(X)$ and $d^* = d - \hat{m}(X)$ represent residualized variables. The only difference is that this familiar formula now applies locally within the adaptive neighborhood around X_i.

This simplification becomes especially valuable when working with high-dimensional covariates. Machine learning methods excel at estimating the nuisance functions $g(X)$ and $m(X)$ globally using the full sample and all available features. Once we have these estimates, the local treatment effect estimation becomes a straightforward application of weighted least squares within each adaptive neighborhood.

Adaptive Weighting: How Forests and Kernels Create Neighborhoods

Different machine learning methods implement the adaptive weighting $K_h(X_i - x)$ through distinct mechanisms, but all achieve the same fundamental goal: creating data-driven neighborhoods that adapt to the local structure of the covariate space.

Causal forests implement adaptive weighting through tree-based partitioning. Each tree in the forest recursively splits the covariate space based on variables that maximize treatment effect heterogeneity (rather than prediction accuracy). For any target point x, observations that frequently appear in the same terminal nodes across multiple trees receive higher weights. The final weight $\alpha_i(x) = \frac{1}{B}\sum_{b=1}^{B} \mathbf{1}\{X_i \in L_b(x)\}$ represents the fraction of trees where observation i and target point x end up in the same leaf.

Kernel methods create weights based on explicit distance functions: $K_h(X_i - x) = K\left(\frac{\|X_i - x\|}{h}\right)$, where common choices include Gaussian kernels $K(u) = \exp(-u^2/2)$ or uniform kernels $K(u) = \mathbf{1}\{|u| \leq 1\}$. The bandwidth h controls the trade-off between bias and variance, with cross-validation typically used for selection.

Neural networks can implement localization through attention mechanisms or local linear approximations. Attention weights effectively learn adaptive similarity measures, while local approximations fit linear models within neighborhoods defined by network activation patterns.

The following table illustrates how heterogeneous treatment effect estimation preserves the essential structure of the scalar PLM while extending it to capture effect heterogeneity:

Component	**Scalar PLM**	**Heterogeneous Effects**
Model	$Y_i = d_i\tau + g(X_i) + \varepsilon_i$	$Y_i = \alpha(X_i) + \tau(X_i)d_i + \varepsilon_i$
Nuisance Functions	$g(X), m(X)$ (global accuracy)	$g(X), m(X)$
Orthogonalization	$Y^* = Y - g_0(X)$, $d^* = d - m_0(X)$	Same residualization, applied with weights
Final Formula	$\tau = \frac{\text{Cov}(Y^*, d^*)}{\text{Var}(d^*)}$	$\tau(x) = \frac{\text{Cov}_{\omega(x)}(Y^*, d^*)}{\text{Var}_{\omega(x)}(d^*)}$
Key Innovation	ML for nuisance estimation	ML for localization + nuisance

This synthesis reveals that heterogeneous treatment effect estimation, despite its surface complexity, achieves the same fundamental simplification as the scalar case. The apparent challenge of estimating treatment effects as unknown functions across high-dimensional spaces transforms into the familiar problem of computing weighted covariances within data-adaptive neighborhoods. The sophistication lies not in abandoning the principles that made linear methods successful, but in extending those principles to capture the rich heterogeneity that characterizes treatment effects in real economic applications.

6 Book Structure and Chapter Overview

The book is organized into ten chapters that progressively build from foundational concepts to advanced applications. Each chapter combines theoretical development with practical implementation guidance and real-world applications.

Contents

6.1 Part I: Introduction and Motivation

Part I establishes a deliberate two-chapter foundation that provides readers with complementary perspectives on causal machine learning methods. This approach recognizes that effective learning occurs through both concrete application and theoretical understanding.

Chapter 1: Looking Ahead serves as the *empirical lens*, directly emulating the "Sneak Peek" approach pioneered in Chernozhukov et al.'s *Applied Causal Inference Powered by* ML *and* AI. Using synthetic Amazon product data with 9,212 observations, this chapter grounds abstract concepts in a tangible e-commerce pricing scenario. Readers immediately see how the fundamental identification challenge manifests—where naive OLS shows a positive price-sales relationship (+1.236) that masks the true negative causal effect (-1.5) due to confounding from unobserved product quality and brand strength. This concrete example threads through the entire book, allowing readers to see each method applied to the same familiar dataset.

Chapter 2: Overview provides the *theoretical lens*, establishing the conceptual architecture through five key "points of departure" that systematically bridge machine learning and causal inference. Rather than diving immediately into technical details, this chapter maps the intellectual landscape, explaining why economists need machine learning, why ML practitioners need causal inference, and how the fundamental tension between prediction and causation drives the need for re-optimized algorithms.

The Strategic Value of This Dual Approach

This two-chapter foundation creates multiple entry points for diverse readers while ensuring comprehensive coverage. The empirical-first approach in Chapter 1 immediately demonstrates practical relevance, while Chapter 2's theoretical framework provides the conceptual scaffolding needed to understand why each subsequent method matters. As later chapters unfold specific techniques—from regularized regression through double machine learning to causal forests—readers can reference both the concrete Amazon example (seeing "how it works") and the theoretical overview (understanding "why it works").

Some readers may find this dual overview sufficient for their immediate needs, while others will use it as reinforcement as they work through the technical chapters. The approach recognizes that modern causal machine learning sits at the intersection of multiple disciplines, requiring both empirical intuition and theoretical rigor to be truly understood and properly applied.

Contents

Chapter 1: Looking Ahead

The opening chapter grounds this discussion through a concrete example using synthetic Amazon product data with 9,212 observations that mirrors real e-commerce pricing complexities. The naive OLS relationship between log-prices and log-sales shows a positive coefficient (+1.236), suggesting higher prices correlate with better sales ranks. However, this observed association masks the true negative causal effect (-1.5), illustrating how confounding from unobserved product quality and brand strength creates the fundamental identification challenge that motivates our entire approach.

Analog to Chernozukov's preview

Chapter 2: Overview

This chapter establishes why economists need machine learning and why machine learning practitioners need causal inference, providing a comprehensive roadmap for the entire book. The central challenge we address is a fundamental tension: ML optimizes for prediction while economics seeks causal understanding. This tension raises the critical question of how and why off-the-shelf ML algorithms cannot simply be applied to locate and make inference on causal effects.

Add reference to objective to create a detailed review of the material

The chapter is structured around five key "points of departure" that provide entry points for readers with little background in machine learning while bridging ML and causal inference. We begin with the Conditional Expectation Function $E[Y|X]$ as the mathematical object that unites both statistical cultures, whether approached through parametric assumptions or algorithmic flexibility. The CEF provides a rigorous framework for understanding relationships between variables and serves as the foundation for both prediction and causal inference.

The key insight is that the CEF grounds our considerations in the familiar base case of the linear regression model. This raises the fundamental question: On what basis do we depart from the linear CEF when it provides the best linear approximation to the CEF? The answer lies in recognizing when the linear specification becomes inadequate for capturing complex, high-dimensional relationships in modern economic data.

Modern economic data increasingly exists in high-dimensional spaces where traditional methods face the curse of dimensionality. When researchers encounter datasets with many potential confounders, control variables, or instruments, parametric approaches become impractical due to overfitting concerns and the sheer computational burden. This practical reality necessitates regularization techniques and motivates the natural progression from LASSO to Double LASSO to Double LASSO with Neyman orthogonality.

The dimensionality challenge extends beyond simple variable selection to include complex transformations. The transformation $X = T(W)$ can quickly expand the feature space from a manageable set of raw variables W to a high-dimensional constructed regressor matrix X through polynomials, interactions, and other nonlinear transformations that multiply the effective dimensionality of the problem.

The Frisch-Waugh-Lovell theorem emerges as a pivotal tool for ML-based causal inference, providing the operational principle that makes complex causal estimation tractable. FWL shows how to transform a multivariate regression problem into a bivariate one through "partialing out" – the process of isolating variation in one variable that is orthogonal to confounding variables. The modern relevance of FWL becomes clear in high-dimensional settings where we cannot simply run a penalized regression of Y on (d, X) and interpret the coefficient on d causally, as regularization bias contaminates our causal estimate.

Instead, we must manually implement FWL logic: first use regularization to model relationships between X and Y, then between X and d, and finally estimate the causal parameter from the residualized regression. This "manual" form of residualization after estimating non-parametric nuisance functions enables "debiased" estimation. The principle extends naturally to instrumental variables settings, where we must partial out confounders from all variables (Y, d, and Z) before applying IV formulas to residualized variables.

The chapter situates this technical apparatus within the treatment effects framework, grounding the discussion in causal reasoning about potential outcomes and counterfactuals. This framework reminds us that behind every regression lies a question about causal effects, and that correlation cannot be transformed into causation without proper identification strategy. Modern applications often involve

heterogeneous treatment effects that vary smoothly as unknown functions of covariates. The search for this heterogeneity requires locating similarly rich sets of comparable units to serve as counterfactuals – a requirement that becomes increasingly challenging in high-dimensional spaces.

The final point of departure addresses the fundamental challenge: ML methods designed for prediction must be fundamentally re-optimized for causal inference. Standard ML methods maximize predictive accuracy through cross-validation and regularization, optimizing for out-of-sample performance. Causal inference demands something different: unbiased estimates of specific parameters with valid standard errors, even at the cost of predictive accuracy.

This re-optimization requires understanding the bias-variance tradeoff in a new light. Traditional econometrics emphasized unbiasedness and consistency, while machine learning forces explicit consideration of finite-sample performance and bias-variance tradeoffs. The related vernacular includes concepts like double-robust methods, double-debiased estimation, orthogonalization, sample splitting, and honest estimation – terms that reflect the careful adaptation of algorithmic tools for causal questions.

Weaving through these points of departure is the principle of Neyman orthogonality, which makes causal estimators locally insensitive to first-order errors in nuisance parameter estimation. This provides robustness when perfect model selection is not achievable – a common situation in high-dimensional settings where we must balance model complexity against overfitting. The distinction between model selection and parameter estimation becomes crucial. While ML excels at the former through techniques like cross-validation, causal inference requires that parameter estimation remain valid even when model selection is imperfect.

The theoretical progression moves systematically from linear benchmark models ($y = \alpha + \tau d + X'\beta + \varepsilon$) through partial linear models ($y = g(X) + \tau d + \varepsilon$) to fully non-parametric specifications accommodating heterogeneous treatment effects ($y = \alpha(X) + \tau(X)d + \varepsilon$). This evolution demonstrates how ML methods must be fundamentally re-optimized for causal estimation rather than prediction, establishing the organizing principle for the entire book.

The genius of modern causal machine learning lies in recognizing these connections. Double Machine Learning uses ML methods to flexibly estimate nuisance functions in the CEF, applies FWL to partial out confounding, and employs orthogonalization to ensure robustness – all while respecting the logic of potential outcomes. High-dimensional methods select relevant confounders, nonparametric methods capture their complex effects, and careful sample splitting preserves valid inference.

By the end of this opening chapter, readers understand that the integration of machine learning and causal inference is not merely about applying new algorithms to old problems. Rather, it requires a fundamental reconceptualization of how we approach causal questions in the age of big data and algorithmic flexibility. The five points of departure provide both the theoretical foundation and practical roadmap for this integration, ensuring that we can harness ML's predictive power while maintaining the rigor and interpretability that define economic science. The challenge is not whether to use these methods, but how to adapt them appropriately within the disciplined framework of causal inference.

6.2 Part II: Foundations

Contents

Chapter 3: The Best Predictor and the Conditional Expectation Function OLD

This chapter introduces the first point of departure: the pivotal role of the Conditional Expectation Function (CEF) as the bridge between prediction and causal inference. We begin by demonstrating the fundamental theorem, namely that the CEF $E[Y|X]$ is also the best linear predictor, providing the mathematical foundation that unifies both statistical cultures. This equivalence raises a crucial question: if the linear CEF already provides the best linear approximation, on what basis do we depart from linearity and consider alternative criterion functions?

We begin by demonstrating that the Conditional Expectation Function $E[Y|X]$ represents the mathematical object that both statistical approaches seek to estimate, whether through parametric

specifications (such as $E[Y|X] = X\beta$) or algorithmic methods that treat the functional form as unknown. This shared target connects the seemingly disparate methodological approaches: traditional econometricians impose structural assumptions about $E[Y|X]$, while machine learning practitioners allow algorithms to discover its form, but both fundamentally seek to characterise the same conditional expectation relationship.

Considering the relationship between wages, education, and gender, Chapter 2 introduces the concept of the covariate space X. We contrast traditional parametric specification $E(\text{Wage}|X) = \beta_0 + \beta_1 \text{Ed} + \beta_2 \text{Fem}$ with a simple nonparametric alternative that computes average wages over local bins defined by all combinations of education levels and gender. This comparison provides a simple illustration of how nonparametric methods can reveal patterns that parametric assumptions might obscure.

Moving from the simple low-dimensional case where the bins are given, we proceed to consider the fundamental challenge of nonparametric estimation: the identification of the appropriate bins when confronted with a high-dimensional covariate spaces. We introduce the nonparametric methods that serve as precursors to modern ML algorithms, including nearest neighbors (with floating bins that adapt to local data density), binned estimators (trading off bias and variance through bandwidth selection), kernel methods such as Nadaraya-Watson estimation, and regression trees.

These methods provide the conceptual foundation for understanding how ML algorithms approximate the CEF in increasingly flexible ways while managing the curse of dimensionality.

Chapter 3: The Best Predictor and the Conditional Expectation Function

This chapter introduces the first point of departure: the pivotal role of the Conditional Expectation Function (CEF) as the theoretical bridge between prediction and causal inference. We begin by demonstrating the fundamental theorem that the Conditional Expectation Function $E[Y|X]$ represents the best predictor of Y given X among all possible functions. Formally, the CEF solves the optimization problem $\min_{m(X)} E[(Y - m(X))^2]$, where the minimization is taken over all measurable functions $m(X)$. This establishes the CEF as the mathematical object that minimizes mean squared error without any functional form restrictions, making it the gold standard for prediction.

However, in practice, we often impose linearity constraints due to computational convenience, interpretability concerns, or modeling traditions. This leads us to the Best Linear Predictor (BLP), which solves the restricted optimization problem $\min_b E[(Y - X'b)^2]$, where the minimization is taken only over linear functions of the form $X'b$. The BLP represents the optimal predictor within the class of linear functions, but this optimality is conditional on accepting the linearity constraint.

The BLP represents the best we can do under a linearity constraint, not the best we can do overall. When the true CEF is linear, the BLP and CEF coincide, making linear methods optimal. However, when the true conditional expectation function exhibits substantial nonlinearity, the optimal linear approximation may provide a poor fit to the underlying relationship. This distinction proves crucial for causal inference because virtually all causal estimators fundamentally rely on the estimation of reduced-form equations, which are essentially prediction problems. Whether we are estimating propensity scores $e(X) = P(D = 1|X)$, outcome functions $\mu_0(X) = E[Y|D = 0, X]$ and $\mu_1(X) = E[Y|D = 1, X]$, or first-stage equations in instrumental variable settings, causal inference reduces to predicting nuisance functions accurately. In such cases, the approximation error from imposing linearity—the gap between the true CEF and its best linear approximation - may be substantial enough to compromise both predictive accuracy and causal inference. This creates opportunities creating opportunities where machine learning's predictive power directly translates to better causal inference.

We know higher predictive performance for reduced forms deceases standard errors of causal estimates. Clarity: functional form issues can create bias in causal estimators?

Parag. below introduces nonparametric estimators and does not link specifically to parag above ...

To illustrate these concepts concretely, we examine the relationship between wages, education, and gender. Consider the traditional parametric specification $E(\text{Wage}|X) = \beta_0 + \beta_1 \text{Ed} + \beta_2 \text{Fem}$, where $X = \{\text{Ed}, \text{Fem}\}$. This linear specification represents the BLP and assumes that the conditional expectation function can be adequately captured through linear relationships. While this specification provides the optimal linear approximation to the wage-education-gender relationship, it may miss important nonlinear patterns such as differential returns to education by gender, nonlinear returns to education, or threshold effects.

We contrast this parametric approach with a simple nonparametric alternative that computes average wages over local bins defined by all combinations of education levels and gender. This binning approach allows the data to reveal the true shape of the conditional expectation function without imposing linearity constraints. By examining conditional means within each education-gender cell, we can directly observe whether the linear specification adequately captures the underlying relationship or whether systematic departures from linearity suggest that more flexible methods might improve both prediction and subsequent causal inference.

Moving beyond the simple low-dimensional case where bins can be manually defined, we confront the fundamental challenge of nonparametric estimation: how to identify appropriate local neighborhoods when dealing with high-dimensional covariate spaces. As the dimension of X increases, the curse of dimensionality makes simple binning approaches impractical, necessitating more sophisticated methods for estimating the CEF nonparametrically.

We introduce the nonparametric methods that serve as conceptual precursors to modern ML algorithms. Nearest neighbor methods create floating bins that adapt to local data density, automatically adjusting the size of neighborhoods based on data availability. Kernel methods such as Nadaraya-Watson estimation provide smooth approximations to the CEF by weighting observations according to their distance from the prediction point. Regression trees partition the covariate space into regions where the conditional expectation is approximately constant, providing a natural way to handle both continuous and discrete variables while revealing interaction effects.

These methods provide the conceptual foundation for understanding how modern ML algorithms approximate the CEF in increasingly flexible ways while managing the curse of dimensionality. Random forests extend regression trees through ensemble methods, neural networks provide highly flexible functional approximations, and regularization techniques like LASSO enable variable selection in high-dimensional settings. Each method represents a different approach to the fundamental trade-off between bias and variance that characterizes nonparametric estimation.

However, applying these methods to causal inference requires careful consideration of how prediction accuracy in nuisance functions translates to valid causal inference. The chapter establishes that while better prediction of the CEF can improve causal inference when nuisance functions are nonlinear, this improvement is not automatic. The methods must be adapted to preserve the identifying assumptions required for causal inference, account for the fact that we care about parameter estimation rather than pure prediction, and ensure that estimation uncertainty is properly characterized for inference.

The theoretical framework developed in this chapter establishes the foundation for understanding when and why machine learning methods can enhance causal inference. When nuisance functions in causal estimators are linear, traditional linear methods provide optimal prediction, and there are limited gains from ML approaches. However, when these functions exhibit important nonlinearities, ML methods' ability to approximate the true CEF more accurately can translate directly into improved causal inference, provided the methods are properly adapted for causal rather than purely predictive applications.

By the end of this chapter, readers understand that the choice between parametric and nonparametric methods is not merely a matter of computational convenience or methodological preference. Rather, it reflects a fundamental decision about whether we believe the linearity constraint implicit in traditional methods adequately captures the complexity of the relationships we seek to understand. When the answer is no, machine learning provides the tools to move beyond these constraints while maintaining the rigor required for credible causal inference. This sets the stage for the detailed exploration of high-dimensional methods and modern causal machine learning techniques that follow in subsequent chapters.

Chapter 4: Estimation and Inference for Causal Effects

This chapter establishes the theoretical framework for applying machine learning methods to causal inference problems. We begin by formalising the potential outcomes framework and demonstrating how causal parameters relate to conditional expectations. The central insight is that causal effects often depend on nuisance functions that must be estimated as intermediate steps, creating opportunities for machine learning to enhance traditional econometric approaches.

We start with simple linear models, examining classical linear estimators such as IV and 2SLS. This analysis reveals a crucial insight: the 2SLS estimator, structured as a ratio of reduced forms, naturally

highlights the potential gains from ML methods for prediction. Given that the first stage of 2SLS is fundamentally a prediction problem, there exists the potential for nonparametric methods to capture nonlinear relationships that linear projections might miss

This chapter establishes the theoretical framework for applying machine learning methods to causal inference problems, using the classic problem of estimating returns to schooling as a unifying example. We begin by formalising the potential outcomes framework and demonstrating how causal parameters relate to conditional expectations. The central insight is that causal effects often depend on nuisance functions that must be estimated as intermediate steps, creating both opportunities and challenges for machine learning to enhance traditional econometric approaches.

We start with the canonical schooling, ability, and wages example to illustrate fundamental identification challenges. Consider the relationship between schooling (s_i), unobserved ability (A_i), and wages (Y_i):

$$Y_i = \alpha + \rho s_i + A_i\gamma + v_i = \alpha + \rho s_i + \eta_i$$

where $\eta_i = A_i\gamma + v_i$ represents the composite error term containing unobserved ability. The endogeneity problem arises because $\text{Cov}(s_i, \eta_i) \neq 0$, rendering OLS estimates of ρ inconsistent.

The chapter develops the instrumental variables solution systematically, introducing covariates X_i and instruments z_i to arrive at the two-stage least squares formulation. The critical equation (11) from the second stage:

$$y_i = \alpha' X_i + \rho\hat{s}_i + [\eta_i + \rho(s_i - \hat{s}_i) \tag{8}$$

reveals how 2SLS purges endogeneity through projection, where $\hat{s}_i$ represents the predicted values from the first-stage regression. This equation serves as the bridge to understanding modern causal machine learning methods.

The Fundamental Challenge: Linear Projections versus Nonlinear Reality

The chapter then addresses a crucial limitation of classical 2SLS: it relies on linear projections in the first stage. What if the true conditional expectation $E[s_i|z_i, X_i]$ is nonlinear? This question motivates the transition to machine learning methods that can capture complex patterns while preserving causal identification.

We explore three key areas where ML enhances traditional IV approaches:

1. **Nonlinear First Stages**: Random forests, neural networks, and other flexible methods can capture $E[d_i|Z_i]$ more accurately than linear projections, but risk inadvertently learning endogeneity patterns
2. **High-Dimensional Instruments**: When Z is high-dimensional, LASSO and other regularization methods provide principled selection from many potential instruments while preventing overfitting
3. **Heterogeneous Effects**: Moving beyond constant treatment effects, methods like IV forests estimate conditional average treatment effects $\tau(x) = E[Y_i(1) - Y_i(0)|X_i = x]$

IV Estimators as Prediction Problems

A key conceptual shift frames the first-stage regression as a prediction problem, opening the door to machine learning methods. This raises critical questions that structure the remainder of the chapter:

- What happens when X is high-dimensional?
- What happens when Z is high-dimensional?
- How do we handle settings where Z is high-dimensional, the optimal instrument $\omega^*(Z_i) = E[s_i|Z_i]$ is nonparametric, and instruments may be weak?

Optimal Instruments for Causal ML

These questions lead naturally to the introduction of optimal instruments and the connection to modern causal ML methods.

We then examine the optimal instruments problem ..... This problem crystallises the key tension in applying ML to causal inference: while flexible methods can better predict optimal instruments, they also risk overfitting to spurious patterns, particularly when instruments are weak. This discussion illuminates the broader overlap between parametric and nonparametric methods, showing how traditional econometric insights about identification remain crucial even as estimation methods become more flexible. Throughout, we emphasize that algorithmic sophistication cannot overcome failures of identification, but given valid identification assumptions, ML can dramatically improve estimation precision and reveal heterogeneous effects.

Trade-offs Between Classical and ML-Enhanced Methods

The chapter provides a balanced assessment of the trade-offs between classical 2SLS and ML-enhanced IV methods:

Classical 2SLS:

- (+) Transparent, well-understood properties
- (+) Clear diagnostics (F-statistic, overidentification tests)
- (-) Misses nonlinear relationships
- (-) Inefficient with many instruments

ML-Enhanced IV:

- (+) Captures complex patterns flexibly
- (+) Handles high-dimensional settings naturally
- (-) Can inadvertently learn confounding patterns
- (-) Requires careful regularization and sample splitting

Binary Response Models with Endogeneity: The Forbidden Regression Problem and ML Solutions

Make cross-reference to Chapters

A particularly challenging econometric problem arises when both the outcome and endogenous regressor are binary variables, exemplified by the Angrist-Evans (1998) fertility study examining women's labor force participation. In this setting, traditional instrumental variable approaches fail due to what is known as the "forbidden regression problem." The core mathematical issue stems from Jensen's inequality: when attempting a 2SLS-style plug-in approach with binary outcomes, we cannot pass expectations through indicator functions, as $\mathbb{E}[\mathbf{1}(x^e\beta_e + x^o\beta_o + \varepsilon \geq 0)] \neq \mathbb{E}[\mathbf{1}(\mathbb{E}[x^e|z]\beta_e + x^o\beta_o + \varepsilon > 0)]$. This inequality renders standard parametric approaches inconsistent—even with strong distributional assumptions, the nonlinearity of the indicator function prevents the exchange of expectation and transformation operations required for valid inference.

Machine learning methods, particularly Random Forests and Generalized Random Forests, elegantly circumvent this fundamental limitation. Rather than transforming linear indices through nonlinear link functions like probit or logit, forest-based methods estimate probabilities directly through empirical frequencies within leaf nodes. As Scornet (2016) demonstrates, this voting mechanism—where each observation's probability estimate emerges from the proportion of positive outcomes in its local neighborhood—completely sidesteps the Jensen's inequality trap.

The IV forest implementation extends this insight to handle endogeneity through local moment conditions, discovering heterogeneous local average treatment effects across the covariate space without requiring any functional form assumptions. This exemplifies a broader principle: machine learning's nonparametric, data-adaptive approaches can solve econometric problems that are mathematically intractable within the parametric framework, opening new frontiers for causal inference with limited dependent variables.

Looking Ahead: Principled Approaches for Causal ML

The chapter concludes by identifying three fundamental challenges that must be addressed when applying machine learning methods to causal inference, setting the stage for the solutions developed in subsequent chapters:

1. **The Orthogonalization Challenge**: Equation (11) reveals that 2SLS achieves consistency through an orthogonality property: the residual term $[\eta_i + \rho(s_i - \hat{s}_i)]$ is orthogonal to $\hat{s}_i$ by construction. When we replace linear projections with flexible ML methods, how do we preserve this crucial orthogonality? This question motivates the development of Neyman orthogonal moment conditions and double/debiased machine learning methods in Chapter 5.

2. **The Overfitting Challenge**: Using the same data to both train flexible models and estimate causal effects creates overfitting bias that can invalidate inference. How do we leverage ML's flexibility without contaminating our causal estimates? Chapter 6 introduces cross-fitting and sample splitting procedures that address this challenge through honest estimation.

3. **The Regularization Challenge**: Unlike pure prediction problems where regularization is chosen solely to minimize prediction error, causal inference requires regularization that respects the problem's causal structure. How do we select regularization parameters when our objective is unbiased causal estimation rather than optimal prediction? Chapter 5 develops frameworks for causal-aware regularization.

These challenges are not merely technical obstacles but fundamental issues at the intersection of prediction and causation. The IV estimator's transition from equation (11)'s linear projection to modern ML methods exemplifies how each challenge arises: the need to maintain orthogonality despite flexibility, to avoid overfitting when learning nuisance functions, and to choose regularization that preserves identification rather than optimizing prediction.

By understanding these challenges through the lens of the schooling example and equation (11), readers are prepared for the rigorous solutions developed in Parts III and IV, where these principles are formalized and implemented across a range of causal inference problems.

6.3 Part III: High-Dimensional Methods

Contents

Chapter 5: High-Dimensional Methods for Linear Models

This chapter develops high-dimensional methods for causal inference through two fundamental points of departure. The first is high-dimensional statistics, where we modify the OLS criterion function to include penalty functions over complexity. This leads to the crucial progression from LASSO for prediction problems to Double LASSO for causal inference, focusing on the problem of model selection: specifically, identifying the union of control variables that are important for both the outcome Y and the endogenous variable d.

The second point of departure is the Frisch-Waugh-Lovell theorem, which becomes essential when we recognise the fundamental distinction between model selection and parameter estimation. While Double LASSO addresses model selection through regularised regression applied to both the outcome and treatment equations, the FWL theorem provides the theoretical foundation for residualisation. This leads to Neyman orthogonality - the simultaneous residualisation of both Y and d, which accounts for local errors in the selection process and ensures that estimation errors in nuisance functions do not contaminate our causal parameter estimates.

The chapter demonstrates how these complementary approaches work together: Double LASSO provides principled variable selection, while Neyman orthogonality provides robustness to selection mistakes. Applications illustrate the practical importance of this dual approach, and the chapter concludes by

signposting the extension of these principles beyond linear models to the flexible, non-parametric methods developed in subsequent chapters.

Chapter 6: Applications of Regularised Regression for Linear Models

This chapter provides a number of applications of the regularisation methods developed in Chapter 4, demonstrating their implementation across both predictive and causal inference problems. We explore two detailed case studies:

(1) predicting Boston house prices using census-level characteristics as a classical prediction problem, comparing OLS, stepwise regression, and various LASSO implementations to illustrate overfitting and the superiority of regularised approaches for out-of-sample prediction.

(2) estimating the impact of institutions on economic growth, addressing measurement error and endogeneity through instrumental variables with high-dimensional controls.

Each application emphasizes practical implementation issues, including the distinction between homoscedastic and heteroscedastic LASSO, the role of data-driven penalty loadings, and implementation using both theory-based and cross-validation approaches. Each application emphasises practical issues: choosing regularization parameters, handling different types of variables, computational considerations, and interpreting results. Code examples in State, R or Python accompany each application.

An important contribution of the chapter is the consideration of "theory-based" regularisation through rigorous LASSO. Rigorous LASSOuses theoretically-derived penalty parameters rather than cross-validation, providing theoretical guarantees against overfitting without requiring sample splitting.

We also consider the important question: what should we expect from machine learning methods when careful parametric causal analyses deliver zero effects? Using the Dale and Krueger elite college study as a case study, we examine whether ML methods can fundamentally overturn well-identified null results, emphasizing that ML methods provide valuable tools for robustness checking and efficiency improvement but cannot create quasi-experimental variation or overcome identification failures.

Question: Is this Chapter the right location for this point?

6.4 Part IV: Modern Causal Machine Learning

Contents

Chapter 7: Double Machine Learning

See commented out section on PLM-IV Estimator, and see if any further detail required here?

This chapter presents the Double/Debiased Machine Learning (DDML) framework by taking as its point of departure the Double LASSO with Neyman orthogonalization developed in Chapter 4. We introduce the partial linear model, maintaining linearity for the effects of endogenous variables while allowing nonparametric components to enter as nuisance functions for both outcome and endogenous variable equations.

The chapter begins by establishing the connection to Chapter 4's regularization bias and addresses the additional challenge of overfitting bias that arises when using the same data to estimate both nuisance functions and the parameter of interest. We then introduce DDML, which generalizes beyond LASSO to accommodate any ML method for nuisance function estimation. The framework employs cross-fitting—a sophisticated form of sample splitting—to eliminate overfitting bias while maintaining the power of Neyman orthogonality for handling regularization bias.

A crucial insight demonstrates the fundamental link between DDML and classical linear estimation. The apparent complexity of the partial linear model $Y_i = d_i\tau + g(X_i) + \varepsilon_i$ dissolves through orthogonalization into the familiar OLS covariance formula $\tau_0 = \text{Cov}(Y^*, d^*)/\text{Var}(d^*)$, where $Y^* = Y - g_0(X)$ and $d^* = d - m_0(X)$ represent residualized variables purged of confounding influences.

The key difference from classical OLS lies in using nonparametric ML methods to estimate the orthogonalization functions $g(\cdot)$ and $m(\cdot)$.

Extension to Instrumental Variables: The PLM-IV Estimator

The chapter extends beyond the unconfoundedness assumption to address settings where endogeneity renders the basic DDML approach invalid. The PLM-IV estimator builds on the DDML foundation by incorporating instrumental variables to handle cases where $\text{Cov}(\varepsilon_i, v_i) \neq 0$—that is, when unobserved confounders affect both treatment assignment and outcomes.

The instrumental variables extension introduces a third nuisance function $r(\cdot)$ to model the instrument's relationship with confounders, creating what is effectively a "triple machine learning" problem. The orthogonal estimating equation becomes $\psi(W, \tau, g, m, r) = (Y - g(X) - \tau(d - m(X)))(Z - r(X))$, where the instrument Z satisfies standard exclusion restriction and relevance conditions.

Remarkably, the PLM-IV estimator maintains the same intuitive covariance structure as classical instrumental variables estimation: $\tau_0 = \text{Cov}(Y^*, Z^*)/\text{Cov}(d^*, Z^*)$, where $Y^* = Y - g_0(X)$, $d^* = d - m_0(X)$, and $Z^* = Z - r_0(X)$ represent residualized versions of all three variables. This connection reveals how the Frisch-Waugh-Lovell theorem extends naturally to the high-dimensional setting: we partial out confounders from *all* variables—outcome, treatment, and instrument—before applying the classical IV covariance formula.

The DDML-IV estimator preserves Neyman orthogonality with respect to all three nuisance functions $g(\cdot)$, $m(\cdot)$, and $r(\cdot)$, ensuring robustness to first-order estimation errors in each. This triple orthogonalization enables "debiased" instrumental variables estimation in high-dimensional settings where traditional IV methods would fail due to the curse of dimensionality.

The application extends the institutions-on-growth analysis using ensemble averaging with model-based weights, leveraging Stata's `ddml` command which implements stacked generalization to combine multiple supervised learners. This demonstrates DDML's practical implementation while showcasing how model averaging can improve both prediction and causal estimation in high-dimensional settings.

7.1 The Partial Linear Model versus Double Lasso

A central pedagogical aim of Chapter 7 is to draw a precise contrast between the Partial Linear Model (PLM) estimated by DDML and the Double LASSO approach developed in Chapter 4. Both methods share the same identification strategy—orthogonalising the treatment variable with respect to a high-dimensional set of controls—yet they differ fundamentally in the nature of the nuisance functions they can accommodate, the mechanism by which debiasing is achieved, and the role assigned to sample splitting.

Functional Form of the Nuisance Functions. Double LASSO is confined to the linear model: it selects controls via two LASSO regressions—Y on X and d on X—and then runs OLS on the union of the selected controls. The nuisance functions $g(X) = E[Y \mid X]$ and $m(X) = E[d \mid X]$ are therefore constrained to be linear in X. DDML imposes no such restriction: these conditional expectations may be estimated by any ML method—random forests, neural networks, gradient boosting, or stacking—accommodating nonlinearity and interactions that LASSO cannot capture. This flexibility is precisely what makes DDML a strict generalisation of Double LASSO.

Orthogonalisation versus Union Selection. Double LASSO debiases $\hat{\tau}$ by taking the *union* of the two variable sets selected by the auxiliary regressions and running a single OLS regression. This works in the linear setting because the Frisch-Waugh-Lovell theorem guarantees that including the selected controls automatically removes their confounding influence. DDML instead achieves debiasing through *explicit orthogonalisation*: the treatment residual $\hat{v}_i = d_i - \hat{m}(X_i)$ is constructed as the component of treatment not explained by X, and τ is recovered from the regression of $Y_i - \hat{g}(X_i)$ on $\hat{v}_i$. This is a direct nonparametric application of the Frisch-Waugh-Lovell logic. Union selection is infeasible when $m(X)$ is nonparametric, since there is no finite set of variables to select.

Sample Splitting and Overfitting Bias. Double LASSO does not require sample splitting. Because the linear LASSO estimator produces an orthogonal score at the parametric level, the post-selection OLS step is not contaminated by overfitting of the nuisance functions. DDML, by contrast, requires cross-fitting because nonparametric ML estimators can overfit the noise in the data. To understand why, consider the treatment equation $d_i = m_0(X_i) + v_i$, where v_i is the structural noise orthogonal to X_i by construction. A linear estimator such as OLS or LASSO cannot absorb v_i into $\hat{m}(X_i)$, since v_i lies outside the column space

of X_i by construction. A nonparametric estimator faces no such constraint: trained on the same sample used to form $\hat{\tau}$, a sufficiently flexible model reduces in-sample prediction error by partially fitting the idiosyncratic noise v_i, producing an estimated residual $\hat{v}_i = d_i - \hat{m}(X_i)$ that is a shrunk and contaminated version of the true residual. The limiting case of a fully saturated tree illustrates the problem starkly—each observation occupies its own leaf, so $\hat{m}(X_i) = d_i$ exactly, $\hat{v}_i = 0$ for all i, and there is no residual variation left to identify τ.

The contamination of $\hat{v}_i$ matters because any component of v_i absorbed into $\hat{m}$ reappears as a spurious correlation between $\hat{v}_i$ and the structural error ε_i, biasing the estimator of τ in a way that does not vanish asymptotically. Cross-fitting severs this dependence by estimating nuisance functions on a held-out fold I^c and evaluating the moment condition only on the complementary sample I. Since $\hat{m}$ has never observed the realisations of (v_i, ε_i) for $i \in I$, the estimation error $\hat{m}(X_i) - m_0(X_i)$ is independent of the structural noise in the main sample, and the problematic cross term

$$E[\hat{v}_i \, \varepsilon_i \mid I^c] = E[v_i \varepsilon_i \mid I^c] + E[(m_0(X_i) - \hat{m}(X_i))\varepsilon_i \mid I^c] = 0$$

vanishes by independence. Cross-fitting extends this to K folds, recovering the full sample for estimation of τ without sacrificing the independence property.

Neyman Orthogonality and the Rate Requirement. The requirement of cross-fitting is closely related to the property of Neyman orthogonality. A naive moment condition is sensitive to nuisance estimation error: a first-order Taylor expansion shows that any bias in $\hat{\eta} = (\hat{g}, \hat{m})$ propagates directly into $\hat{\tau}$. The DDML score $\psi(W_i; \tau, \eta) = (Y_i - \tau d_i - g(X_i))(d_i - m(X_i))$ is constructed to satisfy Neyman orthogonality—the Gateaux derivative of the expected moment condition with respect to η, evaluated at the truth, is zero:

$$\partial_\eta \, E[\psi(W_i; \tau_0, \eta_0)][\eta - \eta_0] = 0 \quad \forall \, \eta.$$

This kills the first-order term in the Taylor expansion, so that nuisance estimation error affects $\hat{\tau}$ only at second order. Orthogonality therefore weakens the rate requirement on the nuisance estimators from the parametric rate $n^{-1/2}$ to the substantially more lenient rate $n^{-1/4}$, since only the *product* of two estimation errors enters the residual bias. Together, Neyman orthogonality (addressing regularisation bias) and cross-fitting (addressing overfitting bias) deliver $\sqrt{n}$-consistency and asymptotic normality of $\hat{\tau}$.

The contrast between Double LASSO and DDML is summarised in the following table.

Dimension	**Double LASSO**	**DDML**
Nuisance function form	Linear (LASSO-selected)	Nonparametric (any ML)
Debiasing mechanism	Union of selected controls + OLS	Explicit orthogonalisation via $\hat{v}_i$
Sample splitting required?	No	Yes (cross-fitting)
Convergence rate required	$n^{-1/2}$ (parametric)	$n^{-1/4}$ per nuisance function
Score orthogonality	Automatic via FWL in linear model	Constructed explicitly

In short, DDML is a strict generalisation of Double LASSO: when the nuisance functions happen to be linear and estimated by LASSO, the two approaches coincide up to the sample-splitting step. The additional machinery of DDML—orthogonalised scores and cross-fitting—purchases validity when $g(X)$ and $m(X)$ are genuinely nonlinear and estimated at slower-than-parametric rates. This comparison underscores a broader theme of the chapter: that the architectural differences between prediction-oriented ML tools and causal estimators are not merely technical refinements but reflect a fundamental reconceptualisation of how estimation error propagates when the target is a causal parameter rather than a predictive function.

Chapter 8: Treatment Effects and Double Robust Estimators

This chapter develops treatment effects as one of our key points of departure, providing a systematic progression through the fundamental estimators for average treatment effects. We begin with regression adjustment and propensity score methods, including balancing scores and inverse probability weighting (IPW), examining their individual strengths and limitations. This foundation leads naturally to the central innovation: double robust (DR) estimation.

Double robust methods, exemplified by the augmented inverse propensity weighting (AIPW) estimator, provide a form of "insurance" against model misspecification by incorporating both outcome modeling

and propensity weighting. These methods remain consistent if either the propensity score or the outcome regression is correctly specified, offering superior robustness compared to approaches that rely on a single modeling strategy.

The chapter then confronts DR estimation with machine learning methods, demonstrating how to use flexible ML techniques to estimate both outcome models $\mu_1(X)$, $\mu_0(X)$ and propensity scores $e(X)$ while implementing cross-fitting to avoid overfitting concerns. We develop approaches for estimating heterogeneous treatment effects using DR principles and address a source of confusion in the literature by clarifying the relationship between DR methods and Double Debiased Machine Learning (DDML). While both approaches combine ML with robustness principles, they represent distinct methodological frameworks with different theoretical foundations and practical implementations.

Chapter 9: The Value Added of Double Machine Learning

Chapter 9 closes Part IV by stepping back from the mechanics of DDML and asking the question that practitioners most need answered: under what conditions does the additional complexity of nonparametric nuisance estimation translate into genuine gains for causal inference, and what does it cost when it does not? The chapter provides a structured assessment of the value added by DDML relative to well-specified parametric benchmarks, drawing the distinction between settings where functional form flexibility is a first-order consideration and settings where it is not.

The central organising argument is that DDML's value is conditional rather than universal. This is not a counsel of caution but a precise diagnostic claim. The method's performance advantage over OLS is concentrated in a specific region of the data-generating process parameter space: where the propensity function $m_0(X) = E[D \mid X]$ or the outcome nuisance function $g_0(X) = E[Y \mid X, D = 0]$ contains genuine nonlinearities that a linear specification would approximate poorly, and where the sample size is large enough for the nonparametric estimator to resolve these nonlinearities against the noise in the data. Outside this region—in the linear regime, or when samples are too small for flexible methods to outperform their parametric competitors—the variance cost of nonparametric estimation is not recovered in bias reduction, and OLS or Double LASSO remain preferable. The chapter formalises this through the residualised treatment variance mechanism: the DDML estimator recovers $\tau_0 = \text{Cov}(Y^*, D^*)/\text{Var}(D^*)$, where the quality of the residualised treatment $D^* = D - \hat{m}(X)$ depends critically on how well the nuisance learner approximates the true propensity surface. When $m_0(X)$ is genuinely nonlinear, a flexible learner produces a residual with higher signal-to-noise ratio than a linear learner, and this improvement flows directly into the causal estimate.

9.1 Value Added: Nuisance Nonlinearity and Learner Divergence

The most informative diagnostic that DDML provides—and one that cannot be recovered from a purely parametric analysis—is learner divergence: systematic differences in the point estimate of τ across learners applied to the same nuisance functions. When a nonparametric learner such as a random forest yields a materially different estimate than a linear learner such as LASSO or elastic net, this divergence is not evidence of instability to be suppressed; it is a signal about the data-generating process. Specifically, it indicates that the propensity or outcome nuisance function contains structure that the linear learner is approximating poorly, and that this approximation error is propagating into the causal estimate. The diagnostic is made precise through the propensity R^2 from each learner: a near-zero R^2 for the linear learner alongside a substantive R^2 for the flexible learner confirms that the treatment assignment mechanism is genuinely nonlinear in the covariates, and that the residualised treatment constructed by the linear learner retains confounding variation that the nonparametric learner has successfully removed. This reframes learner divergence as a structured sensitivity analysis rather than a problem to be resolved by averaging or model selection, and it provides evidence about the DGP that the parametric benchmark alone could not supply. The chapter develops this diagnostic interpretation and shows how a hyperparameter sensitivity analysis across a grid of learner configurations can distinguish genuine DGP-driven divergence from tuning artefacts.

An important corollary is the value of a strong parametric benchmark, constructed before the DDML analysis begins. The chapter emphasises that the parametric specification should be pushed as far as theory and institutional knowledge permit—including interaction terms, polynomial transformations, and theory-motivated functional forms. This provides two things: a ceiling against which to assess the

DDML gains, and a reference point for interpreting learner divergence. If the flexible DDML estimate coincides with the richest parametric specification, the conclusion is informative in its own right: the DGP is well-approximated by the parametric model, and functional form flexibility adds nothing to the average effect estimate. If the two estimates diverge after the richest parametric specification has been constructed, the gap is more credibly attributed to residual nonlinearity than to specification error in the benchmark.

9.2 Value Added: Cross-Fitting Discipline in Heterogeneous Effect Estimation

DDML's cross-fitting requirement, which is the primary architectural cost distinguishing it from Double LASSO, becomes a methodological asset when DDML is embedded within a heterogeneous treatment effect framework. The core issue is the separation between learning and testing heterogeneity. A proxy CATE constructed from the full sample—as a function of the same data used to fit the nuisance learners—will overstate heterogeneity: the learner has seen the outcome residuals for every observation and can exploit this information to create apparent effect variation that reflects noise rather than signal. Cross-fitting addresses this by ensuring that the nuisance predictions for observation i are always formed from a fold that excludes i, so the residuals used to construct the CATE proxy are genuinely out-of-sample. The practical consequence is a sharp reduction in the apparent heterogeneity loading coefficient between the proxy CATE and the outcome residuals—the coefficient collapses from values suggesting strong heterogeneity under the in-sample proxy to near zero under the honest cross-fitted proxy. The chapter argues that this collapse is not a failure but a diagnostic: it correctly identifies that the apparent heterogeneity was an overfitting artefact, and the methods GATE, CLAN, and the best linear predictor (BLP) framework that survive the strict cross-fitted test constitute genuine evidence for heterogeneity. The cross-fitting discipline embedded in DDML is therefore not merely a theoretical requirement for asymptotic validity; it has first-order practical consequences for the reliability of heterogeneous effect findings.

9.3 Value Added: Adaptations for Panel and Structured Data

The standard DDML cross-fitting protocol—random assignment of observations to folds—is valid under independent and identically distributed sampling but requires adaptation when the data have panel or clustered structure. In panel settings, within-unit dependence means that if observations from the same unit appear in both the estimation and evaluation folds, the nuisance learner can exploit within-unit information to reduce its prediction error on the evaluation fold, producing an overly small residual and overstating the precision of the causal estimate. The correct adaptation is block cross-fitting, in which entire units are assigned to folds together, so that the nuisance learner is never evaluated on units it has seen in the estimation fold. This is not a cosmetic modification: the difference between random and block assignment is material in typical panel applications where within-unit variation is a substantial fraction of total variation, and standard implementations of DDML that do not account for the panel structure will overstate precision. The chapter develops the block cross-fitting procedure and shows how it extends naturally to clustered cross-sections. The IV-PLM extension introduces a third nuisance function for the instrument, creating a triple machine learning problem in which block assignment must be applied consistently across all three estimation stages.

9.4 Costs: Variance Inflation and the Insurance Premium

The chapter provides an equally structured account of the costs of DDML. The primary cost is variance inflation relative to OLS when the parametric nuisance specification is correctly specified. This is not a finite-sample peculiarity: it is a consequence of the rate at which nonparametric estimators converge. A correctly specified linear nuisance model estimated by OLS converges at rate $n^{-1/2}$, so the bias from nuisance estimation error is asymptotically negligible. A nonparametric ML estimator converges at a slower rate that depends on the smoothness of g_0 and m_0; under the DDML regularity conditions the product of the two nuisance error rates must be $o(n^{-1/2})$, but the individual nuisance errors are larger than in the parametric case. This larger nuisance estimation error translates into larger variance of the residualised treatment D^* and a wider confidence interval for τ_0. The chapter formalises this variance decomposition: DDML estimation error is inversely proportional to $\text{Var}(D^*)$, the variation in the treatment residual, and flexible nonparametric nuisance estimation systematically reduces $\text{Var}(D^*)$

relative to the linear case when $m_0(X)$ is low-dimensional and well-specified. The result is an "insurance premium" framing: nonparametric DDML offers a performance floor against DGP misspecification at the cost of reduced efficiency when the parametric model is correct. This premium is only recovered when the DGP is sufficiently nonlinear to generate bias in the parametric nuisance estimate that exceeds the additional variance from the nonparametric estimator.

A further cost dimension concerns implementation choices that affect results in ways not always transparent to the practitioner. The clipping of propensity scores for numerical stability removes the extreme-score observations that motivate the use of overlap-reweighted estimators such as FOREST RIESZ; comparisons between overlap-sensitive and standard DDML variants are therefore sensitive to clipping thresholds. The choice of fold number in cross-fitting affects the effective sample size for nuisance estimation at each fold, with five-fold cross-fitting leaving 80% of the sample for nuisance estimation and 20% for evaluation at each step. The interaction between fold number, learner tuning, and sample size determines whether the cross-fitting procedure achieves the rate conditions required for asymptotic validity in a given application, and these interactions cannot be resolved by theory alone in finite samples.

9.5 The Asymptotic–Operative Gap

The chapter closes by confronting the gap between the asymptotic guarantees under which DDML is justified and the finite samples in which it is applied. The regularity conditions for DDML—Neyman orthogonality, nuisance convergence at rate $o(n^{-1/4})$, and Donsker-class restrictions on the nuisance function space—are stated for sequences of sample sizes tending to infinity. In the moderate samples typical of empirical economics, these conditions translate into requirements on sample size, covariate dimensionality, and DGP smoothness that are aspirational rather than verifiable. The consequence is that coverage rates of DDML confidence intervals may fall short of their nominal level in finite samples—not because the procedure is incorrectly implemented but because the approximation from the asymptotic theory to the finite-sample distribution is imperfect. The chapter argues that reporting coverage rates from a specification-matched Monte Carlo exercise alongside the empirical DDML estimates is a methodological standard that applied users of the method should adopt, but rarely do. Where the sample size and covariate structure render the asymptotic approximation credible, DDML provides valid inference; where they do not, confidence intervals should be interpreted with caution irrespective of the nominal coverage implied by the asymptotic theory.

The chapter is supported by a set of illustrative applications using the 401(k) dataset, which provides a setting with a large sample, a binary treatment with a nonlinear assignment mechanism, and a continuous outcome—conditions favourable to DDML—and the food expenditure dataset, which provides a setting with a low-dimensional, well-structured covariate space where the parametric benchmark is competitive with flexible methods. Comparing the two applications makes concrete the conditions under which DDML adds value and those under which it does not. A concluding table, drawn from the empirical literature and from systematic Monte Carlo evidence, provides a practical diagnostic guide mapping data characteristics to the expected relative performance of DDML and its parametric alternatives.

Data characteristic	Implication for DDML	Evidence
Genuine nonlinearity in $m_0(X)$	Learner divergence; flexible learner dominates	401(k) propensity R^2 gap across learners
Low-dimensional, well-structured covariates	ATE converges across methods; no DDML gain	Food expenditure share; ATE $\in [-0.026, -0.024]$
Correctly specified parametric nuisance	Variance inflation; OLS dominates all DDML variants	Monte Carlo: parametric DML dominated by OLS in linear regime
Panel or clustered structure	Block cross-fitting required; standard random splitting overstates precision	China Syndrome application
Small or moderate n	Asymptotic guarantees aspirational; coverage rates may fall short	AJR institutions application ($n = 64$)
Strong DGP nonlinearity, large n	DDML performance frontier shifts against parametric; flexible learner gains	Monte Carlo contour plots (MSE ratio)

Table: Student Project Evidence. The following projects, drawn from the 2026 DS300 cohort, provide the empirical and Monte Carlo evidence synthesised in this chapter. None of the studies cited is referenced individually in the chapter text; all are treated as primary empirical material from which the structured account above is assembled.

Candidate	Title	Primary contribution to chapter
0671Q	When Does Double Machine Learning Outperform OLS? A Monte Carlo Study	Performance frontier via contour plots; Jacobian mechanism for residualised treatment variance
2920H	Double Robustness and Double Machine Learning: Insurance for OLS?	Analytical derivation of DML estimation error; insurance framing; EPS $\times$ TRVAR simulation grid
1947L	When Does Flexibility Add Value? Evidence from 401(k) Eligibility	Learner divergence as DGP diagnostic; R-learner cross-fitting collapse; IHS scale sensitivity
1661J	Revisiting 'The China Syndrome': A Double Debiased Machine Learning Approach	Block K-fold for panel dependence; seven-learner ensemble; PLIV application
1170E	Urban Residence and Food Expenditure Shares: Evidence on Average and Heterogeneous Treatment Effects	ATE convergence under low-dimensional covariates; permutation-adjusted variable importance

6.5 Part V: Tree-Based Methods and Heterogeneity

Contents

Chapter 10: Random Forests

This chapter serves as a key point of departure for understanding machine learning methods for prediction, establishing their fundamental importance for causal inference. Since causal estimators rely on the estimation of reduced-form equations—which are essentially prediction problems—the predictive power of ML methods becomes central to credible causal analysis.

We provide an overview of random forests in anticipation of introducing the architecture of causal and Generalized Random Forests (GRFS) in Chapter 11. Building on the material first introduced in Chapter 2, we examine how nonparametric precursors like KNN, binned estimators, and kernel methods all attempt to estimate regression-like functions by defining neighborhoods around each point of interest. Their key differences lie in how these neighborhoods are defined: KNN fixes the number of observations but allows neighborhood size to vary; binned estimators fix the distance but allow the number of observations to vary; and kernel methods smooth discontinuities by weighting observations based on their distance from the target point.

All these methods are impacted by the curse of dimensionality as the number of covariates increases. Random forests serve as adaptive nonparametric estimators that moderate these challenges through recursive partitioning of the covariate space, ensemble averaging, and implicit feature selection. Unlike fixed kernels which treat all dimensions equally, RFS automatically discover relevant dimensions and interactions, focusing computational resources on the most informative features while creating data-adaptive neighborhoods with variable bandwidth and irregular shapes.

The remainder of the chapter provides a comprehensive overview of random forests for prediction, covering the bias-variance tradeoff in the tree context, bagging as a variance reduction technique, and the two forms of random sampling that underpin random forests: bootstrap sampling of observations and random subsampling of features at each split. We demonstrate how these ensemble methods address the instability and overfitting tendencies of individual trees while maintaining strong predictive performance even in high-dimensional sparse spaces.

The chapter concludes with an application of random forests to predicting house prices using data from the American Housing Survey, comparing performance across multiple algorithms and demonstrating the practical advantages of tree-based ensemble methods in realistic prediction settings with many covariates.

Chapter 11: Causal Forests

This chapter continues the use of ensemble methods but confronts readers with the fundamental challenge of adapting tree architectures from prediction to causal inference. We explore three distinct architectures: random forests (from Chapter 10), causal forests, and generalized random forests (GRFs), with the latter providing a bridge to Chapter 12's extended treatment.

Causal inference presents unique challenges that causal forest architecture must address, building on orthogonalization principles and sample splitting techniques introduced in earlier chapters. The key insight is that causal forests proceed analogously to random forests but with crucial adaptations for the estimand: at each node, the algorithm recursively chooses variables and splitting points, then calculates average treatment effects within partitions. However, the absence of ground truth in causal settings necessitates a fundamental re-optimization of ML algorithms based on variance-penalized mean squared error, which rewards partitions that discover strong heterogeneity in treatment effects while penalizing those that create excessive variance.

The chapter concludes by introducing the architectural evolution represented by Generalized Random Forests, demonstrating a fundamental shift from the original causal tree formulation of Athey and Imbens (2016). While traditional causal trees create discrete partitions with leaf-specific treatment effect estimates, GRFs fundamentally change the approach by using trees to construct continuous weights for each observation. This evolution mirrors the classical progression in nonparametric estimation from binned estimators—which assign discrete zero-one weights to observations within fixed intervals—to kernel estimators that provide smooth, continuous weights based on distance. Just as kernel methods replaced the discontinuous indicator functions of histogram estimators with smooth weighting functions, GRFs replace the discrete leaf assignments of causal trees with forest-based adaptive neighborhoods that assign continuous weights $\alpha_i(\tilde{x})$ representing how similar each observation is to a target point, enabling treatment effects that vary smoothly with covariates.

Chapter 12: Generalised Random Forests

Chapter 11 introduces Generalised Random Forests (GRFs), a fundamental extension of the random forest framework that enables estimation of arbitrary parameters through local moment conditions rather than simple local averages. The key innovation lies in reframing forest-based algorithms as adaptive locally

weighted estimators that construct similarity weights $\alpha_i(x)$ to solve weighted moment conditions of the form $\sum_{i=1}^{N} \alpha_i(x)\psi(O_i;\theta(x)) = 0$, where ψ represents a scoring function that characterizes the parameter of interest $\theta(x)$ through local moment conditions.

The conceptual breakthrough emerges from recognizing that traditional random forests represent merely a special case of this more general framework. Where standard RFs use tree-based partitions to compute local averages $\hat{\mu}(x) = \frac{1}{B}\sum_{b=1}^{B} \hat{\mu}_b(x)$, GRFs construct continuous forest-based weights $\alpha_i(x) = \frac{1}{B}\sum_{b=1}^{B} \mathbf{1}\{X_i \in L_b(x)\}$ that enable estimation of complex parameters beyond simple means. This weight construction creates adaptive neighborhoods where observations appearing in the same terminal nodes as target point x across multiple trees receive higher weights, naturally implementing a data-driven similarity measure.

The framework's power becomes evident in heterogeneous treatment effect estimation, where GRFs implement the R-learner approach through orthogonalisation principles derived from the Frisch-Waugh-Lovell theorem. Rather than requiring causal trees to split simultaneously on variables affecting both treatment propensities and effect heterogeneity—a computationally wasteful approach that risks overfitting—the orthogonalized specification $Y_i - m(X_i) = \tau(X_i)[W_i - p(X_i)] + \varepsilon_i$ separates these estimation tasks. Machine learning methods first estimate nuisance functions $m(X_i) = \mathbb{E}[Y_i|X_i]$ and $p(X_i) = \mathbb{E}[D_i|X_i]$, then GRFs focus exclusively on modeling treatment effect heterogeneity $\tau(X_i)$ using residualized variables.

The resulting local moment condition $\mathbb{E}[(Y_i - m(X_i) - \tau(x)(W_i - p(X_i)))(W_i - p(X_i))|X_i = x] = 0$ yields the weighted estimator $\hat{\tau}(x) = \frac{\sum_i \alpha_i(x)(Y_i - \hat{m}(X_i))(W_i - \hat{p}(X_i))}{\sum_i \alpha_i(x)(W_i - \hat{p}(X_i))^2}$, preserving the familiar covariance structure while enabling heterogeneous effects to vary smoothly across the covariate space. This approach demonstrates how the orthogonalization principle—fundamental to Double Machine Learning—extends naturally to forest-based methods through sample splitting and honest estimation procedures.

Applications to IV estimation reveal GRFs' versatility beyond treatment effects, where the framework handles endogenous binary variables through local moment conditions that circumvent the "forbidden regression" problem plaguing parametric approaches like probit models with 2SLS. The Angrist and Evans fertility instrument study illustrates how IV forests can discover heterogeneous local average treatment effects across different regions of the covariate space, capturing rich patterns of effect variation that parametric specifications constrain to constant average effects. Through nonparametric voting processes for binary outcomes and adaptive neighborhood construction, GRFs provide consistent estimation of complex causal parameters while maintaining the computational advantages and interpretability of tree-based methods.

Chapter 13: From Scalar to Heterogeneous Effects

The evolution from scalar treatment effects to heterogeneous treatment effects represents a fundamental shift in how we understand and estimate causal relationships. This progression encompasses three distinct levels of granularity: scalar effects that provide a single average, group-level effects that capture systematic variation across subpopulations, and fully heterogeneous effects that allow treatment impacts to vary continuously across the covariate space. Each level addresses different policy questions while balancing statistical precision against the richness of the heterogeneity captured.

The Scalar Treatment Effect: A Starting Point

Traditional econometric approaches estimate a single parameter representing the average treatment effect (ATE), providing valuable but limited information about whether an intervention works on average. The scalar treatment effect model takes the familiar form:

$$Y_i = \alpha + \tau d_i + \beta' X_i + \varepsilon_i \tag{9}$$

where τ captures a single, constant effect of treatment d on outcome Y across all individuals. While statistically efficient, this scalar approach obscures potentially rich heterogeneity that is crucial for policy targeting, resource allocation, and economic understanding.

The challenge becomes particularly acute when the distribution of treatment effects is bimodal or highly skewed. An intervention might help 30% of recipients substantially while harming 20% and having no

effect on the remaining 50%. The average effect might appear close to zero, leading to the incorrect conclusion that the intervention is ineffective, when in reality it creates both winners and losers who could potentially be identified ex ante. A job training program might boost earnings dramatically for some workers while having minimal impact on others. Educational interventions may benefit certain student populations while showing no effect for others. Monetary policy might stimulate some sectors while leaving others unchanged.

Group Average Treatment Effects: A Middle Ground

Group Average Treatment Effects (GATEs) provide a practical compromise between the oversimplification of scalar effects and the data requirements of fully individual-level estimation. Following the methodology developed by Chernozhukov et al., the GATE approach uses machine learning methods to first estimate individual treatment effects $S(X_i) = E[Y_i(1) - Y_i(0)|X_i]$, then partitions the population into discrete groups based on these predicted effects.

The standard implementation divides the sample into quartiles, where G_1 represents the 25% of observations with the lowest (most negative) treatment effects, and G_4 represents the 25% with the highest treatment effects. Group-specific treatment effects are then estimated through:

$$Y_i = \alpha_1 B(X_i) + \sum_{k=1}^{K} \gamma_k I(G_k) + \varepsilon_i \tag{10}$$

where $B(X_i)$ represents the baseline outcome function estimated on control observations, $I(G_k)$ indicates group membership, and γ_k captures the average treatment effect for group k.

This approach concentrates statistical power by pooling observations within groups while maintaining meaningful differentiation across the treatment effect distribution. Rather than estimating thousands of individual effects with high variance, or a single average effect that obscures heterogeneity, GATEs identify systematic patterns in how treatments affect different subpopulations. The difference $\gamma_4 - \gamma_1$ provides a powerful test for the presence of heterogeneity, while the individual γ_k parameters reveal whether treatments help some while potentially harming others.

Fully Heterogeneous Effects: The Nonparametric Frontier

The move to fully heterogeneous effects represents the most ambitious extension, defining the conditional average treatment effect (CATE) as a continuous function over the covariate space:

$$\tau(x) = E[Y(1) - Y(0)|X = x] \tag{11}$$

where $Y(1)$ and $Y(0)$ denote potential outcomes under treatment and control. This formulation transforms our estimating equation from the partially linear model to the fully nonparametric specification:

$$Y_i = \alpha(X_i) + \tau(X_i)d_i + \varepsilon_i \tag{12}$$

where both the baseline function $\alpha(\cdot)$ and the treatment effect function $\tau(\cdot)$ vary flexibly with covariates.

The challenges identified for scalar effects become dramatically more severe in this setting. The overlap problem is no longer global but local: we require adequate numbers of treated and control units not just overall, but in each region of the covariate space where we seek to estimate $\tau(x)$. The curse of dimensionality thus manifests doubly: we need sufficient data to estimate complex functions $\alpha(\cdot)$ and $\tau(\cdot)$, while simultaneously requiring strong local overlap to identify treatment effects at each point.

Machine learning methods adapted for heterogeneous effects address these challenges through sophisticated localization strategies. Causal forests, for example, build on the random forest framework but modify the splitting criterion to maximize heterogeneity in treatment effects rather than predictive accuracy. The orthogonalization principles that enabled double machine learning for scalar effects extend to this heterogeneous setting through the R-learner framework, which first estimates and removes the influence of nuisance functions before focusing purely on modeling treatment effect heterogeneity.

Testing and Characterizing Heterogeneity

While it is possible to generate estimates of CATE for arbitrary covariate values with confidence intervals, ultimately we are interested in testing for and describing heterogeneity through interpretable features of the treatment effect function $\tau(x)$. The framework developed by Chernozhukov et al. provides three complementary approaches:

First, an omnibus test for the presence of heterogeneity evaluates whether treatment effects vary systematically across the covariate space, providing a formal statistical test of the null hypothesis that $\tau(x) = \tau$ for all x. This test leverages the difference between the most and least affected groups to detect meaningful variation in treatment response.

Second, the GATE analysis described above provides interpretable summaries of how treatment effects vary across the population while maintaining sufficient statistical power for reliable inference. By grouping individuals based on their predicted treatment effects, we can identify which segments of the population benefit most from treatment and which may be harmed.

Third, classification analysis characterizes which observable characteristics are most associated with treatment effect magnitudes, offering interpretable insights even when the full function $\tau(\cdot)$ is too complex to visualize or summarize. This approach identifies the covariates that best predict whether an individual will have a high or low treatment effect, providing actionable guidance for treatment targeting.

The evolution from scalar to group to fully heterogeneous effects thus represents not merely a technical extension but a fundamental reconceptualization of what we seek to learn from causal analysis. By combining the identification insights of traditional econometrics with the functional approximation capabilities of machine learning, we can move beyond asking "what is the effect?" to understanding "for whom does the effect occur, and why?" – questions that are essential for both economic theory and optimal policy design.

6.6 Part VI: Frontiers

Contents

Chapter 14: An Introduction to Generative AI and Large Language Models

This forward-looking chapter explores emerging applications of generative AI in economics. We discuss how large language models can process unstructured text data, generate synthetic data for privacy-preserving research, and assist in literature review and hypothesis generation.

While maintaining appropriate skepticism about causal claims from correlational patterns in text, we show how LLMs can augment traditional econometric analysis. Applications include measuring economic uncertainty from news text and using GPT models to code policy interventions from legislative documents.