In pursuit of automated statistical inference under minimal assumptions using machine learning tools

Hongxiang Qiu

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2021

Reading Committee:

Marco Carone, Chair

Alex Luedtke, Chair

Noah Simon

Thomas Richardson

Program Authorized to Offer Degree:

Biostatistics

University of Washington

Abstract

In pursuit of automated statistical inference
under minimal assumptions using machine learning tools

Hongxiang Qiu

Co-Chairs of the Supervisory Committee:
Associate Professor Marco Carone
Department of Biostatistics

Assistant Professor Alex Luedtke
Department of Statistics

This dissertation consists of three projects aiming for automated statistical inference under minimal assumptions using machine learning tools. In the first project, we developed two sieve-like methods to construct asymptotically efficient plug-in estimators under nonparametric models. Compared to existing methods, they require less expertise in semiparametric efficiency theory in implementation or rely on weaker smoothness assumptions than traditional sieve estimation and kernel-based methods. In the second project, we studied estimation and evaluation of optimal individualized intervention rules under treatment resource constraints with an instrumental variable (IV). We separately consider intervention on the treatment and the causal IV. We proposed to utilize machine learning tools to estimate optimal rules and efficient plug-in estimators of average causal effects of optimal rules under locally nonparametric models. In the third project, we studied estimation under rich (potentially locally nonparametric) models while utilizing prior information. We proposed to use Gamma-minimax estimators, which minimizes the worst-case Bayes risk over a set of prior distributions that are consistent with prior information. We also proposed to use neural networks to parameterize the class of candidate estimators and developed algorithms to compute an approximate Gamma-minimax estimator with theoretical convergence guarantees.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

It has been a long journey to this dissertation. I would like to thank the great number of people who supported me along this journey and made it an amazing learning experience.

I want to first thank my dissertation advisors, Marco Carone and Alex Luedtke. In the spring quarter of my first year as a PhD student, Marco taught me about semiparametric theory in his special topics course and a weekly one-hour meeting. This led to a summer project, and Alex joined us in this project in the coming summer. Although this project itself did not turn into a paper, I learned a lot from the process, which turned out to be helpful in my first dissertation project. I cannot be more thankful to both Marco and Alex for investing so much time and energy to me, for inspiring me with interesting projects and new ideas, for providing me various opportunities to contribute to the statistical community, and for giving me numerous suggestions on writing, presentation and career. Throughout my graduate studies, they have been great mentors, collaborators and friends.

I thank my dissertation committee, Marco, Alex, Noah Simon, Thomas Richardson and Yanqin Fan. You have all made the exams more bearable. I thank Lurdes Inoue for serving as my academic advisor during my first two years and helping me get comfortable with graduate studies. As a person with a poor sense of direction, I also want to thank her for being patient and twice guiding me from her office in the F-wing through the puzzling Health Sciences Building to Gitana's office in the H-wing, when I first came to UW. I thank Scott Emerson for his inspiring talks on statistics, his conversations with me even if he did not teach me any course, and his insights during

seminars that made the seminars slightly easier to digest.

It is incredible that I am involved in the PROUD trial at KPWRHI as an RA for all these five years. I thank Jennifer Bobb for mentoring me and other collaborators including Denise Boudreau, Kathy Bradley, Onchee Yu, Abisola Idu, Jane Grafton, Rachael Burganowski, Megan Addis, Paige Wartko, Lawrence Madziwa, Chester Pabiniak, Gwen Lapham and others on the PROUD team (no matter currently or previously), for supporting me. You have all made my RAship a great opportunity to gain hands-on experience on pragmatic cluster-randomized trials and electronic health records.

I thank Gitana Garofalo for her various support. She did an extrodinary job in supporting students and made every effort to improve our learning experience. The picture of cherry blossoms at UW she sent to me was one reason for me to choose to come to UW. I also thank my cohort and the cohort that is one year more senior than mine, for supporting me through the coursework stress and teaching me about statistics and life. You all made my first two years of coursework more bearable and fruitful. I thank Parker Xie for being my gym buddy during my third year (and introducing me to Riichi Mahjong!).

Finally, I am thankful to all the people who supported me throughout my life. I thank my math teacher in my middle school, Rui Chen, who made learning math so much fun. I thank Prof. CHAN Hon Fu Raymond for encouraging me to study important (and typically difficult) problems when I was an undergraduate student at CUHK. I am thankful to my mom, dad, grandparents and uncle for their love and constant faith in me. My final thanks go to Tian Zheng for being my life partner and supporting me through my final days as a PhD student.

DEDICATION

To my parents and grandparents:

Wengang Cai

Yuankang Qiu

Lingzhu Hu

Xicong Cai

Chapter 1

INTRODUCTION

It is often of scientific interest to estimate an aspect of the data-generating mechanism underlying the data at hand. Traditionally, parametric or semiparametric models are typically used to model the data-generating mechanism. Although these models may be easy to interpret, they impose strong assumptions on the data-generating mechanism. When these assumptions fail to hold, the corresponding analysis results may no longer be valid or interpretable. Therefore, in this dissertation, we explore methods for statistical inference under minimal assumptions, especially under nonparametric models. There have been numerous statistical methods along this direction, but many are not fully automated. They may require specialized expertise in statistical theory to implement, or provide limited guidance on the choice of tuning parameters. We also explore approaches to automate such statistical procedures in this dissertation. In this chapter, we briefly introduce three projects aiming for automated statistical inference under minimal assumptions using machine learning tools.

1.1 Universal sieve-based strategies for efficient estimation using machine learning tools

In this project, we studied the general problem of efficient estimation under nonparametric models and developed methods that partially overcome shortcomings of existing methods.

Many statistical estimands are summaries of the data-generating mechanism that involve function-valued features of the distribution that cannot be estimated at a parametric rate under a nonparametric model — for example, a regression function or the density function of the distribution. Examples of useful summaries involving such features include average treatment effects (Rubin, 1974), average derivatives (Härdle and Stoker, 1989), moments of

the conditional mean function (Shen, 1997), variable importance measures (Williamson et al., 2020) and treatment effect heterogeneity measures (Levy et al., 2018). For such estimands, it is natural to consider plug-in estimators based on flexible estimates of the functional features. However, in general, such estimators are not asymptotically efficient and not asymptotically normal.

Several methodological frameworks have been proposed to address this issue. The targeted minimum loss-based estimation (TMLE) framework provides a means of constructing efficient plug-in estimators (van der Laan and Rubin, 2006; van der Laan and Rose, 2018) based on an (almost arbitrary) initial machine learning fit. Although it is widely applicable, it is targeted toward the summary of interest and hence different adjusted fits are needed for different estimands. Its implementation requires specialized expertise, namely knowledge of the analytic expression for an influence function of the summary of interest, which may be difficult to derive. Furthermore, even when an influence function is known analytically, additional expertise is needed to construct a TMLE for a given problem.

Other more traditional methods include the use of undersmoothing (e.g., Newey et al., 1998), twicing kernels (Newey et al., 2004), and sieves (e.g., Chen, 2007; Newey, 1997; Shen, 1997). These methods neither require knowing an influence function nor performing any targeting of the function-valued feature estimates, and hence may be easier to implement. However, they all rely on smoothness conditions on derivatives of the functional features that may be overly stringent. In addition, the twicing kernel method requires the use of a higher-order kernel, which may lead to poor performance in small to moderate samples (Marron, 1994).

These methods may also require the user to choose tuning parameters. Though appropriate rates of the tuning parameters that lead to efficient plug-in estimators haven been thoroughly studied (e.g., Newey, 1997; Newey et al., 1998; Shen, 1997; Chen, 2007), these results only provide minimal guidance for applications because there is no indication on how to select the actual tuning parameter for a given data set. For undersmoothing, cross-validation (CV) is known to be invalid; for sieve estimation, CV is often used, but, to the

best of our knowledge, there is no theoretical guarantee that CV leads to efficient plug-in estimators. Among these methods, sieve estimation stands out because CV may select a valid tuning parameter and hence we may expect this method to be automated.

In Chapter 2, we present two sieve-like approaches that lead to efficient plug-in estimators and can partially overcome these shortcomings. Both approaches avoid the need for the analytic expression of the influence function and rely on a more general smoothness assumption than those required by kernel-based and traditional sieve-based methods. The first approach is to estimate the unknown functional feature with Highly Adaptive Lasso (HAL) (Benkeser and van Der Laan, 2016; van Der Laan, 2017) with a carefully chosen tuning parameter. The choice of this tuning parameter in a given data set may be guided by CV. The second approach is to estimate the unknown functional feature with data-adaptive series based on an initial machine learning fit, where the number of terms in the series may be selected by CV.

1.2 Optimal individualized decision rules using instrumental variable methods

In this project, we applied TMLE to the problem of estimating and evaluating optimal individualized decision rules under a locally nonparametric model in the context of instrumental variables.

It is common to evaluate treatment strategies based on the mean of an outcome of interest. However, when several treatment strategies result in qualitatively different effects across subgroups, it may be beneficial to provide individualized treatment strategies based on estimated subgroup-specific treatment effects. Such a treatment strategy is commonly referred to as an individualized treatment rule (ITR). Typically, the objective of an ITR is to optimize the mean outcome, and such an ITR is referred to as an optimal ITR. Estimation of optimal ITRs and of the counterfactual mean outcome under an optimal ITR has been extensively studied (e.g., Murphy, 2003; Robins, 2004; Zhao et al., 2012; Chakraborty and Moodie, 2013; Luedtke and van der Laan, 2016b), but existing approaches mostly require that all confounders are available. Unfortunately, this condition may not hold in observational

studies.

Sometimes, an instrumental variable (IV) — a factor that affects the outcome only via the treatment and is independent of all treatment-outcome confounders — is available. Examples of IVs include encouragement to take a treatment or geographical location that affects a patient's access to treatment (see Baiocchi et al., 2014, for a review). There has been rich literature on estimation of average treatment effect via IVs, existing works focus on fixed *a priori* treatment strategies rather than estimating an optimal ITR and its average effect.

In this project, we studied the problem of estimating an optimal ITR via an IV. We also studied the case where the IV is intervened on when the IV is an encouragement for the treatment. In this case, we studied estimation of an optimal individualized encouragement rule (IER). To account for real-world resource limitations, we also explicitly incorporated treatment resource constraints taking a similar form as in Luedtke and van der Laan (2016a) in the non-IV setting. Incorporating treatment resource constraints is different from Luedtke and van der Laan (2016a) when intervening on encouragement because treatment resources are manipulated only through the encouragement rather than the treatment itself. For each scenario, we also proposed asymptotically linear nonparametric estimators and confidence intervals for the average causal effect of an optimal rule relative to a prespecified reference rule.

1.3 Leveraging vague prior information in general models via iteratively constructed Gamma-minimax estimators

In this project, we shifted away from *asymptotic* theory for estimation under nonparametric/semiparametric models and explored methods to obtain good estimators under rich models in finite samples. Moreover, we aimed for methods that may outperform existing methods by incorporating vague prior knowledge. In terms of Bayesian statistics, such prior knowledge might not correspond to one single prior distribution due to the richness of the model space.

The theoretical framework that motivated us is Gamma-minimax estimation. Gamma-

minimax estimators minimize the worst-case Bayes risk over a set Γ of prior distributions (or priors for short). Such estimators have been studied in a variety of problems. Some explicit forms of Gamma-minimax estimators are given for particular parametric models (e.g., Olman and Shmundak, 1985; Eichenauer-Herrmann, 1990; Eichenauer-Herrmann et al., 1994; Chen et al., 1988, 1991). Unfortunately, these results only apply to certain parametric models and restrictive classes of prior sets, and hence are not general. One possible reason is that it is typically intractable to analytically derive Gamma-minimax estimators. On the other hand, algorithms to compute minimax or Gamma-minimax estimators have been proposed (e.g., Nelson, 1966; Kempthorne, 1987; Bryan et al., 2007; Schafer and Stark, 2009; Noubiap and Seidel, 2001). However, these works focus on parametric models, which may be a stringent assumption itself. More recent works explored such algorithms under more general models. For example, Luedtke et al. (2020a) and Luedtke et al. (2020b) used an approach termed Adversarial Monte Carlo meta-learning to construct minimax estimators.

In this project, we defined Gamma-minimaxity in general models. We proposed to use Gamma-minimax estimation as a means to incorporate prior knowledge by setting Γ to be the set of all priors that are consistent with prior knowledge under general models. We then proposed iterative algorithms to compute Gamma-minimax estimators for general models and a set Γ of priors constrained by generalized moments with theoretical convergence guarantees. We utilized recent advances in neural networks, especially adversarial learning (e.g., Goodfellow et al., 2014; Luedtke et al., 2020a; Luedtke et al., 2020b), when specifying the space of candidate estimators, and theoretically showed that such parameterizations can achieve good performance.

Chapter 2

UNIVERSAL SIEVE-BASED STRATEGIES FOR EFFICIENT ESTIMATION USING MACHINE LEARNING TOOLS

2.1 Introduction

2.1.1 Motivation

A common statistical problem consists of using available data in order to learn about a summary of the underlying data-generating mechanism. In many cases, this summary involves function-valued features of the distribution that cannot be estimated at a parametric rate under a nonparametric model — for example, a regression function or the density function of the distribution. Examples of useful summaries involving such features include average treatment effects (Rubin, 1974), average derivatives (Härdle and Stoker, 1989), moments of the conditional mean function (Shen, 1997), variable importance measures (Williamson et al., 2020) and treatment effect heterogeneity measures (Levy et al., 2018). For ease of implementation and interpretation, in traditional approaches to estimation, these features have typically been restricted to have simple forms encoded by parametric or restrictive semiparametric models. However, when these models are misspecified, both the interpretation and validity of subsequent inferences can be compromised. To circumvent this difficulty, investigators have increasingly relied on machine learning (ML) methods to flexibly estimate these function-valued features.

Once estimates of the function-valued features are obtained, it is natural to consider plug-in estimators of the summary of interest. However, in general, such estimators are not root-n-consistent and asymptotically normal, and hence not asymptotically efficient (referred to as *efficient* henceforth). Lacking this property is problematic since it often forms the basis for constructing valid confidence intervals and hypothesis tests (Bickel and Ritov, 2003; Newey

et al., 2004). When the function-valued features are estimated by ML methods, in order for the plug-in estimator to be CAN, the ML methods must not only estimate the involved function-valued features well, but must also satisfy a small-bias property with respect to the summary of interest (Newey et al., 2004; van der Laan and Rose, 2018). Unfortunately, because ML methods generally seek to optimize out-of-sample performance, they seldom satisfy the latter property.

2.1.2 Existing methodological frameworks

The targeted minimum loss-based estimation (TMLE) framework provides a means of constructing efficient plug-in estimators (van der Laan and Rubin, 2006; van der Laan and Rose, 2018). Given an (almost arbitrary) initial ML fit that provides a good estimate of the function-valued features involved, TMLE produces an adjusted fit such that the resulting plug-in estimator has reduced bias and is efficient. This adjustment process is referred to as targeting since a generic estimate of the function-valued features is modified to better suit the goal of estimating the summary of interest. Though TMLE provides a general template for constructing efficient estimators, its implementation requires specialized expertise, namely knowledge of the analytic expression for an influence function of the summary of interest. Influence functions arise in semiparametric efficiency theory and are key to establishing efficiency, but can be difficult to derive. Furthermore, even when an influence function is known analytically, additional expertise is needed to construct a TMLE for a given problem.

Alternative approaches for constructing efficient plug-in estimators have been proposed in the literature, including the use of undersmoothing (Newey et al., 1998), twicing kernels (Newey et al., 2004), and sieves (Chen, 2007; Newey, 1997; Shen, 1997). These methods neither require knowing an influence function nor performing any targeting of the function-valued feature estimates. Hence, the same fits can be used to simultaneously estimate different summaries of the data-generating distribution, even if these summaries were not pre-specified when obtaining the fit. These approaches also circumvent the difficulties in obtaining an influence function. However, these methods all rely on smoothness conditions

on derivatives of the functional features that may be overly stringent. In addition, undersmoothing provides limited guidance on the choice of the tuning parameter; the twicing kernel method requires the use of a higher-order kernel, which may lead to poor performance in small to moderate samples (Marron, 1994).

In contrast, under some conditions, sieve estimation can produce a flexible fit with the optimal out-of-sample performance while also yielding an efficient — and therefore root-n-consistent and asymptotically normal — plug-in estimator (Shen, 1997). In this chapter, we focus on extensions of this approach. In sieve estimation, we first assume that the unknown function falls in a rich function space, and construct a sequence of approximating subspaces indexed by sample size that increase in complexity as sample size grows. We require that, in the limit, the functions in the subspaces can approximate any function in the rich function space arbitrarily well. These approximating subspaces are referred to as *sieves*. By using an ordinary fitting procedure that optimizes the estimation of the function-valued feature within the sieve, the bias of the plug-in estimator can decrease sufficiently fast as the sieve grows in order for that estimator to be efficient. Thus sieve estimation requires no explicit targeting for the summary of interest.

The series estimator is one of the best known and most widely used sieve techniques. These sieves are taken as the span of the first finitely many terms in a basis that is chosen by the user to approximate the true function well. Common choices of the basis include polynomials, splines, trigonometric series and wavelets, among others. However, series estimators usually require strong smoothness assumptions on derivative of the unknown function in order for the flexible fit to converge at a sufficient rate to ensure the resulting plug-in estimator is efficient. As the dimension of the problem increases, the smoothness requirement may become prohibitive. Moreover, even if the smoothness assumption is satisfied, a prohibitively large sample size may be needed for some series estimators to produce a good fit. For example, if the unknown function is smooth but is a constant over a region, estimation based on a polynomial series can perform poorly in small to moderate samples.

Series estimators may also require the user to choose the number of terms in the series

in such a way that results in a sufficient convergence rate. The rates at which the number of terms should grow with sample size have been thoroughly studied (e.g. Chen (2007); Newey (1997); Shen (1997)). However, these results only provide minimal guidance for applications because there is no indication on how to select the actual number of terms for a given sample size. In practice, the number of terms is the series is often chosen by CV. Upper bounds on the convergence rate of the series estimator as a function of sample size and the number of terms have been derived, and it has been shown that the optimal number of terms that minimizes the bound can also lead to an efficient plug-in estimator (Shen, 1997). However, CV tends to select the number of terms that optimizes the actual convergence rate (van der laan and Dudoit, 2003), which may differ from the number of terms minimizing the derived bound on the convergence rate. Even though the use of CV-tuned sieve estimators has achieved good numerical performance, to the best of our knowledge, there is no theoretical guarantee that they lead to an efficient plug-in estimator.

Two variants of traditional series estimators were proposed in Bickel and Ritov (2003). These methods can use two bases to approximate the unknown function-valued features and the corresponding gradient separately, whereas in traditional series estimators, only one basis is used for both approximations. Consequently, these variants may be applied to more general cases than traditional series estimators. However, like traditional series estimators, they also suffer from the inflexibility of the pre-specified bases.

2.1.3 Contributions and organization of this article

In this chapter, we present two approaches that can partially overcome these shortcomings.

1. *Estimating the unknown function with Highly Adaptive Lasso (HAL)* (Benkeser and van Der Laan, 2016; van Der Laan, 2017).

   If we are willing to assume the unknown functions have a finite variation norm, then they may be estimated via HAL. If the tuning parameter is chosen carefully, then we may obtain an efficient plug-in estimator. This method can help overcome the stringent

smoothness assumptions on derivatives that are required by existing series estimators, as we discussed earlier.

2. *Using data-adaptive series based on an initial ML fit.*

   As long as the initial ML algorithm converges to the unknown function at a sufficient rate, we show that, for certain types of summaries, it is possible to obtain an efficient plug-in estimator with a particular data-adaptive series. The smoothness assumption on the unknown function can be greatly relaxed due to the introduction of the ML algorithm into the procedure. Moreover, for summaries that are highly smooth, we show that the number of terms in the series can be selected by CV.

Although the first approach is not an example of sieve estimation, both approaches are motivated by the sieve literature and can be shown to lead to asymptotically efficient plug-in estimators using the sieve estimation theory derived in Shen (1997). The flexible fits of the functional features from both approaches can be plugged in for a rich class of estimands.

We remark that, although we do not have to restrict ourselves to the plug-in approach in order to construct an asymptotically efficient estimator, other estimators do not overcome the shortcomings described in Section 2.1.2 and can have other undesirable properties. For example, the popular one-step correction approach (also called debiasing in the recent literature on high-dimensional statistics) (Pfanzagl, 1982) constructs efficient estimators by adding a bias reduction term to the plug-in estimator. Thus, it is not a plug-in estimator itself, and as a consequence, one-step estimators may not respect known constraints on the estimand — for example, bounds on a scalar-valued estimand (e.g., the estimand is a probability and must lie in $[0, 1]$) or shape constraints on a vector-valued estimand (e.g., monotonicity constraints). This drawback is also typical for other non-plug-in estimators, such as those derived via estimating equations (van der Laan and Robins, 2003) and double machine learning (Chernozhukov et al., 2017, 2018). Additionally, as with the other procedures described above, the one-step correction approach requires the analytic expression of an influence function.

This chapter is organized as follows. We introduce the problem setup and notation in Section 2.2. We consider plug-in estimators based on HAL in Section 2.3, data-adaptive series in Section 2.4, and its generalized version that is applicable to more general summaries in Section 2.5. Section 2.6 concludes with a discussion. Technical proofs of lemmas and theorems (Appendix A.4), simulation details (Appendix A.5) and other additional details are provided in the Appendix.

2.2 Problem setup and traditional sieve estimation review

Suppose we have independent and identically distributed observations $V_1, \ldots, V_n$ drawn from P_0. Let Θ be a class of functions, and denote by $\theta_0 \in \Theta$ a (possibly vector-valued) functional feature of P_0 — for example, θ_0 may be a regression function. Throughout this chapter, we assume that the generic data unit is $V = (X, Z) \sim P_0$, where X is a (possibly vector-valued) random variable corresponding to the argument of θ_0, and Z may also be a vector-valued random variable. In some cases $V = X$ and Z is trivial. We use $\mathcal{X}$ to denote the support of X. The estimand of interest is a finite-dimensional summary $\Psi(\theta_0)$ of θ_0. We consider a plug-in estimator $\Psi(\hat{\theta}_n)$, where $\hat{\theta}_n$ is an estimator of θ_0, and aim for this plug-in estimator to be asymptotically linear, in the sense that $\Psi(\hat{\theta}_n) = \Psi(\theta_0) + n^{-1}\sum_{i=1}^{n} \mathrm{IF}(V_i) + o_p(n^{-1/2})$ with IF an influence function satisfying $\mathrm{E}_{P_0}[\mathrm{IF}(V)] = 0$ and $\mathrm{E}_{P_0}[\mathrm{IF}(V)^2] < \infty$. This estimator is efficient under a nonparametric model if the estimator is also regular. By the central limit theorem and Slutsky's theorem, it follows that $\Psi(\hat{\theta}_n)$ is a CAN estimator of $\Psi(\theta_0)$, and therefore, $\sqrt{n}[\Psi(\hat{\theta}_n) - \Psi(\theta_0)] \xrightarrow{d} N(0, \mathrm{E}_{P_0}[\mathrm{IF}(V)^2])$. This provides a basis for constructing valid confidence intervals for $\Psi(\theta_0)$.

We now list some examples of such problems.

Example 1. Moments of the conditional mean function (Shen, 1997): Let $\theta_0 : x \mapsto \mathrm{E}_{P_0}[Z|X = x]$ be the conditional mean function. The κ-th moment of $\theta_0(X)$, $X \sim P_0$, namely $\Psi_\kappa(\theta_0) = \mathrm{E}_{P_0}[\theta_0^\kappa(X)]$, can be a summary of interest. The values of $\Psi_1(\theta_0)$ and $\Psi_2(\theta_0)$ are useful for defining the proportion of $\mathrm{Var}_{P_0}(Z)$ that is explained by X, which may be

written as $\mathrm{Var}_{P_0}(\theta_0(X))/\mathrm{Var}_{P_0}(Z)$. This proportion is a measure of variable importance (Williamson et al., 2020). Generally, we may consider $\Psi(\theta_0) = \mathrm{E}_{P_0}[f(\theta_0(X))]$ for a fixed function f.

Example 2. Average derivative (Härdle and Stoker, 1989): Let X follow a continuous distribution on $\mathbb{R}^d$ and $\theta_0 : x \mapsto \mathrm{E}_{P_0}[Z|X = x]$ be the conditional mean function. Let θ_0' denote the vector of partial derivatives of θ_0. Then $\Psi(\theta_0) = \mathrm{E}_{P_0}[\theta_0'(X)]$ summarizes the overall (adjusted) effect of each component of X on Y. Under certain conditions, we can rewrite $\Psi(\theta_0) = \mathrm{E}_{P_0}[\theta_0(X)p_0'(X)/p_0(X)]$, where p_0 is the Lebesgue density of X and p_0' is the vector of partial derivatives of p_0. This expression clearly shows the important role of the Lebesgue density of X in this summary.

Example 3. Mean counterfactual outcome (Rubin, 1974): Suppose that $Z = (A, Y)$ where A is a binary treatment indicator and Y is the outcome of interest. Let $\theta_0 : x \mapsto \mathrm{E}_{P_0}[Y|A = 1, X = x]$ be the outcome regression function under treatment value 1. Under causal assumptions, the mean counterfactual outcome corresponding to the intervention that assigns treatment 1 to the entire population can be nonparametrically identified by the G-computation formula $\Psi(\theta_0) = \mathrm{E}_{P_0}[\theta_0(X)]$.

Example 4. Treatment effect heterogeneity measures (Levy et al., 2018): Similarly to Example 3, suppose that A is a binary treatment indicator and Z is the outcome of interest. Let $\theta_0 = (\mu_{00}, \mu_{01})^\top$, where $\mu_{0a} : x \mapsto \mathrm{E}_{P_0}[Z|A = a, X = x]$ is the outcome regression function for treatment arm $a \in \{0, 1\}$. Then, $\Psi(\theta_0) = \mathrm{Var}_{P_0}(\mu_{01}(X) - \mu_{00}(X))$ is an overall summary of treatment effect heterogeneity.

To obtain an asymptotically linear plug-in estimator, $\hat{\theta}_n$ must converge to θ_0 at a sufficiently fast rate and approximately solve an estimating equation to achieve the small bias property with respect to the summary of interest (Newey et al., 2004; van Der Laan, 2017; van der Laan and Rose, 2018). For simplicity, we assume the estimand to be scalar-valued — when the estimand is vector-valued, we can treat each entry as a separate estimand,

and the plug-in estimators of all entries are jointly asymptotically linear if each estimator is asymptotically linear. Therefore, this leads to no loss in generality if the same fits are used for all entries in the summary of interest.

Sieve estimation allows us to obtain an estimator $\Psi(\hat{\theta}_n)$ with the small bias property with respect to $\Psi(\theta_0)$ while maintaining the optimal convergence rate of $\hat{\theta}_n$ (Chen, 2007; Shen, 1997). The construction of sieve estimators is based on a sequence of approximating spaces Θ_n to Θ. These approximating spaces are referred to as *sieves*. Usually Θ_n is much simpler than Θ to avoid over-fitting but complex enough to avoid under-fitting. For example, Θ_n can be the space of all polynomials with degree K or splines with K knots with $K = K(n) \to \infty$ as $n \to \infty$. In this chapter, with a loss function ℓ such that $\theta_0 \in \operatorname{argmin}_{\theta\in\Theta} \mathrm{E}_{P_0}[\ell(\theta)(V)]$, we consider estimating θ_0 by minimizing an empirical risk based on ℓ, i.e., $\hat{\theta}_n \in \operatorname{argmin}_{\theta\in\Theta_n} n^{-1}\sum_{i=1}^n \ell(\theta)(V_i)$. Under some conditions, the growth rate of Θ_n can be carefully chosen so that $\Psi(\hat{\theta}_n)$ is an asymptotically linear estimator of $\Psi(\theta_0)$ while $\hat{\theta}_n$ converges to θ_0 at the optimal rate.

Throughout this chapter, for a probability distribution P and an integrable function f with respect to P, we define $Pf := \int f(v)dP(v) = \mathrm{E}_P[f(V)]$. We use P_n to denote the empirical distribution. We take $\langle\cdot,\cdot\rangle$ to be the $L^2(P_0)$-inner product, i.e., $\langle\theta_1,\theta_2\rangle = P_0(\theta_1\theta_2)$, where $L^2(P_0)$ is the set of real-valued P_0-squared-integrable functions defined on the support of P_0. When the functions are vector-valued, we take $\langle\theta_1,\theta_2\rangle = P_0(\theta_1^\top\theta_2)$. We use $\|\cdot\|$ to denote the induced norm of $\langle\cdot,\cdot\rangle$. We assume that $\Theta \subseteq L^2(P_0)$. We remark that we have committed to a specific choice of inner product and norm to fix ideas; other inner products can also be adopted, and our results will remain valid upon adaptation of our upcoming conditions. We discuss this explicitly via a case study in Appendix A.1.

For the methods we propose in this article, we assume that Θ is convex. Throughout this chapter, we will further require a set of conditions similar to those in Shen (1997). For any $\theta \in \Theta$, let $\ell_0'[\theta-\theta_0](v) := \lim_{\delta\to 0}[\ell(\theta_0+\delta(\theta-\theta_0))(v) - \ell(\theta_0)(v)]/\delta$ be the Gâteaux derivative of ℓ at θ_0 in the direction $\theta-\theta_0$ and $r[\theta-\theta_0](v) := \ell(\theta)(v) - \ell(\theta_0)(v) - \ell_0'[\theta-\theta_0](v)$ be the corresponding remainder.

Condition A1 (Linearity and boundedness of Gâteaux derivative operator of loss function)**.** For all $\theta \in \Theta$, $\ell_0'[\theta-\theta_0]$ exists and $\ell_0'[\theta-\theta_0](v) - P_0\ell_0'[\theta-\theta_0]$ is linear and bounded in $\theta-\theta_0$.

Condition A2 (Local quadratic behavior of loss function)**.** There exists a constant $\alpha_{0,\ell} \in (0,\infty)$ such that, for all $\theta \in \Theta$ such that $P_0\{\ell(\theta)-\ell(\theta_0)\}$ or $\|\theta-\theta_0\|$ is sufficiently small, it holds that $P_0\{\ell(\theta)-\ell(\theta_0)\} = \alpha_{0,\ell}\|\theta-\theta_0\|^2/2 + o(\|\theta-\theta_0\|^2)$.

Remark 1. We now present an equivalent form of A2 that may be easier to verify in practice. For all $\theta \in \Theta\backslash\{\theta_0\}$, define $h_\theta := (\theta-\theta_0)/\|\theta-\theta_0\|$ and $a_\theta := \frac{d^2}{d\delta^2}P_0\ell(\theta_0+\delta h_\theta)|_{\delta=0}$. Requiring Condition A2 is equivalent to requiring that $a_{\theta_1} = a_{\theta_2}$ for all $\theta_1, \theta_2 \in \Theta\backslash\{\theta_0\}$ and that

$$\sup_{\theta\in\Theta}\left|P_0\ell(\theta_0+\delta h_\theta) - P_0\ell(\theta_0) - \frac{a_\theta}{2}\right| = o(\delta^2).$$

Moreover, if A2 holds, then, for any $\theta \in \Theta\backslash\{\theta_0\}$, it is true that $\alpha_{0,\ell} = a_\theta$.

A large class of loss functions satisfy Conditions A1 and A2. For example, in the regression setting where Z is the outcome, the squared-error loss $\ell(\theta) : v \mapsto [z-\theta(x)]^2$ and the logistic loss $\ell(\theta) : v \mapsto -z\theta(x) + \log\{1+\exp(\theta(x))\}$ both satisfy these conditions; a negative working log-likelihood usually also satisfies these conditions. In Examples 1–4, the unknown functions are all conditional mean functions, which can be estimated with the above loss functions. Thus, Conditions A1 and A2 hold. Examples 3 and 4 require a slight modification discussed in more details in Appendix A.1. We also note that Condition A2 is sufficient for Condition B in Shen (1997).

Condition A3 (Differentiability of summary of interest)**.** $\Psi_{\theta_0}'[\theta-\theta_0] := \lim_{\delta\to 0}[\Psi(\theta_0+\delta(\theta-\theta_0)) - \Psi(\theta_0)]/\delta$ exists for all $\theta \in \Theta$ and is a linear bounded operator.

If Condition A3 holds, then, by the Riesz representation theorem, $\Psi_{\theta_0}'[\theta-\theta_0] = \langle\theta-\theta_0, \dot\Psi\rangle$ for a gradient function $\dot\Psi = \dot\Psi_{\theta_0}$ in the completion of the space spanned by $\Theta-\theta_0 := \{x \mapsto \theta(x)-\theta_0(x) : \theta \in \Theta\}$.

Condition A4 (Locally quadratic remainder)**.** There exists a constant $C > 0$ so that, for all θ with sufficiently small $\|\theta-\theta_0\|$, it holds that

$$|\Psi(\theta) - \Psi(\theta_0) - \Psi_{\theta_0}'[\theta-\theta_0]| \le C\|\theta-\theta_0\|^2.$$

The above condition states that the remainder of the linear approximation to Ψ is locally bounded by a quadratic function.

Conditions A3 and A4 hold for Examples 1–4. For the generalized moment of the conditional mean function in Example 1, it holds that $\dot{\Psi} = f' \circ \theta_0$. For the average derivative of the conditional mean function in Example 2, it holds that $\dot{\Psi} = p_0'/p_0$. For the average treatment effect and the treatment effect heterogeneity measure in Examples 3 and 4, as we show in Appendix A.1, $\dot{\Psi}$ also exists and depends on the propensity score function $x \mapsto P_0(A = 1 \mid X = x)$.

2.3 Estimation with Highly Adaptive Lasso

2.3.1 Brief review of Highly Adaptive Lasso

Recently, the Highly Adaptive Lasso (HAL) was proposed as a flexible ML algorithm that only requires a mild smoothness condition on the unknown function and has a well-described implementation (Benkeser and van Der Laan, 2016; van Der Laan, 2017). In this subsection, we briefly review HAL. We first heuristically introduce its definition and desirable properties, and then introduce the definition and implementation more formally. For ease of presentation, for the moment, we assume that θ_0 is real-valued.

In HAL, θ_0 is assumed to fall in the class of càdlàg functions (right-continuous with left limits) defined on $\mathcal{X} \subseteq \mathbb{R}^d$ with variation norm bounded by a finite constant M. In this section, we denote this function class by $\Theta_{\mathrm{v},M}$. The variation norm of a càdlàg function θ, denoted by $\|\theta\|_{\mathrm{v}}$, characterizes the total variability of θ as its argument ranges over the domain, so $\|\cdot\|_{\mathrm{v}}$ is a global smoothness measure and $\Theta_{\mathrm{v},M}$ is a large function class that even contains functions with discontinuities. Fig. 2.1 presents some examples of univariate càdlàg functions with finite variation norms for illustration. Because $\Theta_{\mathrm{v},M}$ is a rich class, it can be plausible that $\theta_0 \in \Theta_{\mathrm{v},M}$ for some $M < \infty$. The HAL estimator of θ_0 is then $\hat{\theta}_n = \hat{\theta}_{n,M} \in \mathrm{argmin}_{\theta \in \Theta_{\mathrm{v},M}} n^{-1} \sum_{i=1}^n \ell(\theta)(V_i)$. Under this assumption, it has been shown that $\|\hat{\theta}_n - \theta_0\| = o_p(n^{-1/4})$ regardless of the dimension of X under additional mild conditions (van

Der Laan, 2017). Thus, estimation with HAL replaces the usual smoothness requirement on derivatives of traditional series estimators by a requirement on global smoothness, namely $\theta_0 \in \Theta_{\mathrm{v},M}$ for some M.

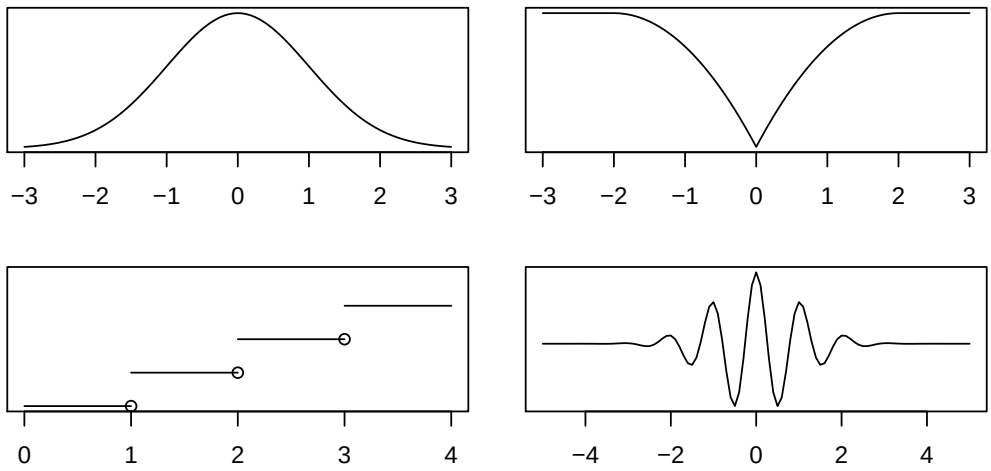

Figure 2.1: Examples of univariate càdlàg functions with finite variation norms. The top-left, top-right, bottom-left and bottom-right plots present the standard normal density function, a minimax concave penalty function (Zhang, 2010), a step function and the real part of a Morlet wavelet (Mallat, 2009) respectively.

We next formally present the definition of variation norm of a càdlàg function $\theta : [x^{(\ell)}, x^{(u)}] \subseteq \mathbb{R}^d \to \mathbb{R}$. Here, $x^{(\ell)}$ and $x^{(u)}$ are vectors in $\mathbb{R}^d$; with $\leq$ being entrywise, $[x^{(\ell)}, x^{(u)}] := \{x \in \mathbb{R}^d : x^{(\ell)} \leq x \leq x^{(u)}\}$.

For any nonempty index set $s \subseteq \{1, 2, \ldots, d\}$ and any $x = (x_1, x_2, \ldots, x_d) \in [x^{(\ell)}, x^{(u)}]$, we define $x_s := \{x_j : j \in s\}$ and $x_{-s} := \{x_j : j \in \{1, 2, \ldots, d\} \setminus s\}$ to be entries of x with indices in and not in s respectively. We defined the s-section of θ as $\theta_s := \theta(x_1 \mathbb{1}(1 \in s), x_2 \mathbb{1}(2 \in s), \ldots, x_d \mathbb{1}(d \in s))$. We can subsequently obtain the following representation of θ at any $x \in [x^{(\ell)}, x^{(u)}]$ in terms of sums and integrals of the variation of s-sections of θ (Gill et al., 1993):

$$\theta(x) = \theta(x^{(\ell)}) + \sum_{s \in \{1,\ldots,d\}, s \neq \emptyset} \int_{(x^{(\ell)}, x]} \theta_s(d\tilde{x}).$$

The variation norm is then subsequently defined as

$$\|\theta\|_{\mathrm{v}} := |\theta(x^{(\ell)})| + \sum_{s\in\{1,\ldots,d\},s\neq\emptyset} \int_{(x^{(\ell)},x^{(u)}]} |\theta_s(d\tilde{x})|.$$

We refer to Benkeser and van Der Laan (2016) and van Der Laan (2017) for more details on variation norm. Notably, this notion of variation norm coincides with that of Hardy and Krause (Owen, 2005).

We finally briefly introduce the algorithm to compute a HAL estimator. It can be shown that an empirical risk minimizer in $\Theta_{\mathrm{v},M}$ is a step function that only jumps at sample points, namely

$$x \mapsto \beta_0 + \sum_{s\subseteq\{1,\ldots,d\},s\neq\emptyset} \sum_{j=1}^{n} \mathbb{1}(X_{j,s} \leq x_s)\beta_{s,j}.$$

Here, β_0 and all $\beta_{s,j}$ are real numbers. To find an empirical risk minimizer in $\Theta_{\mathrm{v},M}$ in the above form, we may solve the following optimization problem:

$$\begin{aligned} &\min_{\theta} \sum_{i=1}^{n} \ell(\theta)(V_i) \\ \text{subject to} \quad &\theta : x \mapsto \beta_0 + \sum_{s\subseteq\{1,\ldots,d\},s\neq\emptyset} \sum_{j=1}^{n} \mathbb{1}(X_{j,s} \leq x_s)\beta_{s,j} \\ &|\beta_0| + \sum_{s\subseteq\{1,\ldots,d\},s\neq\emptyset} \sum_{j=1}^{n} |\beta_{s,j}| \leq M. \end{aligned}$$

The constraint imposes an upper bound on the ℓ_1 norm of a vector. Therefore, for common loss functions, we may use software for LASSO regression (Tibshirani, 1996). For example, if the loss function is the squared-error loss, then we may run a LASSO linear regression to obtain a HAL estimate.

2.3.2 Estimation with an oracle tuning parameter

In this section, we consider plug-in estimators based on HAL. For ease of illustration, for the rest of this section, we consider scalar-valued Ψ, and will discuss vector-valued Ψ only at the end of this subsection. We further introduce the following conditions needed to establish that the HAL-based plug-in estimator is efficient.

Condition B1 (Càdlàg functions)**.** θ_0 and $\dot{\Psi}$ are càdlàg.

Condition B2 (Bound on variation norm)**.** For some $M < \infty$, $\|\theta_0\|_{\mathrm{v}} + \|\dot{\Psi}\|_{\mathrm{v}} \leq M$.

Condition B2 ensures that certain perturbations of θ_0 still lie in $\Theta_{\mathrm{v},M}$, a crucial requirement for proving the asymptotic linearity of our proposed plug-in estimator. In addition, since $\dot{\Psi}$ may depend on components of P_0 other than θ_0 as in Examples 2–4, Conditions B1–B2 may also impose conditions on these components.

In this section, we fix an M that satisfies Condition B2. Additional technical conditions can be found in Appendix A.2.1. Let $\hat{\theta}_n = \hat{\theta}_{n,M} \in \mathrm{argmin}_{\theta \in \Theta_{\mathrm{v},M}} n^{-1} \sum_{i=1}^n \ell(\theta)(V_i)$ denote the HAL fit obtained using the bound M in Condition B2.

We note that $\hat{\theta}_n$ is not a typical sieve estimator because M is fixed and there is no explicit sequence of growing approximating spaces Θ_n. Nevertheless, we may view this method as a special case of sieve estimation with degenerate sieves $\Theta_n = \Theta_{\mathrm{v},M}$ for all n. This allows us to utilize existing results (Shen, 1997) to show the asymptotic linearity and efficiency of the plug-in estimator based on $\hat{\theta}_n$. We next formally present this result.

Theorem 2.1 (Efficincy of plug-in estimator)**.** *Under Conditions A1–A4 and B1–B4,* $\Psi(\hat{\theta}_n)$ *is an asymptotically linear estimator of* $\Psi(\theta_0)$ *with the influence function being* $v \mapsto \alpha_{0,\ell}^{-1}\{-\ell_0'[\dot{\Psi}](v) + \mathrm{E}_{P_0}[\ell_0'[\dot{\Psi}](V)]\}$*, that is,*

$$\Psi(\hat{\theta}_n) = \Psi(\theta_0) + \frac{1}{n}\sum_{i=1}^n \alpha_{0,\ell}^{-1}\left\{-\ell_0'[\dot{\Psi}](V_i) + \mathrm{E}_{P_0}\left[\ell_0'[\dot{\Psi}](V)\right]\right\} + o_p(n^{-1/2}).$$

As a consequence, $\sqrt{n}[\Psi(\hat{\theta}_n) - \Psi(\theta_0)] \xrightarrow{d} N(0, \xi^2)$ *with* $\xi^2 := \mathrm{Var}_{P_0}(\ell_0'[\dot{\Psi}](V))/\alpha_{0,\ell}^2$*. In addition, under Conditions E1 and E2 in Appendix A.2.4,* $\Psi(\hat{\theta}_n)$ *is efficient under a nonparametric model.*

We note that, for HAL to achieve the optimal convergence rate, we only need that $M \geq \|\theta_0\|_{\mathrm{v}}$ (Benkeser and van Der Laan, 2016; van Der Laan, 2017). The requirement of a larger M imposed by Condition B2 resembles undersmoothing (Newey et al., 1998), as using a larger M would result in a fit that is less smooth than that based on the CV-selected bound.

The $L^2(P_0)$-convergence rate of the flexible fit using the larger bound remains the same, but the leading constant may be larger. This is in contrast to traditional undersmoothing, which leads to a fit with a suboptimal rate of convergence.

Under some conditions, the following lemma provides a loose bound on $\|\dot{\Psi}\|_{\mathrm{v}}$ in the case that $\dot{\Psi}$ has a particular structure. Such a bound can be used to select an appropriate bound on variation norm that satisfies Condition B2.

Lemma 2.1. *Suppose that* $\dot{\Psi} = \dot{\psi} \circ \theta_0$, *where* $\dot{\psi} : \mathbb{R} \to \mathbb{R}$ *is differentiable. Let* $x^{(\ell)} = \sup\{x : P_0(X \geq x) = 1\}$ *where* sup *and* $\geq$ *are entrywise. Assume that* θ_0 *is differentiable. If each of* $\|\theta_0\|_{\mathrm{v}}$, $|\dot{\Psi}(x^{(\ell)})|$ *and* $B := \sup_{z:|z| \leq \|\theta_0\|_{\mathrm{v}}} |\dot{\psi}'(z)|$ *is finite, then* $\|\dot{\Psi}\|_{\mathrm{v}} \leq B\|\theta_0\|_{\mathrm{v}} + |\dot{\Psi}(x^{(\ell)})|$. *Hence,* $\|\theta_0\|_{\mathrm{v}} + \|\dot{\Psi}\|_{\mathrm{v}} \leq (B+1)\|\theta_0\|_{\mathrm{v}} + |\dot{\Psi}(x^{(\ell)})| < \infty$.

As we discussed at the end of Section 2.2, such structures as $\dot{\Psi} = \dot{\psi} \circ \theta_0$ are common, especially if we augment θ_0 to include other implicitly relevant components of P_0. For example, in Example 2, we may augment θ_0 with p_0 and p_0'; in Examples 3 and 4, we may augment θ_0 with the propensity score function.

When θ_0 is $\mathbb{R}^q$-valued, θ_0 can often be viewed as a collection of q real-valued variation-independent functions $\eta_{10}, \ldots, \eta_{q0}$. In this case, we can define $\Theta_{\mathrm{v},M} = \{(\eta_1, \ldots, \eta_q) : \eta_j \text{ is càdlàg}, \|\eta_j\|_{\mathrm{v}} \leq M_j, j = 1, \ldots, q\}$ for a positive vector $M = (M_1, \ldots, M_q)$. The subsequent arguments follow analogously, where now each η_j is treated as a separate function.

We remark that an undersmoothing condition such as B2 appears to be necessary for a HAL-based plug-in estimator to be efficient. We illustrate this numerically in Section 2.3.3. The choice of a sufficiently large bound M required by Theorem 2.1 is by no means trivial, since this choice requires knowledge that the user may not have. Nevertheless, this result forms the basis of the data-driven method that we propose in Section 2.3.3 for choosing M. We also remark that, if we wish to plug in the same $\hat{\theta}_n$ based on HAL for a rich estimands, the chosen bound M needs to be sufficiently large for all estimands of interest.

Another method to construct efficient plug-in estimators based on HAL has been independently developed (van der Laan et al., 2019). Unlike our approach based on sieve theory,

in this work, the authors directly analyzed the first-order bias of the plug-in estimator using influence functions. In terms of ease of implementation, their method requires specifying a constant involved in a threshold of the empirical mean of the basis functions, which may be difficult to specify in applications. Our approach in Section 2.3.3 may also require specifying an unknown constant to obtain a valid upper bound on $\|\dot{\Psi}\|_{\mathrm{v}}$, but in some cases the constant may be set to zero, and our simulation suggests that the performance is not sensitive to the choice of the constant.

2.3.3 Data-adaptive selection of the tuning parameter

Since it is hard to prespecify a bound M on the variation norm that is sufficiently large to satisfy Condition B2 but also sufficiently small to avoid overfitting for a given data set, it is desirable to select M in a data-adaptive manner. A seemingly natural approach makes use of k-fold CV. In particular, for each candidate bound M, partition the data into k folds of approximately equal size (k is fixed and does not depend on n), in each fold evaluate the performance of the HAL estimator fitted on all other folds based on this candidate M, and use the candidate bound M_n with the best average performance across all folds to obtain the final fit. It has been shown that $\hat{\theta}_{n,M_n}$ can achieve the optimal convergence rate under mild conditions (van der laan and Dudoit, 2003), but M_n appears not to satisfy Condition B2 in general. In particular, the derived bound on $\|\hat{\theta}_n - \theta_0\|$ relies on an empirical process term, namely $\sup_{\theta\in\Theta_{\mathrm{v},M}} |(P_n - P_0)\{\ell(\theta) - \ell(\theta_0)\}|$, and a larger M implies a larger space $\Theta_{\mathrm{v},M}$. Therefore, the bound on $\|\hat{\theta}_n - \theta_0\|$ grows with M. Because k-fold CV seeks to optimize out-of-sample performance, M_n generally appears to be close to $\|\theta_0\|_{\mathrm{v}}$ and not sufficiently large to obtain an efficient plug-in estimator.

To avoid this issue with the CV-selected bound, we propose a method that takes inspiration from k-fold CV, but modifies the bound so that it is guaranteed to yield an efficient plug-in estimator for $\Psi(\theta_0)$. This method may require the analytic expression for $\dot{\Psi}$. In Sections 2.4 and 2.5, we present methods that do not require this knowledge.

1. Derive an upper bound on $\|\dot{\Psi}\|_{\text{v}}$. This bound is a non-decreasing function of the variation norms of functions that can be learned from data (e.g., using Lemma 2.1). In other words, find a non-decreasing function F such that $\|\dot{\Psi}\|_{\text{v}} \leq F(\|\eta_{10}\|_{\text{v}}, \ldots, \|\eta_{q0}\|_{\text{v}})$ for unknown functions $\eta_{10}, \ldots, \eta_{q0}$ that can be assumed to be càdlàg with finite variation norm and can be estimated with HAL.

2. Estimate $\theta_0, \eta_{10}, \ldots, \eta_{q0}$ by HAL with k-fold CV, and denote the CV-selected bounds for these functions by $M_n, M_{1n}, \ldots, M_{qn}$.

3. For a small $\epsilon > 0$, use the bound $M_n + \epsilon + F(M_{1n} + \epsilon, \ldots, M_{qn} + \epsilon)$ to estimate θ_0 with HAL and plug in the fit. We refer to this step of slightly increasing the bounds as *ϵ-relaxation*.

It follows from Lemma A.1 in the Appendix that this method would yield a sufficiently large bound with probability tending to one. In practice, it is desirable for the bound derived on $\|\dot{\Psi}\|_{\text{v}}$ to be relatively tight to avoid choosing an overly large bound that leads to overfitting in small to moderate samples. We remark that multiplying by $1 + \epsilon$ rater than adding ϵ to each argument also leads to a valid choice for the bound; that is, the bound $M_n(1 + \epsilon) + F(M_{1n}(1 + \epsilon), \ldots, M_{qn}(1 + \epsilon))$ is also sufficiently large with probability tending to one. In practice, the user may increase each CV-selected bound by, for example, 5% or 10%. Although it is more natural and convenient to directly use $M_n + F(M_{1n}, \ldots, M_{qn})$ as the bound, we have only been able to prove the result with a small ϵ-relaxation. However, if the bound is loose and F is continuous, we can show that ϵ-relaxation is unnecessary. The formal argument can be found after Lemma A.1 in the Appendix.

As for methods based on knowledge of an influence function, deriving $\dot{\Psi}$ and a bound for its variation norm requires some expertise, but in some cases this task can be straightforward. The derivation of an influence function is typically based on a fluctuation in the space of distributions, but in many cases, the relation between such fluctuations and the summary of interest is implicit and difficult to handle. In contrast, the derivation of $\dot{\Psi}$ is

based on a fluctuation of θ_0, and the summary of interest explicitly depends on θ_0. As a consequence, it can be simpler to derive $\dot{\Psi}$ than to derive an influence function. For example, for the summary $\Psi_\kappa(\theta_0) = P_0\theta_0^\kappa$ in Example 1, we find that $\dot{\Psi}_\kappa = \kappa\theta_0^{\kappa-1}$ by straightforward calculation, whereas the influence function given in Theorem 2.1 is more difficult to directly derive analytically.

We illustrate the fact that M_n may not be sufficiently large and show that our proposed method resolves this issue via a simulation study in which $\theta_0 : x \mapsto \mathrm{E}_{P_0}[Y|X = x]$ and $\Psi : \theta_0 \mapsto P_0\theta_0^2$. We compare the performance of the plug-in estimators based on the 10-fold CV-selected bound on variation norm (M.cv), the bound derived from the analytic expression of $\dot{\Psi}$ with and without ϵ-relaxation (M.gcv+ and M.gcv respectively), and a sufficiently large oracle choice satisfying Condition B2 (M.oracle). We According to Lemma 2.1, M.oracle is $3\|\theta_0\|_{\mathrm{v}}$ and M.gcv is 3×M.cv. We also investigate the performance of 95% Wald CIs based on the influence function. For each resulting plug-in estimator, we investigate the following quantities: $n \cdot \mathrm{MSE}$, $\sqrt{n} \cdot |\mathrm{bias}|$ and CI coverage. More details of this simulation are provided in Appendix A.5. In theory, for an efficient estimator, we should find that $n \cdot \mathrm{MSE}$ tends to a constant (the variance of the influence function $\xi^2 := P_0\mathrm{IF}^2$), $\sqrt{n} \cdot |\mathrm{bias}|$ tends to 0, and 95% Wald CIs have approximately 95% coverage.

We report performance summaries in Fig 2.2 and Table 2.1 with this criterion, from which it appears that the plug-in estimators with M.oracle and M.gcv+ achieve efficiency, while the plug-in estimator based on M.cv does not. The desirable performance of M.oracle and M.gcv+ agrees with the available theory, whereas the poor performance of M.cv suggests that cross-validation may not yield a valid choice of variation norm in general. Interestingly, M.gcv performs similarly to M.oracle and M.gcv+. We conjecture that using an ϵ-relaxation is unnecessary in this setting. In Fig 2.3, we can also see that M.cv tends to $\|\theta_0\|_{\mathrm{v}}$ and has a high probability of being less than M.oracle. Therefore, this simulation suggests that using a sufficiently large bound — in particular, a bound larger than the CV-selected bound — may be necessary and sufficient for the plug-in estimator to achieve efficiency.

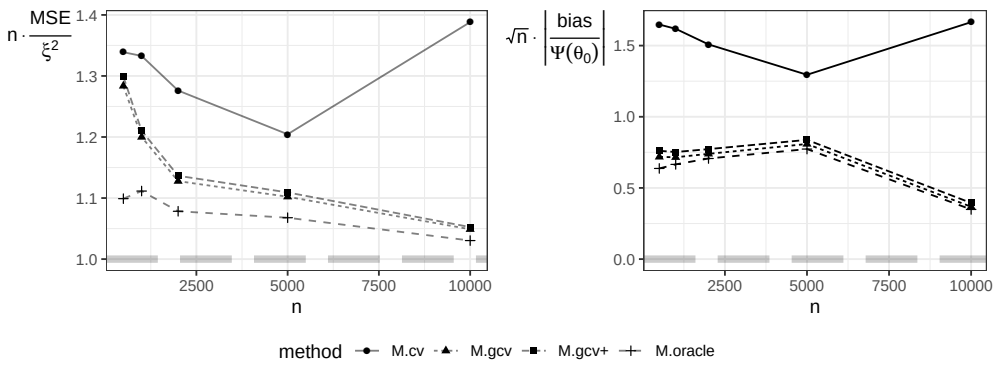

Figure 2.2: The relative MSE, $n \cdot \text{MSE}/\xi^2$, and the relative absolute bias, $\sqrt{n} \cdot |\text{bias}/\Psi(\theta_0)|$, of the plug-in estimator of $\Psi(\theta_0) = P_0\theta_0^2$ based on HAL for an oracle choice of the bound on variation norm (M.oracle), the 10-fold CV-selected bound (M.cv), a bound based on M.cv and analytic expression of $\dot{\Psi}$ without and with ϵ-relaxation (M.gcv and M.gcv+ respectively). $\xi^2 := P_0\text{IF}^2$ is the asymptotic variance that the $n \cdot \text{MSE}$ of an AL estimator should converge to. Note that the $n \cdot \text{MSE}$ for M.oracle, M.gcv and M.gcv+ tends to ξ^2 but that for M.cv does not.

Table 2.1: Coverage probability of 95% Wald CI of the plug-in estimator of $\Psi(\theta_0) = P_0\theta_0^2$ based on HAL for an oracle choice of the bound on variation norm (M.oracle), the 10-fold CV-selected bound (M.cv), a bound based on M.cv and analytic expression of $\dot{\Psi}$ without and with ϵ-relaxation (M.gcv and M.gcv+ respectively). The CI is constructed based on the influence function. The coverage for M.oracle, M.gcv and M.gcv+ is approximately 95%, but that for M.cv is not.

n	M.cv	M.gcv	M.gcv+	M.oracle
500	0.87	0.96	0.96	0.97
1000	0.87	0.97	0.97	0.97
2000	0.90	0.95	0.95	0.96
5000	0.93	0.95	0.95	0.95
10000	0.89	0.95	0.95	0.95

2.4 Data-adaptive series

2.4.1 Proposed method

For ease of illustration, we consider the case that Ψ is scalar-valued in this section. As we will describe next, our proposed estimation procedure for function-valued features does not

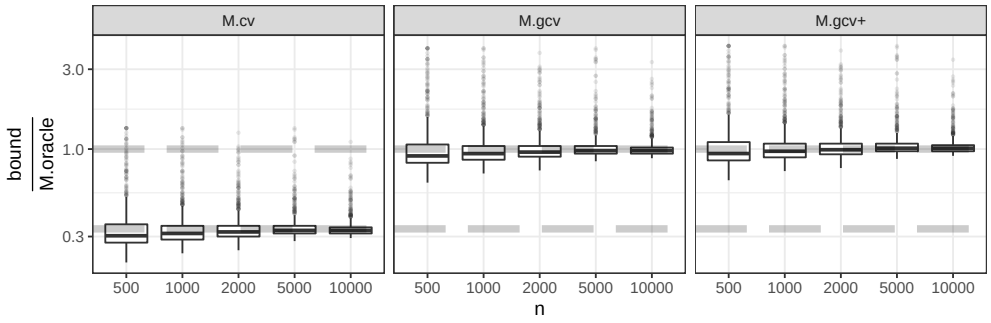

Figure 2.3: A boxplot of the ratio of bounds based on 10-fold CV and M.oracle. The horizontal gray thick dashed lines are 1 and 1/3. The y-axis is scaled based on logarithm for readability. There is a high probability that M.cv is much smaller than M.oracle; M.cv tends to the variation norm of the function being estimated, $\|\theta_0\|_{\mathrm{v}}$, corresponding to 1/3 of M.oracle. Enlarging M.cv according to the analytic expression of $\dot{\Psi}$ with ϵ-relaxation results in sufficiently large bounds. The enlargement without ϵ-relaxation appears to have similar performance.

rely on Ψ and hence can be used for a class of summaries.

Suppose that Θ is a vector space of $\mathbb{R}^q$-valued functions equipped with the $L^2(P_0)$-inner product. Further, suppose that $\dot{\Psi} = \dot{\psi} \circ \theta_0$ for some function $\dot{\psi} : \mathbb{R}^q \to \mathbb{R}^q$. This holds, for example, when $\Psi : \theta \mapsto P_0(f \circ \theta)$ for a fixed differentiable function f in Example 1. In this case, $\dot{\Psi} = f' \circ \theta_0$ and hence $\dot{\psi} = f'$. Particularly useful examples include Examples 1 and 4. For now we assume that the marginal distribution of X is known so that we only need to estimate θ_0 for this summary. We will address the more difficult case in which the marginal distribution of X is unknown in Section 2.4.3.

Let θ_n^0 be a given initial flexible ML fit of θ_0 and consider the data-adaptive sieve-like subspaces based on θ_n^0, $\Theta_n := \Theta_{n,\theta_n^0} := \mathrm{Span}\{\phi_1, \phi_2, \ldots, \phi_K\} \circ \theta_n^0$, where $\phi_1, \phi_2, \ldots$ are $\mathbb{R}^q$-valued basis functions in a series defined on $\mathbb{R}^q$ and $K = K(n)$ is a deterministic number of terms in the series — we will consider selecting K via CV in Section 2.4.4. Let $\theta_n^* = \theta_n^*(\theta_n^0) \in \mathrm{argmin}_{\theta \in \Theta_n} n^{-1} \sum_{i=1}^n \ell(\theta)(V_i)$ denote the series estimator within this data-adaptive sieve-like subspace that minimizes the empirical risk. We propose to use $\Psi(\theta_n^*)$ to estimate $\Psi(\theta_0)$.

2.4.2 Results for a deterministic number of terms

Following Chen (2007); Shen (1997), our proofs of the validity of our data-adaptive series approach make heavy use of projection operators. We use $\pi_n := \pi_{n,\theta_n^0}$ to denote the projection operator for functions in Θ onto $\Theta_n = \Theta_{n,\theta_n^0}$ with respect to $\langle \cdot, \cdot \rangle$. For any function $\theta \in \Theta$, let $\Pi_{n,\theta}$ denote the operator that takes as input a function $g : \mathbb{R}^q \to \mathbb{R}^q$ for which $g \circ \theta \in L^2(P_0)$ and outputs a function $\Pi_{n,\theta}(g) : \mathbb{R}^q \to \mathbb{R}^q$ such that $\Pi_{n,\theta}(g) \circ \theta = \pi_{n,\theta}(g \circ \theta)$. In other words, letting β_j be the quantity that depends on g and θ such that $\pi_{n,\theta}(g \circ \theta) = (\sum_{j=1}^K \beta_j \phi_j) \circ \theta$, we define $\Pi_{n,\theta}(g) := \sum_{j=1}^K \beta_j \phi_j$. The operator $\Pi_{n,\theta}$ may also be interpreted as follows: letting P_θ be the distribution of $\theta(X)$ with $V = (X, Z) \sim P_0$, then $\Pi_{n,\theta}$ is the projection operator of functions $\mathbb{R}^q \to \mathbb{R}^q$ with respect to the $L^2(P_\theta)$-inner product. We use $\mathcal{I}$ to denote the identity function in $\mathbb{R}^q$.

We now present additional conditions we will require to ensure that $\Psi(\theta_n^*)$ is an efficient estimator of $\Psi(\theta_0)$.

Condition C1 (Sufficient convergence rate of initial ML fit)**.** $\|\theta_n^0 - \theta_0\| = o_p(n^{-1/4})$.

Condition C2 (Sufficiently small estimation error)**.** $\|\theta_n^* - \pi_n(\theta_0)\| = o_p(n^{-1/4})$.

Condition C3 (Sufficiently small approximation error to $\mathcal{I}$ for Θ_{n,θ_0})**.** $\|\theta_0 - \Pi_{n,\theta_0}(\mathcal{I}) \circ \theta_0\| = o(n^{-1/4})$.

Condition C4 (Sufficiently small approximation error to $\dot{\psi}$ for Θ_{n,θ_0} and convergence rate of θ_n^*)**.** $\|[\dot{\psi} - \Pi_{n,\theta_0}(\dot{\psi})] \circ \theta_0\| \cdot \|\theta_n^* - \theta_0\| = o_p(n^{-1/2})$.

Appendix A.2.2 contains further technical conditions and Appendix A.3 discusses their plausibility. As discussed in Appendix A.3, Conditions C2–C4 typically imply restrictions on the growth rate of K: if K grows too fast with n, then Condition C2 may be violated; if K instead grows too slow, then Conditions C3 and C4 may be violated. For the generalized moment $\Psi : \theta \mapsto P_0(f \circ \theta)$ with a fixed known function f in Example 1, Condition C4 typically also imposes a smoothness condition f so that f' can be approximated by the series well. Our conditions are closely related to the conditions in Theorem 1 of Shen (1997).

Conditions C1–C3 and C6 serve as sufficient conditions for the condition on the smoothness of Ψ and the convergence rate of θ_n^* in Theorem 1 of Shen (1997). Together with Conditions C4 and C7, we can derive Lemma A.3, which is similar to the first part of Condition C of Shen (1997). The empirical process condition C8 is sufficient for Conditions A, D and the second part of C in Theorem 1 in Shen (1997).

We now present a theorem ensuring the asymptotic linearity and efficiency of the plug-in estimator based on θ_n^*.

Theorem 2.2 (Efficiency of plug-in estimator)**.** *Under Conditions A1–A4 and C1–C9, $\Psi(\theta_n^*)$ is an asymptotically linear estimator of $\Psi(\theta_0)$ with the influence function being $v \mapsto \alpha_{0,\ell}^{-1}\{-\ell_0'[\dot{\psi} \circ \theta_0](v) + \mathrm{E}_{P_0}[\ell_0'[\dot{\psi} \circ \theta_0](V)]\}$, that is,*

$$\Psi(\theta_n^*) = \Psi(\theta_0) + \frac{1}{n}\sum_{i=1}^{n} \alpha_{0,\ell}^{-1}\left\{-\ell_0'[\dot{\psi} \circ \theta_0](V_i) + \mathrm{E}_{P_0}\left[\ell_0'[\dot{\psi} \circ \theta_0](V)\right]\right\} + o_p(n^{-1/2}).$$

As a consequence, $\sqrt{n}[\Psi(\theta_n^*) - \Psi(\theta_0)] \xrightarrow{d} N(0, \xi^2)$ with $\xi^2 := \mathrm{Var}_{P_0}(\ell_0'[\dot{\psi} \circ \theta_0](V))/\alpha_{0,\ell}^2$. In addition, under Conditions E1 and E2 in Appendix A.2.4, $\Psi(\hat{\theta}_n)$ is efficient under a non-parametric model.

Remark 2. Consider the general case in which it may not be true that $\dot{\Psi}$ can be represented as $\dot{\psi} \circ \theta_0$ for some $\dot{\psi} : \mathbb{R}^q \to \mathbb{R}^q$. If the analytic expression of $\dot{\Psi}$ can be derived and $\dot{\Psi}$ can be estimated by $\dot{\Psi}_n$ such that $\|\dot{\Psi}_n - \dot{\Psi}\| \cdot \|\theta_n^0 - \theta_0\| = o_p(n^{-1/2})$, then our data-adaptive series can take a special form that is targeted towards Ψ. Specifically, letting $\vartheta_0 := (\theta_0, \dot{\Psi})^\top$ and $\Psi(\vartheta_0) := \Psi(\theta_0)$, it is straightforward to show that the gradient of Ψ is $\dot{\Psi} = (\dot{\Psi}, 0)^\top = (e_2, \mathbf{0})^\top \vartheta_0$ with $\mathbf{0} = (0,0)^\top$ and $e_2 = (0,1)^\top$, which is a function composed with ϑ_0. We can set $\vartheta_n^0 = (\theta_n^0, \dot{\Psi}_n)^\top$ and $\Theta_n = \mathrm{Span}\{\theta_n^0, \dot{\Psi}_n\}$ in our data-adaptive series. This approach does not have a growing number of terms in Θ_n and is not similar to sieve estimation, but can be treated as a special case of data-adaptive series. It can be shown that Conditions C1–C4 are still satisfied for ϑ and Ψ with this choice of Θ_n, and hence our data-adaptive series estimator leads to an efficient plug-in estimator. We remark that the introduction of ϑ and Ψ is a purely theoretical device, and this targeted approach to

estimation is quite similar to that used in the context of TMLE (van der Laan and Rubin, 2006; van der Laan and Rose, 2018).

2.4.3 Summaries involving the marginal distribution of X

We now generalize the setting considered thus far by allowing the parameter to depend both on θ_0 and on P_0, i.e., estimating $\Psi(\theta_0, P_0)$. The example given at the beginning of Section 2.4.1, namely that of estimating $\Psi(\theta_0) = P_0(f \circ \theta_0)$, is a special case of this more general setting. In what follows, we will make use of the following conditions:

Condition D1 (Conditions with P_0 fixed)**.** When we regard $\Psi(\theta_0, P_0)$ as the mapping $\theta \mapsto \Psi(\theta, P_0)$ evaluated at θ_0, Conditions A1–A4, C1–C4 and C6–C9 are satisfied for estimating $\Psi(\theta_0, P_0)$.

Condition D2 (Hadamard differentiability with θ_0 fixed)**.** The mapping $P \mapsto \Psi(\theta_0, P)$ is Hadamard differentiable at P_0.

By the functional delta method, it follows that $\Psi(\theta_0, P_n) = \Psi(\theta_0, P_0) + P_n \mathrm{IF}_0 + o_p(n^{-1/2})$ for a function IF_0 satisfying $P_0 \mathrm{IF}_0 = 0$ and $P_0 \mathrm{IF}_0^2 < \infty$.

Condition D3 (Negligible second-order difference)**.**

$$[\Psi(\theta_n^*, P_n) - \Psi(\theta_0, P_n)] - [\Psi(\theta_n^*, P_0) - \Psi(\theta_0, P_0)] = o_p(n^{-1/2}).$$

This condition usually holds, for example, when $\Psi(\theta_0, P_0) = P_0(f \circ \theta_0)$, as in this case the left-hand side is equal to $(P_n - P_0)(f \circ \theta_n^* - f \circ \theta_0)$, which is $o_p(n^{-1/2})$ under empirical process conditions.

Theorem 2.3 (Asymptotic linearity of plug-in estimator)**.** *Under Conditions D1–D3,* $\Psi(\theta_n^*, P_n)$ *is an asymptotically linear estimator of* $\Psi(\theta_0, P_0)$ *with influence function*

$$v \mapsto \alpha_{0,\ell}^{-1} \left\{ -\ell_0'[\dot{\psi} \circ \theta_0](v) + \mathrm{E}_{P_0}[\ell_0'[\dot{\psi} \circ \theta_0](V)] \right\} + \mathrm{IF}_0(V),$$

that is,

$$\Psi(\theta_n^*, P_n) = \Psi(\theta_0, P_0) + \frac{1}{n}\sum_{i=1}^{n}\left\{-\alpha_{0,\ell}^{-1}\ell_0'[\dot{\psi}\circ\theta_0](V_i) + \alpha_{0,\ell}^{-1}\,\mathrm{E}_{P_0}\left[\ell_0'[\dot{\psi}\circ\theta_0](V)\right] + \mathrm{IF}(V_i)\right\} + o_p(n^{-1/2}).$$

As a consequence, $\sqrt{n}[\Psi(\theta_n^*, P_n) - \Psi(\theta_0, P_0)] \xrightarrow{d} N(0, \xi^2)$ *with* $\xi^2 := \mathrm{Var}_{P_0}(\alpha_{0,\ell}^{-1}\ell_0'[\dot{\psi}\circ\theta_0](V) + \mathrm{IF}(V))$.

This result is easy to verify by decomposing $\Psi(\theta_n^*, P_n) - \Psi(\theta_0, P_0)$ as

$$[\Psi(\theta_n^*, P_0) - \Psi(\theta_0, P_0)] + [\Psi(\theta_0, P_n) - \Psi(\theta_0, P_0)] + \{[\Psi(\theta_n^*, P_n) - \Psi(\theta_0, P_n)] - [\Psi(\theta_n^*, P_0) - \Psi(\theta_0, P_0)]\}$$

Moreover, under conditions similar to the conditions E1 and E2 given in Appendix A.2.4, we can show that $\Psi(\theta_n^*, P_n)$ is efficient under a nonparametric model.

Remark 3. Conditions D2 and D3 can be relaxed. Specifically, if $\hat{P}_n$ is an estimator of P_0 that satisfies that $\Psi(\theta_0, \hat{P}_n) = \Psi(\theta_0, P_0) + P_n\mathrm{IF}_0 + o_p(n^{-1/2})$ for an influence function IF_0 and Condition D3 holds with P_n replaced by $\hat{P}_n$, then $\Psi(\theta_n^*, \hat{P}_n)$ is an asymptotically linear estimator of $\Psi(\theta_0, P_0)$.

2.4.4 CV selection of the number of terms in data-adaptive series

In the preceding subsections, we established the efficiency of the plug-in estimator based on suitable rates of growth for K relative to the sample size n. In this subsection, we show that, under some conditions, such a K can be selected by k-fold CV: after obtaining θ_n^0, for each K in a range of candidates, we can calculate the cross-validated risk from k folds and choose the value of K with the smallest CV risk. We denote the number of terms in the series that CV selects by K^*. In this section, we use K in the subscripts for notation related to data-adaptive sieves-like spaces and projections; this represents a slight abuse of notation because, in Sections 2.4.1 and 2.4.2, these subscripts were instead used for sample size n.

That is, we use $\Theta_{K,\theta}$ to denote $\mathrm{Span}\{\phi_1, \phi_2, \ldots, \phi_K\} \circ \theta$, $\pi_{K,\theta}$ to denote the projection onto $\Theta_{K,\theta}$, $\Pi_{K,\theta}$ to denote the operator such that $\Pi_{K,\theta}(g) \circ \theta = \pi_{K,\theta}(g \circ \theta)$ for all $g : \mathbb{R}^q \to \mathbb{R}^q$ with $g \circ \theta \in L^2(P_0)$, and $\theta_n^\sharp := \theta_{K^*}^*(\theta_n^0)$ to be the data-adaptive series estimator based on θ_n^0 and K^*.

Condition C5 (Bounded approximation error of $\dot{\psi}$ relative to $\mathcal{I}$)**.** There exists a constant $C > 0$ such that, with probability tending to one, $\|\dot{\psi} \circ \theta_n^0 - \Pi_{K,\theta_n^0}(\dot{\psi}) \circ \theta_n^0\| \leq C\|\theta_n^0 - \Pi_{K,\theta_n^0}(\mathcal{I}) \circ \theta_n^0\|$ for all K.

This condition is equivalent to

$$\|\dot{\psi} - \Pi_{K,\theta_n^0}(\dot{\psi})\|_{L^2(P_{\theta_n^0})} \leq C\|\mathcal{I} - \Pi_{K,\theta_n^0}(\mathcal{I})\|_{L^2(P_{\theta_n^0})}$$

for all K with probability tending to one, which may be interpreted in terms of two simultaneous requirements. The first requirement is that the identity function $\mathcal{I}$ is not exactly contained in the span of $\phi_1, \ldots, \phi_K$ for any K, since otherwise, the right-hand side would be zero for all sufficiently large K. Therefore, common series such as polynomial and spline series are not permitted for general summaries. In contrast, other series such as trigonometric series and wavelets satisfy this requirement. The second requirement is that the approximation error of the chosen series for the identity function $\mathcal{I}$ is not much larger than $\dot{\psi}$. If a trigonometric or wavelet series is used, then this condition imposes a strong smoothness condition on derivatives of $\dot{\psi}$. Nonetheless, this may not be stringent in some interesting examples. For example, if $\Psi(\theta) = P_0(f \circ \theta)$ for a fixed function f in Example 1, then $\dot{\psi}$ equals f' and hence can be expected to satisfy this strong smoothness condition provided that f is infinitely differentiable with bounded derivatives. The estimands encountered in many applications involve f satisfying this smoothness condition.

The following theorem justifies the use of k-fold CV to select K under appropriate conditions.

Theorem 2.4 (Efficiency of CV-based plug-in estimator)**.** *Assume that Conditions A1–A4, C1–C3, C5, C8 and C9 hold for a deterministic $K = K(n)$. Suppose part (a) of Condition C7*

holds, then, with $\theta_n^\sharp := \theta_{K^*}(\theta_n^0)$, $\Psi(\theta_n^\sharp)$ *is an asymptotically linear estimator of* $\Psi(\theta_0)$ *with influence function* $v \mapsto \alpha_{0,\ell}^{-1}\{-\ell_0'[\dot{\psi} \circ \theta_0](v) + \mathrm{E}_{P_0}[\ell_0'[\dot{\psi} \circ \theta_0](V)]\}$, *that is,*

$$\Psi(\theta_n^\sharp) = \Psi(\theta_0) + \frac{1}{n}\sum_{i=1}^{n} \alpha_{0,\ell}^{-1}\left\{-\ell_0'[\dot{\psi} \circ \theta_0](V_i) + \mathrm{E}_{P_0}\left[\ell_0'[\dot{\psi} \circ \theta_0](V)\right]\right\} + o_p(n^{-1/2}).$$

As a consequence, $\sqrt{n}[\Psi(\theta_n^\sharp) - \Psi(\theta_0)] \xrightarrow{d} N(0, \xi^2)$ *with* $\xi^2 := \mathrm{Var}_{P_0}(\ell_0'[\dot{\psi} \circ \theta_0](V))/\alpha_{0,\ell}^2$. *In addition, under Conditions E1 and E2 in Appendix A.2.4,* $\Psi(\hat{\theta}_n)$ *is efficient under a nonparametric model.*

2.4.5 Simulation

Demonstration of Theorem 2.4

We illustrate our method in a simulation in which we take $\theta_0(x) = \mathrm{E}_{P_0}[Z|X = x]$ and $\Psi(\theta_0) = P_0\theta_0^2$. This is a special case of Example 1. The true function θ_0 is chosen to be discontinuous, which violates the smoothness assumptions commonly required in traditional series estimation. In this case, $\dot{\psi} = 2\mathcal{I}$ and so the constant in Condition C5 is 2. We compare the performance of plug-in estimators based on three different nonparametric regressions: (i) polynomial regression with degree selected by 10-fold CV (poly), which results in a traditional sieve estimator, (ii) gradient boosting (xgb) (Friedman, 2001, 2002; Mason et al., 1999, 2000), and (iii) data-adaptive trigonometric series estimation with gradient boosting as the initial ML fit and 10-fold CV to select the number of terms in the series (xgb.trig). We also compare these plug-in estimators with the one-step correction estimator (Pfanzagl, 1982) based on gradient boosting (xgb.1step). Further details of this simulation can be found in Appendix A.5.

Fig 2.4 presents $n \cdot \mathrm{MSE}$ and $\sqrt{n} \cdot |\mathrm{bias}|$ for each estimator, whereas Table 2.2 presents the coverage probability of 95% Wald CIs based on these estimators. We find that xgb.trig and xgb.1step estimators perform well, while poly and xgb plug-in estimators do not appear to be efficient. Since polynomial series estimators only work well when estimating smooth functions, in this simulation, we would not expect the fit from the polynomial series estimator

to converge sufficiently fast, and consequently, we would not expect the resulting plug-in estimator to be efficient. In contrast, gradient boosting is a flexible ML method that can learn discontinuous functions, so we can expect an efficient plug-in estimator based on this ML method. However, gradient boosting is not designed to approximately solve the estimating equation that achieves the small-bias property for this particular summary, so we would not expect its naïve plug-in estimator to be efficient. Based on gradient boosting, our estimator and the one-step corrected estimator both appear to be efficient, but our method has the advantage of being a plug-in estimator. Moreover, the construction of our estimator does not require knowledge of the analytic expression of an influence function.

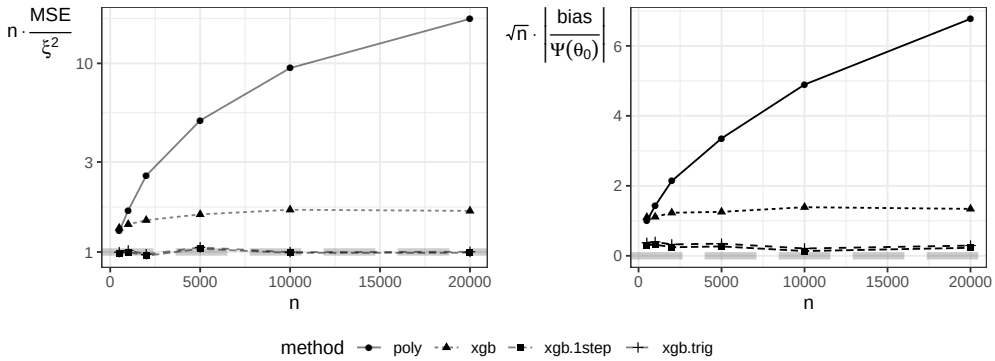

Figure 2.4: The relative MSE, $n \cdot \text{MSE}/\xi^2$, and the relative absolute bias, $\sqrt{n} \cdot |\text{bias}/\Psi(\theta_0)|$, of estimators of $\Psi(\theta_0) = P_0\theta_0^2$. $\xi^2 := P_0\text{IF}^2$ is the asymptotic variance that the $n \cdot \text{MSE}$ of an AL estimator should converge to. poly: plug-in estimator based on polynomial sieve estimation. xgb: plug-in estimator based on gradient boosting. xgb.1step: one-step correction (debiasing) of the plug-in estimator based on gradient boosting. xgb.trig: data-adaptive series with trigonometric series composed with gradient boosting. All tuning parameters are CV-selected. The y-axis for relative MSE is scaled based on logarithm for readability. Note that the $n \cdot \text{MSE}$ for xgb.trig and xgb.1step tend to ξ^2, but those for poly and xgb do not.

We also investigate the effect of the choice of K on the performance of our method. Fig 2.5 presents $n \cdot \text{MSE}$ for the data-adaptive series estimator with different choices of K. We can see that our method is insensitive to the choice of K in this simulation setting. Although a relatively small K performs better, choosing a much larger K does not appear

Table 2.2: Coverage probability of 95% Wald CI based on estimators of $\Psi(\theta_0) = P_0\theta_0^2$. poly: plug-in estimator based on polynomial sieve estimation. xgb: plug-in estimator based on gradient boosting. xgb.1step: one-step correction (debiasing) of the plug-in estimator based on gradient boosting. xgb.trig: data-adaptive series with trigonometric series composed with gradient boosting. All tuning parameters are CV-selected. The CI is constructed based on the influence function. The coverage probabilities for xgb.trig and xgb.1step are approximately 95%, but those for poly and xgb are not.

n	poly	xgb	xgb.1step	xgb.trig
500	0.90	0.90	0.95	0.95
1000	0.86	0.89	0.95	0.95
2000	0.74	0.88	0.96	0.96
5000	0.47	0.88	0.94	0.94
10000	0.16	0.87	0.95	0.96
20000	0.02	0.86	0.96	0.96

to substantially harm the behavior of the estimator. This insensitivity to the selected tuning parameter suggests that in some applications, without using CV, an almost arbitrary choice of K that is sufficiently large might perform well.

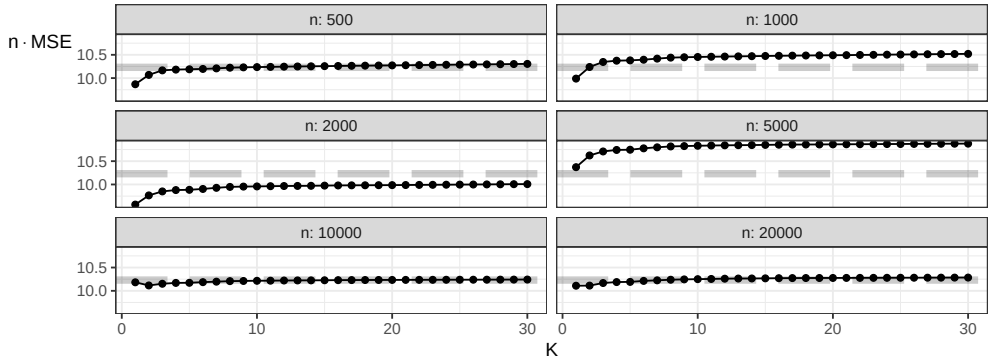

Figure 2.5: $n \cdot \text{MSE}$ of estimators of $\Psi(\theta_0) = P_0\theta_0^2$ based on data-adaptive series with different choices of K. The horizontal gray thick dashed line is the asymptotic variance that the $n \cdot \text{MSE}$ of an AL estimator should converge to, $\xi^2 := P_0\text{IF}^2$. Note that $n \cdot \text{MSE}$ is not sensitive to the choice of K over a wide range of K.

Violation of Condition C5

For the k-fold CV selection of K in our method to yield an efficient plug-in estimator, Ψ must be highly smooth in the sense that $\dot{\psi}$ can be approximated by the series about as well as can the identity function (see Condition C5). Although we have argued that this condition is reasonable, in this section, we explore via simulation the behavior of our method based on CV when $\dot{\psi}$ is rough. We again take $\theta_0 : x \mapsto \mathrm{E}_{P_0}[Z|X = x]$ and an artificial summary $\Psi(\theta_0) = P_0(f \circ \theta_0)$, where f is an element of $C^1[-1, 1]$ but not of $C^2[-1, 1]$. In this case, $\dot{\psi} = f'$ is very rough, so we do not expect it to be approximated by a trigonometric series as well as the identity function. However, it is sufficiently smooth to allow for the existence of a deterministic K that achieves efficiency. Further simulation details are provided in Appendix A.5.

Table 2.3 presents the performance of our estimator based on 10-fold CV. We note that it performs reasonably well in terms of the $n \cdot \mathrm{MSE}$ criterion. However, it is unclear whether its scaled bias converges to zero for large n, so our method may be too biased. The coverage of 95% Wald CIs is close to the nominal level, suggesting that the bias is fairly small relative to the standard error of the estimator at the sample sizes considered. One possible explanation for the good performance observed is that the $L^2(P_0)$-convergence rate of θ_n^* is much faster than $n^{-1/4}$, which allows for a slower convergence rate of the approximation error $\|\dot{\psi} \circ \theta_0 - \Pi_{n,\theta_0}(\dot{\psi}) \circ \theta_0\|$ (see Appendix A.3). This simulation shows that our proposed method may still perform well even if Condition C5 is violated, especially when the initial ML fit is close to the unknown function.

2.5 Generalized data-adaptive series

2.5.1 Proposed method

As in Section 2.4, we consider the case that Ψ is scalar-valued in this section. The assumption that $\dot{\Psi} = \dot{\psi} \circ \theta_0$ may be too restrictive for general summaries as in Examples 2–4, especially if $\dot{\Psi}$ is not derived analytically (see Remark 2). In this section, we generalize the method

Table 2.3: Performance of the plug-in estimator of $\Psi(\theta_0) = P_0(f \circ \theta_0)$ based on data-adaptive series. Here f is not infinitely differentiable. The relative MSE is $n \cdot \text{MSE}/\xi^2$ where $\xi^2 := P_0\text{IF}^2$ is the asymptotic variance that the $n \cdot \text{MSE}$ of an AL estimator should converge to; the root-n abs relative bias is $\sqrt{n}|\text{bias}/\Psi(\theta_0)|$. The performance appears to be acceptable in view of the small MSE and reasonable CI coverage.

n	relative MSE	root-n absolute relative bias	95% Wald CI coverage
500	0.88	3.95	0.97
1000	0.89	3.73	0.96
2000	0.79	3.15	0.97
5000	0.78	2.02	0.97
10000	0.88	2.57	0.97
20000	0.88	1.75	0.96

in Section 2.4 to deal with these summaries. Letting $\mathcal{I}_x$ be the identity function defined on $\mathcal{X}$, we can readily generalize the above method to the case where $\dot{\Psi}$ can be represented as $\dot{\psi} \circ (\theta_0, \mathcal{I}_x)$ for a function $\dot{\psi} : \mathbb{R}^q \times \mathcal{X} \to \mathbb{R}^q$; that is, $\dot{\Psi}(x) = \dot{\psi}(\theta_0(x), x)$. This form holds trivially if we set $\dot{\psi}(t, x) = \dot{\Psi}(x)$, i.e., $\dot{\psi}$ is independent of its first argument, but we can utilize flexible ML methods if $\dot{\psi}$ is nontrivial. Again, we assume Θ is a vector space of $\mathbb{R}^q$-valued function equipped with the $L^2(P_0)$-inner product. We assume $\dot{\psi}$ can be approximated well by a basis $\phi_1, \phi_2, \ldots : \mathbb{R}^q \times \mathcal{X} \to \mathbb{R}^q$, and consider the data-adaptive sieve-like subspace $\Theta_n := \Theta_{n,\theta_n^0} := \text{Span}\{\phi_1, \ldots, \phi_K\} \circ (\theta_n^0, \mathcal{I}_x)$. We propose to use $\Psi(\theta_n^*)$ to estimate $\Psi(\theta_0)$, where $\theta_n^* = \theta_n^*(\theta_n^0) \in \text{argmin}_{\theta \in \Theta_n} n^{-1} \sum_{i=1}^n \ell(\theta)(V_i)$ denotes the series estimator within Θ_n minimizing the empirical risk.

2.5.2 Results for proposed method

With a slight abuse of notation, in this section we use $\mathcal{I}$ to denote the function $(t, x) \mapsto t$ where $t \in \mathbb{R}^q$ and $x \in \mathcal{X}$. Again, we use $\pi_n := \pi_{n,\theta_n^0}$ to denote the projection operator onto Θ_{n,θ_n^0}. Let $\Pi_{n,\theta}$ be defined such that, for any function $g : \mathbb{R}^q \times \mathcal{X} \to \mathbb{R}^q$ with $g \circ$

$(\theta, \mathcal{I}_x) \in L^2(P_0)$, it holds that $\Pi_{n,\theta}(g) \circ (\theta, \mathcal{I}_x) = \pi_{n,\theta}(g \circ (\theta, \mathcal{I}_x))$; that is, letting β_j be the quantity that depends on g and θ such that $\pi_{n,\theta}(g \circ (\theta, \mathcal{I}_x)) = (\sum_{j=1}^K \beta_j \phi_j) \circ (\theta, \mathcal{I}_x)$, we define $\Pi_{n,\theta}(g) := \sum_{j=1}^K \beta_j \phi_j$.

We introduce conditions and derive theoretical results that are parallel to those in Section 2.4.

Condition C3* (Sufficiently small approximation error to $\mathcal{I}$ for Θ_{n,θ_0})**.** $\|\theta_0 - \Pi_{n,\theta_0}(\mathcal{I}) \circ (\theta_0, \mathcal{I}_x)\| = o(n^{-1/4})$.

Condition C4* (Sufficiently small approximation error to $\dot{\psi}$ for Θ_{n,θ_0} and convergence rate of θ_n^*)**.** $\|[\dot{\psi} - \Pi_{n,\theta_0}(\dot{\psi})] \circ (\theta_0, \mathcal{I}_x)\| \cdot \|\theta_n^* - \theta_0\| = o_p(n^{-1/2})$.

Additional regularity conditions can be found in Appendix A.2.3. Note that $\dot{\Psi}$ may depend on components of P_0 other than θ_0, Condition C4* may impose smoothness conditions on these components so that $\dot{\psi}$ can be well approximated by the chosen series. For example, in Example 2, Condition C4* requires that p_0'/p_0 and the propensity score can be approximated by the series well; in Examples 3 and 4, Condition C4* imposes the same requirement on the propensity score. We now present a theorem that establishes the efficiency of the plug-in estimator based on θ_n^*.

Theorem 2.5 (Efficiency of plug-in estimator)**.** *Under Conditions A1–A4, C1, C2, C3***, C4***, C6***, C7***, C8 and C9, $\Psi(\theta_n^*)$ is an asymptotically linear estimator of $\Psi(\theta_0)$ with influence function $v \mapsto \alpha_{0,\ell}^{-1}\{-\ell_0'[\dot{\psi} \circ (\theta_0, \mathcal{I}_x)](v) + \mathrm{E}_{P_0}[\ell_0'[\dot{\psi} \circ (\theta_0, \mathcal{I}_x)](V)]\}$, that is,*

$$\Psi(\theta_n^*) = \Psi(\theta_0) + \frac{1}{n}\sum_{i=1}^n \alpha_{0,\ell}^{-1}\left\{-\ell_0'[\dot{\psi} \circ (\theta_0, \mathcal{I}_x)](V_i) + \mathrm{E}_{P_0}\left[\ell_0'[\dot{\psi} \circ (\theta_0, \mathcal{I}_x)](V)\right]\right\} + o_p(n^{-1/2}).$$

As a consequence, $\sqrt{n}[\Psi(\theta_n^*) - \Psi(\theta_0)] \xrightarrow{d} N(0, \xi^2)$ with $\xi^2 := \mathrm{Var}_{P_0}(\ell_0'[\dot{\psi} \circ (\theta_0, \mathcal{I}_x)](V))/\alpha_{0,\ell}^2$. In addition, under Conditions E1 and E2 in Appendix A.2.4, $\Psi(\hat{\theta}_n)$ is efficient under nonparametric models.

Remark 4. When Ψ depends on both θ_0 and P_0, we can readily adapt this method as in Section 2.4.3.

We now present a condition for selecting K via k-fold CV, in parallel with Condition C5 from Section 2.4.4.

Condition C5* (Bounded approximation error of $\dot{\psi}$ relative to $\mathcal{I}$)**.** There exists a constant $C > 0$ such that, with probability tending to one, $\|\dot{\psi} \circ (\theta_n^0, \mathcal{I}_x) - \Pi_{K,\theta_n^0}(\dot{\psi}) \circ (\theta_n^0, \mathcal{I}_x)\| \leq C\|\theta_n^0 - \Pi_{K,\theta_n^0}(\mathcal{I}) \circ (\theta_n^0, \mathcal{I}_x)\|$ for all K.

Remark 5. Similarly to Condition C5, Condition C5* requires that the identity function $\mathcal{I}$ is not contained in the span of finitely many terms of the chosen series and that $\dot{\psi}$ is sufficiently smooth so that $\dot{\psi}$ can be approximated well by the chosen series. However, Condition C5* may be far more stringent than Condition C5. In fact, it may be overly stringent in practice. Since $\dot{\Psi}$ may depend on components of P_0 other than θ_0, Condition C5* may require these components to be sufficiently smooth. When a common candidate series such as the trigonometric series is used, a sufficient condition for Condition C5* is that $\dot{\psi}$ is infinitely differentiable with bounded derivatives, which further imposes assumptions on the smoothness of other components of P_0. For example, in Example 2, a sufficient condition for Condition C5* is that p_0'/p_0 is infinitely differentiable with bounded derivatives; in Examples 3 and 4, a sufficient condition for Condition C5* is that Condition C5* is that the propensity score function satisfies the same requirement. Due to the stringency of Condition C5*, we conduct a simulation in Section 2.5.3 to understand the performance of our proposed method when this condition is violated. The simulation appears to indicate that our method may be robust against violation of Condition C5*.

The following theorem shows that k-fold CV can be used to select K under certain conditions.

Theorem 2.6 (Efficiency of CV-based plug-in estimator)**.** *Assume Conditions A1–A4, C1, C2, C3*, C6*, C7*, C8 and C9 hold for a deterministic $K = K(n)$. Suppose that part (a) of Condition C7* holds. With $\theta_n^\sharp := \theta_{K^*}(\theta_n^0)$, $\Psi(\theta_n^\sharp)$ is an asymptotically linear estimator of $\Psi(\theta_0)$ with influence function $v \mapsto \alpha_{0,\ell}^{-1}\{-\ell_0'[\dot{\psi} \circ (\theta_0, \mathcal{I}_x)](v) + \mathrm{E}_{P_0}[\ell_0'[\dot{\psi} \circ (\theta_0, \mathcal{I}_x)](V)]\}$, that*

is,

$$\Psi(\theta_n^\sharp) = \Psi(\theta_0) + \frac{1}{n}\sum_{i=1}^{n} \alpha_{0,\ell}^{-1} \left\{ -\ell_0'[\dot{\psi} \circ (\theta_0, \mathcal{I}_x)](V_i) + \mathrm{E}_{P_0}\left[\ell_0'[\dot{\psi} \circ (\theta_0, \mathcal{I}_x)](V)\right] \right\} + o_p(n^{-1/2}).$$

Therefore, $\sqrt{n}[\Psi(\theta_n^\sharp) - \Psi(\theta_0)] \xrightarrow{d} N(0, \xi^2)$ *with* $\xi^2 := \mathrm{Var}_{P_0}(\ell_0'[\dot{\psi} \circ \theta_0](V))/\alpha_{0,\ell}^2$. *In addition, under Conditions E1 and E2 in Appendix A.2.4,* $\Psi(\hat{\theta}_n)$ *is efficient under a nonparametric model.*

2.5.3 Simulation

In the following simulations, we consider the problem in Example 4. As we show in Appendix A.1, letting $g_0 : x \mapsto P_0(A = 1|X = x)$ be the propensity score and setting $\theta = (\mu_0, \mu_1)$, with $\ell(\theta) : v \mapsto a[z - \mu_1(x)]^2 + (1-a)[z - \mu_0(x)]^2$, the generalized data-adaptive series methodology may be used to obtain an efficient estimator. As in Section 2.4.5, we conduct two simulation studies, the first demonstrating Theorem 2.6 and the other exploring the robustness of CV against violation of Condition C5*.

Demonstration of Theorem 2.6

We choose θ_0 to be a discontinuous function while g_0 is highly smooth. We compare the performance of plug-in estimators based on three different nonparametric regressions: (i) polynomial regression with the degree selected by 5-fold CV (poly), which results in a traditional sieve estimator, (ii) gradient boosting (xgb) (Friedman, 2001, 2002; Mason et al., 1999, 2000), and (iii) generalized data-adaptive trigonometric series estimation with gradient boosting as the initial ML fit and 5-fold CV to select the number of terms in the series (xgb.trig). Further details of the simulation setting are provided in Appendix A.5.

Fig 2.6 presents $n \cdot \mathrm{MSE}$ and $\sqrt{n} \cdot |\mathrm{bias}|$ for each estimator, whereas Table 2.4 presents the coverage probability of 95% Wald CIs based on these estimators. There are a few runs in the simulation with noticeably poor behavior, so we trimmed the most extreme values when computing MSE and bias in Fig 2.6 (1% of all Monte Carlo runs). The outliers may be

caused by the performance of gradient boosting and the instability of 5-fold CV. In practice, the user may ensemble more ML methods and use 10-fold CV to mitigate such behavior. We note that xgb.trig and xgb.1step estimators perform well, while poly and xgb plug-in estimators do not appear to be efficient. Based on gradient boosting, our estimator and the one-step corrected estimator both appear to be efficient, but the construction of our estimator has the advantage of not requiring the analytic expression of an influence function.

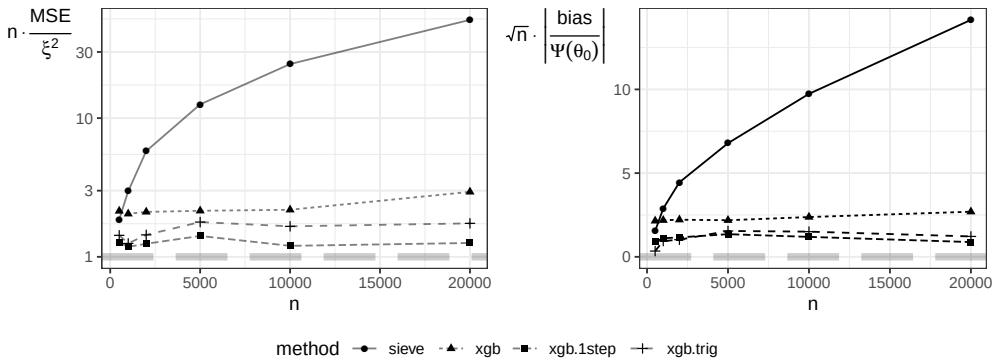

Figure 2.6: The relative MSE, $n \cdot \text{MSE}/\xi^2$, and the relative absolute bias, $\sqrt{n} \cdot |\text{bias}/\Psi(\theta_0)|$, of estimators of $\Psi(\theta_0) = \text{Var}_{P_0}(\mu_{0,1}(X) - \mu_{0,0}(X))$ where $\mu_{0,a} : x \mapsto \text{E}_{P_0}[Y|A = a, X = x]$. $\xi^2 := P_0\text{IF}^2$ is the asymptotic variance that the $n \cdot \text{MSE}$ of an AL estimator should converge to. poly: plug-in estimator based on polynomial sieve estimation. xgb: plug-in estimator based on gradient boosting. xgb.1step: one-step correction (debiasing) of the plug-in estimator based on gradient boosting. xgb.trig: data-adaptive series with trigonometric series composed with gradient boosting. All tuning parameters are CV-selected. The y-axis for relative MSE is scaled based on logarithm for readability. Note that the $n \cdot \text{MSE}$ for xgb.trig and xgb.1step tend to ξ^2, but those for poly and xgb do not.

*Violation of Condition C5**

We also study via simulation the behavior of our estimator when Condition C5* is violated. We note that whether Condition C5* holds depends on the smoothness of g_0. We choose g_0 to be rougher than $\mathcal{I}$ with g_0 being an element of $C^2[-1, 1]$ but not of $C^3[-1, 1]$. Consequently, $\dot{\Psi}$ cannot be approximated by our generalized data-adaptive series as well as $\mathcal{I}$,

Table 2.4: Coverage probability of 95% Wald CI based on estimators of $\Psi(\theta_0) = \mathrm{Var}_{P_0}(\mu_{0,1}(X) - \mu_{0,0}(X))$ where $\mu_{0,a} : x \mapsto \mathrm{E}_{P_0}[Y|A = a, X = x]$. poly: plug-in estimator based on polynomial sieve estimation. xgb: plug-in estimator based on gradient boosting. xgb.1step: one-step correction (debiasing) of the plug-in estimator based on gradient boosting. xgb.trig: data-adaptive series with trigonometric series composed with gradient boosting. All tuning parameters are CV-selected. The CI is constructed based on the influence function. The coverage probabilities for xgb.trig and xgb.1step are relatively close to 95%, but those for poly and xgb are not.

n	poly	xgb	xgb.1step	xgb.trig
500	0.85	0.76	0.89	0.90
1000	0.68	0.78	0.93	0.93
2000	0.44	0.81	0.93	0.92
5000	0.11	0.80	0.89	0.87
10000	0.00	0.79	0.92	0.90
20000	0.00	0.67	0.91	0.88

but its smoothness is sufficient for the existence of a deterministic K to achieve efficiency. Appendix A.5 describes further details of this simulation setting.

Table 2.5 presents the performance of our estimator based on 5-fold CV. We observe that its scaled MSE appears to converge to one, but it is unclear whether its scaled bias converges to zero for large n, and so our method may be overly biased.. The coverage of 95% Wald CIs is close to the nominal level, suggesting that the bias may be fairly small relative to the standard error of the estimator at the sample sizes considered. Therefore, according to this simulation, our generalized data-adaptive series methodology appears to be robust against violation of Condition C5*.

2.6 Discussion

Numerous methods have been proposed to construct efficient estimators for statistical parameters under a nonparametric model, but each of them has one or more of the following

Table 2.5: Performance of the plug-in estimator of $\Psi(\theta_0) = \mathrm{Var}_{P_0}(\mu_{0,1}(X) - \mu_{0,0}(X))$ where $\mu_{0,a} : x \mapsto \mathrm{E}_{P_0}[Z|A = a, X = x]$ based on data-adaptive series. Here the propensity score $g_0 : x \mapsto \mathrm{E}_{P_0}[A|X = x]$ is rough. The relative MSE is $n \cdot \mathrm{MSE}/\xi^2$ where $\xi^2 := P_0\mathrm{IF}^2$ is the asymptotic variance that the $n \cdot \mathrm{MSE}$ of an AL estimator should converge to; the root-n abs relative bias is $\sqrt{n}|\mathrm{bias}/\Psi(\theta_0)|$. The performance appears to be acceptable in view of small MSE and reasonable CI coverage.

n	relative MSE	root-n absolute relative bias	95% Wald CI coverage
500	1.02	0.28	0.92
1000	1.13	0.26	0.91
2000	1.10	0.19	0.94
5000	1.03	0.02	0.93
10000	0.96	0.23	0.95
20000	0.99	0.24	0.94

undesirable limitations: (i) their construction may require specialized expertise that is not accessible to most statisticians; (ii) for any given data set, there may be little guidance, if any, on how to select a key tuning parameter; and (iii) they may require stringent smoothness conditions, especially on derivatives. In this chapter, we propose two sieve-like methods that can partially overcome these difficulties.

Our first approach, namely that based on HAL, can be further generalized to the case in which the flexible fit is an empirical risk minimizer over a function class assumed to contain the unknown function. The key condition B2 may be modified in that case as long as it ensures that certain perturbations of the unknown function still lie in that function class. We note that our methods may also be applied under semiparametric models.

A major direction for future work is to construct valid CIs without the knowledge of the influence function of the resulting plug-in estimator. The nonparametric bootstrap is in general invalid when the overall summary is not Hadamard differentiable and especially when the method relies on CV (Bickel et al., 1997; Hall, 2013), but a model-based bootstrap

is a possible solution (Chapter 28 of van der Laan and Rose (2018)). In many cases only certain components of the true data-generating distribution must be estimated to obtain a plug-in estimator, while its variance may depend on other components that are not explicitly estimated. Therefore, generating valid model-based bootstrap samples is generally difficult.

There have been recent advances in the sieve estimation framework (e.g., Ai and Chen, 2003; Chen et al., 2006, 2008; Chen and Pouzo, 2009; Chen and Liao, 2014; Chen and Pouzo, 2015; Chen and Liao, 2015), either for nonparametric or semiparametric models. In particular, Chen and Liao (2014) developed a sieve estimation procedure that leads to valid statistical inference even for summaries that are not differentiable; Chen and Liao (2015) developed a sieve estimation procedure for weakly dependent data. It would be interesting to incorporate flexible ML estimators into these procedures.

Our proposed sieve-like methods may be used to construct efficient plug-in estimators for new applications in which the relevant theoretical results are difficult to derive. They may also inspire new methods to construct such estimators under weaker conditions.

Chapter 3

OPTIMAL INDIVIDUALIZED DECISION RULES USING INSTRUMENTAL VARIABLE METHODS

3.1 Introduction

Candidate treatment strategies are often compared by their effect on the mean of an outcome of interest. When several treatment strategies exist and result in qualitatively different effects across subgroups of individuals, it may be beneficial to provide individualized treatment strategies based on estimated subgroup-specific treatment effects. Any such treatment strategy is commonly referred to as an individualized treatment rule (ITR). Typically, the objective of the treatment rule is to optimize the mean outcome via an intervention on the treatment, and the resulting ITR is referred to as an optimal ITR. While estimation of optimal ITRs and of the counterfactual mean outcome under an optimal ITR has been extensively studied (e.g., Murphy, 2003; Robins, 2004; Zhao et al., 2012; Chakraborty and Moodie, 2013; Luedtke and van der Laan, 2016b), existing approaches mostly require that the available set of covariates be sufficiently rich to allow deconfounding of the treatment-outcome relationship. Unfortunately, this condition may fail to hold in observational studies and in randomized trials in which noncompliance occurs because important confounding factors may not have been recorded.

If an instrumental variable (IV) — a cause of the outcome that only operates via the treatment of interest and is independent of all treatment-outcome confounders — exists and has been recorded, it can be leveraged to infer treatment effects even when there is unmeasured confounding. Common IVs include encouragement to take a certain treatment, assignment in a randomized trial with possible noncompliance, and geographical location as a determinant of patient access to treatment (see Baiocchi et al., 2014, for a review). Despite

the rich literature on the use of IVs in causal inference, existing works focus primarily on estimation of marginal effects for static (e.g., Abadie, 2003) or dynamic (Mauro et al., 2018) interventions that are fixed *a priori* rather than learned from data to optimize individual-level outcomes.

Our interest in this problem is motivated by an ongoing study of the effect of combat deployment on suicide among US Army soldiers. It is of great policy importance not only to understand the effect of deployment on suicide but also to determine how to make deployment decisions that minimize the risk of suicide (LeardMann et al., 2013). These questions are complicated by the fact that deployment is not randomized and its relationship to suicide is determined by a host of unmeasured confounders. Nevertheless, initial duty assignment can be used as a plausible binary IV since (i) it is largely random with respect to the characteristics of individual soldiers, who are assigned sequentially based on unit needs; and (ii) it is a strong predictor of deployment because deployments largely occur at the unit level.

In our general treatment of this problem, we focus primarily on encouragement to undergo one particular treatment (treatment 1) instead of another (treatment 0) as the IV. The interventions we consider operate directly on the treatment strategy (case I) or instead on the encouragement strategy (case II). In case I, the average treatment effect (ATE) is the causal estimand of interest, which we identify from the observed data under conditions that are slightly weaker than those recently introduced in Wang and Tchetgen Tchetgen (2018). We are then interested in determining an optimal ITR and the resulting ATE of this individualized strategy. In case II, we consider two causal estimands, the average encouragement effect (AEE) and the local average treatment effect (LATE), which can be identified under alternative conditions. The latter effect represents the ATE among compliers, who are defined as those individuals who adopt treatment 1 only when encouraged to do so. We are interested in determining an optimal individualized encouragement rule (IER) and the resulting LATE of this individualized strategy. To account for real-world resource limitations, the methodology we develop explicitly incorporates investigator-specified constraints on the proportion of individuals who can receive treatment 1, much as in Luedtke and van der Laan (2016a)

in the non-IV setting. Incorporating constraints is not straightforward when intervening on encouragement since treatment resources are then manipulated only indirectly through the IV. For each scenario considered, we propose (i) nonparametric estimators of an optimal IER or ITR, and (ii) asymptotically linear nonparametric estimators and confidence intervals for the ATE, AEE or LATE of an optimal IER or ITR. These estimators are constructed using the general framework of targeted minimum loss-based estimation (TMLE) (van der Laan and Rubin, 2006; van der Laan and Rose, 2018) and explicitly allow the use of machine learning tools for nuisance estimation.

As in Luedtke and van der Laan (2016a), throughout this chapter, we consider stochastic individualized rules. Such rules output a probability of providing (or encouraging) a given treatment as a function of individual characteristics. However, we will see that the optimal ITR (IER) among all stochastic rules is in fact deterministic whenever there is heterogeneity in the average treatment (encouragement) effect across subgroups defined by measured covariates in the population. When optimal stochastic and deterministic rules do not coincide, an optimal stochastic rule may actually be more sensible than an optimal deterministic rule. For example, under a resource constraint, an optimal stochastic treatment rule may output a probability of providing treatment 1 between zero and one for a given subgroup, whereas an optimal deterministic rule must suggest treatment 0 to this subgroup and provide treatment 1 to other subgroups with smaller ATE. In this situation, optimal stochastic rules can be interpreted as distributing available resources among individuals with greatest ATE, even though not all individuals in the same subgroup might be able to receive the same treatment due to resource constraints. In contrast, deterministic rules insist on providing the same treatment or encouragement to all individuals in any given subgroup, and as such, can lead to a suboptimal mean outcome.

Some authors have used IV methods to assess treatment effect heterogeneity and to construct optimal individualized rules. For example, Abadie (2003) studied inference on the counterfactual mean outcome among compliers as a function of baseline covariates. Basu (2014) studied inference on the so-called person-centered treatment effect using IV methods,

although they required more stringent causal assumptions than usually required to identify the ATE. Neither of these two works explicitly considered the construction of optimal individualized rules. Extending the approach of Luedtke and van der Laan (2016b), Toth and van der Laan (2018) considered the use of IVs to estimate an optimal ITR under resource constraints and the resulting counterfactual mean outcome under stronger causal conditions.

This chapter is organized as follows. In Section 3.2, we describe the setup of the problem, outline basic assumptions on the IV, and introduce the causal estimands that we consider. In Section 3.3, we establish the nonparametric identifiability of these causal estimands. We present and study our proposed inferential procedures in Section 3.4. In Sections 3.5, we demonstrate our methods via simulation, and we present results from the analysis of the motivating data on suicide among US Army soldiers in Section 3.6. In Appendix B.1, we outline additional technical conditions required in our main results along with proofs. In Appendix B.2, we provide a more in-depth discussion of two particular conditions that may appear difficult to verify.

3.2 Setup and objectives

3.2.1 Observed and counterfactual data structures

We denote by $W \in \mathcal{W} \subseteq \mathbb{R}^p$ the vector of baseline covariates, by $Z \in \{0,1\}$ the value of binary IV, by $A \in \{0,1\}$ the treatment received, and by $Y \in \mathbb{R}$ the outcome of interest. The prototypical observed data unit consists of $O = (W, Z, A, Y)$, and we observe independent draws $O_1, O_2, \ldots, O_n$ from a distribution P_0 with support $\mathcal{O} = \mathcal{W} \times \{0,1\} \times \{0,1\} \times \mathbb{R}$. As we will see in Section 3.2.2, W should be chosen so that Z is a valid instrument after conditioning on W. We assume that larger values of outcome Y are preferable.

We will denote by $V = V(W) \in \mathcal{V}$ a fixed transformation of W — for example, V may only include a subset of variables in W. Encouragement or treatment recommendations will be made based on V. In practice, V may be chosen based on prior knowledge of the causal mechanism and the cost of measurements. In the remainder, we will use the

shorthand notation V, V_i and v to refer to $V(W)$, $V(W_i)$ and $V(w)$, respectively. We define an individualized encouragement (IER) or treatment (ITR) rule $d : \mathcal{V} \to [0, 1]$ as a stochastic rule that prescribes encouragement or administration of treatment 1 with probability $d(v)$ for a person with covariate value v. A deterministic rule results from choosing d to be an indicator function.

To discuss causal conditions and define causal estimands of interest, we will adopt the potential outcome framework in Neyman (1923) and Rubin (1974). For each person, we denote by $A(z)$ the potential treatment received if the person's IV were z, and by $Y(z, a)$ the potential outcome if the person's IV were z and the person's treatment received were a. We use $\mathbb{E}$ to denote an expectation over the counterfactual observations and the exogenous random mechanism defining a rule. When taking expectations over observables alone under sampling from P_0, we will instead use the symbol E_0.

3.2.2 Causal conditions

As a starting point, we require the counterfactual data unit to be well-defined.

Condition A1 (Stable Unit Treatment Value Assumption)**.** The counterfactual data unit is unaffected by the IV or treatment assigned to other individuals, and there is only a single version of the IV and treatment, so that $Z = z$ implies that $A = A(z)$ and $(Z, A) = (z, a)$ implies that $Y = Y(z, a)$.

Several conditions must be imposed on the IV. We make use of common conditions found in the IV literature (e.g., Imbens and Angrist, 1994; Abadie, 2003; Tchetgen Tchetgen and Vansteelandt, 2013; Wang and Tchetgen Tchetgen, 2018).

Condition A2 (Strong IV relevance)**.** There exists a constant $\delta^A > 0$ such that

$$|P_0\left(A = 1 \mid Z = 1, W = w\right) - P_0\left(A = 1 \mid Z = 0, W = w\right)| > \delta^A$$

holds for P_0-almost every $w \in \mathcal{W}$.

Condition A3 (Strong IV positivity)**.** There exists a constant $\delta^Z > 0$ such that $\mu_0^Z(W) \in (\delta^Z, 1 - \delta^Z)$ almost surely, where we have defined pointwise $\mu_0^Z(w) := P_0(Z = 1 \mid W = w)$.

For the purpose of identifiability, it will suffice for Conditions A2 and A3 to hold with $\delta^A = \delta^Z = 0$. However, the stronger requirements stated will prove useful when constructing estimators of the causal effects of interest.

Condition A4 (Exclusion restriction)**.** $Y(z, a) = Y(z', a)$ for all z, z' and a in $\{0, 1\}$.

Because we assume Conditions A1–A4 throughout this chapter, we do not explicitly refer to them again when stating our theorems. In light of Condition A4, we omit the z argument in the counterfactual outcome, that is, we write $Y(a)$ to refer to the common value of $Y(0, a)$ and $Y(1, a)$.

3.2.3 Causal estimands

Case I: intervention on treatment

We first introduce the case where the treatment is intervened upon. For a treatment rule $t : \mathcal{V} \to [0, 1]$, we define $Y(t)$ as the counterfactual outcome observed under an exogenous random mechanism that assigns treatment 1 with probability $t(V)$ and treatment 0 otherwise. If t were implemented in the population, then the population mean outcome would be $\mathbb{E}[Y(t)]$. We consider settings in which the proportion of people taking treatment 1 in the population is constrained to be at most $\kappa \in (0, 1]$. The constrained optimal ITR t_0 is the solution in $t : \mathcal{V} \to [0, 1]$ to

$$\text{maximize } \mathbb{E}[Y(t)] \quad \text{subject to} \quad \mathbb{E}[t(V)] \leq \kappa \ .$$

While the above optimization problem on the counterfactual data structure is the same as that studied in Luedtke and van der Laan (2016a), we focus here on settings in which the typical no unmeasured confounding assumption is implausible. As such, alternative assumptions are required to establish identifiability — we provide details in Section 3.3.1.

Once the optimal ITR t_0 has been found, we will evaluate its performance relative to that of an investigator-specified reference treatment rule t_r. Specifically, we will consider the average treatment effect (ATE) defined for an arbitrary ITR t as $\mathbb{E}\left[Y(t) - Y(t_r)\right]$. Any number of reference rules are possible, but ones likely to be of special interest are the rules that always assigns treatment 0 and rules that assign treatment 1 at random with some prespecified probability κ (e.g., the value observed in current practice or values in a plausible range of possibilities) irrespective of individual characteristics.

Remark 6. In contrast to existing works for estimating and evaluating optimal treatment rules, the conditions in this chapter allow for the presence of unmeasured modifiers of the additive effect of treatment on outcome. As a consequence, it may be that no ITR based on measured covariates alone achieves a mean outcome as favorable as the observed mean outcome. For example, clinicians may have assigned treatment based on their clinical expertise and information not recorded in the data set. Nonetheless, if the treatment strategy will be used in settings in which this additional information or clinical expertise is unavailable, an algorithm for recommending treatment would likely still prove useful in improving the population mean outcome. See the upcoming Remark 10 for further discussion.

Case II: intervention on encouragement

We now introduce the case in which the IV will be intervened upon. In this setting, after receiving encouragement or not, the individual may select a treatment at will. For an encouragement rule $e : \mathcal{V} \to [0, 1]$, we define $A(e)$ as the counterfactual outcome observed under an exogenous random mechanism that assigns encouragement with probability $e(V)$. If e were implemented in the population, then the population mean outcome would be $\mathbb{E}\left[Y(A(e))\right]$. We again consider settings in which the proportion of people taking treatment 1 is constrained to be at most $\kappa \in (0, 1]$. The constrained optimal IER e_0 is the solution in $e : \mathcal{V} \to [0, 1]$ to

$$\text{maximize } \mathbb{E}[Y(A(e))] \quad \text{subject to} \quad \mathbb{E}[A(e)] \leq \kappa \, . \tag{3.1}$$

We will evaluate the performance of e_0 relative to a reference rule $e_r : \mathcal{V} \to [0,1]$ based upon the average encouragement effect (AEE), which is defined for an arbitrary IER e as $\mathbb{E}\left[Y(A(e)) - Y(A(e_r))\right]$. We explicitly consider three reference rules. As in the ITR setting, the first rule is any fixed, investigator-specified reference rule e^{FR}. The second rule, which we denote by e_0^{RD}, encourages treatment at random uniformly across individuals but is designed to ensure either that the resource constraint is saturated or that all individuals in the population are encouraged. Symbolically, this rule is given by $e_0^{\mathrm{RD}} : v \mapsto \min\left\{(\kappa - \mathbb{E}[A(0)])/\,\mathbb{E}[A(1) - A(0)], 1\right\}$. This rule is most sensible when it is known *a priori* that encouragement is not harmful. The third rule involves assigning encouragement according to the true propensity $e_0^{\mathrm{TP}} : w \mapsto P_0\left(Z = 1 \mid W = w\right)$ used in the observed population. This reference rule is interesting because, if resource constraints allow the status quo encouragement strategy to be maintained, then it is useful to know the extent to which implementation of an optimal IER improves on the mean outcome under this status quo strategy. In view of the fact that $e_0^{\mathrm{TP}} = \mu_0^Z$, combined with an exchangeability assumption that we formalize below, we have that $\mathbb{E}\left[Y(e_r)\right] = \mathrm{E}_0\left(Y\right)$. In other words, under the conditions we consider, the counterfactual mean outcome under the status quo encouragement strategy coincides with the mean outcome in the observed population. It follows that, in observational studies, this third reference rule leads to an AEE that contrasts the mean outcome under e with the population mean outcome that occurs in the absence of intervention.

The AEE only measures the effect of treatment indirectly via the impact of encouragement upon eventual treatment receipt. To quantify the impact of treatment on outcome, we also study an alternative quantity motivated by the following two observations. First, the observed outcome for so-called always-takers (individuals who choose treatment 1 irrespective of encouragement) and never-takers (individuals who choose treatment 0 irrespective of encouragement) does not depend on the assigned encouragement, and thus, no IER can possibly lead to an improvement in the outcome of these individuals. Second, among compliers, an individual receives treatment 1 if and only if encouragement is provided. Consequently, among these individuals, the counterfactual outcome $Y(A(e))$ under an encouragement rule e has the

same distribution as the counterfactual outcome $Y(e)$ under the *treatment* rule $t : v \mapsto e(v)$. Therefore, any encouragement effect observed among compliers is in fact a treatment effect. This motivates the definition of the *local average treatment effect* (LATE) that compares any given IER e to a reference rule e_r, defined as $\mathbb{E}\left[Y(e) - Y(e_r) \mid A(1) > A(0)\right]$. It is worth noting that, in settings in which there are multiple IVs available, complier status generally relies on the particular choice of IV — this can complicate interpretation of the resulting LATEs.

Remark 7. In situations in which the cost of encouragement is non-negligible, it may be useful to consider a resource constraint that limits the proportion of encouraged people to no more than κ. The associated optimal IER is the solution in $e : \mathcal{V} \to [0, 1]$ to

$$\text{maximize} \;\; \mathbb{E}\left[Y(A(e))\right] \qquad \text{subject to} \qquad \mathbb{E}\left[e(V)\right] \leq \kappa \; .$$

This decision problem is equivalent to the decision problem addressed in Luedtke and van der Laan (2016a). The equivalence can be seen by treating the encouragement in our context as the usual treatment in Luedtke and van der Laan (2016a). We refer interested readers to that work for details.

3.3 Identification of causal estimands

We now provide sets of additional conditions under which the causal estimands described above are nonparametrically identified by parameters of the observed data distribution P_0. These identification results, which have been established before in the literature, serve as the starting point for the development of the inferential procedures presented in Section 3.4.

We begin by introducing further notation. For any observed data distribution P, we define the conditional means $\mu_P^A : (z, w) \mapsto \mathrm{E}_P\left(A \mid Z = z, W = w\right)$ and $\mu_P^Y : (z, w) \mapsto \mathrm{E}_P\left(Y \mid Z = z, W = w\right)$ and corresponding contrasts $\Delta_P^A : w \mapsto \mu_P^A(1, w) - \mu_P^A(0, w)$ and $\Delta_P^Y : w \mapsto \mu_P^Y(1, w) - \mu_P^Y(0, w)$. Additionally, we define the *conditional Wald estimand* $\Delta_P : w \mapsto \Delta_P^Y(w)/\Delta_P^A(w)$ as well as $\bar{\mu}_P^A : v \mapsto \mathrm{E}_P\left[P\left(A = 1 \mid Z = 1, W\right) \mid V = v\right]$, which we will show to identify the probability of assigning treatment 1 for an individual with covariate

level v under an appropriate condition. We also define $\underline{\bar{\mu}}_P^A : (z, v) \mapsto \mathrm{E}_P[P(A = 1 \mid Z = z, W) \mid V = v]$, $\varphi_P := \Phi(P) := \mathrm{E}_P[\mu_P^A(0, W)]$. Throughout the chapter, for ease of notation, if f_P is a quantity or operation indexed by distribution P, we denote f_{P_0} by f_0. As an example, we have that $\Delta_0 = \Delta_{P_0}$.

3.3.1 Case I: identifying assumptions for the ATE

Our identification of the ATE relies on arguments first given in Wang and Tchetgen Tchetgen (2018). Specifically, we assume that there exists a random variable U defined on the same probability space as the counterfactual random variables of interest that satisfies the following two assumptions:

Condition A5a (Exchangeability given W and U)**.** $Y(a)$ and (A, Z) are independent given (W, U).

Condition A5b. At least one of the following holds:

(1) Both conditions below hold:

   (a) (Uncorrelated IV) $\mathrm{Cov}(Y(0), Z \mid W) = 0$ almost surely;

   (b) (No unmeasured treatment-outcome effect modification)

$$\mathbb{E}\left[Y(1) - Y(0) \mid W, U\right] = \mathbb{E}\left[Y(1) - Y(0) \mid W\right] \text{ almost surely;}$$

(2) Both conditions below hold:

   (a) (Independent IV) Z and U are independent given W;

   (b) (Independent compliance)

$$\mathbb{E}\left[A(Z) \mid Z = 1, W, U\right] - \mathbb{E}\left[A(Z) \mid Z = 0, W, U\right] = \mu_0^A(1, W) - \mu_0^A(0, W) \text{ almost surely.}$$

The variable U is typically viewed as an unobserved confounder of the relationship between A and Y. Figure 3.1 shows a directed acyclic graph (DAG) satisfying these assumptions.

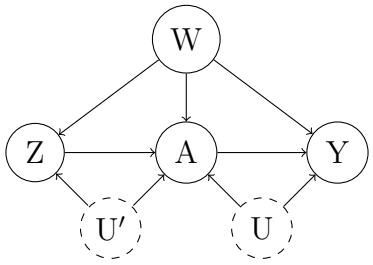

Figure 3.1: Directed acyclic graph (DAG) used to identify the ATE under Conditions A1–A4 and A5a–A5b. The dashed node U' denotes the potential unobserved confounders of the effect of Z on A. Note that Condition A5a is represented in this DAG: conditionally on (W, U), Y and (Z, A) are d-connected only through paths through A; on the contrary, Condition A5b cannot be represented on a DAG. Though not indicated by the DAG, under part (1) of A5b, it is possible that Z and U are dependent given W but nonetheless satisfy the lack of correlation assumption in part (1)a of A5b.

Notably, the causal assumptions in Toth and van der Laan (2018) automatically imply the *no unmeasured treatment-outcome effect modifier* condition, namely part (1) of A5b. When Y is not binary, our stated conditions are also slightly weaker than those presented by Wang and Tchetgen Tchetgen (2018). Similarly as in this chapter, Wang and Tchetgen Tchetgen (2018) relied on Conditions A1–A4, but that work relied on a stronger version of A5b. Specifically, they assumed that the independence assumption in part (2)a of A5b always holds, and that at least one of part (1)b and (2)b also holds. In contrast, when part (1)b holds, we only need to assume a lack of correlation (part (1)a) rather than independence.

Theorem S2 in the Supplement establishes that, under Conditions A5a–A5b, the *conditional average treatment effect* within levels of W can be identified by the conditional Wald estimand (Wald, 1940), namely yielding that $\mathbb{E}\left[Y(1) - Y(0) \mid W = w\right] = \Delta_0(w)$ for almost every w. Moreover, this result shows that the ATE contrasting rules t and t_r is given by

$$\mathbb{E}\left[Y(t) - Y(t_r)\right] = \mathrm{E}_0\left[\{t(V) - t_r(V)\}\,\Delta_0(W)\right].$$

It immediately follows that the constrained optimal ITR t_0 is identified as a solution in

$t : \mathcal{V} \to [0, 1]$ to

$$\text{maximize} \;\; \mathrm{E}_0\left[t(V)\Delta_{b,0}(V)\right] \qquad \text{subject to} \qquad \mathrm{E}_0\left[t(V)\right] \leq \kappa \;, \tag{3.2}$$

where we define pointwise $\Delta_{b,0}(v) := \mathrm{E}_0\left[\Delta_0(W) \mid V = v\right]$. This corresponds to a fractional knapsack problem (Dantzig, 1957), which is known to have a closed-form solution constructed as follows. For each covariate level $V = v$, this solution makes decisions based on the expected benefit of treatment 1, namely $\Delta_{b,0}$. Treatment resources are distributed deterministically to individuals in decreasing order of expected benefit up until all resources have been exhausted or treatment 1 is expected to be harmful (negative expected benefit). That is, the solution assigns resources of treatment 1 in a manner that prioritizes the subgroups (defined by V) with the highest expected benefit, and hence, for most subgroups, takes the form of an indicator that $\Delta_{b,0}$ is greater than a threshold. If not all subjects on the positive expected benefit boundary can receive treatment 1 due to resource constraints, the remaining resources are distributed randomly among them. That is, for individuals with expected benefit lying exactly on the threshold, the optimal ITR randomly assigns the remaining resources to them and takes a value in $[0, 1)$. Formally, defining $\eta_0^T := \inf\left\{\eta : P_0\left(\Delta_{b,0}(V) > \eta\right) \leq \kappa\right\}$ and the threshold $\tau_0^T := \max\left\{\eta_0^T, 0\right\}$, an explicit solution is given by

$$t_0(v) := \begin{cases} \frac{\kappa - P_0\left\{\Delta_{b,0}(V) > \tau_0^T\right\}}{P_0\left\{\Delta_{b,0}(V) = \tau_0^T\right\}} & : \; \Delta_{b,0}(v) = \tau_0^T, \; \tau_0^T > 0 \text{ and } P_0\{\Delta_{b,0}(V) = \tau_0^T\} > 0 \\ I\left\{\Delta_{b,0}(v) > \tau_0^T\right\} & : \; \text{otherwise} \; . \end{cases}$$

This result is formally established in Theorem S3 found in the Supplement. We remark that Condition B1 in Section 3.4.1 states that $P_0\left\{\Delta_{b,0}(V) = \tau_0^T\right\} = 0$, which is closely related to the non-exceptional law condition that first appeared in the optimal treatment literature in Robins (2004). This condition is necessarily satisfied if the distribution of expected benefits $\Delta_{b,0}(V)$ is continuous. However, we note that this may not necessarily be the case even if the distribution of V is continuous. For example, when there is no treatment effect heterogeneity, that is, $\Delta_{b,0}$ is a constant function, then the condition fails irrespective of the distribution of V.

When discussing inferential procedures in Section 3.4, it will be convenient to refer to the quantities involved above at a generic observed data distribution $P \neq P_0$. We therefore define $\Delta_{b,P}$, ξ_P, η_P^T, τ_P^T and t_P similarly to their P_0-specific analogues — this is consistent with our use of subscripting by zero to denote indexing by P_0.

Remark 8. Condition A5b requires that there is no additive interaction for at least one of the $Z \to A$ and $A \to Y$ effects. The first set of assumptions in A5b are the traditional IV assumptions that lead to point identification of the ATE — these can be stringent since they exclude unmeasured $A \to Y$ additive effect modification (Martens et al., 2006). As a union of two distinct sets of identifying assumptions, our Condition A5b, based on Wang and Tchetgen Tchetgen (2018), is strictly weaker, notably allowing unmeasured $A \to Y$ effect modification while excluding unmeasured $Z \to A$ additive effect modification. Still, this may be a strong assumption in practice. In future work, it would be interesting to develop a framework for sensitivity analyses for Condition A5b, somewhat along the lines of the work done to scrutinize the typical unmeasured confounding assumption (e.g. Robins et al., 2000; Arah et al., 2008; Groenwold et al., 2009; VanderWeele and Arah, 2011).

Remark 9. As noted earlier, we consider ITRs that are possibly stochastic. Given the result just established, it is straightforward to identify an optimal deterministic ITR using the observed data distribution, defined as the solution in $t : \mathcal{V} \to \{0, 1\}$ to (3.2). As we are optimizing over a smaller collection of rules, the value attained by an optimal deterministic ITR is no larger than that obtained by an optimal stochastic ITR. When the non-exceptional law condition $P_0\{\Delta_{b,0}(V) = \tau_0^T\} = 0$ holds (Robins, 2004), these rules almost surely recommend the same treatment. When this condition does not hold, finding an expression for an optimal deterministic rule can be challenging, even in the extreme case that the true distribution P_0 is known. Specifically, the resulting optimization problem is a 0-1 knapsack problem, which is weakly NP-hard (Martello and Toth, 1990). In contrast, as shown above, the optimal stochastic ITR t_0 admits a simple closed form.

Remark 10. We now present an example in which the mean outcome $\mathrm{E}_0(Y)$ in the observed population is larger than the observed outcome $\mathbb{E}[Y(t_0)]$ under an optimal ITR, as was suggested possible in Remark 6. This example focuses on the most favorable setting for the ITR based on measured covariates: letting $\kappa = 1$ so that there is no resource constraint, and letting $V = W$ so that the rule can make use of all measured baseline covariates when making a treatment decision. In this example, W, U and Z are generated independently from a Bern(0.5) distribution. Conditionally on (W, U, Z), the counterfactual treatment variables $A(0)$ and $A(1)$ are generated independently from $\mathrm{Bern}(0.3U + 0.4)$ and $\mathrm{Bern}(0.3U + 0.6)$ distributions, respectively. Finally, conditionally on $(W, U, Z, A(0), A(1))$, $Y(0)$ and $Y(1)$ are generated independently from $\mathrm{N}(1 - 2U, 1)$ and $\mathrm{N}(2U - 1, 1)$ distributions, respectively. We further suppose that the observed data arises under the stable unit treatment value assumption (Condition A1). Conditions A2–A4, A5a and part (2) of A5b can be shown to hold for this data-generating mechanism. Straightforward calculations show that $\mathbb{E}[Y(t)] = 0$ for any rule $t : \mathcal{W} \to [0, 1]$. In contrast, the population mean outcome $\mathrm{E}_0(Y) = \mathbb{E}[Y(A(Z))]$ can be shown to equal 0.3. Thus, implementing an optimal ITR would actually *worsen* the population mean outcome relative to the currently employed treatment assignment strategy.

Remark 11. The phenomenon illustrated in Remark 10, wherein $\mathbb{E}[Y(t_0)] < \mathrm{E}_0[Y]$, cannot occur provided both (i) $\mathrm{E}_0[A] \leq \kappa$ and (ii) part (1) of Condition A5b hold. Condition (i) ensures that t_0 is not constrained to allocate treatment in a manner that is sparser than occurred naturally in the population. Under this condition, it can be deduced that the observed mean outcome in the population is necessarily no better than the mean outcome under a (W, U)-optimal ITR, that is, under an ITR optimal among the class of rules based on both W (measured) and U (unmeasured). Because any (W, U)-optimal ITR is determined by the (W, U)-conditional ATE, Condition (ii) implies that t_0 is in fact a (W, U)-optimal ITR. Consequently, we find that $\mathbb{E}[Y(t_0)] \geq \mathrm{E}_0[Y]$.

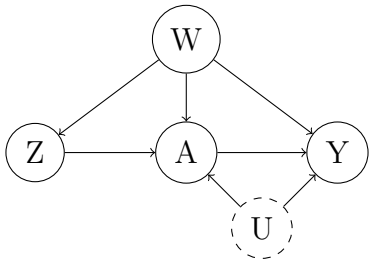

Figure 3.2: Directed acyclic graph (DAG) used to identify the AEE and LATE under Conditions A1–A4 and A6a. The dashed node U denotes the potential unobserved confounder of the effect of A on Y.

3.3.2 Case II: identifying assumptions for the AEE and LATE

We now turn to estimation of the AEE and LATE, which does not require Condition A5b but relies instead on alternative conditions. Identification of the LATE requires two such conditions whereas one suffices for the AEE. The condition that we rely upon to estimate the AEE is stated below and is illustrated in a directed acyclic graph (DAG) along with previous conditions in Figure 3.2.

Condition A6a (Unconfoundedness of encouragement with treatment and outcome)**.** Z and $(A(z), Y(a))$ are independent given W for each $z, a \in \{0, 1\}$ (Imbens and Angrist, 1994; Angrist et al., 1996).

Importantly, the above assumption does not require that the mechanism generating the encouragement Z be known, as would be the case in a randomized encouragement design. Instead, the data could have arisen from an observational study in which all confounders of the effect of encouragement on treatment and outcome are measured.

As shown in Theorem S4 in the Supplement, Condition A6a allows the identification of the *conditional average encouragement effect* within levels of W as $\mathbb{E}\left[Y(A(1)) - Y(A(0)) \mid W = w\right] = \Delta_0^Y(w)$ for each w. This readily yields that the marginal AEE contrasting encouragement

rules e and e_r can be identified as

$$\mathbb{E}\left[Y(A(e)) - Y(A(e_r))\right] = \mathrm{E}_0\left[\{e(V) - e_r(V)\}\,\Delta_0^Y(W)\right].$$

In the rest of this chapter, we consider two settings in terms of the treatment resource used by the individuals that are not encouraged. In the first setting, the individuals that are not encouraged never take the treatment. This may be the case, for example, when most individuals are unaware of the treatment (e.g., because it is novel). An equivalent setting is where the resource constraint is on the individuals who receive both encouragement and treatment. This is the case, for example, in the context of voucher-based encouragement programs in which there is a budget constraint on the total redeemed value (e.g., Peterson et al., 1999; Angrist et al., 2002). In the second more general setting, encouraged individuals may nevertheless take the treatment.

Note that Condition A6a implies that $\mathbb{E}\left[A(1) \mid V = v\right] = \bar{\mu}_P^A(v) = \underline{\bar{\mu}}_P^A(1, v)$ and $\mathbb{E}\left[A(0) \mid V = v\right] = \underline{\bar{\mu}}_P^A(0, v)$. In the first setting, any constrained optimal encouragement rule e_0 is a solution in $e : \mathcal{V} \to [0, 1]$ to

$$\text{maximize} \;\; \mathrm{E}_0\left[e(V)\Delta_{b,0}^Y(V)\right] \qquad \text{subject to} \qquad \mathrm{E}_0\left[e(V)\bar{\mu}_0^A(V)\right] \leq \kappa\,, \tag{3.3}$$

where we define pointwise $\Delta_{b,0}^Y(v) := \mathrm{E}_0\left[\Delta_0^Y(W) \mid V = v\right]$. Similarly to (3.2), the above is a fractional knapsack problem. Consequently, a solution takes an analogous form to the ITR setting in (3.2), except that here we consider the ratio between the expected benefit and the expected resource expenditure resulting from encouraging treatment 1, namely $\xi_0(v) := \Delta_{b,0}^Y(v)/\bar{\mu}_0^A(v)$, rather than the expected benefit of treatment as in the ITR setting. The intuition behind using ξ_0 rather than $\Delta_{b,0}$ is that, in this fractional knapsack problem, $\Delta_{b,0}(v)$ and $\bar{\mu}_0^A(v)$ can be viewed as the value and the weight of an item, respectively. To maximize the total value in the knapsack while constraining the total weight, a greedy algorithm that considers the unit value of each item, namely $\xi_0(v)$, can be adopted. This algorithm states that units with the highest value should be put into the knapsack. In particular, letting $\eta_0^E := \inf\left\{\eta : \mathrm{E}_0\left[I(\xi_0(V) > \eta)\bar{\mu}_0^A(V)\right] \leq \kappa\right\}$ and $\tau_0^E := \max\left\{\eta_0^E, 0\right\}$, an optimal IER

is explicitly given by

$$e_0(v) = \begin{cases} \frac{\kappa - \mathrm{E}_0\left[I(\xi_0(V) > \tau_0^E)\bar{\mu}_0^A(V)\right]}{\mathrm{E}_0\left[I(\xi_0(V) = \tau_0^E)\bar{\mu}_0^A(V)\right]} & : \text{ if } \tau_0^E > 0,\ \xi_0(v) = \tau_0^E \text{ and } \mathrm{E}_0\left[I(\xi_0(V) = \tau_0^E)\bar{\mu}_0^A(V)\right] > 0 \\ I\left(\xi_0(v) > \tau_0^E\right) & : \text{ otherwise .} \end{cases}$$

We formally show that e_0 takes this form in Theorem S5 in the Supplement.

Similarly, in the second setting, any constrained optimal encouragement rule $\underline{e}_0$ is a solution in $e : \mathcal{V} \to [0,1]$ to

$$\text{maximize } \mathrm{E}_0\left[e(V)\Delta_{b,0}^Y(V)\right] \quad \text{subject to} \quad \mathrm{E}_0\left[e(V)\underline{\bar{\mu}}_0^A(1,V) + (1-e(V))\underline{\bar{\mu}}_0^A(0,V)\right] \leq \kappa \ ,$$

Simple manipulation on the constraint shows that the above is equivalent to the following problem:

$$\text{maximize } \mathrm{E}_0\left[e(V)\Delta_{b,0}^Y(V)\right] \quad \text{subject to} \quad \mathrm{E}_0\left[e(V)\Delta_{b,0}^A(V)\right] + \varphi_0 \leq \kappa \ , \tag{3.4}$$

where we define pointwise $\Delta_{b,0}^A(v) := \mathrm{E}_0\left[\Delta_0^A(W) \mid V = v\right]$. We additionally assume two conditions so that (3.4) is also a fractional knapsack problem.

Condition A6b (Strong encouragement)**.** $\Delta_0^A(w) > \delta^A$ holds for P_0-almost every $w \in \mathcal{W}$. Here δ^A is introduced in Condition A2.

This condition differs from Condition A2 only in that the absolute value is removed, and holds when the IV is indeed an encouragement of treatment.

Condition A6c (Exsistence of nontrivial feasible IER)**.** $\varphi_0 < \kappa$.

Under Condition A6b, this condition is necessary, because if $\varphi_0 > \kappa$, then no IER can satisfy the treatment resource constraint in view of the fact that $\mathrm{E}_0\left[e(V)\Delta_{b,0}^A(V)\right] \geq 0$; if $\mathrm{E}_0[\mu_0^A(1,W)] = \kappa$, then only the trivial IER $v \mapsto 0$ satisfies the constraint and there is no need to estimate an optimal IER.

Under these two additional conditions, (3.4) is a fractional knapsack problem. With $\underline{\xi}_0(v) := \Delta_{b,0}^Y(v)/\Delta_{b,0}^A(v)$, $\underline{\eta}_0^E := \inf\left\{\eta : \mathrm{E}_0\left[I(\underline{\xi}_0(V) > \eta)\Delta_{b,0}^A(V)\right] \leq \kappa - \varphi_0\right\}$ and $\underline{\tau}_0^E :=$

$\max\left\{\underline{\eta}_0^E, 0\right\}$, an optimal IER is explicitly given by

$$\underline{e}_0(v) = \begin{cases} \frac{\kappa - \varphi_0 - \mathrm{E}_0\left[I(\underline{\xi}_0(V) > \underline{\tau}_0^E)\Delta_{b,0}^A(V)\right]}{\mathrm{E}_0\left[I(\underline{\xi}_0(V) = \underline{\tau}_0^E)\Delta_{b,0}^A(V)\right]} & : \text{ if } \underline{\tau}_0^E > 0,\ \underline{\xi}_0(v) = \underline{\tau}_0^E \text{ and } \mathrm{E}_0\left[I(\underline{\xi}_0(V) = \underline{\tau}_0^E)\Delta_{b,0}^A(V)\right] > 0 \\ I\left(\underline{\xi}_0(v) > \underline{\tau}_0^E\right) & : \text{ otherwise .} \end{cases}$$

In addition to studying the AEE of e_0, we are interested in studying the LATE of this rule. Identification of the latter requires the following additional condition.

Condition A6d (Monotonicity)**.** $A(1) \geq A(0)$ given W almost surely (Imbens and Angrist, 1994; Angrist et al., 1996; Abadie, 2003).

We remark that Condition A2 and A6d imply Condition A6b. As we show in Theorem S6 in the Supplement, under this additional condition, the LATE contrasting rules e and e_r is identified as

$$\mathbb{E}\left[Y(A(e)) - Y(A(e_r)) \mid A(1) > A(0)\right] = \frac{\mathrm{E}_0\left[\{e(V) - e_r(V)\}\,\Delta_0^Y(W)\right]}{\mathrm{E}_0\left[\Delta_0^A(W)\right]}\ .$$

In the special case where $\kappa = 1$, it is straightforward to show that $\tau_0^E = 0$ and $e_0(v) = I(\xi_0(v) > 0)$. Moreover, under Condition A6d, since $\mu_0^A > 0$, we have that $e_0(v) = I(\Delta_{b,0}^Y(v) > 0)$. Therefore, the optimal IER can be determined by the conditional LATE, which has been thoroughly studied (e.g., Imbens and Angrist, 1994; Angrist et al., 1996; Ogburn et al., 2015).

As before, for use in the next section, we define $\Delta_{b,P}^Y$, η_P^E, τ_P^E, e_P, $\Delta_{b,P}^A$, $\underline{\eta}_P^E$, $\underline{\tau}_P^E$ and $\underline{e}_P$ computed under a generic observed data distribution P analogously to the quantities defined above under P_0.

Remark 12. Under Condition A6a, the counterfactual mean outcome under e_0^{TP} is the same as the observed mean outcome: $\mathbb{E}\left[Y(A(e_0^{\mathrm{TP}}))\right] = \mathrm{E}\left(Y\right)$. As such, the AEE of e relative to this reference rule measures the increase $\mathbb{E}\left[Y(A(e))\right] - \mathrm{E}\left(Y\right)$ in population mean outcome if the rule e is implemented in the population instead of retaining the status quo encouragement strategy seen in the population.

It is natural to ask whether the conditions we outline for identification of the ATE similarly allow the identification of the expected change in population mean outcome if t were implemented, that is, the identification of $\beta := \mathbb{E}[Y(t)] - \mathrm{E}(Y)$. To investigate this question, we start by noting that

$$\beta = \mathbb{E}[Y(t) - Y(0) \mid A = 0] P_0(A = 0) - \mathbb{E}[Y(1) - Y(t) \mid A = 1] P_0(A = 1).$$

As the marginal probabilities of treatment assignment are identifiable from the observed data, a sufficient condition for the identifiability of the right-hand side is that the average effect of implementing t instead of the observed treatment among the untreated and the treated both be identified. Though this sufficient condition is useful for establishing identifiability, a *necessary* condition would be needed to establish a *lack of* identifiability of β. The aforementioned sufficient condition turns out to be necessary when t is a static treatment rule. Specifically, if t assigns treatment 0 (1) to everyone, then the right-hand side is identified if and only if the ATE among the treated (untreated), namely $\mathbb{E}[Y(1) - Y(0) \mid A = a]$ for $a = 0$ (1), is identified. As neither of these quantities is generally identified under Conditions A5a–A5b, β cannot generally be identified unless stronger conditions hold.

Remark 13. Under the above conditions, an IER that maximizes the counterfactual mean outcome is also a treatment rule that maximizes the counterfactual mean outcome among compliers. As such, these two criteria lead to the same optimal IER.

Remark 14. Similarly as with the ITRs considered in Section 3.3.1, the IERs that we consider are allowed to be stochastic. We refer to Remark 9 for a comparison between stochastic and deterministic rules.

3.4 Estimating and evaluating optimal individualized decision rules

3.4.1 Case I: optimal individualized treatment rules

Throughout this section, we assume Conditions A5a and A5b, under which the ATE is identified when the treatment is intervened upon. We aim to estimate an optimal ITR and

its ATE relative to a reference treatment rule. To construct and study our estimator, it will be convenient to define $\Psi_t^T(P) := \mathrm{E}_P\left[t(V)\Delta_P(W)\right]$ for each ITR t and distribution $P \in \mathcal{M}$, where the model $\mathcal{M}$ is locally nonparametric at P_0 (see, e.g., Chapter 1 of Pfanzagl, 1990). For $P \in \mathcal{M}$, we then define the ATE of an optimal ITR relative to a reference rule as $\Psi_r^T(P) := \Psi_{t_0}^T(P) - \Psi_{t_r}^T(P)$. Our goal here is to make inference about $\psi_0^T := \Psi_r^T(P_0)$. Our proposed estimator will be of the form $\psi_n^T := \Psi_{t_n}^T(\hat{P}_n) - \Psi_{t_r}^T(\hat{P}_n)$, where t_n is an estimator of t_0 and $\hat{P}_n$ is an estimator of P_0 obtained using the TMLE framework.

Before presenting the proposed estimator, we must establish conditions under which the parameter of interest is pathwise differentiable and compute its canonical gradient. The latter will be a key ingredient in the construction of a regular and asymptotically linear point estimator of ψ_0^T and an asymptotically valid confidence interval for ψ_0^T. Throughout this chapter, for ease of notation, if f_P is a quantity or operation f_P indexed by distribution P, we denote $f_{\hat{P}_n}$ by $\hat{f}_n$. As an example, we have that $\hat{\Delta}_n := \Delta_{\hat{P}_n}$. Additionally, we use the notation P_n to refer to the empirical distribution.

Pathwise differentiability of the ATE

To establish the pathwise differentiability of the ATE at P_0 relative to the locally nonparametric model $\mathcal{M}$, we require two additional conditions:

Condition B1 (Non-exceptional law)**.** $P_0\left\{\Delta_{b,0}(V) = \tau_0^T\right\} = 0$.

Condition B2 (Nonzero continuous density of $\Delta_{b,0}(V)$ at η_0^T)**.** If $\eta_0^T > -\infty$, then the distribution of $\Delta_{b,0}(V)$ has positive, finite and continuous Lebesgue density at η_0^T.

When $V = W$ and $\eta_0^T > -\infty$, Condition B2 states that the distribution of the conditional ATE $\Delta_0(W)$ is continuous in a neighborhood of η_0^T. When $V \neq W$, Condition B2 can be interpreted similarly for the ATE conditional on V. This may be reasonable if, for example, the distribution of $\Delta_0(W)$ is continuous and the mapping $w \mapsto V(w)$ is differentiable.

For each candidate distribution $P \in \mathcal{M}$, functions μ^Y and μ^A, ITR t, and decision thresh-

old $\tau \in \mathbb{R}$, we define pointwise

$$D^T(P, \mu^Y, \mu^A, t, \tau)(o) := \frac{t(v)}{\Delta^A(w)\left[z + \mu_P^Z(w) - 1\right]} \left\{ y - \mu_P^Y(z,w) - \Delta(w)\left[a - \mu_P^A(z,w)\right] \right\} + \left[t(v)\Delta_P(w) - \Psi_t^T(P) \right] - \tau\left[t(v) - \kappa\right],$$

where we have defined $\Delta^A : w \mapsto \mu^A(1,w) - \mu^A(0,w)$, $\Delta^Y : w \mapsto \mu^Y(1,w) - \mu^Y(0,w)$ and $\Delta := \Delta^Y/\Delta^A$, and suppressed in our notation the dependence of D^T on κ.

Theorem 3.1 (Pathwise differentiability of ATE)**.** *Under Conditions B1 and B2, the parameters $P \mapsto \Psi_{t_P}^T(P)$ and $P \mapsto \Psi_{t_r}^T(P)$ are pathwise differentiable at P_0 relative to $\mathcal{M}$ with respective canonical gradients $D^T(P_0, \mu_0^Y, \mu_0^A, t_0, \tau_0^T)$ and $D^T(P_0, \mu_0^Y, \mu_0^A, t_r, 0)$.*

The above implies that the ATE parameter Ψ_r^T of an optimal ITR has nonparametric canonical gradient $D_r^T(P_0) := D^T(P_0, \mu_0^Y, \mu_0^A, t_0, \tau_0^T) - D^T(P_0, \mu_0^Y, \mu_0^A, t_r, 0)$ at P_0.

Description of proposed estimator

We propose the following procedure to estimate an optimal ITR and its ATE:

1. Use the empirical distribution $\hat{P}_{W,n}$ as estimate of the marginal distribution of W under P_0. Compute estimates μ_n^Y, μ_n^A, μ_n^Z and $\Delta_{b,n}$ of μ_0^Y, μ_0^A, μ_0^Z and $\Delta_{b,0}$ using flexible regression methods.

2. Let $\Delta_n^A : w \mapsto \mu_n^A(1,w) - \mu_n^A(0,w)$, $\Delta_n^Y : w \mapsto \mu_n^Y(1,w) - \mu_n^Y(0,w)$ and $\Delta_n : w \mapsto \Delta_n^Y(w)/\Delta_n^A(w)$.

3. Estimate an optimal ITR:

   (a) let $\eta_n^T := \inf\{\eta : P_n\{\Delta_{b,n}(V) > \eta\} \leq \kappa\}$ and $\tau_n^T := \max\{\eta_n^T, 0\}$;

   (b) estimate t_0 by

$$t_n : v \mapsto \begin{cases} \frac{\kappa - P_n\{\Delta_{b,n}(V) > \tau_n^T\}}{P_n\{\Delta_{b,n}(V) = \tau_n^T\}} & : \text{ if } \Delta_{b,n}(v) = \tau_n^T,\ \tau_n^T > 0 \text{ and } P_n\{\Delta_{b,n}(V) = \tau_n^T\} > 0 \\ I(\Delta_{b,n}(v) > \tau_n^T) & : \text{ otherwise.} \end{cases}$$

4. Estimate the ATE relative to reference rule t_r as follows:

   (a) obtain a targeted estimate $\hat{\mu}_n^Y$ of μ_0^Y by running an ordinary least-squares linear regression with outcome Y, covariate $h(Z,W) := \frac{t_n(V)-t_r(V)}{[Z+\mu_n^Z(W)-1]\Delta_n^A(W)}$, offset $\mu_n^Y(Z,W)$ and no intercept;

   (b) obtain a targeted estimate $\hat{\mu}_n^A$ of μ_0^A by running a logistic regression with outcome A, covariate $h(Z,W)\Delta_n(W)$, offset $\operatorname{logit}\mu_n^A(Z,W)$ and no intercept;

   (c) letting $\hat{P}_n$ denote any distribution with components $\hat{\mu}_n^Y$, $\hat{\mu}_n^A$ and $\hat{P}_{W,n}$, estimate the ATE of t_0 relative to t_r with $\psi_n^T := \Psi_{t_n}^T(\hat{P}_n) - \Psi_{t_r}^T(\hat{P}_n)$.

Remark 15. To obtain the targeted estimate $\hat{\mu}_n^Y$, we could use any exponential family regression model with canonical link function fitted via maximum likelihood. If g is the canonical link for the chosen exponential family, the offset should be taken to be $g(\mu_n^Y(Z,W))$ but the covariate and outcome remain the same as with the linear regression model, and the fitted mean function gives $\hat{\mu}_n^Y$. The choice of exponential family can affect the finite-sample performance of the resulting estimator. For example, when Y is binary, use of a logistic regression may lead to better performance. This approach can also be used when it is known that $Y \in [a,b]$, in which case the outcome would first be scaled to fall in $[0,1]$.

Properties of the proposed estimator

We now provide conditions under which our proposed estimator of the ATE is asymptotically nonparametric efficient (Bickel et al., 1993a). For convenience in presenting these results, we suppose that $\hat{P}_n$ has component μ_n^Z, even though μ_n^Z is not explicitly needed to evaluate ψ_n^T.

Condition B3 (Consistency of strong IV positivity)**.** With probability tending to one over the sample used to define μ_n^Z, it holds that $\int I\{\delta^Z < \mu_n^Z(w) < 1-\delta^Z\}dP_0(w) = 1$.

Condition B4 (Consistency of strong IV relevance)**.** With probability tending to one over the sample used to define Δ_n^A and $\hat{\Delta}_n^A$, it holds that $\int I\{\delta^A < |\Delta_n^A(w)| < 1-\delta^A\}dP_0(w) = 1$ and $\int I\{\delta^A < |\hat{\Delta}_n^A(w)| < 1-\delta^A\}dP_0(w) = 1$.

Condition B5 (Fast rate of estimated optimal ITR)**.** As sample size n tends to infinity, it holds that

$$\int \{t_n(v) - t_0(v)\}\{\Delta_{b,0}(v) - \tau_0^T\} dP_0(w) = \mathrm{o}_p(n^{-1/2}).$$

Because Condition B5 may seem difficult to verify, we discuss it in detail in Appendix B.2.

A few additional technical conditions are needed below, and these are stated in Appendix B.1.1. Condition B6 requires that the estimators μ_n^Z, μ_n^A and μ_n^Y have a sufficiently fast rate of convergence. Condition B7 requires asymptotic boundedness of the initial and targeted estimators of the conditional ATE function. Condition B8 requires the consistency of the influence function estimator used. Conditions B9 and B10 are empirical process conditions that place some restrictions on the flexibility of the algorithms used to obtain μ_n^Z, μ_n^A and μ_n^Y.

Theorem 3.2 (Asymptotic linearity of ATE estimator)**.** *Under Conditions B1–B10, it holds that*

$$\psi_n^T - \psi_0^T = \frac{1}{n}\sum_{i=1}^{n} D_r^T(P_0)(O_i) + \mathrm{o}_p(n^{-1/2}) \ .$$

In particular, ψ_n^T *is consistent and* $\sqrt{n}\left(\psi_n^T - \Psi_0^T\right) \xrightarrow{d} \mathrm{N}\left(0, \sigma_0^2\right)$*, where* $\sigma_0^2 := \mathrm{E}_0\left[D_r^T(P_0)(O)^2\right]$.

If σ_n^2 is a consistent estimator of σ_0^2 and $z_{\alpha/2}$ is the $(1-\alpha/2)^{th}$ standard normal quantile, then $(\psi_n^T - z_{\alpha/2}\sigma_n/\sqrt{n}, \psi_n^T + z_{\alpha/2}\sigma_n/\sqrt{n})$ is an asymptotically valid $(1-\alpha)$-level confidence interval for ψ_0^T. By inspecting the remainder of the expansion of ψ_n^T, we note that ψ_n^T is doubly robust in the sense that if t_n is consistent for t_0, then ψ_n^T is consistent for ψ_0^T if either μ_n^Z or (μ_n^Y, μ_n^A) is consistent.

Remark 16. In traditional ITR settings, it has been observed that using the same data to both estimate and evaluate the performance of an individualized rule can lead to overly optimistic estimates of performance, where this optimism manifests as positive bias for the performance estimator (Zhang et al., 2012; Chakraborty et al., 2014; van der Laan and Luedtke, 2014). Though this bias is often asymptotically negligible, it can have a meaningful impact on the conclusions drawn in finite samples. Van der Laan and Luedtke (2014)

presented a cross-validated estimator of the mean outcome and formally showed that it yields negligible bias under reasonable conditions. Moreover, they showed that when these conditions fail, their estimator will typically yield *negative* bias for the performance measure of interest. Luedtke and van der Laan (2016b) presented a related sample splitting estimator and showed that this negative bias property causes the lower confidence bound derived from this estimator to be valid under only mild conditions. The IV setting that we consider in this chapter leads to a different estimation setting than the one considered previously. Nonetheless, in preliminary simulation studies, we have seen that the same general positive-bias phenomenon holds in our case. We therefore present cross-validated versions of the above method in Supplementary Appendix S2.1, and recommend it for use in practice. As in the setting considered by van der Laan and Luedtke (2014), the theoretical arguments justifying the use of this method are entirely analogous to the arguments for the method proposed above, though the additional notation needed to properly reflect the choice of sample splits is burdensome. These arguments are therefore omitted.

3.4.2 Case II: optimal individualized encouragement rules

In this section, we aim to estimate an optimal individualized encouragement rule and its AEE/LATE relative to a reference encouragement rule. Throughout this section, we assume Condition A6a, that is, that Z and $(A(z), Y(a))$ are independent given W for each $z, a \in \{0, 1\}$. This assumption identifies the AEE. To construct and study our estimators, it will be convenient to define the parameter $\Psi^E_e(P) := \mathrm{E}_P\left[e(V)\Delta^Y_P(W)\right]$ for each IER e and distribution $P \in \mathcal{M}$. For $P \in \mathcal{M}$, the AEE of an optimal IER relative to a reference rule $e^{\mathcal{R}}_P$ is thus given by $\Psi^E_{\mathcal{R}}(P) := \Psi^E_{e_P}(P) - \Psi^E_{e^{\mathcal{R}}_P}(P)$ ($\underline{\Psi}^E_{\mathcal{R}}(P) := \Psi^E_{\underline{e}_P}(P) - \Psi^E_{\underline{e}^{\mathcal{R}}_P}(P)$ resp., where $\underline{e}^{\mathcal{R}}_P$ is the reference rule corresponding the second setting that will be defined later). As the notation indicates, the reference rule may itself depend on P — for example, the reference rule e^{RD}_P involves the probability that an individual receives treatment 1 given encouragement to treatment 1 and covariates under P. We are interested in inference about $\psi^E_0 := \Psi^E_{\mathcal{R}}(P_0)$, where we have suppressed dependence on $\mathcal{R}$ from our shorthand notation. Our estimator

of ψ_0^E will have the form $\psi_n^E := \Psi_{e_n}^E(\hat{P}_n) - \Psi_{e_n^{\mathcal{R}}}^E(\hat{P}_n)$, where e_n and $e_n^{\mathcal{R}}$ are estimators of e_0 and $e_0^{\mathcal{R}}$, respectively, and $\hat{P}_n$ is a TMLE-based estimator of P_0. In this section, we consider $\mathcal{R}$ to be any element of $\{\text{FR}, \text{RD}, \text{TP}\}$. We note that $\underline{e}_0^{\text{FR}} = e_0^{\text{FR}} = e^{\text{FR}}$ since the fixed reference rule e_P^{FR} does not depend on P; $\underline{e}_0^{\text{TP}} = e_0^{\text{TP}}$ by definition; in the first setting, $e_0^{\text{RD}} : v \mapsto \min\left\{\kappa / \text{E}_0\left[\mu_0^A(1,W)\right], 1\right\}$, while $\underline{e}_0^{\text{RD}} : v \mapsto \min\left\{(\kappa - \varphi_0) / \text{E}_0\left[\Delta_0^A(1,W)\right], 1\right\}$ in the second setting. When estimating the LATE of an optimal IER relative to a reference rule, we also assume that Condition A6d holds.

For the rest of this chapter, when the conditions and results for both settings are similar, we state them simultaneously with the difference for the second setting in brackets. When discussing the conditions and results, we focus on the first setting whenever the second setting is similar.

Pathwise differentiability of the AEE and LATE

In order to establish the pathwise differentiability of the AEE and LATE, and to derive their respective canonical gradients, additional conditions are needed.

Condition C1 (Non-exceptional law)**.** In the first setting, $P_0(\xi_0(V) = \tau_0^E) = 0$; in the second setting, $P_0(\underline{\xi}_0(V) = \underline{\tau}_0^E) = 0$.

Condition C2 (Nonzero continuous density of $\xi_0(V)$ ($\underline{\xi}_0(V)$ resp.) around η_0^E ($\underline{\eta}_0^E$ resp.))**.** If η_0^E($\underline{\eta}_0^E$ resp.) $> -\infty$, then the distribution of $\xi_0(V)$ ($\underline{\xi}_0(V)$ resp.) has positive, finite and continuous Lebesgue density in a neighborhood of η_0^E ($\underline{\eta}_0^E$ resp.).

Condition C3 (Smooth resource expenditure function or lack of constraint)**.** If η_0^E($\underline{\eta}_0^E$ resp.) $> -\infty$, then the function $\eta \mapsto \text{E}_0\left[I\left(\xi_0(V) > \eta\right)\mu_0^A(1,W)\right]$ ($\eta \mapsto \text{E}_0\left[I\left(\underline{\xi}_0(V) > \eta\right)\Delta_{b,0}^A(V)\right]$ resp.) is continuously differentiable with nonzero derivative in a neighborhood of η_0^E ($\underline{\eta}_0^E$); and if η_0^E($\underline{\eta}_0^E$ resp.) $= -\infty$ and $\kappa < 1$, then $\text{E}_0\left[\mu_0^A(1,W)\right] < \kappa$ ($\text{E}_0\left[\Delta_0^A(W)\right] < \kappa - \varphi_0$ resp.).

Under Condition C2, if $\eta_0^E > -\infty$, that is, if treatment resources do not suffice for everyone to receive encouragement, then Condition C3 is guaranteed if $\eta \mapsto \text{E}_0\left[\mu_0^A(1,W) \mid \xi_0(V) = \eta\right]$

is continuous in a neighborhood of η_0^E. If $\eta_0^E = -\infty$, then $\mathrm{E}_0\left[\mu_0^A(1,W)\right] \leq \kappa$ and Condition C3 only excludes the boundary case in which treatment resources just barely suffice for encouragement to be provided to everyone (except in the trivial case in which $\kappa = 1$ and everyone is either a complier or always-taker).

Condition C4 (Strong positivity of proportion of always-takers and compliers)**.** In the first setting, there exists some $\delta_{\bar{\mu}^A} > 0$ such that $\bar{\mu}_0^A(V) > \delta_{\bar{\mu}^A}$ almost surely.

Condition C4 holds if both Conditions A2 and A6d are true. Note that the counterpart for the second setting has been stated in Condition A6b. Below, we will refer to the following assumption whenever $\mathcal{R} = \mathrm{RD}$.

Condition C5 (Active constraint)**.** If $\mathcal{R} = \mathrm{RD}$, then it holds that $\kappa / \mathrm{E}_0[\mu_0^A(1,W)] < 1$ ($(\kappa - \varphi_0)/\mathrm{E}_0\left[\mu_0^A(1,W)\right]$ resp.).

For a distribution $P \in \mathcal{M}$, a function μ^A, an IER e, and a decision threshold $\tau \in \mathbb{R}$, we define pointwise

$$
\begin{aligned}
D^E(P,e,\tau,\mu^A)(o) \;:=\; & \left[\frac{e(v)}{z+\mu_P^Z(w)-1}\right]\left[y-\mu_P^Y(z,w)\right] + e(v)\Delta_P^Y(w) - \Psi_e^E(P) \\
& - \tau\left[e(v)\left\{\frac{z}{\mu_P^Z(w)}\left[a-\mu^A(1,w)\right]+\mu^A(1,w)\right\}-\kappa\right]; \\
D(P,\mu^A)(o) \;:=\; & \frac{z}{\mu_P^Z(w)}\left[a-\mu^A(1,w)\right]+\mu^A(1,w)-\mathrm{E}_P\left[\mu^A(1,W)\right]\;; \\
G_{\mathrm{RD}}^E(P)(o) \;:=\; & D^E(P,e_P^{\mathrm{RD}},0,\mu_P^A) - \left\{\mathrm{E}_P\left[\mu_P^A(1,W)\right]\right\}^{-1}\Psi_{e_P^{\mathrm{RD}}}^E(P)D(P,\mu_P^A)\;; \\
\underline{D}^E(P,e,\tau,\mu^A)(o) \;:=\; & \left[\frac{e(v)}{z+\mu_P^Z(w)-1}\right]\left[y-\mu_P^Y(z,w)\right] + e(v)\Delta_P^Y(w) - \Psi_e^E(P) \\
& - \tau\Bigg[e(v)\left\{\frac{1}{z+\mu_P^Z(w)-1}\left[a-\mu^A(z,w)\right]+\Delta^A(w)\right\} \\
& \qquad + \frac{1-z}{1-\mu_P^Z(w)}[a-\mu^A(0,w)]+\mu^A(0,w)-\kappa\Bigg]; \\
D_1(P,\mu^A)(o) \;:=\; & \frac{1-z}{1-\mu_P^Z(w)}\left[a-\mu^A(0,w)\right]+\mu^A(0,w)-\mathrm{E}_P\left[\mu^A(0,W)\right]\;; \\
D_2(P,\mu^A)(o) \;:=\; & \frac{1}{z+\mu_P^Z(w)-1}\left[a-\mu^A(z,w)\right]+\Delta^A(w)-\mathrm{E}_P\left[\Delta^A(W)\right]\;;
\end{aligned}
$$

$$\begin{aligned}
\underline{G}^E_{\mathrm{RD}}(P)(o) &:= D^E(P, \underline{e}^{\mathrm{RD}}_P, 0, \mu^A_P)(o) - (\kappa - \varphi_P)^{-1}\Psi^E_{\underline{e}^{\mathrm{RD}}_P}(P)D_1(P, \mu^A_P) \\
&\quad - \left\{\mathrm{E}_P\left[\Delta^A_P(1, W)\right]\right\}^{-1}\Psi^E_{\underline{e}^{\mathrm{RD}}_P}(P)D_2(P, \mu^A_P)\ ; \\
G^E_{\mathrm{TP}}(P)(o) &:= \frac{z}{\mu^Z_P(w)}\left[y - \mu^Y_P(z, w)\right] + z\Delta^Y_P(w) - \Psi^E_{e_{P,r}}(P)\ .
\end{aligned}$$

We also define pointwise $G^E(P)(o) := D^E(P, e_P, \tau^E_P, \mu^A_P)(o)$, $\underline{G}^E(P)(o) := \underline{D}^E(P, \underline{e}_P, \underline{\tau}^E_P, \mu^A_P)(o)$ and $G^E_{\mathrm{FR}}(P)(o) := D^E(P, e^{\mathrm{FR}}, 0, \mu^A_P)(o)$. We also define $\underline{G}^E_{\mathrm{TP}} := G^E_{\mathrm{TP}}$ and $\underline{G}^E_{\mathrm{FR}} := G^E_{\mathrm{FR}}$.

Theorem 3.3 (Pathwise differentiability of AEE)**.** *Let $\mathcal{R} \in \{FR, RD, TP\}$. Under Conditions A3 and C1–C5, the parameters $P \mapsto \Psi^E_{e_P}(P)$ ($\Psi^E_{\underline{e}_P}(P)$ resp.) and $P \mapsto \Psi^E_{e^{\mathcal{R}}_P}(P)$ ($\Psi^E_{\underline{e}^{\mathcal{R}}_P}(P)$ resp.) are pathwise differentiable at P_0 relative to $\mathfrak{M}$ with respective canonical gradients $G^E(P_0)$ ($\underline{G}^E(P_0)$ resp.) and $G^E_{\mathcal{R}}(P_0)$ ($\underline{G}^E_{\mathcal{R}}(P_0)$ resp.).*

The above implies that the AEE parameter $P \mapsto \Psi^E_{\mathcal{R}}(P)$ ($\underline{\Psi}^E_{\mathcal{R}}(P)$ resp.) of an optimal IER is pathwise differentiable with nonparametric canonical gradient $D^E_{\mathcal{R}}(P_0) := G^E(P_0) - G^E_{\mathcal{R}}(P_0)$ ($\underline{D}^E_{\mathcal{R}}(P_0) := \underline{G}^E(P_0) - \underline{G}^E_{\mathcal{R}}(P_0)$ resp.) at P_0 for $\mathcal{R} \in \{\mathrm{FR}, \mathrm{RD}, \mathrm{TP}\}$.

Proposed estimator

We now present our procedure for estimating an optimal IER and the corresponding AEE. It will turn out that, when $e^{\mathcal{R}}_0$ is the rule underlying the data-generating mechanism, namely $e^{\mathrm{TP}}_0 : w \mapsto P_0(Z = 1 \mid W = w)$, it will be convenient to take Z_i as estimate of $e^{\mathcal{R}}_0(W_i)$, even though Z_i is not a function of W_i. Therefore, to be able to handle all cases we consider for $e^{\mathcal{R}}_0$ using a common notation, we allow $e^{\mathcal{R}}_n$ to be a function of the full observation o rather than only as a function of w or v. As we show in Theorem 3.4 and Corollary 3.4.1 below, the estimators of $e^{\mathcal{R}}_n$ that we use result in the asymptotic linearity of our estimator of the AEE (or LATE) under reasonable conditions.

We first describe the estimator in the first setting:

1. Use the empirical distribution $\hat{P}_{W,n}$ as estimate of the marginal distribution of W under P_0. Compute estimates μ^Y_n, μ^Z_n, $\Delta^Y_{b,n}$ and $\bar{\mu}^A_n$ of μ^Y_0, μ^Z_0, $\Delta^Y_{b,0}$ and $\bar{\mu}^A_0$ using flexible regression methods.

2. Estimate an optimal IER:

   (a) let $\xi_n := \Delta^Y_{b,n}/\bar{\mu}^A_n$. Set $\Gamma_n : \tau \mapsto \frac{1}{n}\sum_{i:\xi_n(V_i)>\tau} \mu^A_n(1, W_i)$ and $\gamma_n : \tau \mapsto \frac{1}{n}\sum_{i:\xi_n(V_i)=\tau} \mu^A_n(1, W_i)$. For any $k \in [0, 1]$, define $\eta^E_n(k) := \inf\{\tau : \Gamma_n(\tau) \leq k\}$, $\tau^E_n(k) := \max\left\{\eta^E_n(k), 0\right\}$ and

$$d_{n,k} : v \mapsto \begin{cases} \frac{k-\Gamma_n(\eta^E_n(k))}{\gamma_n(\eta^E_n(k))} & : \text{ if } \xi_n(v) = \eta^E_n(k) \text{ and } \gamma_n(\eta^E_n(k)) > 0 \ , \\ I\{\xi_n(v) > \eta^E_n(k)\} & : \text{ otherwise.} \end{cases}$$

   (b) compute k_n, which is used to define an estimate of e_0 for which the plug-in estimator is asymptotically linear, as follows:

   - if $\tau^E_n(\kappa) > 0$ *and* there is a solution in $k \in [0, 1]$ to

$$\frac{1}{n}\sum_{i=1}^{n} d_{n,k}(V_i)\left\{\mu^A_n(1, W_i) + \frac{I(Z_i = 1)}{\mu^Z_n(W_i)}\left[I(A_i = 1) - \mu^A_n(1, W_i)\right]\right\} = \kappa \ , \quad (3.5)$$

     then take k_n be this solution.

   - else, set $k_n = \kappa$.

   (c) estimate e_0 with

$$e_n : v \mapsto \begin{cases} \frac{k_n-\Gamma_n(\tau^E_n(k_n))}{\gamma_n(\tau^E_n(k_n))} & : \text{ if } \xi_n(v) = \tau^E_n(k_n) \text{and } \gamma_n(\tau^E_n(k_n)) > 0 \\ I\{\xi_n(v) > \tau^E_n(k_n)\} & : \text{ otherwise.} \end{cases}$$

3. estimate the reference rule $e^{\mathcal{R}}_0$ as follows (depending on which reference rule is used):

   - For $\mathcal{R} = \text{FR}$, take $e^{\mathcal{R}}_n : o \mapsto e^{\text{FR}}(v)$.
   - For $\mathcal{R} = \text{RD}$,
     (a) obtain a targeted estimate $\hat{\mu}^A_n(1, \cdot)$ of $\mu^A_0(1, \cdot)$ by running a logistic regression with outcome A, covariate $Z/\mu^Z_n(W)$, offset $\text{logit}\,\mu^A_n(1, W)$ and no intercept, and let $\hat{\mu}^A_n$ be the fitted mean model;
     (b) let $e^{\mathcal{R}}_n : o \mapsto \min\left\{1, \kappa/\frac{1}{n}\sum_{i=1}^{n} \hat{\mu}^A(1, W_i)\right\}$, a function that is constant in its input.

- For $\mathcal{R} = \text{TP}$, take $e_n^{\mathcal{R}} : o \mapsto z$.

4. estimate AEE of e_0 relative to the reference rule $e_0^{\mathcal{R}}$:

   (a) obtain a targeted estimate $\hat{\mu}_n^Y$ of μ_0^Y by running an ordinary least-squares linear regression with outcome Y, covariate $[e_n(V) - e_n^{\mathcal{R}}(O)]/[Z + \mu_n^Z(W) - 1]$, offset $\mu_n^Y(Z, W)$ and no intercept, and take $\hat{\mu}_n^Y$ to be the fitted mean function.

   (b) letting $\hat{P}_n$ be any distribution with components $\hat{\mu}_n^Y$ and $\hat{P}_{W,n}$, estimate the AEE of e_0 relative to $e_0^{\mathcal{R}}$ with $\psi_n^E := \Psi_{e_n}^E(\hat{P}_n) - \Psi_{e_n^{\mathcal{R}}}^E(\hat{P}_n)$.

The estimator in the second setting is similar.

1. Almost identical to the first setting, except that $\Delta_{b,n}^A$ needs to be estimated.

2. Estimate an optimal IER:

   (a) Estimate φ_0 with a one-step correction estimator $\varphi_n := \frac{1}{n}\sum_{i=1}^n \{\mu_n^A(0, W_i) + \frac{1-Z_i}{1-\mu_n^Z(W_i)}[A_i - \mu_n^A(0, W_i)]\}$. Let $\underline{\xi}_n := \Delta_{b,n}^Y / \Delta_{b,n}^A$. Set $\underline{\Gamma}_n : \tau \mapsto \frac{1}{n}\sum_{i:\underline{\xi}_n(V_i) > \tau} \Delta_n^A(W_i)$ and $\underline{\gamma}_n : \tau \mapsto \frac{1}{n}\sum_{i:\underline{\xi}_n(V_i) = \tau} \Delta_n^A(W_i)$. For any $k \in [0, 1]$, define $\underline{\eta}_n^E(k) := \inf\{\tau : \underline{\Gamma}_n(\tau) \leq k - \varphi_n\}$, $\underline{\tau}_n^E(k) := \max\left\{\underline{\eta}_n^E(k), 0\right\}$ and

$$\underline{d}_{n,k} : v \mapsto \begin{cases} \frac{k - \varphi_n - \underline{\Gamma}_n(\underline{\eta}_n^E(k))}{\underline{\gamma}_n(\underline{\eta}_n^E(k))} & : \text{ if } \underline{\xi}_n(v) = \underline{\eta}_n^E(k) \text{ and } \underline{\gamma}_n(\underline{\eta}_n^E(k)) > 0 \ , \\ I\{\underline{\xi}_n(v) > \underline{\eta}_n^E(k)\} & : \text{ otherwise.} \end{cases}$$

   (b) Compute $\underline{k}_n$, which is used to define an estimate of $\underline{e}_0$ for which the plug-in estimator is asymptotically linear, as follows:

   - if $\underline{\tau}_n^E(\kappa) > 0$ *and* there is a solution in $k \in [0, 1]$ to

$$\frac{1}{n}\sum_{i=1}^n \underline{d}_{n,k}(V_i)\left\{\Delta_n^A(W_i) + \frac{1}{Z_i + \mu_n^Z(W_i) - 1}\left[A_i - \mu_n^A(Z_i, W_i)\right]\right\} + \varphi_n = \kappa \ , \tag{3.6}$$

     then take $\underline{k}_n$ be this solution.

- else, set $\underline{k}_n = \kappa$.

(c) Estimate $\underline{e}_0$ with

$$\underline{e}_n : v \mapsto \begin{cases} \frac{\underline{k}_n - \varphi_n - \underline{\Gamma}_n(\underline{\tau}_n^E(\underline{k}_n))}{\underline{\gamma}_n(\underline{\tau}_n^E(\underline{k}_n))} & : \text{ if } \underline{\xi}_n(v) = \underline{\tau}_n^E(\underline{k}_n) \text{and } \underline{\gamma}_n(\underline{\tau}_n^E(\underline{k}_n)) > 0 \\ I\{\underline{\xi}_n(v) > \underline{\tau}_n^E(k_n)\} & : \text{ otherwise.} \end{cases}$$

3. Estimate the reference rule $\underline{e}_0^{\mathcal{R}}$ as follows (depending on which reference rule is used):

   - For $\mathcal{R} =$ FR or TP, $\underline{e}_n^{\mathcal{R}}$ is identical to $e_n^{\mathcal{R}}$ in the first setting.
   - For $\mathcal{R} =$ RD,
     (a) obtain a targeted estimate $\underline{\hat{\mu}}_n^A(1, \cdot)$ of $\mu_0^A(1, \cdot)$ by running a logistic regression with outcome A, covariate $1/(Z + \mu_n^Z(W) - 1)$, offset $\operatorname{logit} \mu_n^A(Z, W)$ and no intercept, and let $\hat{\mu}_n^A$ be the fitted mean model;
     (b) let $\underline{e}_n^{\mathcal{R}} : o \mapsto \min\left\{1, (\kappa - \varphi_n)/\frac{1}{n}\sum_{i=1}^n \underline{\hat{\Delta}}_n^A(W_i)\right\}$, a function that is constant in its input. Here, we define pointwise $\underline{\hat{\Delta}}_n^A(w) := \underline{\hat{\mu}}_n^A(1, w) - \underline{\hat{\mu}}_n^A(0, w)$.

4. Same as the first setting with $(e_n, e_n^{\mathcal{R}})$ replaced by $(\underline{e}_n, \underline{e}_n^{\mathcal{R}})$. In this second setting, we denote the targeted estimate of μ_0^Y by $\underline{\hat{\mu}}_n^Y$, the distribution to be plugged in by $\underline{\hat{P}}_n$, and the estimator of the AEE by $\underline{\psi}_n^E$.

If instead the LATE is of interest, estimation of $\phi_0 := \mathrm{E}_0\left[\Delta_0^A(W)\right]$ is also needed. However, we note that ϕ_0 corresponds to the evaluation of the standard G-computation parameter at P_0. We can use the approach outlined above to estimate the AEE, use existing methods (e.g., van der Laan and Rose, 2018) to estimate ϕ_0, and then finally take the ratio of these estimators to obtain an estimator of the LATE.

Remark 17. In the first setting, the fact that k_n solves (3.5) is key to ensuring that ψ_n^E is asymptotically linear. If we naïvely use κ when estimating the optimal IER as in Section 3.4.1, then the resulting ψ_n^E might not be asymptotically linear. Moreover, k_n may be viewed as a revised estimate of κ aiming to obtain a better estimate $\tau_n^E(k_n)$ of τ_0^E in the

sense that resource constraints are better respected. A similar remark applies to the second setting.

Remark 18. Similarly as discussed for the ITR setting in Remark 15, the targeting steps used to obtain $\hat{\mu}_n^Y$ ($\underline{\hat{\mu}}_n^Y$ resp.) and $\hat{\mu}_n^A$ ($\underline{\hat{\mu}}_n^A$ resp.) need not be framed as ordinary least-squares regression problems. Instead, any exponential family GLM with the canonical link function fitted via maximum likelihood could be employed.

Results about the proposed estimator

We now present results on the asymptotic linearity of our proposed estimator under some conditions. For convenience in presenting the results, we suppose that $\hat{P}_n$ ($\underline{\hat{P}}_n$ resp.) have component μ_n^Z and $\hat{\mu}_n^A$ ($\underline{\hat{\mu}}_n^A$ resp.) (if $\hat{\mu}_n^A$ or $\underline{\hat{\mu}}_n^A$ is obtained in the procedure, i.e., when $\mathcal{R} = \text{RD}$), even though μ_n^Z and $\hat{\mu}_n^A$ ($\underline{\hat{\mu}}_n^A$ resp.) are not explicitly used in the plug-in estimator. We let e_n^{TP} and e_n^{RD} be the estimators of e_0^{TP} and e_0^{RD}, respectively, even though e_0^{TP} is not formally an IER according to our definition but simply a decision for each observation in the data. We also use ψ_0^E ($\underline{\psi}_n^E$ resp.) to denote the true AEE corresponding to the chosen reference rule.

Condition C6 (Consistency of strong IV positivity)**.** Condition B3 holds.

Condition C7 (Consistency of strong positivity of proportion of always-takers and compliers (strong encouragement resp.))**.** With probability tending to one, it holds that $\int I(\bar{\mu}_n^A(v) \geq \delta_{\bar{\mu}^A}) dP_0(v) = 1$ ($\int I(\Delta_n^A(w) > \delta_A) dP_0(w) = 1$ and $\int I(\Delta_{b,n}^A(v) > \delta_A) dP_0(v) = 1$ resp.).

Condition C8 (Fast rate of estimated optimal IER)**.** As sample size n tends to infinity, it holds that

$$\int \{e_n(v) - e_0(v)\} \left\{\Delta_{b,0}^Y(v) - \tau_0^E \bar{\mu}_0^A(v)\right\} dP_0(w) = \text{o}_p(n^{-1/2})$$

in the first setting, and

$$\int \{\underline{e}_n(v) - \underline{e}_0(v)\} \left\{\Delta_{b,0}^Y(v) - \underline{\tau}_0^E \Delta_{b,0}^A(v)\right\} dP_0(v) = \text{o}_p(n^{-1/2})$$

in the second setting.

As in the case of the ATE, Condition C8 is discussed in detail in Appendix B.2.

Additional technical conditions are required for the theorem below, and these are stated in Appendix B.1.2. Condition C9 requires that the flexible nuisance estimators μ_n^Z, μ_n^A and μ_n^Y converge to their respective true targets at a sufficient rate. Condition C10 requires the consistency of the influence function estimator used. Conditions C11 and C12 are empirical process conditions that place some restrictions on the flexibility of the algorithms used to obtain μ_n^Z, μ_n^A and μ_n^Y.

Theorem 3.4 (Asymptotic linearity of AEE estimator)**.** *Suppose that Conditions C1–C4 and C6–C12 hold and let $\mathcal{R} \in \{FR, RD, TP\}$. Furthermore, if $\mathcal{R} = RD$, suppose that Condition C5 holds. If the reference rule is $e_0^{\mathcal{R}}$ ($\underline{e}_0^{\mathcal{R}}$ resp.), then it holds that*

$$\psi_n^E - \psi_0^E = \frac{1}{n}\sum_{i=1}^{n} D_{\mathcal{R}}^E(P_0)(O_i) + \mathrm{o}_p(n^{-1/2})$$

in the first setting, and

$$\underline{\psi}_n^E - \underline{\psi}_0^E = \frac{1}{n}\sum_{i=1}^{n} \underline{D}_{\mathcal{R}}^E(P_0)(O_i) + \mathrm{o}_p(n^{-1/2})$$

in the second setting. In particular, $\sqrt{n}\left(\psi_n^E - \psi_0^E\right) \xrightarrow{d} \mathrm{N}\left(0, \sigma_0^2\right)$ ($\sqrt{n}\left(\underline{\psi}_n^E - \underline{\psi}_0^E\right) \xrightarrow{d} \mathrm{N}\left(0, \sigma_0^2\right)$ resp.), where $\sigma_0^2 := \mathrm{E}_0\left[D_{\mathcal{R}}^E(P_0)(O)^2\right]$ ($\mathrm{E}_0\left[\underline{D}_{\mathcal{R}}^E(P_0)(O)^2\right]$ resp.).

If σ_n^2 is a consistent estimator of σ_0^2 and $z_{\alpha/2}$ is the $(1-\alpha/2)^{th}$ standard normal quantile, then, in the first setting, $(\psi_n^E - z_{\alpha/2}/\sqrt{n}\sigma_n, \psi_n^E + z_{\alpha/2}\sigma_n/\sqrt{n})$ is an asymptotically valid $(1-\alpha)$-level confidence interval. By inspecting the remainder of the expansion of ψ_n^E ($\underline{\psi}_n^E$ resp.), we note that ψ_n^E ($\underline{\psi}_n^E$ resp.) is doubly robust in the sense that if e_n ($\underline{e}_n$) is consistent for e_0 ($\underline{e}_0$ resp.), then ψ_n^E ($\underline{\psi}_n^E$ resp.) is consistent for ψ_0^E ($\underline{\psi}_0^E$ resp.) if either μ_n^Z or $\hat{\mu}_n^Y$ ($\underline{\hat{\mu}}_n^Y$ resp.) is consistent; if $\mathcal{R} = \mathrm{RD}$, at least one of $w \mapsto \hat{\mu}_n^A(1, w)$ ($\underline{\hat{\mu}}_n^A$ resp.) and μ_n^Z also needs to be consistent for ψ_n^E ($\underline{\psi}_n^E$ resp.) to be consistent.

Corollary 3.4.1 (Asymptotic linearity of LATE estimator)**.** *Let ϕ_n be a nonparametric RAL estimator of $\phi_0 := \mathrm{E}_0\left[\Delta_0^A(W)\right]$. Let ψ_0^E ($\underline{\psi}_0^E$ resp.) and $D_{\mathcal{R}}^E(P_0)$ ($\underline{D}_{\mathcal{R}}^E(P_0)$ resp.),*

$\mathcal{R} \in \{FR, RD, TP\}$, denote the reference rule-specific AEE and influence function. Suppose that Conditions A2, A6d, C1–C4 and C6–C12 hold and let $\mathcal{R} \in \{FR, RD, TP\}$. Furthermore, if $\mathcal{R} = RD$, suppose that Condition C5 holds. If the reference rule is $e_0^{\mathcal{R}}$ ($\underline{e}_0^{\mathcal{R}}$ resp.), then it holds that $\phi_0 > 0$,

$$\frac{\psi_n^E}{\phi_n} - \frac{\psi_0^E}{\phi_0} = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{D_{\mathcal{R}}^E(P_0)(O_i)}{\phi_0} - \frac{\psi_0^E}{\phi_0^2}D_A(P_0)(O_i)\right] + \mathrm{o}_p(n^{-1/2})$$

in the first setting, and

$$\frac{\underline{\psi}_n^E}{\phi_n} - \frac{\underline{\psi}_0^E}{\phi_0} = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{\underline{D}_{\mathcal{R}}^E(P_0)(O_i)}{\phi_0} - \frac{\underline{\psi}_0^E}{\phi_0^2}D_A(P_0)(O_i)\right] + \mathrm{o}_p(n^{-1/2})$$

in the second setting, where for each $P \in \mathcal{M}$ we define

$$D_A(P) : o \mapsto \frac{a - \mu_P^A(z,w)}{z + \mu_P^Z(w) - 1} + \Delta_P^A(w) - P\Delta_P^A \ .$$

In particular, ψ_n^E/ϕ_n ($\underline{\psi}_n^E/\phi_n$ resp.) is a consistent estimator of the LATE parameter value. Furthermore,

$$\sqrt{n}\left(\frac{\psi_n^E}{\phi_n} - \frac{\psi_0^E}{\phi_0}\right) \xrightarrow{d} \mathrm{N}\left(0,\ \mathrm{E}_0\left[\frac{D_{\mathcal{R}}^E(P_0)(O)}{\phi_0} - \frac{\psi_0^E}{\phi_0^2}D_A(P_0)(O)\right]^2\right)$$

in the first setting, and

$$\sqrt{n}\left(\frac{\underline{\psi}_n^E}{\phi_n} - \frac{\underline{\psi}_0^E}{\phi_0}\right) \xrightarrow{d} \mathrm{N}\left(0,\ \mathrm{E}_0\left[\frac{\underline{D}_{\mathcal{R}}^E(P_0)(O)}{\phi_0} - \frac{\underline{\psi}_0^E}{\phi_0^2}D_A(P_0)(O)\right]^2\right)$$

in the second setting.

As before, a Wald-type confidence interval may thus be readily constructed. In this corollary, we note that $D_A(P)$ can be shown to be the nonparametric canonical gradient of $P \mapsto \mathrm{E}_P\left[\Delta_P^A(W)\right]$.

Remark 19. Similarly to the discussion given in Remark 16, in practice, it may be desirable to use sample splitting to estimate an optimal IER to avoid the positive bias that can result from estimating and evaluating an IER on the same data set. This algorithm is described in Supplementary Appendix S2.2.

3.5 Simulation

In the simulation, we generated data from an IV model that simultaneously satisfies Conditions A1–A4, A5a–A5b and A6a–A6d. We evaluated the performance of our methods for estimating:

(i) RC-AEE: the AEE of an optimal IER relative to reference rule

$$e_0^{\mathrm{RD}} : v \mapsto \min\left\{\frac{\kappa}{\mathrm{E}_0\left[\mu_0^A(1,W)\right]}, 1\right\},$$

under treatment resource constraints, in the first setting;

(ii) RC-AEE2: the AEE of an optimal IER relative to reference rule

$$\underline{e}_0^{\mathrm{RD}} : v \mapsto \min\left\{\frac{\kappa - \mathrm{E}_0\left[\mu_0^A(0,W)\right]}{\mathrm{E}_0\left[\Delta_0^A(W)\right]}, 1\right\},$$

under treatment resource constraints, in the second setting;

(iii) ATE: the ATE of an optimal ITR relative to the reference rule that assigns treatment 0 to everyone, under no treatment resource constraint;

(iv) RC-ATE: the ATE of an optimal ITR relative to the reference rule that assigns each individual to treatment 1 at random with probability κ, under treatment resource constraints.

We exclude evaluation of optimal IERs without resource constraints because our method is algorithmically equivalent to that described in Luedtke and van der Laan (2016a). We also exclude estimation of the LATE since it follows simply from estimation of the AEE.

The data was generated as follows. First, a trivariate covariate $W := (W_1, W_2, W_3)$ was simulated with $W_1 \sim \mathrm{Unif}(-1, 1)$, $W_2 \sim \mathrm{Bern}(0.5)$ and $W_3 \sim \mathrm{N}(0, 1)$ jointly independent. Next, U was simulated from a $\mathrm{Bern}(0.5)$ distribution independently of W. An IV Z was simulated as a $\mathrm{Bern}(\mathrm{expit}\{2.5W_1 + 0.5W_2W_3\})$ conditional upon (W, U). We considered the following distribution of the compliance types: conditionally on (W, U), the proportions of compliers and always-takers are $0.9 \times \mathrm{expit}\left(2 + W_1 - W_2 + 0.7W_3\right) - 0.9 \times \mathrm{expit}\left(W_1 - W_2 + 0.7W_3\right)$

and $0.9 \times \text{expit}\,(W_1 - W_2 + 0.7W_3) + 0.05 + 0.1\,(U - 0.5)$, respectively; the rest are never-takers. This implies that $P_0(A = 1 \mid Z, W, U) = 0.9\text{expit}\,\{2Z + W_1 - W_2 + 0.7W_3\} + 0.05 + 0.1\,(U - 0.5)$ for the observable A. Conditionally on (W, U), the counterfactual outcome $Y(a)$ was taken to be distributed as a Bern(expit$\{2aW_1 + 0.2W_2 - 0.5W_3 + 0.5aU\}$) random variable. We simulated A and Y from the implied distribution on observables, namely

$$\begin{aligned} A \mid Z, W, U &\sim \text{Bern}\left(0.9\text{expit}\,\{2Z + W_1 - W_2 + 0.7W_3\} + 0.05 + 0.1\,(U - 0.5)\right), \\ Y \mid Z, A, W, U &\sim \text{Bern}\left(\text{expit}\,\{2AW_1 + 0.2W_2 - 0.5W_3 + 0.5AU\}\right). \end{aligned}$$

Here, U is an unmeasured confounder of the effect of treatment on outcome. Nevertheless, under this data-generating mechanism. the ATE and AEE are identified with the aid of the IV.

We considered $V = W$ in the construction of individualized rules. Since the distribution of $\Delta_0(W)$ is continuous, Condition B1 is satisfied.

We estimated nuisance functions using the Super Learner (van der Laan et al., 2007) with library including logistic regression, generalized additive model with logit link (Hastie and Tibshirani, 1990) and gradient boosting (Mason et al., 1999, 2000; Friedman, 2001, 2002). Because none of μ_0^Y, μ_0^A and μ_0^Z follow a logistic regression model, the resulting ensemble learner were known not to achieve the parametric convergence rate. In evaluating rules under resource constraints, we set $\kappa = 0.25$ ($\kappa = 0.5$ for RC-AEE2), which is an active constraint. We considered sample sizes of 200, 500, 1000 and 2000. Our implementation incorporated the sample splitting described in Remark 16 and Supplementary Appendix S2.

For the estimands ATE and RC-ATE, we compare our results with those based on the more naïve method in Luedtke and van der Laan (2016a). This method assumes no unmeasured confounding and does not use the IV to assess average causal effects. An R package of this method can be found at `https://github.com/alexluedtke12/sg`. For the estimand RC-AEE, we do not compare with any other method because, to the best of our knowledge, there is no comparable method to estimate the optimal IER and the corresponding AEE under treatment resource constraints.

For all scenarios, we investigated the bias, root mean squared error (RMSE) and root median squared error of the estimator. We also computed the coverage probability and the width of nominal 95% Wald CIs constructed using influence function-based standard error estimates. We further investigated the probability that our confidence limit was below the true average effect, that is, the coverage probability of the 97.5% Wald confidence lower bound. We note that due to the fact that the Wald estimand is a ratio, our proposed plug-in estimators do suffer from the fact that the estimated denominator may at times be close to zero, thus generating extreme estimates. When calculating the bias and RMSE of our proposed method, we truncated the estimates to lie in $[-1, 1]$ because the outcome Y is binary. We also computed the proportion of runs in which this truncation was required. We also note that for IER in the second setting, if $\varphi_n \geq \kappa$, the estimator is undefined. The probability of this extreme event should tend to zero as sample size tends to infinity, and we computed the proportion of runs in which this event happened.

Table 3.1 presents the performance of our proposed estimator (IV) and of the method in Luedtke and van der Laan (2016a), which assumes no unmeasured confounder (Con). For our proposed method, the coverage of CIs was lower than 95% for RC-AEE and RC-AEE2 but close to 95% for ATE and RC-ATE, especially for large sample size; the confidence lower bounds were all below the truth with high probability. Therefore, our Wald CI appears provides a meaningful confidence lower bound for the average effect of the true optimal individualized rule in all four scenarios. In contrast, the method in Luedtke and van der Laan (2016a) does not use the IV and makes the (invalid) no unmeasured confounders assumption – its 95% Wald CI coverage was lower than 95%. Because our method has a negative bias, our estimator is a conservative estimator of the average causal effect. The bias and RMSE of our proposed method appear to tend to zero as sample size grows; our proposed method appears to have higher bias and higher variance than the method in Luedtke and van der Laan (2016a). In this simulation setting, the overall performance of the method in Luedtke and van der Laan (2016a) appears desirable because U is not a strong confounder: without treatment resource constraints, the true ATE is 0.123 while the incorrect limit of the

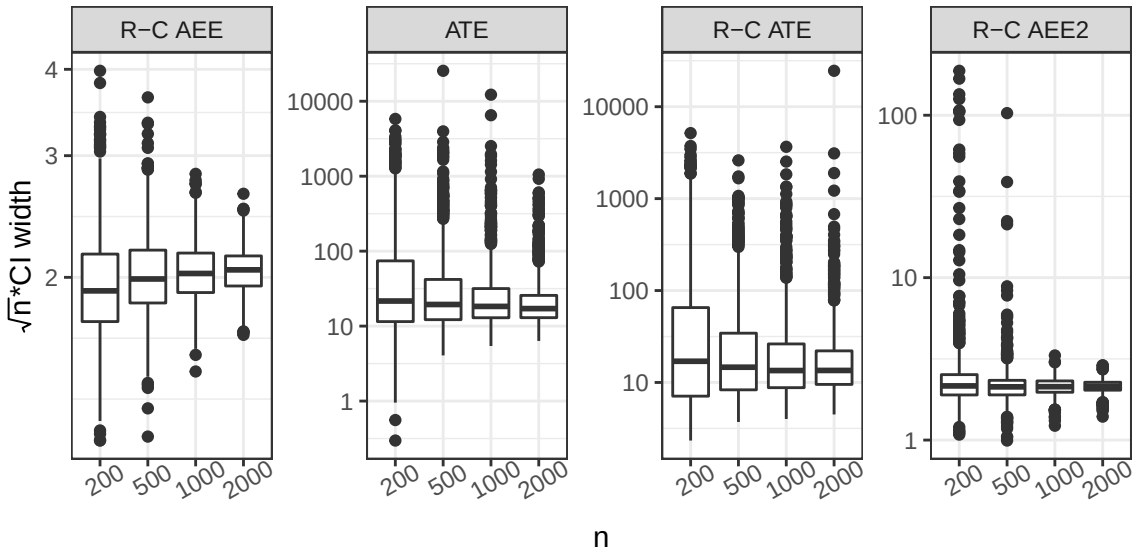

Figure 3.3: Boxplot of $\sqrt{n}\times$ CI width for each estimand.

estimator of Luedtke and van der Laan (2016a) is 0.125; with treatment resource constraints, these two quantities are 0.073 and 0.072. In both cases, the intrinsic confounding bias is very small. The proportion of truncated estimates for our proposed method was 4.5% for ATE and 3.4% for RC-ATE respectively. The proportions of event $\varphi_n \geq \kappa$ for RC-AEE2 are 6.8%, 0.6%, 0% and 0% for n =200, 500, 1000 and 2000, respectively, so the probability of this extreme event appears to tend to zero. The coverage was poor for RC-AEE and RC-AEE2. There are several possible explanations for this: the potential negative bias induced by sample splitting, the instability of the estimator of a ratio of two functions, the slower rate of convergence achieved by the machine learning nuisance estimators, and the relatively small sample sizes. Figure 3.3 presents the width of our 95% Wald CI scaled by the square root of sample size n. As theory suggests, the width of these CIs appears to shrink at a root-n rate. There are, however, rare outlying cases in which the CI is quite wide. In practice, this issue can be mitigated by fine-tuning the machine learning algorithms or using a one-step correction procedure.

Table 3.1: Performance of estimators of average causal effects in Simulation 1.

		RC-AEE	RC-AEE2	ATE		RC-ATE	
Performance measure	Sample size	IV	IV	IV	Con	IV	Con
95% Wald CI coverage	200	71%	80%	94%	89%	94%	83%
	500	74%	75%	94%	91%	92%	85%
	1000	77%	78%	94%	94%	94%	90%
	2000	77%	80%	95%	94%	96%	93%
97.5% confidence lower	200	98%	98%	99%	98%	$> 99\%$	98%
bound coverage	500	98%	98%	$> 99\%$	98%	99%	98%
	1000	98%	98%	99%	98%	$> 99\%$	99%
	2000	99%	99%	$> 99\%$	98%	$> 99\%$	99%
bias	200	-0.033	-0.027	-0.011	-0.015	-0.012	-0.024
	500	-0.022	-0.020	-0.022	-0.008	-0.021	-0.061
	1000	-0.014	-0.013	-0.017	-0.003	-0.007	-0.007
	2000	-0.011	-0.009	-0.013	-0.002	-0.001	-0.005
RMSE	200	0.063	0.056	0.425	0.072	0.339	0.050
	500	0.040	0.039	0.324	0.046	0.293	0.033
	1000	0.027	0.027	0.281	0.031	0.236	0.021
	2000	0.019	0.018	0.200	0.022	0.214	0.014
root median SE	200	0.045	0.039	0.217	0.013	0.126	0.034
	500	0.027	0.028	0.139	0.030	0.094	0.021
	1000	0.019	0.019	0.096	0.021	0.069	0.014
	2000	0.013	0.013	0.066	0.014	0.047	0.009

We performed another simulation study to compare the two methods under a setting with stronger unmeasured confounding. The data-generating mechanism remains the same except that U is a stronger confounder of the $A \to Y$ effect:

$$\begin{aligned} A \mid Z, W, U &\sim \text{Bern}\left(0.6\text{expit}\left\{2Z + W_1 - W_2 + 0.7W_3\right\} + 0.2 + 0.4\left(U - 0.5\right)\right), \\ Y \mid Z, A, W, U &\sim \text{Bern}\left(\text{expit}\left\{AW_1 + 0.2W_2 - 0.5W_3 + 4(U - 0.5)\right\}\right). \end{aligned}$$

The true ATE and the limit of the method in Luedtke and van der Laan (2016a) are 0.028 and 0.329, respectively, without treatment resource constraints, and 0.021 and 0.030, respectively, with treatment resource constraints.

Table 3.2 presents the performance of the two methods in this scenario. Our method performed similarly as in the previous simulation. The proportions of event $\varphi_n \geq \kappa$ for RC-AEE2 are 18%, 7.6%, 1.5% and 0% for n =200, 500, 1000 and 2000, respectively. The proportions are higher than the previous simulation, but κ is also closer to φ_0 in this simulation setting, making it more difficult to conclude that $\varphi_0 < \kappa$ from data. The proportion of truncated estimates for our proposed method was 2.3% for ATE and 4.2% for RC-ATE, respectively. However, for the method that does not make use of the IV, the coverage of 95% Wald CI and 97.5% confidence lower bound is lower than the nominal level. When seeking to estimate the ATE, the coverage of the method in Luedtke and van der Laan (2016a) was nearly zero. Therefore, our method appears clearly preferable whenever unmeasured confounding is suspected to exist and an IV is available.

3.6 Results from the study of suicide in US Army soldiers

3.6.1 Study objectives, characteristics and data

Policy makers are interested in characterizing the impact of combat deployment on the risk of suicide. However, doing so is not straightforward since decisions regarding deployment of eligible soldiers are generally informed by soldier characteristics. For example, unit members considered to be less fit for deployment due to physical or emotional problems are typically

Table 3.2: Performance of estimators of average causal effects in Simulation 2.

		RC-AEE	RC-AEE2	ATE		RC-ATE	
Performance measure	Sample size	IV	IV	IV	Con	IV	Con
95% Wald CI coverage	200	78%	85%	97%	3%	96%	83%
	500	76%	81%	95%	$< 1\%$	95%	84%
	1000	74%	78%	93%	$< 1\%$	93%	87%
	2000	78%	78%	92%	$< 1\%$	94%	88%
97.5% confidence lower	200	96%	99%	$> 99\%$	3%	$> 99\%$	97%
bound coverage	500	96%	97%	$> 99\%$	$< 1\%$	$> 99\%$	97%
	1000	96%	97%	$> 99\%$	$< 1\%$	$> 99\%$	96%
	2000	98%	98%	$> 99\%$	$< 1\%$	$> 99\%$	92%
bias	200	-0.023	-0.022	0.074	0.290	0.008	-0.021
	500	-0.017	-0.014	0.009	0.297	0.028	-0.010
	1000	-0.013	-0.013	-0.021	0.300	-0.009	-0.003
	2000	-0.010	-0.010	-0.050	0.302	-0.015	0.002
RMSE	200	0.061	0.057	0.470	0.301	0.365	0.048
	500	0.042	0.037	0.383	0.301	0.319	0.028
	1000	0.031	0.028	0.286	0.302	0.283	0.018
	2000	0.022	0.020	0.227	0.303	0.214	0.012
root median SE	200	0.043	0.035	0.223	0.290	0.158	0.030
	500	0.027	0.025	0.165	0.169	0.123	0.018
	1000	0.020	0.018	0.119	0.121	0.099	0.011
	2000	0.014	0.012	0.084	0.085	0.066	0.008

not deployed. This has led to speculation about the existence of a so-called *healthy warrior* effect, whereby potentially harmful effects of deployment may not be apparent due to the greater physical or emotional health of deployed soldiers (Haley, 1998). Recent research has shown that deployment during the Afghanistan and Iraq wars was not strongly associated with risk of suicide overall, but that subsequent suicide risk was significantly elevated among soldiers who left service before they were eligible for deployment (Reger et al., 2015). These two results have not yet been integrated, though, into a cohesive description of the effects of deployment on suicide adjusting for the healthy warrior effect and of the heterogeneity of this effect as a function of baseline individual differences in soldier vulnerability factors.

To investigate this question, we analyzed data on all 55,272 soldiers with a Direct Combat Arms military occupation in the Regular US Army who began their first duty assignment after completing training between January 2004 and December 2007. These soldiers were assigned across a total of 405 units. Each unit was characterized as having either a low (or high) probability-of-deployment (POD) at the time a given soldier joined the unit based on whether less (or more) than 50% of the soldiers who joined the unit in the prior six months were deployed within 18 months of joining the unit. This definition is dynamic, with a single unit possibly defined as having a low and high POD at different times. This dichotomous characterization was used as an IV based on the fact that initial unit assignment after graduating from Army training is largely random, strongly predictive of deployment, and its impact on suicide risk, if any, is expected to be mediated by deployment. Furthermore, a unit's POD status at any given time depends on its position on the unit deployment schedule and not on any soldier characteristic. The existence of an IV in this problem is fortunate since there is likely unmeasured confounding of the relationship between deployment and suicide risk. We applied our method for ITR and IER in the first setting, and have not applied our method for the second setting yet.

The month before the start of deployment was used as the 'start month' for each deployed soldier. For each soldier who did not deploy, a start month was selected using a probability mechanism that made the distribution of time from joining until the start month equal to

the observed distribution among soldiers in the same unit who deployed. Information from a range of administrative records (e.g., medical and pharmacy claims, job performance, criminal justice) was obtained for each soldier as of the end of the start month. This information was used to extract a series of 17 baseline binary variables that we used to develop optimal individualized rules. These variables included two indicators of time in service (13+ months; 18+ months); two indicators of rank (E4+; E5+); one indicator of demotion in the past 12 months; four indicators of the number of days hospitalized for mental health or substance abuse in the past 12 months (1+; 3+; 6+; 11+ days); four indicators of the number of outpatient mental health visits in the past 12 months (1+; 3+; 6+; 11+ visits); two indicators of disqualification on the Army rating system for fitness to deploy (physical; mental); one indicator of suicide attempt in the past 12 months; and one indicator of having been accused of a major violent crime in the past 12 months.

Using the generic notation employed throughout this chapter, the relevant variables in this analysis of this data set are the following:

- the baseline covariate vector W = vector of the 17 variable enumerated above;
- the IV Z = indicator of assignment to a low POD unit;
- the treatment (exposure) A = indicator of non-deployment;
- the outcome Y = indicator of the soldier having neither died by suicide, made a nonfatal suicide attempt, been hospitalized for a psychiatric disorder, or had medically-reported suicidal ideation in the first 24 months post-deployment.

We now discuss the plausibility of key conditions in this setting. The fact that $V = W$ is discrete makes Condition C1 implausible when $\kappa < 1$. However, since the support of V contains 2^{17} points, we believe that Condition C1 is likely to be approximately valid with our sample size, that is, that the probability discussed in that condition is very close to zero. Based on our earlier discussion, Condition A6a is plausible in this setting. As our

encouragement is defined as belonging to a unit with a low or high POD, the monotonicity condition, namely A6d, is also plausible. To identify the optimal individualized deployment strategy, we require that Conditions A5a and A5b hold. Condition A5a is similar to A6a, and is therefore plausible in this setting. Condition A5b states that the indicator of assignment to a low POD unit is a valid instrument conditional on the 17 baseline covariates, and that at least one of the effect of deployment on suicide and the effect of low POD unit assignment on deployment has no unmeasured additive effect modifier. Though it is not possible to be certain that this assumption strictly holds, we believe that it is more likely to (approximately) hold than the traditional no unmeasured confounder assumption in this setting. Therefore, we believe that the IV approach represents the best option for estimating an optimal deployment strategy based on these data.

3.6.2 Implementation and results

Of the $55,272$ soldiers in this data set, $38,404$ were in low POD units while the remaining $16,868$ were in high POD units. Overall, the observed prevalence of a negative outcome ($Y = 0$; either suicide death, nonfatal suicide attempt, or psychiatric hospitalization in the 18 months after the start month) was 2.3%. In high-POD units ($Z = 0$), a negative outcome was seen in 2.4% of those who deployed ($A = 0$) versus 2.5% of those who did not ($A = 1$). In contrast, in low-POD units ($Z = 1$), a negative outcome was seen in 1.6% of those who did deploy versus 2.4% of those who did not. We used the methods developed in this chapter to construct optimal individualized treatment (deployment vs no deployment) or encouragement (assignment to high vs low POD unit) rules and to make inference about the benefits derived from using these rules. Since any rule must result in a sufficient large number of deployed soldiers, we constrained the proportion of soldiers not deployed at level $\kappa \in \{0.50, 0.25, 0.10\}$. All nuisance functions involved in the inferential methods described in Section 3.4 were estimated using a Super Learner (SL) with library consisting of generalized linear models with various regularization penalties (Lasso, ridge, elastic net), with shrinkage parameters selected via cross-validation, and with appropriate

link functions. In all of our analyses, the inferential method was run five times with different pre-specified random seeds, and results were averaged. This averaging reduces the random noise inherited from the use of random folds in our method. We also investigated how each covariate contributes to the estimated optimal rule, whose functional form may be complex. To summarize the contribution of the various covariates to the rule, we fitted a lasso-logistic model (Tibshirani, 1996) with the outcome being the estimated individualized recommended encouragement or treatment for each person.

We now report the results of our analysis. When constraining the resulting proportion of soldiers deployed to be no smaller than 50%, we find that the estimated counterfactual probability of suicide under an optimal ITR is 0.00021 (95% CI: -0.00211, 0.00254) less than under the reference rule that assigns deployment randomly across all soldiers until a contingent of sufficient size is achieved. We also find that, among soldiers who could not get deployed in a low POD unit unless they got deployed in a high POD, the estimated counterfactual probability of suicide under an optimal IER is 0.00069 (95% CI: -0.00135, 0.00273) less than under the reference rule that assigns soldiers to low vs high POD units randomly in a way to satisfy the constraints on the proportion deployed. These results indicate that the available soldier characteristics do not appear to provide sufficient information in order to devise either unit assignment or deployment rules that meaningfully reduce the risk of suicide. Qualitatively, the results obtained using different constraints on the proportion of soldiers deployed (75+; 90+ % deployed) are similar. Point estimates are provided in Table 3.3 along with approximate 95% confidence intervals.

As for the contribution of the covariates to the individualized rules, we found that (i) the indicator of having ≤ 18 months in service, (ii) the indicator of having 3+ days hospitalized for mental health or substance abuse in the past 12 months, and (iii) the indicator of having 6+ days hospitalized for mental health or substance abuse in the past 12 months, contribute most to both the individualized unit-assignment and deployment rules across all considered values of κ. When intervening on deployment, the indicator of demotion in the past 12 months also strongly affects the individualized recommendation. Our finding is consistent

Table 3.3: Estimates and approximate 95% confidence intervals for the ATE, AEE and LATE of optimal ITRs and IERs at different constraint levels on the proportion of soldiers deployed in the unit.

min deployment	ATE	AEE	LATE
50%	0.00021 (-0.00211, 0.00254)	0.00045 (-0.00088, 0.00178)	0.00069 (-0.00135, 0.00273)
75%	-0.00056 (-0.00271, 0.00160)	0.00087 (-0.00052, 0.00227)	0.00134 (-0.00083, 0.00352)
90%	0.00053 (-0.00132, 0.00238)	-0.00042 (-0.00162, 0.00078)	-0.00064 (-0.00249, 0.00120)

with earlier evidence from the literature that having $\leq$ 18 months in service is associated with high suicide risk.

3.7 Conclusion

There has been extensive work on estimation of optimal individualized treatment rules and on evaluation of their performance when there are no unmeasured confounders. In this chapter, we have proposed novel methods to address these problems in settings in which there may be unmeasured confounders, but an instrumental variable is available. We focused specifically on binary treatments and instrumental variables. We also incorporated into our framework how resource constraints can be taken into account in the definition of individualized rules. Despite their importance in practice, such constraints have seldom been considered in the literature. In future work, it would be interesting to extend our approach to be able to handle longitudinal settings in which treatment is assigned over multiple time points, thereby making our approach relevant to the treatment of chronic conditions.

Chapter 4

LEVERAGING VAGUE PRIOR INFORMATION IN GENERAL MODELS VIA ITERATIVELY CONSTRUCTED GAMMA-MINIMAX ESTIMATORS

4.1 Introduction

It is often of scientific interest to estimate an aspect of the data-generating mechanism underlying the data at hand. To obtain a sensible estimator, we often use certain principles to guide the search for a good estimator. Asymptotic efficiency (Pfanzagl, 1990), minimaxity and Bayes optimality (Berger, 1985) are popular examples of such principles. Defining the performance criteria underlying these principles requires specifying a *model space*, that is, a collection of possible data-generating mechanisms that contains the true, underlying distribution.

It is often desirable to incorporate prior information about the true data-generating mechanism into a statistical procedure. This might be done differently in different statistical paradigms. For example, prior information is often incorporated by specifying the model space. For frequentist methods such as those based on the asymptotic efficiency or minimax principle, this is the primary way to incorporate prior information. However, in some applications, there is more vague prior information that cannot be accurately represented in this manner. In the Bayesian paradigm, such information may be represented by further specifying a *prior distribution* (or *prior* for short) over the model space and aiming for an estimator that minimizes the induced Bayes risk. However, in many cases, there may be several priors that are compatible with the available information; this is especially likely to occur if the model space is rich. The Gamma-minimax paradigm provides a principled means to overcome this challenge. Under this paradigm, the statistician first specifies the

set Γ of all priors that are all consistent with available prior information and subsequently aims for an estimator that minimizes the worst-case Bayes risk over the set of priors. The Gamma-minimax paradigm may be viewed as a robust version of the Bayesian paradigm (Vidakovic, 2000). Moreover, the Gamma-minimax paradigm is closely related to Bayes and minimax paradigms: when the set of priors consists of one prior, a Gamma-minimax estimator is Bayes with respect to that prior; when the set Γ of priors is the entire set of possible prior distributions, under mild conditions, a Gamma-minimax estimator is minimax (Wald, 1945). In this chapter, we focus on Gamma-minimax estimation.

Gamma-minimax estimators have been studied in a variety of problems. Some explicit forms of Gamma-minimax estimators have been obtained. For example, Olman and Shmundak (1985) studied Gamma-minimax estimation of the mean of a normal distribution for the set of symmetric and unimodal priors on an interval and obtained an explicit form when this interval is sufficiently small. Eichenauer-Herrmann (1990) generalized this result to more general parametric models and Eichenauer-Herrmann et al. (1994) obtained a further generalization with the requirement of symmetry on the priors dropped. Chen et al. (1988) studied Gamma-minimax estimation for multinomial distributions and the set of priors with bounded mean. Chen et al. (1991) studied Gamma-minimax estimation for one-parameter exponential families and the set of priors that place certain bounds on the first two moments. These results do not deal with general model spaces, such as semiparametric or nonparametric models, and general forms of the set of priors that may not directly impose bounds on prior moments of the parameters of interest. One reason for this lack of generality might be that, in the existing literature, Gamma-minimaxity is usually defined only for parametric models. Another possible explanation is that it is typically intractable to analytically derive Gamma-minimax estimators.

To overcome this lack of analytical tractability, algorithms to compute a minimax or Gamma-minimax estimator have been proposed. Still, most of these works focus on parametric models. For example, Nelson (1966) and Kempthorne (1987) each proposed an algorithm to compute a minimax estimator. Bryan et al. (2007) and Schafer and Stark (2009) proposed

an algorithm to compute an approximate confidence region of optimal expected size in the minimax sense. Noubiap and Seidel (2001) proposed an iterative algorithm to compute a Gamma-minimax decision for the set of priors constrained by generalized moment conditions. More recent works explored computing estimators under more general models. For example, Luedtke et al. (2020a) introduced an approach, termed Adversarial Monte Carlo meta-learning (AMC), for constructing minimax estimators. In the special case of prediction problems with mean-squared error, Luedtke et al. (2020b) studied the invariance properties of the decision problem and their implications for AMC.

In this chapter, we make the following contributions.

1. We define Gamma-minimaxity in general models. Our general definition suggests an approach for leveraging potentially-vague prior information even when the statistical model is infinite dimensional.

2. We propose iterative algorithms to compute Gamma-minimax estimators for a general model space and a set of priors constrained by generalized moments. Such constraints provide a natural means to represent prior information (Berger, 1990). To the best of our knowledge, this is the first algorithm to compute Gamma-minimax estimators under general models, including infinite-dimensional models. We also show that, for certain problems, there is a unique Gamma-minimax estimator and, moreover, our computed estimator converges to this estimator as the number of iteration increases to infinity.

3. Similarly to the approach proposed in Noubiap and Seidel (2001), our proposed iterative algorithm involves solving a minimax optimization problem in each intermediate step. However, we explicitly describe algorithms to solve these minimax problems. Moreover, these algorithms are not nested optimizations and therefore may take less time to converge. When the space of estimators can be parameterized by a Euclidean space and gradients are available, we propose to use a gradient-based algorithm or a stochastic

variant thereof. When gradients are unavailable, we propose to instead use fictitious play (Brown, 1951; Robinson, 1951) and provide a convergence result that, unlike the results in Robinson (1951), is applicable even when the space of estimators is an infinite set.

4. Like the approach proposed in Noubiap and Seidel (2001), our proposed iterative algorithm relies on increasingly fine finite grids over the model space. However, since we allow the model space to be high or even infinite dimensional, randomly adding grid points to the grid may lead to unacceptably slow convergence. To overcome this challenge, we propose an algorithm that is similar to a Markov chain Monte Carlo (MCMC) method to efficiently construct such grids.

5. We utilize recent advances in neural networks, especially adversarial learning (e.g., Goodfellow et al., 2014; Luedtke et al., 2020a; Luedtke et al., 2020b), when specifying the space of estimators. We also discuss an alternative parameterization using extreme learning machines (Huang et al., 2006b) and show that, if the Gamma-minimax estimator is unique, our computed estimator converges to the Gamma-minimax estimator under this parameterization. Thanks to the universal approximation properties of neural networks (e.g., Hornik, 1991; Csáji, 2001) and extreme learning machines (Huang et al., 2006a), we also show that both of these parameterizations can achieve good performance for sufficiently large networks. Furthermore, inspired by pre-training (e.g., Erhan et al., 2010) and transfer learning (e.g., Torrey and Shavlik, 2009), we recommend leveraging knowledge of existing estimators as inputs to the network in settings where this is possible. Under such choices of the space of estimators, we can expect to obtain a reasonably good estimator even if the associated nonconvex-concave minimax problems prove to be difficult.

This chapter is organized as follows. In Section 4.2, we introduce the framework of Gamma-minimax estimation and regularity conditions that we assume throughout the chap-

ter. In Section 4.3, we describe our proposed algorithms. Our proposal involves two layers of iterations, and we describe algorithms for the first and second layer in Sections 4.3.1 and 4.3.2, respectively. In Section 4.4, we discuss considerations when choosing hyperparameters in the algorithms. In Section 4.5, we demonstrate our method in three simulation studies. We conclude with a discussion in Section 4.6. Proof sketches of key results are provided in the main text, and complete proofs can be found in the appendix. All simulation codes are available at `https://github.com/QIU-Hongxiang-David/Gamma-minimax-learninng`.

4.2 Problem setup

Let $\mathcal{M}$ be a space of data-generating mechanisms P that contains the truth, P_0, and let ρ be a metric on $\mathcal{M}$. Under a data-generating mechanism $P \in \mathcal{M}$, let $\mathbf{X}^* \in \mathcal{X}^*$ denote the random data being generated. Let $\mathcal{C}$ denote a known coarsening mechanism such that the observed data $\mathbf{X} = \mathcal{C}(\mathbf{X}^*)$ belongs to $\mathcal{X}$. In some cases, the coarsening mechanism will be the identity map, whereas in other settings, such as those in which there is censored or missing data, the coarsening mechanism will be more involved (e.g., Birmingham et al., 2003; Gill et al., 1997; Heitjan and Rubin, 1991; Heitjan, 1993, 1994). Let $\mathcal{D}$ denote the space of estimators (or decision functions) equipped with a metric ϱ. Let $R : \mathcal{D} \times \mathcal{M} \to \mathbb{R}$ denote a risk function that measures the performance of an estimator under a data-generating mechanism such that smaller risks are preferable. We suppose throughout that $\mathcal{M}$ and $\mathcal{D}$ are equipped with the topologies induced by ρ and ϱ, respectively.

We now present two examples in which we formulate statistical decision problems in the above form.

Example 5. (Point estimation) Suppose that $\mathcal{M}$ statistical model, which may be parametric, semiparametric, or locally nonparametric (Bickel et al., 1993b). The data $\mathbf{X}^*$ consists of independently and identically distributed random variables following the true distribution $P_0 \in \mathcal{M}$. We set $\mathcal{C}$ to be the identity function so that $\mathbf{X} = \mathbf{X}^*$. We wish to estimate an aspect $\Psi(P_0) \in \mathbb{R}$ of P_0. Then, we can consider $\mathcal{D}$ being a set of functions $\mathcal{X} \to \mathbb{R}$ and set

the risk to be induced by the quadratic loss, that is, $R(d, P) = \mathrm{E}_P[\{d(\mathbf{X}) - \Psi(P)\}^2]$.

Example 6. (Predicting the expected number of new categories) Suppose that $\mathcal{M}$ consists of multinomial distributions with an unknown number of categories. Let an independent and identically distributed (iid) random sample of size n be generated from the true multinomial distribution, so that $\mathbf{X}^*$ is a multiset containing the number of observations X_k in each category k. Let the observed data be the multiset containing only the nonzero entries of $\mathbf{X}^*$, so that $\mathbf{X} = \mathcal{C}(\mathbf{X}^*) = \{X_k : X_k \neq 0\}$. Hence, only categories with nonzero occurrence are observed. Then, we may wish to predict the number of new categories that would be observed if a new sample of size m were collected. This problem has been extensively studied in the literature with applications in microbiome data, species taxonomic surveys, assessment of vocabulary size, etc. (e.g., Shen et al., 2003; Bunge et al., 2014; Orlitsky et al., 2016). We now show how to formulate this prediction problem in our framework. For each $P \in \mathcal{M}$, let p_k $(k = 1, \ldots, K)$ be the probability of category k, and let $\Psi(P) : \mathbf{X}^* \mapsto \sum_{k=1}^{K} I(X_k = 0)(1 - (1 - p_k)^m)$ be the expected number of new observed categories given the current full data $\mathbf{X}^*$. We consider $\mathcal{D}$ to be a set of $\mathcal{X} \to \mathbb{R}$ functions and set the risk to be the mean-squared error, that is, $R(d, P) = \mathrm{E}_P[\{d(\mathbf{X}) - \Psi(P)(\mathbf{X}^*)\}^2]$.

We now define Gamma-minimaxity within our decision theoretic framework. We assume that $\mathcal{M}$ is equipped with the Borel σ-field and let Π denote the set of all probability distributions on the measurable space $(\mathcal{M}, \mathcal{B})$. We also assume that, for any $d \in \mathcal{D}$ and any $\pi \in \Pi$, $P \mapsto R(d, P)$ is π-integrable. The Bayes risk corresponding to an estimator d and a prior π is defined as $r : (d, \pi) \mapsto \int R(d, P)\, \pi(\mathrm{d}P)$. Let $\Gamma \subseteq \Pi$ be the set of priors such that all $\pi \in \Gamma$ are consistent with prior information. An estimator is called a Γ-minimax estimator if it is in the set $\operatorname{argmin}_{d \in \mathcal{D}} \sup_{\pi \in \Gamma} r(d, \pi)$.

In this chapter, we consider the case in which Γ is characterized by finitely many generalized moment conditions, that is, $\Gamma = \{\pi \in \Pi : \Phi_k \in L^1(\pi), \int \Phi_k(P)\, \pi(\mathrm{d}P) \leq c_k, k = 1, \ldots, K\}$ where each $\Phi_k : \mathcal{M} \to \mathbb{R}$ is a prespecified function that extracts an aspect of a data-generating mechanism and $c_k \in \mathbb{R}$ is a prespecified constant. Such constraints can

represent a variety of forms of prior information. For example, with $\Phi_k = \pm\Psi^\kappa$ for some $\kappa \geq 1$, Γ imposes bounds on prior moments of $\Psi(P)$; with $\Phi_k(P) = \pm\mathbb{1}(\Psi(P) \in I)$ for some known interval I, Γ imposes bounds on the prior probability of $\Psi(P)$ lying in A. Similar prior information on aspects of P_0 other than $\Psi(P_0)$ can also be represented. In addition, note that an equality can be equivalently expressed by two inequalities, Γ may also impose equality constraints on prior generalized moments.

We assume that the following conditions hold throughout the rest of the chapter.

Condition 1. $\mathcal{M}$ is separable.

Condition 2. $\mathcal{D}$ is compact.

Condition 3. (i) $R : \mathcal{D} \times \mathcal{M} \to \mathbb{R}$ is a bounded function and (ii) $d \mapsto R(d, P)$ is Lipschitz continuous with a universal Lipschitz constant $L \in (0, \infty)$ independent of $P \in \mathcal{M}$, that is, there exists an L so that $|R(d_1, P) - R(d_2, P)| \leq L\varrho(d_1, d_2)$ for any $d_1, d_2 \in \mathcal{D}$ and any $P \in \mathcal{M}$.

Condition 2 is satisfied by many interesting classes of estimators. For example, we may choose $\mathcal{D}$ to be a space of neural networks whose indexing parameters fall in some specified compact set. We now illustrate the plausibility of the other two conditions in Example 5. For Condition 1, if the metric ρ on $\mathcal{M}$ is chosen as the supremum norm of the difference in cumulative distribution functions, then a countable dense subset of $\mathcal{M}$ can be the set of all empirical distributions with support contained in a countable dense subset of $\mathcal{X}$. If we instead assume that $\mathcal{X}$ is contained in a Euclidean space and all distributions in $\mathcal{M}$ have a differentiable Lebesgue density, then we may choose the metric to be the supremum norm of the difference of density functions. A countable dense subset of $\mathcal{M}$ is then the set of all kernel densities with locations being rational points in $\mathcal{X}$ and scales being positive rational numbers. For Condition 3, suppose that all distributions in $\mathcal{M}$ are dominated by a measure μ and their density functions are uniformly bounded. If $\int d(\mathbf{X})^2\mu(\mathrm{d}\mathbf{X})$ is uniformly bounded and Ψ is bounded, then $\mathrm{E}_P[d(\mathbf{X})^2]$ is uniformly bounded and hence R is bounded. In addition,

it holds that $|R(d_1, P) - R(d_2, P)| = |\mathrm{E}_P[(d_1(\mathbf{X}) - d_2(\mathbf{X}))(d_1(\mathbf{X}) + d_2(\mathbf{X}) - 2\Psi(P))]| \lesssim \mathrm{E}_P[(d_1(\mathbf{X}) - d_2(\mathbf{X}))^2] \lesssim \|d_1 - d_2\|_{P,2} \lesssim \|d_1 - d_2\|_{\mu,2}$ where $\lesssim$ stands for less than or equal to up to a multiplicative constant and $\|\cdot\|_{P,2}$ and $\|\cdot\|_{\mu,2}$ denote the $L^2(P)$- and $L^2(\mu)$-distance, respectively. Therefore, Condition 3 holds for ϱ being the $L^2(\mu)$-distance. Example 6 is similar.

4.3 Proposed algorithm to compute a Γ-minimax estimator

Our propose iterative algorithm consists of two layers of iterations: the outer layer described in Section 4.3.1 is used to approximate the minimax problem with its discretized version on an increasingly fine grid; the inner layer described in Section 4.3.2 is used to solve the discretized minimax problem constructed in the outer layer. We now describe these two layers separately.

4.3.1 Grid-based approximation of Γ-minimax estimators

In this section, we present an algorithm that is similar to that in Noubiap and Seidel (2001) but can be applied to richer model spaces, including to nonparametric models. Let $\{\mathcal{M}_\ell\}_{\ell=1}^\infty$ be an increasing sequence of finite subsets of $\mathcal{M}$ such that $\bigcup_{\ell=1}^\infty \mathcal{M}_\ell$ is dense in $\mathcal{M}$. That is, $\{\mathcal{M}_\ell\}_{\ell=1}^\infty$ is an increasingly fine grid over $\mathcal{M}$. By Condition 1, such an $\{\mathcal{M}_\ell\}_{\ell=1}^\infty$ necessarily exists. Define $\Gamma_\ell := \{\pi \in \Gamma : \pi \text{ has support in } \mathcal{M}_\ell\}$. We also define $r_{\sup}(d, \Gamma') := \sup_{\pi \in \Gamma'} r(d, \pi)$ for any $d \in \mathcal{D}$ and $\Gamma' \subseteq \Pi$.

In this section, we propose an algorithm (Algorithm 1) that approximates a Γ-minimax estimator with a Γ_ℓ-minimax estimator. We will show that the approximation error decays to zero as ℓ grows to infinity. We note that, under Condition 3, $d \mapsto r_{\sup}(d, \Gamma_\ell)$ is continuous for all ℓ by Lemma C.2, and hence d_ℓ^* exists. Here and in the rest of the algorithms in the chapter, for any real-valued function f, when we assign $\operatorname{argmin}_x f(x)$ or $\operatorname{argmax}_x f(x)$ to a variable, we arbitrarily pick a minimizer or maximizer if there are multiple optimizers. We note that the minimax problem in Line 3 of Algorithm 1 is nontrivial to solve, and therefore we propose two algorithms to solve it in Section 4.3.2.

Algorithm 1 Iteratively approximate a Γ-minimax estimator over an increasingly fine grid.

1: **for** $\ell = 1, 2, \ldots$ **do**

2: Construct a grid $\mathcal{M}_\ell \subseteq \mathcal{M}$ such that $\mathcal{M}_{\ell-1} \subsetneq \mathcal{M}_\ell$

3: $d_\ell^* \leftarrow \operatorname{argmin}_{d \in \mathcal{D}} \sup_{\pi \in \Gamma_\ell} r(d, \pi)$

We will present algorithms to find Γ_ℓ-minimax estimators d_ℓ^* in Section 4.3.2.

Let $d^* \in \mathcal{D}$ be an accumulation point of the sequence $\{d_\ell^*\}_{\ell=1}^\infty$, which is guaranteed to exist by Condition 2. We next present a sufficient condition to ensure that d^* is Γ-minimax, so that d_ℓ^* is approximately Γ-minimax for sufficiently large ℓ.

Condition 4. We assume that there exists an increasing sequence $\{\Omega_\ell\}_{\ell=1}^\infty$ of subsets of $\mathcal{M}$ such that

1. $\bigcup_{\ell=1}^\infty \Omega_\ell = \mathcal{M}$;

2. for all $\ell = 1, 2, \ldots$ and all $d \in \mathcal{D}$, it holds that

$$\lim_{i \to \infty} r_{\sup}(d, \Gamma_{i|\ell}) = r_{\sup}(d, \tilde{\Gamma}_\ell),$$

where $\tilde{\Gamma}_\ell := \{\pi \in \Gamma : \pi \text{ has support in } \Omega_\ell\}$ and $\Gamma_{i|\ell} := \{\pi \in \Gamma : \pi \text{ has support in } \mathcal{M}_i \bigcap \Omega_\ell\}$.

We note that, in contrast to $\mathcal{M}_\ell$, Ω_ℓ may be an infinite set. We may expect Condition 4 to hold in many cases. Exceptions may be caused by $P \mapsto R(d, P)$ being discontinuous. Another cause could be that Γ imposes a constraint on $\mathcal{M}$ such that no prior in Γ has support contained in $\mathcal{M}_i$ and hence $\Gamma_{i|\ell} = \emptyset$, but this can be resolved by rewriting the problem such that the constraint is incorporated into the specification of $\mathcal{M}$. We now present the theorem on Γ-minimaxity of d^*.

Theorem 4.1. *Under Conditions 1–4, d^* is Γ-minimax and*

$$r_{\sup}(d_\ell^*, \Gamma_\ell) \nearrow \min_{d \in \mathcal{D}} r_{\sup}(d, \Gamma)$$

as $\ell \to \infty$.

To prove Theorem 4.1, we utilize a result in Pinelis (2016) to establish that $r_{\sup}(d, \Gamma)$ can be approximated arbitrarily well by a discrete prior in Γ for any $d \in \mathcal{D}$. This is a key ingredient in the proof of Lemma C.1, which states that, for any $d \in \mathcal{D}$, $r_{\sup}(d, \tilde{\Gamma}_\ell)$ converges to $r_{\sup}(d, \Gamma)$. Then, we show that the sequence $\{r_{\sup}(d^*_\ell, \Gamma_\ell)\}_{\ell=1}^{\infty}$ is nondecreasing and upper bounded by $\inf_{d \in \mathcal{D}} r_{\sup}(d, \Gamma)$, which is less than or equal to the Γ-maximal Bayes risk $r_{\sup}(d^*, \Gamma)$ of the earlier-defined accumulation point d^* of $\{d^*_\ell\}_{\ell=1}^{\infty}$. Therefore, $r_{\sup}(d^*_\ell, \Gamma_\ell)$ converges to a limit. We finally use a contradiction argument to prove that this limit is greater than or equal to $r_{\sup}(d^*, \Gamma)$, which implies Theorem 4.1.

We have the following corollary on the uniqueness of the Γ-minimax estimator and the convergence of $\{d^*_\ell\}_{\ell=1}^{\infty}$ for certain problems.

Corollary 4.1.1. *Suppose that $\mathcal{D}$ is a convex subset of a vector space, $d \mapsto R(d, P)$ is strictly convex for each $P \in \mathcal{M}$, and $r_{\sup}(d, \Gamma)$ is attainable for each $d \in \mathcal{D}$ in the sense that, for all $d \in \mathcal{D}$, there exists a $\pi \in \Gamma$ such that $r(d, \pi) = r_{\sup}(d, \Gamma)$. Under Conditions 1–4, d^* is the unique Γ-minimax estimator and $d^*_\ell \to d^*$ as $\ell \to \infty$.*

We prove Corollary 4.1.1 by establishing that $d \mapsto r_{\sup}(d, \Gamma)$ is strictly convex.

In practice, the user also needs to specify a stopping criterion for Algorithm 1. In Noubiap and Seidel (2001), the authors proposed to compute or approximate $r_{\sup}(d^*_\ell, \Gamma)$ and stop if $r_{\sup}(d^*_\ell, \Gamma)$ is sufficiently close to $r_{\sup}(d^*_\ell, \Gamma_\ell)$. However, the procedure to approximate $r_{\sup}(d^*_\ell, \Gamma)$ in that work relies on the compactness of $\mathcal{M}$, but we do not want to assume this condition because it may restrict the applicability of the method. Therefore, we propose to use the following alternative criterion: stop if $r_{\sup}(d^*_\ell, \Gamma_{\ell+1}) - r_{\sup}(d^*_\ell, \Gamma_\ell) \leq \epsilon$ for a prespecified tolerance level $\epsilon > 0$. Note that this criterion does not guarantee that $r_{\sup}(d^*_\ell, \Gamma_\ell)$ is close to $r_{\sup}(d^*, \Gamma)$. For example, if $\mathcal{M}_{\ell+1} \setminus \mathcal{M}_\ell$ is small, it is even possible that $r_{\sup}(d^*_\ell, \Gamma_{\ell+1}) - r_{\sup}(d^*_\ell, \Gamma_\ell) = 0$, but d^*_ℓ is far from being Γ-minimax. We discuss this issue in more detail in Section 4.4.1.

We finally remark that $r_{\sup}(d, \Gamma_\ell)$ may be difficult to evaluate exactly. Since the risk is often an expectation, we recommend approximating $r_{\sup}(d, \Gamma_\ell)$ for any given d via Monte

Carlo as follows: first, estimate risks $R(d, P)$ for all $P \in \mathcal{M}_\ell$ with a large number of Monte Carlo runs; second, estimate the corresponding least favorable prior $\pi_{d,\ell} \in \operatorname{argmax}_{\pi \in \Gamma_\ell} r(d, \pi)$ using the estimated risks; third, estimate the risks $R(d, P)$ ($P \in \mathcal{M}_\ell$) again with independent Monte Carlo runs, and, finally, calculate $r(d, \pi_{d,\ell})$ with the estimated risks and the estimated least favorable prior. Using two independent estimates of the risk can remove the positive bias that would otherwise arise due to using the same data to estimate the risks and the least favorable prior.

4.3.2 Computation of a Γ_ℓ-minimax estimator

In this section, we present two candidate algorithms to compute a Γ_ℓ-minimax estimator, which corresponds to Line 3 in Algorithm 1. One is based on gradient and generally more computationally feasible, but requires differentiability of R in the parameters indexing the estimators; the other is more computationally intensive by requiring computing a sequence of Bayes estimators for a sequence of priors, but is also more general in that it does not rely on any differentiability conditions.

(Stochastic) gradient descent with max-oracle

Gradient descent with max-oracle (GDmax) and its stochastic variant (SGDmax). which were presented in Lin et al. (2019), can be used to solve general minimax problems in Euclidean spaces. To apply these algorithms to find a Γ_ℓ-minimax estimator, we need to assume that $\mathcal{D}$ can be parameterized by a subset of a Euclidean space, that is, that for any $d \in \mathcal{D}$, there exists a real vector-valued coefficient β in a compact set $\mathcal{H} \subseteq \mathbb{R}^D$ such that d may be written as $d(\beta)$. For example, $\mathcal{D}$ may be a neural network class. More discussions on the parameterization of $\mathcal{D}$ are in Section 4.4.2. In this section, in a slight abuse of notation, we define $R(\beta, P) := R(d(\beta), P)$, $r(\beta, \pi) := r(d(\beta), \pi)$ and $r_{\sup}(\beta, \Gamma_\ell) := r_{\sup}(d(\beta), \Gamma_\ell)$ for a coefficient $\beta \in \mathbb{R}^D$, data-generating mechanism $P \in \mathcal{M}$ and prior $\pi \in \Gamma$. We assume that $\beta \mapsto R(\beta, P)$ is differentiable for all $P \in \mathcal{M}$, and hence so is $\beta \mapsto r(\beta, \pi)$ for all $\pi \in \Gamma$. We also assume that a coefficient $\beta_\ell^* \in \operatorname{argmin}_{\beta \in \mathcal{H}} r_{\sup}(\beta, \Gamma_\ell)$ also minimizes the same function

over $\mathbb{R}^D$, so that we may solve the minimax problem over the unbounded space $\mathbb{R}^D$ ignoring the specification of $\mathcal{H}$.

We now present GDmax and SGDmax in out context of finding a Γ_ℓ-minimax estimator. If we can evaluate $R(\beta, P)$ exactly for all $\beta \in \mathcal{H}$ and $P \in \mathcal{M}_\ell$, then the GDmax algorithm (Algorithm 2) may be used. Note that the Line 3 can be formulated into a linear program, which can always be solved in polynomial time with an interior point method (e.g., Jiang et al., 2020) and often be solved in polynomial time with a simplex method (Spielman and Teng, 2004).

Algorithm 2 Gradient descent with max-oracle (GDmax) to compute a Γ_ℓ-minimax estimator

1: Initialize $\beta_{(0)} \in \mathbb{R}^D$. Set learning rate $\eta > 0$ and max-oracle accuracy $\zeta > 0$.
2: **for** $t = 1, 2, \ldots$ **do**
3: Maximization: find $\pi_{(t)} \in \Gamma_\ell$ such that $r(\beta_{(t-1)}, \pi_{(t)}) \geq \max_{\pi \in \Gamma_\ell} r(\beta_{(t-1)}, \pi) - \zeta$
4: Gradient descent: $\beta_{(t)} \leftarrow \beta_{(t-1)} - \eta \nabla_\beta r(\beta, \pi_{(t)})|_{\beta = \beta_{(t-1)}}$

In many cases, it is difficult to evaluate $R(\beta, P)$ exactly. When $R(\beta, P)$ is expressed as an expectation, $R(\beta, P)$ may instead be approximated using Monte Carlo techniques. With ξ being an exogenous source of randomness according to law Ξ, let $\hat{R}(\beta, P, \xi)$ be an unbiased approximation of $R(\beta, P)$ with $\mathrm{E}[\|\nabla_\beta \hat{R}(\beta, P, \xi) - R(\beta, P)\|^2] \leq \sigma^2 < \infty$, where $\|\cdot\|$ denotes the ℓ_2-norm in Euclidean spaces. Let $\hat{r}(\beta, \pi, \xi) := \int \hat{R}(\beta, P, \xi)\, \pi(\mathrm{d}P)$ for $\pi \in \Gamma_\ell$. In this case, SGDmax (Algorithm 3) may be used to find a (locally) Γ_ℓ-minimax estimator. Note that Algorithm 3 represents a generalization of the nested minimax AMC strategy in Luedtke et al. (2020a) to Γ_ℓ-minimax problems.

We now present further conditions needed for the convergence result for Algorithms 2 and 3.

Condition 5. For each $\ell = 1, 2, \ldots$, $\beta \mapsto R(\beta, P)$ is Lipschitz continuous with a universal Lipschitz constant L_1 independent of $P \in \mathcal{M}_\ell$.

Algorithm 3 Stochastic gradient descent with max-oracle (SGDmax) to compute a Γ_ℓ-minimax estimator

1: Initialize $\beta_{(0)} \in \mathbb{R}^D$. Set learning rate $\eta > 0$, max-oracle accuracy $\zeta > 0$ and batch size J.
2: **for** $t = 1, 2, \ldots$ **do**
3: Stochastic maximization: use a stochastic procedure to find $\pi_{(t)} \in \Gamma_\ell$ such that $\mathrm{E}[r(\beta_{(t-1)}, \pi_{(t)})] \geq \max_{\pi \in \Gamma_\ell} r(\beta_{(t-1)}, \pi) - \zeta$, where the expectation is over the randomness in stochastic maximization (e.g., variants of stochastic gradient ascent).
4: Generate iid copies $\xi_1, \ldots, \xi_J$ of ξ.
5: Stochastic gradient descent: $\beta_{(t)} \leftarrow \beta_{(t-1)} - \frac{\eta}{J} \sum_{j=1}^{J} \nabla_\beta \hat{r}(\beta, \pi_{(t)}, \xi_j)|_{\beta=\beta_{(t-1)}}$.

Note that Condition 5 differs from Condition 3 in that the former relies on the parameterization of $\mathcal{D}$ in a Euclidean space equipped with the Euclidean norm, while the latter may rely on a different metric on $\mathcal{D}$ such as an L^2-distance. In addition, the Lipschitz constant in Condition 5 may depend on ℓ, while that in Condition 3 must not.

Condition 6. For each $\ell = 1, 2, \ldots$, $\nabla_\beta R(\beta, P)$ is bounded; $\beta \mapsto \nabla_\beta R(\beta, P)$ is Lipschitz continuous with a universal Lipschitz constant L_2 independent of $P \in \mathcal{M}_\ell$.

Under these conditions, using the results in Lin et al. (2019), we can show that, in general, GDmax and SGDmax can yield a local minimum of $\beta \mapsto r_{\sup}(\beta, \Gamma_\ell)$ when the algorithms' hyperparameters are suitably chosen. Before we formally present the theorem, we introduce some definitions related to locally optimality of a potentially nondifferentiable and nonconvex function. A real-valued function f is called q-weakly convex if $x \mapsto f(x) + (q/2)\|x\|^2$ is convex ($q > 0$). The Moreau envelope of a real-valued function f with parameter $q > 0$ is $f_q : x \mapsto \min_{x'} f(x') + \|x' - x\|^2/(2q)$. A point x is an ϵ-stationary point ($\epsilon \geq 0$) of a q-weakly convex function f if $\|\nabla f_{1/(2q)}(x)\| \leq \epsilon$. Similarly, a random point x is an ϵ-stationary point ($\epsilon \geq 0$) of a q-weakly convex function f in expectation if $\mathrm{E}[\|\nabla f_{1/(2q)}(x)\|] \leq \epsilon$. If x is an ϵ-stationary point in expectation, we may conclude that it is an ϵ-stationary point with high

probability by Markov's inequality. Lemma 3.8 in Lin et al. (2019) shows that an ϵ-stationary point of f is close to a point x' at which f has at least one small subgradient for small ϵ.

We next present the convergence result for Algorithms 2 and 3.

Theorem 4.2. *Suppose that Conditions 1–3 and 5–6 hold. Let $\epsilon > 0$ be fixed and define $\Delta := (r_{\sup})_{1/(2L_1)}(\beta_{(0)}) - \min_{\beta \in \mathbb{R}^D}(r_{\sup})_{1/(2L_1)}(\beta)$, where we recall that $(r_{\sup})_{1/(2L_1)}$ is the Moreau envelope of $r_{\sup}$ with parameter $1/(2L_1)$.*

- *In Algorithm 2, with $\eta = \epsilon^2/(L_1 L_2^2)$ and $\zeta = \epsilon^2/(24L_1)$, $\beta_{(t)}$ is an ϵ-stationary point of $\beta \mapsto r_{\sup}(\beta, \Gamma_\ell)$ for $t = O(L_1 L_2 \Delta/\epsilon^4)$.*

- *In Algorithm 3, with $\eta = \epsilon^2/[L_1(L_2^2 + \sigma^2)]$, $\zeta = \epsilon^2/(24L_1)$ and $J = 1$, $\beta_{(t)}$ is an ϵ-stationary point of $\beta \mapsto r_{\sup}(\beta, \Gamma_\ell)$ in expectation for $t = O(L_1(L_2^2 + \sigma^2)\Delta/\epsilon^4)$.*

It may be inconvenient to implement Line 3 in Algorithm 3 because linear program solvers often do not use stochastic optimization. Therefore, we propose a variant (Algorithm 4) by replacing this line with Lines 3–4 so that ordinary linear program solvers can be directly applied. The following theorem justifies this variant.

Algorithm 4 Convenient variant of SGDmax (Algorithm 3) to compute a Γ_ℓ-minimax estimator

1: Initialize $\beta_{(0)} \in \mathbb{R}^D$. Set learning rate $\eta > 0$ and batch sizes J, J'.
2: **for** $t = 1, 2, \ldots$ **do**
3: Generate iid copies $\xi_1, \ldots, \xi_{J'}$ of ξ.
4: Stochastic maximization: $\pi_{(t)} \leftarrow \operatorname{argmax}_{\pi \in \Gamma_\ell} \frac{1}{J'} \sum_{j=1}^{J'} \hat{r}(\beta_{(t-1)}, \pi, \xi_j)$.
5: Generate iid copies of $\xi_{J'+1}, \ldots, \xi_{J'+J}$ of ξ.
6: Stochastic gradient descent: $\beta_{(t)} \leftarrow \beta_{(t-1)} - \frac{\eta}{J} \sum_{j=J'+1}^{J'+J} \nabla_\beta \hat{r}(\beta, \pi_{(t)}, \xi_j)|_{\beta=\beta_{(t-1)}}$.

Theorem 4.3. *Suppose that $\{\xi \mapsto \hat{r}(\beta, \pi, \xi) : \beta \in \mathbb{R}^D, \pi \in \Gamma_\ell\}$ is a Ξ-Glivenko-Centelli class. For any $\zeta > 0$, there exists a sufficiently large J' such that $\mathrm{E}[r(\beta_{(t-1)}, \pi_{(t)})] \geq$*

$\max_{\pi\in\Gamma_\ell} r(\beta_{(t-1)},\pi) - \zeta$ *for all* t*, where the expectation is taken over* $\pi_{(t)}$ *and* $\beta_{(t-1)}$ *is fixed. Therefore, with the chosen parameters in Theorem 4.2, we may choose a sufficiently large* J' *so that* $\beta_{(t)}$ *is an* ϵ*-stationary point of* $\beta \mapsto r_{\sup}(\beta,\Gamma_\ell)$ *in expectation for* $t = O(L_1(L_2^2+\sigma^2)\Delta/\epsilon^4)$.

We prove Theorem 4.3 by showing that $\max_{\pi\in\Gamma_\ell} r(\beta_{(t-1)},\pi) - \mathrm{E}[r(\beta_{(t-1)},\pi_{(t)})]$ converges to 0 as $J'\to\infty$. The proof is essentially an application of empirical process theory to the study of an M-estimator.

Fictitious play

Brown (1951) introduced fictitious play as a means to find the value of a zero-sum game. Robinson (1951) then proved that fictitious play can be used to iteratively solve a two-player zero-sum game for a saddle point that is a pair of mixed strategies where both players have finitely many pure strategies. Our problem of finding a Γ-minimax estimator may also be viewed as a two-player zero-sum game where one player chooses a prior from Γ and the other player chooses an estimator from $\mathcal{D}$. If we assume that, for the Γ-minimax problem at hand, the pair of both players' optimal strategies is a saddle point, which holds in many minimax problems (e.g., v. Neumann, 1928; Fan, 1953; Sion, 1958), then fictitious play may also be used to find a Γ-minimax estimator. Since Γ may be too rich to allow for feasible implementation of fictitious play, we propose to use this algorithm to find a Γ_ℓ-minimax estimator. In this section, we present how to use this algorithm to find a Γ_ℓ-minimax estimator and its convergence results.

In the fictitious play algorithm in Robinson (1951), the two players take turns to play the best pure strategy against the mixture of the opponent's historic pure strategies, and the final output is a pair of mixtures of the two players' historic pure strategies. Since this algorithm aims to find minimax mixed strategies, we consider stochastic estimators. That is, consider the Borel σ-field $\mathcal{F}$ over $\mathcal{D}$ and let $\mathit{\Pi}$ denote the set of all probability distributions on the measurable space $(\mathcal{D},\mathcal{F})$. We define $\overline{\mathcal{D}}$ to be the space of stochastic estimators

with each element taking the following form: first draw an estimator from $\mathcal{D}$ according to a distribution $\varpi \in \mathit{\Pi}$ with an exogenous random mechanism and then use the estimator to obtain an estimate based on the data. Note that we may write any $\overline{d} \in \overline{\mathcal{D}}$ as $\overline{d}(\varpi)$ for some $\varpi \in \mathit{\Pi}$. We consider estimators in $\overline{\mathcal{D}}$ throughout this section, with the definition of Γ-minimaxity extended in the natural way, so that $\overline{d}^* = \overline{d}(\varpi^*) \in \overline{\mathcal{D}}$ is Γ-minimax if $r_{\sup}(\overline{d}^*, \Gamma) = \min_{\overline{d} \in \overline{\mathcal{D}}} r_{\sup}(\overline{d}, \Gamma)$; we similarly extend all other definitions from Section 4.2. We assume that there exists $\pi_\ell^* \in \Gamma_\ell$ $(\ell = 1, 2, \ldots)$ such that

$$r(\overline{d}^*, \pi_\ell^*) = \sup_{\pi \in \Gamma_\ell} \inf_{\overline{d} \in \overline{\mathcal{D}}} r(\overline{d}, \pi) = \inf_{\overline{d} \in \overline{\mathcal{D}}} \sup_{\pi \in \Gamma_\ell} r(\overline{d}, \pi). \tag{4.1}$$

In other words, $(\overline{d}^*, \pi_\ell^*)$ is a saddle point of r in $\overline{\mathcal{D}} \times \Gamma_\ell$. Under this condition and the further conditions that $\mathcal{D}$ is convex and $d \mapsto R(d, P)$ is convex for all $P \in \mathcal{M}$, it is possible to use a Γ-minimax estimator over the richer class $\overline{\mathcal{D}}$ of stochastic estimators to derive a Γ-minimax estimator over the original class $\mathcal{D}$. Indeed, for any $\overline{d}(\varpi) \in \overline{\mathcal{D}}$ and $P \in \mathcal{M}$, by Jensen's inequality, $R(\overline{d}(\varpi), P) = \int R(d, P)\, \varpi(\mathrm{d}d) \geq R(\underline{\overline{d}}(\varpi), P)$ where $\underline{\overline{d}}(\varpi) := \int d\, \varpi(\mathrm{d}d) \in \mathcal{D}$ is the average of the stochastic estimator $\overline{d}(\varpi)$; that is, the risk of $\underline{\overline{d}}(\varpi)$ is never greater than that of $\overline{d}(\varpi)$. Therefore, we may use the fictitious play algorithm to compute $\overline{d}(\varpi_\ell^*)$ for each ℓ and further apply Algorithm 1 to compute $\overline{d}(\varpi^*)$. After that, we may take $\underline{\overline{d}}(\varpi^*)$ as the final output deterministic estimator.

Algorithm 5) presents the fictitious play algorithm for finding a Γ_ℓ-minimax estimator in $\overline{\mathcal{D}}$. Note that Γ_ℓ is convex, and hence π always lies in Γ_ℓ throughout the iterations. In practice, we may initialize ϖ as a point mass at an initial estimator in $\mathcal{D}$. In addition, similarly to Robinson (1951), we may replace Line 5 with $d_{(t)}^\dagger \leftarrow \operatorname{argmin}_{d \in \mathcal{D}} r(d, \pi_{(t)})$, that is, minimizing the Bayes risk with the most recently updated prior rather than with the previous prior.

We next present a convergence result for this algorithm.

Theorem 4.4. *Using Algorithm 5, under Conditions 1–3, it holds that*

$$r(d_{(t)}^\dagger, \pi_{(t-1)}) \leq r(\overline{d}(\varpi_\ell^*), \pi_\ell^*) \leq r(\overline{d}(\varpi_{(t-1)}), \pi_{(t)}^\dagger)$$

Algorithm 5 Fictitious play to compute a Γ_ℓ-minimax stochastic estimator

1: Initialize $\varpi_{(0)} \in \mathit{\Pi}$ and $\pi_{(0)} \in \Gamma_\ell$.

2: **for** t=1,2,... **do**

3: $\quad \pi^\dagger_{(t)} \leftarrow \text{argmax}_{\pi \in \Gamma_\ell} \, r(\overline{d}(\varpi_{(t-1)}), \pi)$

4: $\quad \pi_{(t)} \leftarrow \frac{t-1}{t}\pi_{(t-1)} + \frac{1}{t}\pi^\dagger_{(t)}$

5: $\quad d^\dagger_{(t)} \leftarrow \text{argmin}_{d \in \mathcal{D}} \, r(d, \pi_{(t-1)})$

6: $\quad \varpi_{(t)} \leftarrow \frac{t-1}{t}\varpi_{(t-1)} + \frac{1}{t}\delta(d^\dagger_{(t)})$, where $\delta(d)$ denotes a point mass at $d \in \mathcal{D}$.

for all t and

$$\lim_{t\to\infty} \left[r(\overline{d}(\varpi_{(t-1)}), \pi^\dagger_{(t)}) - r(d^\dagger_{(t)}, \pi_{(t-1)}) \right] = 0.$$

Consequently, the Γ_ℓ-maximal risk of $\overline{d}(\varpi_{(t)})$ converges to the Γ_ℓ-minimax risk, that is, $r_{\sup}(\overline{d}(\varpi_{(t-1)}), \Gamma_\ell) \to r_{\sup}(\overline{d}(\varpi^*_\ell), \Gamma_\ell)$ as $t \to \infty$.

Robinson (1951) proved a similar case for two-player zero-sum games where each player has finitely many pure strategies. In contrast, in our problem, each player may have infinitely many pure strategies. A natural approach to use to attempt to prove Theorem 4.4 would be to consider finite covers of $\mathcal{D}$ and Γ_ℓ, i.e., $\mathcal{D} = \bigcup_{i=1}^I \mathcal{D}_i$ and $\Gamma_\ell = \bigcup_{j=1}^J \Pi_j$, such that the range of $r(d, \pi)$ in each $\mathcal{D}_i$ and Π_j is small (say less than ϵ), bin pure strategies into these subsets, and then apply the argument in Robinson (1951) to these bins. The collection of $\mathcal{D}_i$ and Π_j may be viewed as finitely many approximated pure strategies to Γ_ℓ and $\mathcal{D}$ up to accuracy ϵ, respectively. Unfortunately, we found that this approach fails. The problem arises because Robinson (1951) inducted on I and J, and, after each induction step, the corresponding upper bound becomes twice as large. Unlike the case with finitely many pure strategies that was considered in Brown (1951) and Robinson (1951), as the desired approximation accuracy ϵ approaches zero, the numbers of approximated pure strategies, I and J, may diverge to infinity, and so does the number of induction steps. Therefore, the resulting final upper bound is of order $2^{I+J}\epsilon$ and generally does not converge to zero as ϵ tends to zero. To overcome this challenge, we instead control the increase in the relevant upper bound

after each induction step more carefully so that the final upper bound converges to zero as ϵ decreases to zero, despite the fact that I and J may diverge to infinity.

We remark that, because Line 5 of Algorithm 5 typically involves another layer of iteration in addition to that over t, this algorithm will often be more computationally intensive than are Algorithms 2–4. Nevertheless, Algorithm 5 provides an approach to construct Γ_ℓ-minimax estimators in cases where these other algorithms cannot be applied, for example, in settings where the risk is not differentiable in the parameters indexing the estimator.

4.4 Considerations in implementation

4.4.1 Considerations when constructing the grid over the model space

By Theorem 4.1, $r_{\sup}(d^*_\ell, \Gamma_\ell) \nearrow \min_{d\in\mathcal{D}} r_{\sup}(d, \Gamma)$ whenever Conditions 1–4 hold and the increasing sequence $\{\mathcal{M}_\ell\}_{\ell=1}^\infty$ is such that $\bigcup_{\ell=1}^\infty \mathcal{M}_\ell$ is dense in $\mathcal{M}$. Though this guarantee holds for all such sequences $\{\mathcal{M}_\ell\}_{\ell=1}^\infty$, in practice, judiciously choosing this sequence of grids of distributions can lead to faster convergence. In particular, it is desirable that the least favorable prior Γ_ℓ puts mass on some of the distributions in $\mathcal{M}_\ell \backslash \mathcal{M}_{\ell-1}$ since, if this is not the case, then d^*_ℓ will be the same as $d^*_{\ell-1}$. While we may try to arrange for this to occur by adding many new points when enlarging $\mathcal{M}_{\ell-1}$ to $\mathcal{M}_\ell$, it may not be likely that any of these points will actually modify the least favorable prior unless they are carefully chosen.

To better address this issue, we propose to add grid points using a method that is similar to Markov chain Monte Carlo (MCMC). Our intuition is that, given an estimator d, the maximal Bayes risk is likely to significantly increase if we add distributions that (i) have high risk for d, and (ii) are consistent with prior information so that there exists some prior such that these distributions lie in a high-probability region. We propose to use an MCMC-like algorithm to bias the selection of distributions in favor of those with the above characteristics. Let $\tau : \mathcal{M} \to [0, \infty)$ denote a function such that $\tau(P) > \tau(P')$ if P is more consistent with prior information than P'. For example, given a prior mean μ of some real-valued summary $\Psi(P)$ of P and an interval I that contains $\Psi(P)$ with prior probability at least 95%, we may

choose $\tau : P \mapsto \phi(\Psi(P))$, where ϕ is the density of a normal distribution that has mean μ and places 95% of its probability mass in I. We call τ a *pseudo-prior*. Then, with the current estimator being d, we wish to select distributions P for which $R(d, P)\tau(P)$ is large. We may use the Metropolis-Hastings-Green algorithm (Metropolis et al., 1953; Hastings, 1970; Green, 1995) to draw samples from a density proportional to $P \mapsto R(d, P)\tau(P)$. We then let $\mathcal{M}_\ell$ be equal to the union of $\mathcal{M}_{\ell-1}$ and the set containing all unique distributions in this sample.

Details of the proposed scheme are provided in Algorithm 6. To use this proposed algorithm, we rely on it being possible to define a sequence of parametric models $\{\tilde{\Omega}_\ell\}_{\ell=1}^{\infty}$ such that $\tilde{\mathcal{M}} := \cup_{\ell=1}^{\infty}\tilde{\Omega}_\ell$ is dense in $\mathcal{M}_\ell$ — this is possible in many interesting examples (see, e.g., Chen, 2007). When combined with Condition 1, this condition enables the definition of an increasing sequence of grids of distributions $\{\mathcal{M}_\ell\}_{\ell=1}^{\infty}$ such that, for each ℓ, $\mathcal{M}_\ell \subseteq \tilde{\mathcal{M}}$.

Algorithm 6 MCMC-like algorithm to construct $\mathcal{M}_\ell$

Require: Previous grid $\mathcal{M}_{\ell-1}$, current estimator $d^*_{\ell-1}$ and number T of iterations. We define $\mathcal{M}_{-1} := \emptyset$. An initial estimator d^*_0 must be available if $\ell = 1$.

1: Initialize $P_{(0)} \in \tilde{\mathcal{M}}$.
2: **for** $t = 1, 2, \ldots, T$ **do**
3: Propose a distribution $P' \in \tilde{\mathcal{M}}$ from $P_{(t-1)}$
4: Calculate the MCMC acceptance probability p_{accept} of P' for target density $P \mapsto R(d^*_{\ell-1}, P)\tau(P)$
5: With probability p_{accept}, accept P' and $P_{(t)} \leftarrow P'$
6: **if** P' is not accepted **then**
7: $P_{(t)} \leftarrow P_{(t-1)}$
8: $\mathcal{M}_\ell \leftarrow$ unique elements of $\mathcal{M}_{\ell-1} \bigcup \{P_{(1)}, P_{(2)}, \ldots, P_{(T)}\}$

The following theorem on distributional convergence follows from that for Metropolis-Hastings-Green algorithm (see Section 3.2 and 3.3 of Green, 1995).

Theorem 4.5. *Suppose that $P \mapsto R(d^*_{\ell-1}, P)\tau(P)$ is bounded and integrable with respect to*

some measure μ on $\tilde{\mathfrak{M}}$ and let $\mathcal{L}$ denote the probability law on $\tilde{\mathfrak{M}}$ whose density function with respect to μ is proportional to this function. Then, in Algorithm 6, $P_{(t)}$ converges weakly to $\mathcal{L}$ as $t \to \infty$.

Implementing Algorithm 6 relies on the user making several decisions. These decisions include the choice of the pseudo-prior τ and the technique used to approximate the risk $R(d, P)$ to a reasonable accuracy. Fortunately, regardless of the decisions made, Theorem 4.1 suggests that $r_{\sup}(d^*_\ell, \Gamma_\ell) \nearrow \min_{d \in \mathcal{D}} r_{\sup}(d, \Gamma)$ for a wide range of sequences $\{\mathcal{M}_\ell\}_{\ell=1}^{\infty}$. Indeed, all that theorem requires on this sequence is that the grid $\mathcal{M}_\ell$ become arbitrarily fine as ℓ increases. Though the final decisions made are not important when ℓ is large, we still comment briefly on the decisions that we have made in our experiments, First, we have found it effective to approximate $R(d, P)$ via a large number of Monte Carlo draws. Second, in a variety of settings, we have also identified, via numerical experiments, candidate pseudo-priors that balance high risk and consistency with prior information (see Sections 4.5.2 and 4.5.3 for details).

4.4.2 Considerations when choosing the space of estimators

It is desirable to consider a rich space $\tilde{\mathcal{D}}$ of estimators to obtain an estimator with low maximal Bayes risk, and thus good general performance. However, to make numerically constructing these estimators computationally feasible, we usually have to consider a restricted space $\mathcal{D}$ of estimators. In the upcoming theorem, we provide an upper bound on the increment of the maximal Bayes risk induced by making this restriction. This result shows that, if estimators in $\mathcal{D}$ can approximate estimators in $\tilde{\mathcal{D}}$ well, then the resulting excess maximal Bayes risk is small. This result relies on what we call Condition 3', which is the same as Condition 3 except that each instance of $\mathcal{D}$ in that condition is replaced by $\tilde{\mathcal{D}}$.

Theorem 4.6. *Fix $\mathcal{D} \subseteq \tilde{\mathcal{D}}$. Let d^* be a Γ-minimax estimator in $\mathcal{D}$ and $\tilde{d}^*$ be a Γ-minimax estimator in $\tilde{\mathcal{D}}$, so that $r_{\sup}(d^*, \Gamma) = \min_{d \in \mathcal{D}} r_{\sup}(d, \Gamma)$ and $r_{\sup}(\tilde{d}^*, \Gamma) = \min_{d \in \tilde{\mathcal{D}}} r_{\sup}(d, \Gamma)$. Under Condition 3', $r_{\sup}(d^*, \Gamma) - r_{\sup}(\tilde{d}^*, \Gamma) \leq L \inf_{d' \in \mathcal{D}} \varrho(d', \tilde{d}^*)$.*

Proof of Theorem 4.6. By the definition of d^*, for any $d' \in \mathcal{D}$, $r_{\sup}(d^*, \Gamma) - r_{\sup}(\tilde{d}^*, \Gamma) \leq r_{\sup}(d', \Gamma) - r_{\sup}(\tilde{d}^*, \Gamma)$, and so $r_{\sup}(d^*, \Gamma) - r_{\sup}(\tilde{d}^*, \Gamma) \leq \inf_{d' \in \mathcal{D}}[r_{\sup}(d', \Gamma) - r_{\sup}(\tilde{d}^*, \Gamma)]$. By Lemma C.2 in Appendix C.1.1, $d \mapsto r_{\sup}(d, \Gamma)$ is Lipschitz continuous with Lipschitz constant L. Therefore, the right hand side is upper bounded by $L \inf_{d' \in \mathcal{D}} \varrho(d', \tilde{d}^*)$. □

Feedforward neural networks (or neural networks for short) are natural options for the space of estimators because of their universal approximation property (e.g., Hornik, 1991; Csáji, 2001; Hanin and Sellke, 2017; Kidger and Lyons, 2020). However, training commonly used neural networks can be computationally intensive. Moreover, a space of neural networks is typically nonconvex, and hence it may be difficult to find a global minimizer of the maximal Bayes risk even if the risk is convex in the estimator. Therefore, the learned estimator might not perform well.

To help overcome this challenge, we advocate for utilizing available statistical knowledge when designing the space of estimators. We call estimators that take this form *statistical knowledge networks*. In particular, if a sensible simple estimator is already available, we propose to use neural networks with such an estimator as a node connected to the output node. An example of such an architecture is presented in Fig 4.1. In this sample architecture, each node is an activation function such as the sigmoid or the rectified linear unit (ReLU) (Glorot et al., 2011) function applied to an affine transformation of the vector containing the ancestors of the node. The only exception is the output node, which is again an affine transformation of its ancestors, but uses the identity activation function. When training the neural network, we may initialize the affine transformation in the output layer to only give weight to the simple estimator. Under this approach, the space of estimators is a set of perturbations of a sensible simple estimator. Although we may still face the challenge of nonconvexity and local optimality, we can at least expect to improve the initial simple estimator.

We note that we might overcome the challenge of nonconvexity and local optimality by using an extreme learning machine (ELM) (Huang et al., 2006b) to parameterize the

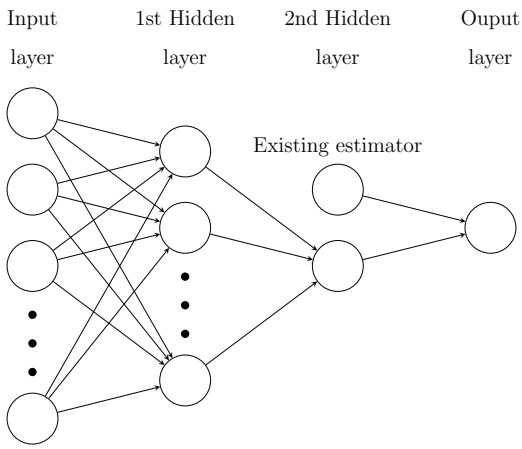

Figure 4.1: Example of neural network estimator architecture utilizing an existing estimator. The arrows from the input nodes to the existing estimator are omitted from this graph.

estimator. ELMs are neural networks for which the weights in hidden nodes are randomly generated and are held fixed, and only the weights in the output layer are trained. Thus, the space of ELMs with a fixed architecture and fixed hidden layer weights is convex. Like traditional neural networks, ELMs have the universal approximation property (Huang et al., 2006a). In addition, Corollary 4.1.1 may be applied to an ELM so that the Γ_ℓ-minimax estimator may converge to the Γ-minimax estimator. As for traditional neural networks, we may incorporate knowledge of existing statistical estimators into an ELM.

Next, we present a corollary of Theorem 4.6 for some special cases of neural networks and ELMs based on their universal approximation results. We expect similar results to hold for more general architectures of neural networks and ELMs, for example, with other activation functions, more hidden layers or more complicated architectures. Indeed, whenever universal approximation results are available over the space $\tilde{\mathcal{D}}$, Theorem 4.6 can be immediately applied to obtain an upper bound for the excess maximal Bayes risk $r_{\sup}(d^*, \Gamma) - r_{\sup}(\tilde{d}^*, \Gamma)$ due to restriction of the space of estimators.

Corollary 4.6.1. *Suppose that $\mathcal{X}$ is a compact subset of a Euclidean space $\mathbb{R}^\alpha$. Let $\tilde{\mathcal{D}}$ be the collection of all continuous functions defined on $\mathcal{X}$ that are square-integrable with respect*

to Lebesgue measure. Let the metric ϱ on $\tilde{\mathcal{D}}$ be the L^2 distance with respect to Lebesgue measure. Suppose that Condition 3 holds.

1. *Suppose that $\mathcal{D}$ is a space of estimators parameterized as neural networks with identity activation for the output layer and ReLU activation for all hidden layers. Then, for any $\epsilon > 0$, it holds that $\inf_{d\in\mathcal{D}} r_{\sup}(d,\Gamma) - \inf_{\tilde{d}\in\tilde{\mathcal{D}}} r_{\sup}(\tilde{d},\Gamma) \leq \epsilon$ provided that networks in $\mathcal{D}$ have a sufficiently large number of hidden layers and a sufficiently large number of hidden nodes in each hidden layer.*

2. *Suppose that $\mathcal{D}$ is a space of estimators parameterized as ELMs with one hidden layer, identity activation for the output layer and a bounded nonconstant piecewise continuous $\mathbb{R} \to \mathbb{R}$ activation function for the hidden layer. Suppose that the values of the hidden weights and hidden biases in the ELM are independently drawn from a continuous distribution with support $\mathbb{R}^{\alpha+1}$. Then, for any $\epsilon > 0$, $\mathbb{P}\{\inf_{d\in\mathcal{D}} r_{\sup}(d,\Gamma) - \inf_{\tilde{d}\in\tilde{\mathcal{D}}} r_{\sup}(\tilde{d},\Gamma) \leq \epsilon\} \to 1$ as the number of hidden nodes tends to infinity.*

Proof. The result follows from the universal approximation results (Theorem 4.16 in Kidger and Lyons (2020) and Theorem II.1 in Huang et al. (2006a), respectively) and Theorem 4.6. □

Under Condition 3, the above result can be generalized to a variety of collections of estimators $\mathcal{D}_1$ that are richer than the space $\tilde{\mathcal{D}}$ of continuous functions considered in the above lemma. Indeed, if $\mathcal{D}_1$ is such that $\tilde{\mathcal{D}}$ is dense in $\mathcal{D}_1$, then Lemma C.2 in Appendix C.1.1 shows that the same conclusion will hold. This shows that the same conclusions of the above theorem hold when the collection of estimators $\tilde{\mathcal{D}}$ is enriched to contain all $\mathcal{X} \to \mathbb{R}$ functions that are square integrable with respect to Lebesgue measure (e.g., Theorem 1.15 in Evans and Gariepy, 2015).

We finally remark that, besides computational intensity when constructing (i.e., learning) a Γ-minimax estimator, another important factor to be considered when choosing $\mathcal{D}$ is the computational intensity to evaluate the learned estimator at the observed data set. This

is another reason for our choosing neural networks or ELMs as the space of estimators. Indeed, existing software packages (e.g., Paszke et al., 2019) make it easy to leverage graphics processing units to efficiently evaluate the output of neural networks for any given input. Therefore, if the existing estimator being used is not too difficult to compute, then estimators parameterized using similar architectures to that displayed in Figure 4.1 will be able to be computed efficiently in practice. This efficiency may be especially important in settings where the estimator will be applied to many datasets, so that the cost of learning the estimator is amortized and the main computational expense is evaluating the learned estimator.

4.5 Simulation

4.5.1 Estimation of the mean

We start by illustrating our proposed method in a special case of Example 5, namely for estimating the mean of a distribution. We assume that $\mathcal{M}$ consists of all probability distributions defined on the Borel σ-algebra on $[0, 1]$ and we observe $\mathbf{X} = (X_1, X_2, \ldots, X_n)$, where $X_1, \ldots, X_n \overset{\text{iid}}{\sim} P_0 \in \mathcal{M}$. Here we take $n = 10$. The estimand is $\Psi(P_0) = \int x \, P_0(\mathrm{d}x)$. We use the mean squared error risk introduced in Example 5. Suppose that we represent the prior information by $\Gamma = \{\pi \in \Pi : \int \Psi(P) \, \pi(\mathrm{d}P) = 0.3\}$, which corresponds to the set of prior distributions in Π that satisfy an equality constraint on the prior mean of $\Psi(P)$.

We apply our method to three spaces of estimators separately. The first space, $\mathcal{D}_{\text{linear}}$, is the set of affine transformations of the sample mean, that is, $\mathcal{D}_{\text{linear}} = \{d : d(\mathbf{X}) = \beta_0 + \beta_1 \sum_{i=1}^{n} X_i/n, \beta_0, \beta_1 \in \mathbb{R}\}$. As shown in Proposition C.1 in Appendix C.1.5, there is an estimator d^* in $\mathcal{D}_{\text{linear}}$ that is Γ-minimax in the space of all estimators that are square-integrable with respect to all $P \in \mathcal{M}$, so we consider this simple space to better compare our computed estimator with that theoretical Γ-minimax estimator. When computing a Γ-minimax estimator in $\mathcal{D}_{\text{linear}}$, we initialize the estimator to be the sample mean, that is, we let $\beta_0 = 0$ and $\beta_1 = 1$.

The second space, $\mathcal{D}_{\text{skn}}$ (statistical knowledge network), is a set of neural networks de-

signed based on statistical knowledge that includes the sample mean as an input. We consider this space to illustrate our proposal in Section 4.4.2. More precisely, we use an architecture in Fig 4.2 that is similar to the deep set architecture (Zaheer et al., 2017; Maron et al., 2019), which is a permutation invariant neural network. We use such an architecture to account for the fact that the sample is iid. In this architecture, the sample mean node is used as an augmenting node to an ordinary deep set network and is combined with the output of that ordinary network in the fourth hidden layer to obtain the final output. Note that $\mathcal{D}_{\text{skn}} \supset \mathcal{D}_{\text{linear}}$. When computing a Γ-minimax estimator for this class, we also initialize the network to be exactly the sample mean, which is a reasonable choice given that the sample mean is known to be sensible estimator. In this simulation experiment, we choose the dimensionality of nodes in each hidden layer in Fig 4.2 as follows: each node in the first, second, third and fourth hidden layer represents a vector in $\mathbb{R}^{10}$, $\mathbb{R}^5$, $\mathbb{R}^{10}$ and $\mathbb{R}$, respectively. We do not use larger architectures because usually the sample mean is already a good estimator, and we expect to obtain a useful estimator as a small perturbation of this estimator. We also use the ReLU as the activation function. We did not use ELMs in this and the following simulations because we found that neural networks perform well.

The third space, $\mathcal{D}_{\text{nn}}$, is a set of neural networks that does not utilize knowledge of the sample mean. We consider this space to illustrate our method without utilizing existing sensible estimators. These estimators are also deep set networks with a similar architecture as $\mathcal{D}_{\text{skn}}$ in Fig 4.2. The main difference is that the explicit sample mean node and the fourth hidden layer are removed. When computing a Γ-minimax estimator in $\mathcal{D}_{\text{nn}}$, we also randomly initialize the network, unlike $\mathcal{D}_{\text{linear}}$ and $\mathcal{D}_{\text{skn}}$, in order not to input statistical knowledge. Because the ReLU activation function is used, $\mathcal{D}_{\text{nn}} \supset \mathcal{D}_{\text{linear}}$, and we do not expect that optimizing over $\mathcal{D}_{\text{nn}}$ should not lead to a Γ-minimax estimator with worse performance than those in $\mathcal{D}_{\text{linear}}$ and $\mathcal{D}_{\text{skn}}$.

To construct the grid $\mathcal{M}_\ell$ for this problem, we use a simpler method than Algorithm 6. As indicated by Lemma C.5 in Appendix C.1.5, for estimators in $\mathcal{D}_{\text{linear}}$, Bernoulli distributions tend to have high risks since all probability weights lie on the boundary of $[0, 1]$; in addition,

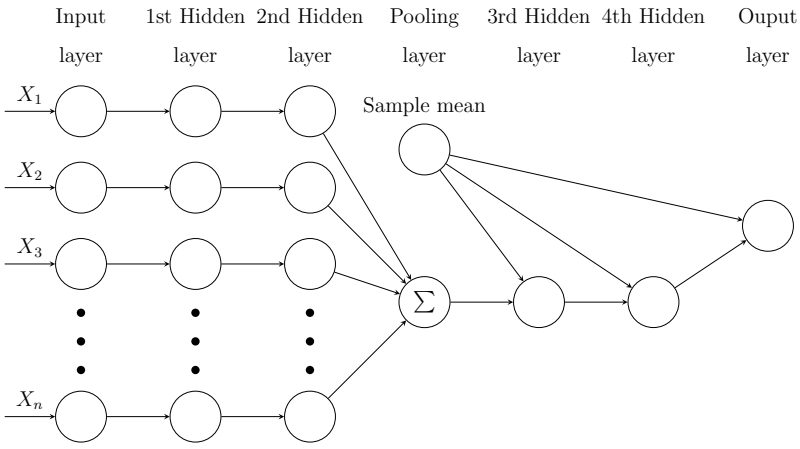

Figure 4.2: Architecture of the permutation invariant neural network estimator of the mean in $\mathcal{D}_{\text{skn}}$. X_i: observation i in the sample; $\sum$: the node that sums up all ancestor nodes. In the first two hidden layers, all inputs nodes are transformed by the same function. The arrows from the input nodes to the sample mean estimator are omitted from this graph. Each node in the hidden layers represent a vector.

a prior π^* for which d^* is Bayes is a Beta prior over Bernoulli distributions. Therefore, we randomly generate 2,000 Bernoulli distributions as grid points in $\mathcal{M}_1$. We also include two degenerate distributions in this grid, namely the distribution that places all of its mass at 0 and that which places all of its mass at 1. When constructing $\mathcal{M}_\ell$ from $\mathcal{M}_{\ell-1}$, we still add in more complicated distributions to make the grid dense in the limit: we first randomly generate 500 discrete distributions with support being those in $\mathcal{M}_{\ell-1}$; then we randomly generate 10 new support points in $[0, 1]$ and 1,000 distributions with support points being the union of the new support points and the existing support points in $\mathcal{M}_{\ell-1}$.

When computing the Γ-minimax estimator, for each grid $\mathcal{M}_\ell$, we compute the Γ_ℓ-minimax estimator for all three estimator spaces with Algorithm 4. We set the learning rate $\eta = 0.005$, the batch size $J = 50$ and the number of iterations to be 200 for Γ_ℓ ($\ell > 1$). The number of iterations for Γ_1 is larger because, in our experiments, we saw that a Γ_1-minimax estimator is already close to a Γ-minimax estimator, and using a large number of iterations in this step can improve the initial estimator substantially. For $\mathcal{D}_{\text{linear}}$ and $\mathcal{D}_{\text{skn}}$, the number of iterations

for Γ_1 is 2,000; the corresponding number for $\mathcal{D}_{\text{nn}}$ is 6,000 to account for the lack of human knowledge input. We also use Algorithm 5 with 10,000 iterations to compute a Γ_ℓ-minimax estimator for $\mathcal{D}_{\text{linear}}$ for illustration. In this setup, as described in Section 4.3.2, we take the average of the computed Γ-minimax stochastic estimator as the final output estimator in $\mathcal{D}_{\text{linear}}$. We do not apply Algorithm 5 to $\mathcal{D}_{\text{skn}}$ or $\mathcal{D}_{\text{nn}}$ because it is computationally intractable.

We set the stopping criterion in Algorithm 1 as follows. When Algorithm 4 is used to compute Γ_ℓ-minimax estimators, we estimate $r_{\sup}(d^*_{\ell-1}, \Gamma_\ell)$ and $r_{\sup}(d^*_{\ell-1}, \Gamma_{\ell-1})$ with 2,000 Monte Carlo runs as described in Section 4.3.1; when Algorithm 5 is used, $r_{\sup}(d^*_{\ell-1}, \Gamma_\ell)$ and $r_{\sup}(d^*_{\ell-1}, \Gamma_{\ell-1})$ are computed exactly because $R(d, P)$ has a closed-form expression for all $d \in \mathcal{D}_{\text{linear}}$ and $P \in \mathcal{M}_\ell$. We set the tolerance ϵ to be equal to 0.0001 so that we stop Algorithm 1 if $r_{\sup}(d^*_{\ell-1}, \Gamma_\ell) - r_{\sup}(d^*_{\ell-1}, \Gamma_{\ell-1}) \leq \epsilon$.

After computation, we report the Bayes risk of the computed and theoretical Γ-minimax estimators under π^*, the prior such that $r(d^*, \pi^*) = \inf_{d \in \mathcal{D}} r_{\sup}(d, \Gamma)$. For the estimators in $\mathcal{D}_{\text{linear}}$, we further report their coefficients. We also report two coefficients of the computed estimator in $\mathcal{D}_{\text{skn}}$ as follows. Since $\mathcal{D}_{\text{linear}} \subseteq \mathcal{D}_{\text{skn}}$ and we initialize the estimator to be the sample mean for $\mathcal{D}_{\text{skn}}$, we would expect that the bias β_0 and the weight of the sample mean β_1 in the output layer for the computed Γ-minimax estimator in $\mathcal{D}_{\text{skn}}$ may correspond to those in $\mathcal{D}_{\text{linear}}$. Therefore, we also report these two coefficients β_0 and β_1 for $\mathcal{D}_{\text{skn}}$. This may not be the case for $\mathcal{D}_{\text{nn}}$ because sample mean is not explicit in its parameterization and all coefficients are randomly initialized, so we do not report any coefficients for $\mathcal{D}_{\text{nn}}$.

Table 4.1 presents the computation results. By Theorem C.1 in Appendix C.1.5, these computed estimators are all approximately Γ-minimax since their Bayes risks for π^* are all close to that of a theoretical Γ-minimax estimator. The coefficients β_0 and β_1 of the computed estimators in $\mathcal{D}_{\text{linear}}$ and $\mathcal{D}_{\text{skn}}$ are also close to a theoretically derived estimator. For the computed estimator in $\mathcal{D}_{\text{skn}}$, the weight of the other ancestor node in the output layer (i.e., the node in the 4th hidden layer in Fig 4.2) is 0.000. Therefore, our computed Γ-minimax estimator in $\mathcal{D}_{\text{skn}}$ is also close to a theoretically derived Γ-minimax estimator.

In our experiments, Algorithm 1 converged after computing a Γ_1-minimax estimator

Table 4.1: Coefficients and Bayes risks of estimators of the mean. Unrestricted space: the space of all estimators that are square-integrable with respect to all $P \in \mathcal{M}$.

Estimator space	Method to obtain d^*	β_0	β_1	$r(d, \pi^*)$
Unrestricted space	Theoretical derivation	0.072	0.760	0.012
$\mathcal{D}_{\text{linear}}$	Algorithms 1 & 4	0.072	0.763	0.012
$\mathcal{D}_{\text{skn}}$	Algorithms 1 & 4	0.071	0.767	0.012
$\mathcal{D}_{\text{nn}}$	Algorithms 1 & 4	—	—	0.012
$\mathcal{D}_{\text{linear}}$	Algorithms 1 & 5	0.072	0.760	0.012

except when using Algorithm 4 for $\mathcal{D}_{\text{linear}}$. Even in this exceptional case, the computed Γ_1-minimax estimator is still approximately Γ-minimax. We think the algorithm does not stop then in these cases because of Monte Carlo errors when computing $r_{\sup}(d^*_{\ell-1}, \Gamma_\ell)$ and $r_{\sup}(d^*_{\ell-1}, \Gamma_{\ell-1})$.

Fig 4.3 presents the Bayes risks (or its unbiased estimates) over iterations when computing a Γ_1-minimax estimator. In all cases using Algorithm 4, the Bayes risks appear to decrease and converge. When using Algorithm 5, the upper and lower bounds both converge to the same limit. The limiting values of the Bayes risks in all cases are close to $r(d^*, \pi^*)$ because Γ_1 can approximate π^* well.

4.5.2 Prediction of the expected number of new categories

We apply our proposed method to Example 6. We set the true population to be an infinite population with the same categories and same proportions as the sample studied in Miller and Wiegert (1989), which consists of 1088 observations in 188 categories. This is the same as the simulation setting in Shen et al. (2003). We set the sample size to be $n = 100$ and the size of the new sample to be $m = 200$. In this setting, the expected number of new categories in the new sample unconditionally on the observed sample, namely $\Phi(P_0) := \mathrm{E}_{P_0}[\Psi(P_0)(\mathbf{X}^*)]$, can be analytically computed and equals 48.02. We note that this quantity can also be

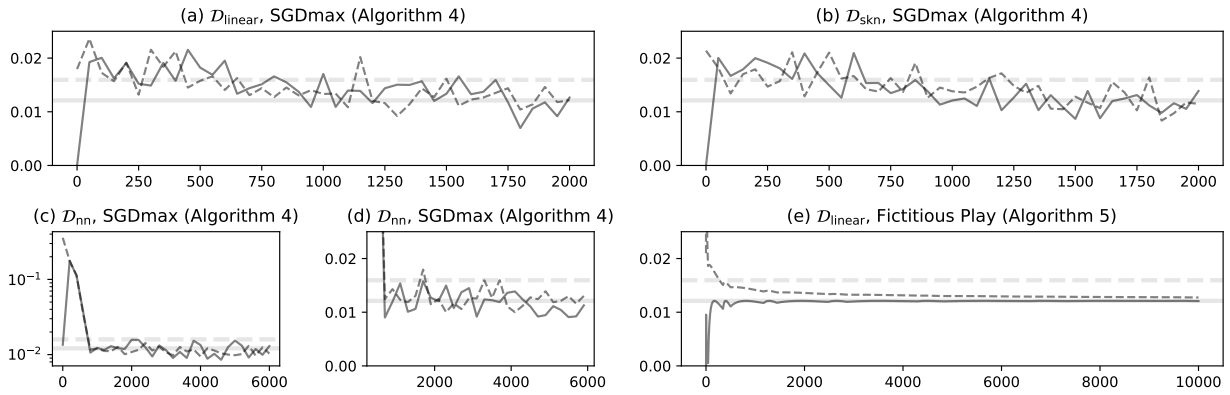

Figure 4.3: Estimated Bayes risks of the estimator over iterations when computing a Γ_1-minimax estimator. The lines are the current Bayes risks (y-axis) over iterations (x-axis) (unbiased estimates with 50 Monte Carlo runs for Algorithm 4; exact values for Algorithm 5). The solid lines are the Bayes risks after an update in the estimator to decrease the Bayes risk. The dashed lines are the Bayes risks after an update in the prior to increase the Bayes risk. The two horizontal lines are the Bayes risk of the sample mean (dashed) and d^* (solid), respectively, for π^*. For ease of visualization, in subfigures (a) and (b), the Bayes risks are plotted every 50 iterations; in subfigures (c) and (d), the Bayes risks are plotted every 200 iterations; subfigure (d) contains the part in subfigure (c) after 500 iterations.

computed via simulation: (i) sample n and m individuals with replacement from the data set in Miller and Wiegert (1989), (ii) count the number of new categories in the second sample, and (iii) repeat steps (i) and (ii) many times and compute the average.

We consider three sets of prior information:

1. strongly informative: prior mean of $\Phi(P)$ in $[45, 50]$, $\geq 95\%$ prior probability that $\Phi(P)$ lies in $[40, 55]$;

2. weakly informative: prior mean of $\Phi(P)$ in $[40, 55]$, $\geq 95\%$ prior probability that $\Phi(P)$ lies in $[30, 65]$; and

3. almost noninformative: prior mean of $\Phi(P)$ in $[35, 60]$, $\geq 95\%$ prior probability that $\Phi(P)$ lies in $[20, 75]$.

We note that a traditional Bayesian approach would require specifying a prior on $\mathcal{M}$, including the total number of categories and the proportion of each category, which may be difficult in practice.

We design the architecture of the neural network estimator as in Fig 4.4. We choose two existing estimators (referred to as OSW and SCL estimators, respectively) proposed by Orlitsky et al. (2016) and Shen et al. (2003) as human knowledge inputs to the architecture. As in Section 4.5.1, we use the ReLU activation function. There are 50 hidden nodes in the first hidden layer. We initialize the neural network that we train to output the average of these two existing estimators.

We use Algorithm 6 to construct $\mathcal{M}_\ell$. There are 2,000 grid points in $\mathcal{M}_1$, and we add 1,000 grid points each time we enlarge the grid. When generating $\mathcal{M}_1$, we chose the starting point to be a distribution $P_{(0)}$ with 146 categories and $\Phi(P_{(0)}) = 49.9$. We selected the log pseudo-prior as a weighted sum of two log density functions: (i) a normal distribution with mean being the midpoint of the interval constraint on prior mean of $\Phi(P)$ and central 95% probability interval being the interval with at least 95% prior probability, (ii) a negative-binomial distribution of the total number of categories with success probability 0.995 and 2

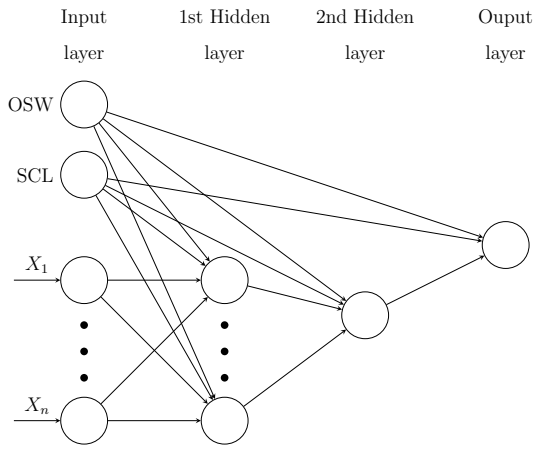

Figure 4.4: Architecture of the neural network estimator of the expected number of new categories. X_k: number of categories with k observations; OSW: estimator proposed in Orlitsky et al. (2016); SCL: estimator proposed in Shen et al. (2003). The arrows from data $(X_1, \ldots, X_n)$ to OSW and SCL estimators are omitted from this graph.

failures until the Bernoulli trial is stopped so that the mode and the variance are approximately 200 and 8×10^4, respectively. These log-densities are provided weight 30 and 10, respectively. We selected the weights based on the empirical observation that distributions with only a few categories tend to have high risks, but these distributions are relatively inconsistent with prior information and may well be given almost negligible probability weight in a computed least favorable prior, thus contributing little to computing a Γ-minimax estimator. We chose the aforementioned weights so that Algorithm 6 can explore a fairly large range of distributions and does not generate too many distributions with too few categories.

We use Algorithm 4 with learning rate $\eta = 0.005$ and batch size $J = 30$ to compute Γ_ℓ-minimax estimators. The number of iterations is 4,000 for Γ_1 and 200 for Γ_ℓ ($\ell > 1$). The stopping criterion in Algorithm 1 is that the estimated maximal Bayes risk with 2,000 Monte Carlo runs does not relatively increase by more than 2% or absolutely increase by more than 0.0001.

We finally examine the performance of OSW estimator, SCL estimator and our trained Γ-minimax estimator by comparing their risks under our set data-generating mechanism

Table 4.2: Risks and Bayes risks of estimators. $R(d, P_0)$: risk of the estimator under the true data-generating mechanism P_0. $r(d, \hat{\pi}^*)$: Bayes risk under prior $\hat{\pi}^*$, the computed prior from Algorithm 4 in the last and finest grid in the computation.

Strength of prior	Estimator	$R(d, P_0)$	$r(d, \hat{\pi}^*)$
strong	OSW	265	300
	SCL	146	179
	Γ-minimax	22	36
weak	OSW	265	252
	SCL	146	142
	Γ-minimax	56	85
almost none	OSW	265	220
	SCL	146	119
	Γ-minimax	76	108

computed with 20,000 Monte Carlo runs. We also compare their Bayes risks under the computed prior from Algorithm 4 using the last and finest grid in the computation with 20,000 Monte Carlo runs. We present the results in Table 4.2. In this simulation experiment, our Γ-minimax estimator significantly reduces the risk compared to two existing estimators. The Γ-minimax estimator also has the lowest Bayes risk in all cases. Therefore, incorporating the fairly informative prior knowledge into the estimator may lead to a significant improvement in performance.

Fig 4.5 presents the unbiased estimated of Bayes risks over iterations when computing a Γ_1-minimax estimator. The Bayes risks appear to have a decreasing trend and to approach a liming value. Over iterations, the Bayes risks decrease by a considerable amount. The limiting value of the Bayes risks appears to be slightly higher than the risk of the computed Γ-minimax estimator under P_0. This might indicate that P_0 is not an extreme distribution that yields a high risk.

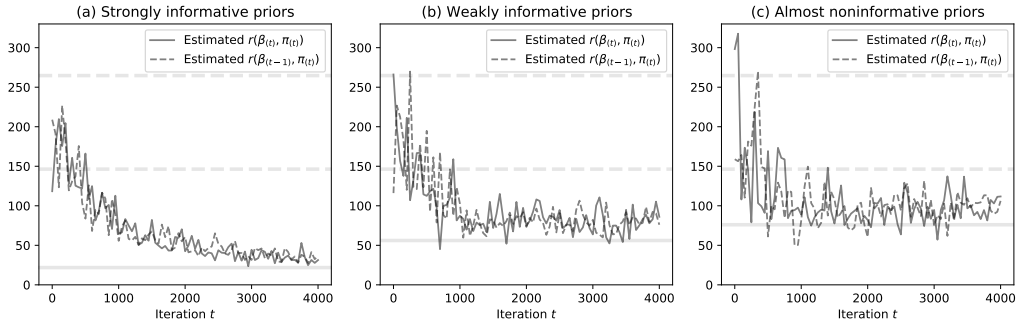

Figure 4.5: Estimated Bayes risks of the estimator over iterations when computing a Γ_1-minimax estimator. The lines are unbiased estimates of the current Bayes risks (y-axis) with 30 Monte Carlo runs over iterations (x-axis). The two dashed horizontal lines are the risks of OSW (upper) and SCL (lower) estimators, respectively, under P_0 in the simulation. The solid horizontal line is the risk of the computed Γ-minimax estimator under P_0. For clearness of visualization, the estimated Bayes risks are plotted every 50 iterations.

4.5.3 Estimation of the entropy

We also apply our method to estimate the entropy of a multinomial distribution. The setup of data-generating mechanism is the same as in Example 6, and the estimand of interest is the entropy, that is, $\Psi(P_0) = \sum_{k=1}^{K} -p_k \log p_k$. We choose the same true population and the same sample size $n = 100$ as in Section 4.5.2. We take the same risk function as in Example 5. The true entropy $\Psi(P_0)$ is 4.57. As a reference, the entropy of the uniform distribution with the same number of categories — which corresponds to the maximum entropy of multinomial distributions with the same total number of categories — is 5.24.

As in Section 4.5.2, we consider three sets of prior information:

1. Strongly informative: Prior mean of $\Psi(P)$ in $[4.3, 4.7]$, $\geq 95\%$ probability that $\Psi(P)$ lies in $[4, 5]$;

2. Weakly informative: Prior mean of $\Psi(P)$ in $[4, 5]$, $\geq 95\%$ probability that $\Psi(P)$ lies in $[3.5, 5.5]$;

3. Almost noninformative: Prior mean of $\Psi(P)$ in $[3.7, 5.3]$, $\geq 95\%$ probability that $\Psi(P)$

lies in $[3, 6]$.

The architecture of our neural network estimator is almost identical to that in Section 4.5.2 except that the existing estimator being used is the one proposed in Jiao et al. (2015) (referred to as JVHW estimator), and we initialize the network to return the JVHW estimator. We use Algorithm 6 to construct $\mathcal{M}_\ell$ and Algorithm 4 to compute a Γ_ℓ-minimax estimator. The tuning parameters in the algorithms are identical to those used in Section 4.5.2 except that, in Algorithm 4, (i) the learning rate is $\eta = 0.001$, and (ii) the number of iterations is 6,000 for Γ_1. We change these tuning parameters because JVHW estimator is already minimax in terms of its convergence rate (Jiao et al., 2015), and we think we need to update the estimator more carefully in Algorithm 4 to obtain any possible improvement.

We finally compare the risk of JVHW and our trained Γ-minimax estimator estimator under our set data-generating mechanism computed with 20,000 Monte Carlo runs. We also compare their Bayes risk under the computed prior from Algorithm 4 using the last and finest grid in the computation with 20,000 Monte Carlo runs. The results are summarized in Table 4.3. In this simulation experiment, our Γ-minimax estimator reduces the risk by a fair percentage compared with JVHW estimator with somewhat informative prior knowledge. With almost noninformative prior knowledge, the risk of our Γ-minimax under P_0 is slightly higher than JVHW estimator, but the Bayes risk is still lower. The elevated risk under P_0 in this case is not surprising given that Γ-minimax estimators generally do not achieve optimal performance under every data-generating mechanism, but rather achieve optimal performance under the least favorable prior that is consistent with available knowledge. According to these simulation results, we conclude that incorporating weakly or strongly informative prior knowledge into the estimator may result in some improvement.

Fig 4.6 presents the unbiased estimated of Bayes risks over iterations when computing a Γ_1-minimax estimator. With somewhat informative prior information present, the Bayes risks appear to fluctuate without an increasing or decreasing trend at the beginning and decrease after several thousand iterations. With almost no prior information, the Bayes

Table 4.3: Risks and Bayes risks of estimators. $R(d, P_0)$: risk of the estimator under the true data-generating mechanism P_0. $r(d, \hat{\pi}^*)$: Bayes risk under prior $\hat{\pi}^*$, the computed prior from Algorithm 4 in the last and finest grid in the computation.

Strength of prior	Estimator	$R(d, P_0)$	$r(d, \hat{\pi}^*)$
strong	JVHW	0.041	0.045
	Γ-minimax	0.033	0.033
weak	JVHW	0.041	0.056
	Γ-minimax	0.040	0.048
almost none	JVHW	0.041	0.063
	Γ-minimax	0.046	0.055

risks appear to fluctuate with no trend. A reason may be that JVHW estimator is already minimax rate optimal (Jiao et al., 2015). The computed Γ-minimax estimators also appear to be somewhat similar to JVHW estimator: in the output layer of the three settings with different prior information, the coefficients for JVHW estimator are 0.96, 0.95 and 0.95, respectively; the coefficients for the previous hidden layer are 0.17, 0.09 and 0.02, respectively; the intercepts are 0.09, 0.13 and 0.16, respectively.

4.6 Discussion

We mainly focus on estimation. Nevertheless, our framework can be immediately applied to prediction. In this setup, an estimator may take in the observed data and output a function (e.g., coefficients of a feedforward neural network), whose prediction performance may be evaluated by an appropriately chosen risk function. Studying the performance of our algorithms in this setting is an interesting area for future work.

We propose algorithms to compute a Gamma-minimax estimator with theoretical guarantees under fairly general settings. These algorithms still leave room for improvement. As we discussed in Section 4.3.1, the stopping criterion we employ does not necessarily indi-

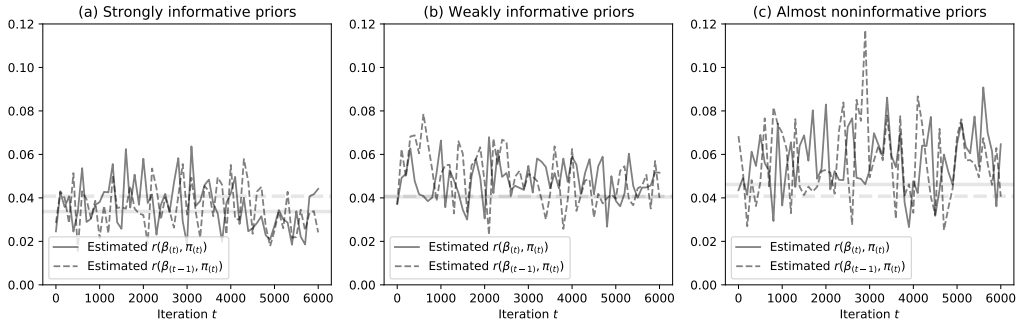

Figure 4.6: Estimated Bayes risks of the estimator over iterations when computing a Γ_1-minimax estimator. The lines are unbiased estimates of the current Bayes risks (y-axis) with 30 Monte Carlo runs over iterations (x-axis). The horizontal lines are the risks of JVHW (dashed) and the computed Γ-minimax (solid) estimators, respectively, under P_0 in the simulation. For clearness of visualization, the estimated Bayes risks are plotted every 100 iterations.

cate that the maximal Bayes risk is close to the true minimax Bayes risk. In future work, it would be interesting to derive a better criterion that necessarily does indicate this near optimality. Our algorithms also require the user to choose increasingly fine approximating grids to the model space. Although we propose a heuristic algorithm for this procedure that performed well in our experiments, at this point, we have not provided optimality guarantees for this scheme. It may also possible to improve our proposed algorithms to solve intermediate minimax problems in Section 4.3.2 by utilizing recent and ongoing advances from the machine learning literature that can be used to improve the training of generative adversarial networks.

We do not explicitly consider uncertainty quantification such as confidence intervals or credible intervals under a Gamma-minimax framework. Uncertainty quantification is important in practice since it provides more information than a point estimator and can be used for decision making. In theory, our method may be directly applied if such a problem can be formulated into a Gamma-minimax problem. However, such a formulation remains unclear. The most challenging part is to identify a suitable risk function that correctly balances the

level of uncertainty and the size of the output interval/region. Though the risk function used in Schafer and Stark (2009) appears to provide one possible starting point, it is not clear how to extend this approach to nonparametric settings.

In conclusion, we propose algorithms to compute a Gamma-minimax estimator under general models that can incorporate prior information in the form of generalized moment conditions. They can be useful when a parametric model is undesirable, semi-parametric efficiency theory does not apply, or we wish to utilize prior information to improve estimation.

Chapter 5
CONCLUDING REMARKS

5.1 Summary

In the three projects of this dissertation, we explored automated methods to efficiently estimate an aspect of the underlying data-generating mechanism of the data under minimal assumptions using machine learning tools. In Chapter 2, we proposed two novel sieve-like efficient plug-in estimators. We utilized the convenience of sieve estimation and the flexibility of machine learning tools in the proposed procedures, so that these procedures may be automated with weaker smoothness assumptions on the unknown functional feature. In Chapter 3, in the context of instrumental variables/encouragement, we considered individualized decision rules under treatment resource constraints in two cases. In one case, the treatment is intervened on; in the other case, the encouragement is intervened on. For both cases, we proposed nonparametric estimators for an optimal individualized rule and asymptotically linear estimators of its average causal effect relative to a prespecified reference rule. In Chapter 4, we proposed to use Gamma-minimax procedures in order to incorporate vague prior knowledge into the estimation under general and possibly rich models. We also proposed iterative algorithms to compute a Gamma-minimax estimator with theoretical convergence guarantees. We further proposed to use neural networks as the set of candidate estimators so that the resulting estimator achieves good performance while computation may be tractable.

5.2 Future research

As stated in Chapter 2, although our proposed estimators are efficient and asymptotically linear, we do not have an automated procedure to estimate the asymptotic variance and subsequently provide an asymptotically valid confidence interval. Unfortunately, the non-

parametric bootstrap is in general invalid when the summary is not Hadamard differentiable and especially when the method relies on cross-validation. It would be ideal if such a procedure can be developed so that the statistical inference procedure may be automated.

As stated in Chapter 3, a natural direction is to extend this work to longitudinal settings. We expect that identification results and techniques to construct asymptotic linear estimators that are similar to those in Chapter 3 can be applied to this setting. In addition, our method for intervention on the encouragement in Chapter 3 can be applied to non-IV settings where the treatment cost may stochastically depend on baseline covariates and there is a constraint on the average treatment cost.

For Chapter 4, we are interested in quantifying uncertainty under general models with prior knowledge incorporated. In theory, our method may be directly applied if such a problem can be formulated into a Gamma-minimax problem, but such a formulation remains unclear. We are also interested in utilizing recent advances in machine learning to accelerate the iterative training algorithm.

BIBLIOGRAPHY

Alberto Abadie. Semiparametric instrumental variable estimation of treatment response models. *Journal of Econometrics*, 113(2):231–263, 2003. ISSN 03044076.

Chunrong Ai and Xiaohong Chen. Efficient estimation of models with conditional moment restrictions containing unknown functions. *Econometrica*, 71(6):1795–1843, 2003. ISSN 00129682.

Joshua Angrist, Eric Bettinger, Erik Bloom, Elizabeth King, and Michael Kremer. Vouchers for private schooling in Colombia: Evidence from a randomized natural experiment. *American Economic Review*, 92(5):1535–1558, 2002. ISSN 00028282.

Joshua D. Angrist, Guido W. Imbens, and Donald B. Rubin. Identification of Causal Effects Using Instrumental Variables. *Journal of the American Statistical Association*, 91(434): 444–455, jun 1996. ISSN 1537274X.

Onyebuchi A Arah, Yasutaka Chiba, and Sander Greenland. Bias formulas for external adjustment and sensitivity analysis of unmeasured confounders. *Annals of epidemiology*, 18(8):637–646, 2008.

Jean Yves Audibert and Alexandre B. Tsybakov. Fast learning rates for plug-in classifiers. *Annals of Statistics*, 35(2):608–633, 2007. ISSN 00905364.

Michael Baiocchi, Jing Cheng, and Dylan S Small. Instrumental variable methods for causal inference. *Statistics in medicine*, 33(13):2297–2340, 2014.

Anirban Basu. Estimating person-centered treatment (pet) effects using instrumental variables: An application to evaluating prostate cancer treatments. *Journal of Applied Econometrics*, 29(4):671–691, 2014.

David Benkeser and Mark van Der Laan. The Highly Adaptive Lasso Estimator. In *Data Science and Advanced Analytics (DSAA), 2016 IEEE International Conference on*, pages 689–696. IEEE, 2016.

James O Berger. *Statistical Decision Theory and Bayesian Analysis*. Springer Series in Statistics. Springer New York, New York, NY, 1985. ISBN 978-1-4419-3074-3.

James O. Berger. Robust Bayesian analysis: sensitivity to the prior. *Journal of Statistical Planning and Inference*, 25(3):303–328, 1990. ISSN 03783758.

P J Bickel, C A J Klaassen, Y Ritov, and J A Wellner. *Efficient and adaptive estimation for semiparametric models*. Johns Hopkins University Press, Baltimore, 1993a.

Peter J. Bickel and Ya'acov Ritov. Nonparametric estimators which can be "plugged-in". *Annals of Statistics*, 31(4):1033–1053, 2003. ISSN 00905364.

Peter J Bickel, Chris AJ Klaassen, Peter J Bickel, Ya'acov Ritov, J Klaassen, Jon A Wellner, and YA'Acov Ritov. *Efficient and adaptive estimation for semiparametric models*, volume 4. Johns Hopkins University Press Baltimore, 1993b.

PJ Bickel, Friedrich Götze, and WR van Zwet. Resampling fewer than n observations: Gains, losses, and remedies for losses. *STATISTICA SINICA*, 7(1), 1997.

Jolene Birmingham, Andrea Rotnitzky, and Garrett M. Fitzmaurice. Pattern-mixture and selection models for analysing longitudinal data with monotone missing patterns. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 65(1):275–297, 2003. ISSN 13697412.

George W Brown. Iterative solution of games by fictitious play. *Activity analysis of production and allocation*, 13(1):374–376, 1951.

Brent Bryan, H. Brendan McMahan, Chad M. Schafer, and Jeff Schneider. Efficiently computing minimax expected-size confidence regions. *ACM International Conference Proceeding Series*, 227:97–104, 2007.

John Bunge, Amy Willis, and Fiona Walsh. Estimating the Number of Species in Microbial Diversity Studies. *Annual Review of Statistics and Its Application*, 1(1):427–445, 2014. ISSN 2326-8298.

Bibhas Chakraborty and Erica E.M. Moodie. *Statistical Methods for Dynamic Treatment Regimes*. Statistics for Biology and Health. Springer New York, New York, NY, 2013. ISBN 978-1-4614-7427-2.

Bibhas Chakraborty, Eric B Laber, and Ying-Qi Zhao. Inference about the expected performance of a data-driven dynamic treatment regime. *Clinical Trials*, 11(4):408–417, 2014.

LanXiang Chen, J. Eichenauer-Herrmann, and J. Lehn. Gamma-minimax estimators for the parameters of a multinomial distribution. *Applicationes Mathematicae*, 20(4):561–564, 1988. ISSN 1233-7234.

Lanxiang Chen, Jürgen Eichenauer-Herrmann, Heike Hofmann, and Jürgen Kindler. *Gamma-minimax estimators in the exponential family*. Polska Akademia Nauk, Instytut Matematyczny, 1991.

Xiaohong Chen. Chapter 76 Large Sample Sieve Estimation of Semi-Nonparametric Models. *Handbook of Econometrics*, 6(SUPPL. PART B):5549–5632, 2007. ISSN 15734412.

Xiaohong Chen and Zhipeng Liao. Sieve M inference on irregular parameters. In *Journal of Econometrics*, volume 182, pages 70–86. Elsevier Ltd, 2014.

Xiaohong Chen and Zhipeng Liao. Sieve semiparametric two-step GMM under weak dependence. *Journal of Econometrics*, 189(1):163–186, 2015. ISSN 18726895.

Xiaohong Chen and Demian Pouzo. Efficient estimation of semiparametric conditional moment models with possibly nonsmooth residuals. *Journal of Econometrics*, 152(1):46–60, 2009. ISSN 03044076.

Xiaohong Chen and Demian Pouzo. Sieve Wald and QLR Inferences on Semi/Nonparametric Conditional Moment Models. *Econometrica*, 83(3):1013–1079, 2015. ISSN 00129682.

Xiaohong Chen, Yanqin Fan, and Viktor Tsyrennikov. Efficient estimation of semiparametric multivariate copula models, sep 2006. ISSN 01621459. URL `https://www.tandfonline.com/action/journalInformation?journalCode=uasa20`.

Xiaohong Chen, Han Hong, and Alessandro Tarozzi. Semiparametric Efficiency in GMM Models of Nonclassical Measurement Errors, Missing Data and Treatment Effects. *Cowles Foundation Discussion Paper*, (1644), 2008.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, and Whitney Newey. Double/debiased/Neyman machine learning of treatment effects. *American Economic Review*, 107(5):261–265, may 2017. ISSN 00028282.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *Econometrics Journal*, 21(1):C1–C68, feb 2018. ISSN 1368423X.

B. Csáji. Approximation with artificial neural networks. Technical report, 2001.

George B Dantzig. Discrete-variable extremum problems. *Operations research*, 5(2):266–288, 1957.

J. Eichenauer-Herrmann. A gamma-minimax result for the class of symmetric and unimodal priors. *Statistical Papers*, 31(1):301–304, 1990. ISSN 09325026.

J. Eichenauer-Herrmann, K. Ickstadt, and E. Weiß. Gamma-minimax results for the class of unimodal priors. *Statistical Papers*, 35(1):43–56, 1994. ISSN 09325026.

Dumitru Erhan, Aaron Courville, Yoshua Bengio, and Pascal Vincent. Why does unsupervised pre-training help deep learning? Technical report, 2010. URL http://proceedings.mlr.press/v9/erhan10a.html.

Lawrence C. Evans and Ronald F. Gariepy. *Measure theory and fine properties of functions, revised edition*. CRC Press, 2015. ISBN 9781482242393.

Ky Fan. Minimax theorems. *Proceedings of the National Academy of Sciences of the United States of America*, 39(1):42, 1953.

Jerome H Friedman. Greedy Function Approximation: A Gradient Boosting Machine. Technical Report 5, 2001.

Jerome H. Friedman. Stochastic gradient boosting. *Computational Statistics and Data Analysis*, 38(4):367–378, feb 2002. ISSN 01679473.

Richard D Gill, Mark J van der Laan, and Jon A Wellner. *Inefficient estimators of the bivariate survival function for three models*. Rijksuniversiteit Utrecht. Mathematisch Instituut, 1993.

Richard D. Gill, Mark J. van der Laan, and James M. Robins. Coarsening at Random: Characterizations, Conjectures, Counter-Examples. pages 255–294. Springer, New York, NY, 1997.

Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Deep sparse rectifier neural networks. Technical report, 2011.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. Technical Report January, 2014. URL http://www.github.com/goodfeli/adversarial.

Peter J. Green. Reversible Jump Markov Chain Monte Carlo Computation and Bayesian Model Determination. *Biometrika*, 82(4):711, 1995. ISSN 00063444.

Rolf HH Groenwold, David B Nelson, Kristin L Nichol, Arno W Hoes, and Eelko Hak. Sensitivity analyses to estimate the potential impact of unmeasured confounding in causal research. *International journal of epidemiology*, 39(1):107–117, 2009.

Robert W Haley. Point: Bias from the "healthy-warrior effect" and unequal follow-up in three government studies of health effects of the gulf war. *American journal of epidemiology*, 148(4):315–323, 1998.

Peter Hall. *The bootstrap and Edgeworth expansion*. Springer Science & Business Media, 2013.

Boris Hanin and Mark Sellke. Approximating Continuous Functions by ReLU Nets of Minimal Width. *arXiv preprint arXiv:1710.11278v2*, oct 2017. URL `http://arxiv.org/abs/1710.11278`.

Wolfgang Härdle and Thomas M. Stoker. Investigating smooth multiple regression by the method of average derivatives. *Journal of the American Statistical Association*, 84(408): 986–995, dec 1989. ISSN 1537274X.

Trevor. Hastie and Robert. Tibshirani. *Generalized additive models*. Chapman and Hall, 1990. ISBN 9780412343902.

W. K. Hastings. Monte Carlo Sampling Methods Using Markov Chains and Their Applications. *Biometrika*, 57(1):97, 1970. ISSN 00063444.

Daniel F. Heitjan. Ignorability and Coarse Data: Some Biomedical Examples. *Biometrics*, 49(4):1099, 1993. ISSN 0006341X.

Daniel F. Heitjan. Ignorability in general incomplete-data models. *Biometrika*, 81(4):701–708, 1994. ISSN 00063444.

Daniel F. Heitjan and Donald B. Rubin. Ignorability and Coarse Data. Technical Report 4, 1991.

Kurt Hornik. Approximation capabilities of multilayer feedforward networks. *Neural Networks*, 4(2):251–257, 1991. ISSN 08936080.

Guang Bin Huang, Lei Chen, and Chee Kheong Siew. Universal approximation using incremental constructive feedforward networks with random hidden nodes. *IEEE Transactions on Neural Networks*, 17(4):879–892, 2006a. ISSN 10459227.

Guang Bin Huang, Qin Yu Zhu, and Chee Kheong Siew. Extreme learning machine: Theory and applications. *Neurocomputing*, 70(1-3):489–501, 2006b. ISSN 09252312.

Guido W. Imbens and Joshua D. Angrist. Identification and Estimation of Local Average Treatment Effects. *Econometrica*, 62(2):467–475, 1994. ISSN 0012-9682.

Shunhua Jiang, Zhao Song, Omri Weinstein, and Hengjie Zhang. Faster Dynamic Matrix Inverse for Faster LPs. *arXiv preprint arXiv:2004.07470v1*, 2020. URL `http://arxiv.org/abs/2004.07470`.

Jiantao Jiao, Kartik Venkat, Yanjun Han, and Tsachy Weissman. Minimax Estimation of Functionals of Discrete Distributions. *IEEE Transactions on Information Theory*, 61(5): 2835–2885, 2015. ISSN 00189448.

Peter J. Kempthorne. Numerical Specification of Discrete Least Favorable Prior Distributions. *SIAM Journal on Scientific and Statistical Computing*, 8(2):171–184, 1987. ISSN 0196-5204.

Edward H Kennedy. Semiparametric theory and empirical processes in causal inference. In *Statistical causal inferences and their applications in public health research*, pages 141–167. Springer, 2016.

Patrick Kidger and Terry Lyons. Universal Approximation with Deep Narrow Networks. Technical report, 2020.

Michael R. Kosorok. *Introduction to Empirical Processes and Semiparametric Inference*, volume 77 of *Springer Series in Statistics*. Springer New York, 2008. ISBN 978-0-387-74977-8.

Cynthia A LeardMann, Teresa M Powell, Tyler C Smith, Michael R Bell, Besa Smith, Edward J Boyko, Tomoko I Hooper, Gary D Gackstetter, Mark Ghamsary, and Charles W Hoge. Risk factors associated with suicide in current and former us military personnel. *Jama*, 310(5):496–506, 2013.

E. L. Lehmann and George Casella. *Theory of Point Estimation*. Springer, 1998.

Jonathan Levy, Mark van der Laan, Alan Hubbard, and Romain Pirracchio. A Fundamental Measure of Treatment Effect Heterogeneity. nov 2018.

Tianyi Lin, Chi Jin, and Michael I. Jordan. On Gradient Descent Ascent for Nonconvex-Concave Minimax Problems. *arXiv preprint arXiv:1906.00331v6*, 2019. URL `http://arxiv.org/abs/1906.00331`.

Alex Luedtke, Marco Carone, Noah Simon, and Oleg Sofrygin. Learning to learn from data: Using deep adversarial learning to construct optimal statistical procedures. *Science Advances*, 6(9):eaaw2140, 2020a. ISSN 23752548.

Alex Luedtke, Incheoul Chung, and Oleg Sofrygin. Adversarial Monte Carlo Meta-Learning of Optimal Prediction Procedures. *arXiv preprint arXiv:2002.11275v1*, 2020b. URL `http://arxiv.org/abs/2002.11275`.

Alexander R. Luedtke and Mark J. van der Laan. Optimal Individualized Treatments in Resource-Limited Settings. *International Journal of Biostatistics*, 12(1):283–303, 2016a. ISSN 15574679.

Alexander R. Luedtke and Mark J. van der Laan. Statistical inference for the mean outcome under a possibly non-unique optimal treatment strategy. *Annals of Statistics*, 44(2):713–742, 2016b. ISSN 00905364.

Stephane Mallat. *A Wavelet Tour of Signal Processing*. Elsevier, 2009. ISBN 9780123743701.

Haggai Maron, Ethan Fetaya, Nimrod Segol, and Yaron Lipman. On the Universality of Invariant Networks. 2019.

J. S. Marron. Visual understanding of higher-order kernels. *Journal of Computational and Graphical Statistics*, 3(4):447–458, dec 1994. ISSN 15372715.

Silvano. Martello and Paolo. Toth. *Knapsack problems : algorithms and computer implementations*. J. Wiley & Sons, 1990. ISBN 9780471924203.

Edwin P. Martens, Wiebe R. Pestman, Anthonius De Boer, Svetlana V. Belitser, and Olaf H. Klungel. Instrumental variables: Application and limitations. *Epidemiology*, 17(3):260–267, 2006. ISSN 10443983.

Llew Mason, Jonathan Baxter, Peter Bartlett, and Marcus Frean. Boosting Algorithms as Gradient Descent in Function Space. Technical report, 1999.

Llew Mason, Jonathan Baxter, Peter L. Bartlett, and Marcus Frean. Boosting Algorithms as Gradient Descent. Technical report, 2000.

Jacqueline A Mauro, Edward H Kennedy, and Daniel Nagin. Instrumental variable methods using dynamic interventions. *arXiv preprint arXiv:1811.01301*, 2018.

Nicholas Metropolis, Arianna W. Rosenbluth, Marshall N. Rosenbluth, Augusta H. Teller, and Edward Teller. Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6):1087–1092, 1953. ISSN 00219606.

R. I. Miller and R. G. Wiegert. Documenting completeness, species-area relations, and the species-abundance distribution of a regional flora. *Ecology*, 70(1):16–22, 1989. ISSN 00129658.

S. A. Murphy. Optimal dynamic treatment regimes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(2):331–355, 2003.

Elizbar A Nadaraya. On estimating regression. *Theory of Probability & Its Applications*, 9 (1):141–142, 1964.

Wayne Nelson. Minimax Solution of Statistical Decision Problems by Iteration. *The Annals of Mathematical Statistics*, 37(6):1643–1657, 1966. ISSN 0003-4851.

Whitney Newey, Fushing Hsieh, and James Robins. Undersmoothing and Bias Corrected Functional Estimation. *Working papers*, 1998.

Whitney K. Newey. Convergence rates and asymptotic normality for series estimators. *Journal of Econometrics*, 79(1):147–168, jul 1997. ISSN 0304-4076.

Whitney K. Newey, Fushing Hsieh, and James M. Robins. Twicing Kernels and a Small Bias Property of Semiparametric Estimators. *Econometrica*, 72(3):947–962, may 2004.

J Neyman. Sur les applications de la théorie des probabilités aux expériences agricoles: Essay des principles. (Excerpts reprinted and translated to English, 1990). *Statistical Science*, 5: 463–472, 1923.

Roger Fandom Noubiap and Wilfried Seidel. An algorithm for calculating Γ-minimax decision rules under generalized moment conditions. *Annals of Statistics*, 29(4):1094–1116, 2001. ISSN 00905364.

Elizabeth L Ogburn, Andrea Rotnitzky, and James M Robins. Doubly robust estimation of the local average treatment effect curve. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 77(2):373–396, 2015.

V Olman and A Shmundak. Minimax Bayes estimation of mean of normal law for the class of unimodal a priori distributions. *Proc. Acad. Sci. Estonian Physics Math*, 34:148–153, 1985.

Alon Orlitsky, Ananda Theertha Suresh, and Yihong Wu. Optimal prediction of the number

of unseen species. *Proceedings of the National Academy of Sciences of the United States of America*, 113(47):13283–13288, 2016. ISSN 10916490.

Art B Owen. Multidimensional variation for quasi-monte carlo. In *Contemporary Multivariate Analysis And Design Of Experiments: In Celebration of Professor Kai-Tai Fang's 65th Birthday*, pages 49–74. World Scientific, 2005.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019. URL `http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-`

Paul E Peterson, William G Howell, and Jay P Greene. An Evaluation of the Cleveland Voucher Program after Two Years. Technical report, 1999.

J. Pfanzagl. *Contributions to a General Asymptotic Statistical Theory*, volume 13 of *Lecture Notes in Statistics*. Springer New York, New York, NY, 1982. ISBN 978-0-387-90776-5.

Johann Pfanzagl. Estimation in semiparametric models. In *Estimation in Semiparametric Models*, pages 17–22. Springer, 1990.

Iosif Pinelis. On the extreme points of moments sets. *Mathematical Methods of Operations Research*, 83(3):325–349, 2016. ISSN 14325217.

Min Qian and Susan A Murphy. Performance guarantees for individualized treatment rules. *Annals of statistics*, 39(2):1180, 2011.

Mark A Reger, Derek J Smolenski, Nancy A Skopp, Melinda J Metzger-Abamukang, Han K Kang, Tim A Bullman, Sondra Perdue, and Gregory A Gahm. Risk of suicide among us

military service members following operation enduring freedom or operation iraqi freedom deployment and separation from the us military. *JAMA psychiatry*, 72(6):561–569, 2015.

James M. Robins. Optimal Structural Nested Models for Optimal Sequential Decisions. pages 189–326. Springer, New York, NY, 2004.

James M Robins, Andrea Rotnitzky, and Daniel O Scharfstein. Sensitivity analysis for selection bias and unmeasured confounding in missing data and causal inference models. In *Statistical models in epidemiology, the environment, and clinical trials*, pages 1–94. Springer, 2000.

Julia Robinson. An Iterative Method of Solving a Game. *The Annals of Mathematics*, 54 (2):296, 1951. ISSN 0003486X.

Donald B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701, 1974. ISSN 0022-0663.

Chad M. Schafer and Philip B. Stark. Constructing confidence regions of optimal expected size. *Journal of the American Statistical Association*, 104(487):1080–1089, 2009. ISSN 01621459.

Tsung Jen Shen, Anne Chao, and Chih Feng Lin. Predicting the number of new species in further taxonomic sampling. *Ecology*, 84(3):798–804, 2003. ISSN 00129658.

Xiaotong Shen. On methods of sieves and penalization. *Annals of Statistics*, 25(6):2555–2591, 1997. ISSN 00905364.

Maurice Sion. On general minimax theorems. *Pacific Journal of Mathematics*, 8(1):171–176, 1958. ISSN 00308730.

Daniel A. Spielman and Shang Hua Teng. Smoothed analysis of algorithms: Why the simplex algorithm usually takes polynomial time. *Journal of the ACM*, 51(3):385–463, 2004. ISSN 00045411.

E. J. Tchetgen Tchetgen and S. Vansteelandt. Alternative Identification and Inference for the Effect of Treatment on the Treated with an Instrumental Variable. *Harvard University BiostatisticsWorking Paper Series*, 2013.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

Lisa Torrey and Jude Shavlik. Transfer learning. *Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques*, 11:242–264, 2009.

Boriska Toth and Mark van der Laan. *Targeted Learning of Optimal Individualized Treatment Rules Under Cost Constraints*, pages 1–22. Springer Singapore, Singapore, 2018. ISBN 978-981-10-7820-0.

J. v. Neumann. Zur Theorie der Gesellschaftsspiele. *Mathematische Annalen*, 100(1):295–320, 1928. ISSN 00255831.

M J van der laan and S Dudoit. Unified cross-validation methodology for selection among estimators and a general cross-validated adaptive epsilon-net estimator: Finite sample oracle inequalities and examples. *U.C. Berkeley Division of Biostatistics Working Paper*, 2003.

Mark van Der Laan. A Generally Efficient Targeted Minimum Loss Based Estimator based on the Highly Adaptive Lasso. *International Journal of Biostatistics*, 13(2), 2017. ISSN 15574679.

Mark van der Laan and Daniel Rubin. Targeted Maximum Likelihood Learning. *U.C. Berkeley Division of Biostatistics Working Paper Series*, oct 2006.

Mark J. van der Laan and Alexander R. Luedtke. Targeted Learning of the Mean Outcome under an Optimal Dynamic Treatment Rule. *Journal of Causal Inference*, 3(1), 2014. ISSN 2193-3677.

Mark J. van der Laan and James M. Robins. *Unified Methods for Censored Longitudinal Data and Causality*. Springer Series in Statistics. Springer New York, New York, NY, 2003. ISBN 978-1-4419-3055-2.

Mark J. van der Laan and Sherri Rose. *Targeted Learning in Data Science*. 2018. ISBN 978-3-319-65303-7.

Mark J. van der Laan, Eric C Polley, and Alan E. Hubbard. Super Learner. *Statistical Applications in Genetics and Molecular Biology*, 6(1), jan 2007. ISSN 2194-6302.

Mark J. van der Laan, David Benkeser, and Weixin Cai. Efficient Estimation of Pathwise Differentiable Target Parameters with the Undersmoothed Highly Adaptive Lasso. *arXiv preprint arXiv:1908.05607v1*, 2019. URL `http://arxiv.org/abs/1908.05607`.

Aad van der Vaart and Jon Wellner. *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer Series in Statistics. Springer, 2000. ISBN 0387946403.

Tyler J VanderWeele and Onyebuchi A Arah. Bias formulas for sensitivity analysis of unmeasured confounding for general outcomes, treatments, and confounders. *Epidemiology (Cambridge, Mass.)*, 22(1):42–52, 2011.

Vladimir Vapnik. Principles of risk minimization for learning theory. In *Advances in neural information processing systems*, pages 831–838, 1992.

Brani Vidakovic. Γ-Minimax: A Paradigm for Conservative Robust Bayesians. pages 241–259. Springer, New York, NY, 2000.

Abraham Wald. The fitting of straight lines if both variables are subject to error. *The Annals of Mathematical Statistics*, 11(3):284–300, 1940.

Abraham Wald. Statistical Decision Functions Which Minimize the Maximum Risk. *The Annals of Mathematics*, 46(2):265, 1945. ISSN 0003486X.

Linbo Wang and Eric Tchetgen Tchetgen. Bounded, efficient and multiply robust estimation of average treatment effects using instrumental variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(3):531–550, 2018.

Brian D. Williamson, Peter B. Gilbert, Marco Carone, and Noah Simon. Nonparametric variable importance assessment using machine learning techniques. *Biometrics*, 2020. ISSN 15410420.

S Yang and P Ding. Asymptotic inference of causal effects with observational studies trimmed by the estimated propensity scores. *Biometrika*, 105(2):487–493, 03 2018. ISSN 0006-3444.

Manzil Zaheer, Satwik Kottur, Siamak Ravanbhakhsh, Barnabás Póczos, Ruslan Salakhutdinov, and Alexander J. Smola. Deep sets. *Advances in Neural Information Processing Systems*, 2017-Decem(ii):3392–3402, 2017. ISSN 10495258.

Baqun Zhang, Anastasios A Tsiatis, Eric B Laber, and Marie Davidian. A robust method for estimating optimal treatment regimes. *Biometrics*, 68(4):1010–1018, 2012.

Cun Hui Zhang. Nearly unbiased variable selection under minimax concave penalty. *Annals of Statistics*, 38(2):894–942, apr 2010. ISSN 00905364.

Yingqi Zhao, Donglin Zeng, A. John Rush, and Michael R. Kosorok. Estimating Individualized Treatment Rules Using Outcome Weighted Learning. *Journal of the American Statistical Association*, 107(499):1106–1118, 2012. ISSN 0162-1459.

Appendix A

SUPPORTING INFORMATION FOR CHAPTER 2

A.1 Modification of chosen norm for evaluating the conditions: case study of mean counterfactual outcome

In this appendix, we consider a parameter that requires a modification in the chosen norm for evaluating the conditions. In particular, we discuss estimating counterfactual mean outcome in Example 3.

Let $g_0 : x \mapsto P_0(A = 1|X = x)$ be the propensity score function. A natural choice of the loss function is $\ell(\theta) : v \mapsto a[y - \theta(x)]^2$. Indeed, learning a function with this loss function is equivalent to fitting a function within the stratum of observations that received treatment 1. Unfortunately, this loss function does not satisfy Condition A2 with $L^2(P_0)$-norm, because $P_0\{\ell(\theta) - \ell(\theta_0)\} = P_0\{g_0 \cdot (\theta - \theta_0)^2\}$ cannot be well approximated by $\alpha_{0,\ell} P_0\{(\theta - \theta_0)^2\}/2$ for any constant $\alpha_{0,\ell} > 0$ unless g_0 is a constant. One way to overcome this challenge is to choose the alternative inner product $\langle \theta_1, \theta_2 \rangle_{g_0} := P_0\{g_0\theta_1\theta_2\}$ and its induced norm $\|\cdot\|_{g_0}$. In this case, Condition A2 is satisfied once $\|\cdot\|$ is replaced by $\|\cdot\|_{g_0}$ in the condition statement. Under this choice, $\Psi'_{\theta_0} = P_0(\theta - \theta_0) = \langle 1/g_0, \theta - \theta_0 \rangle_{g_0}$. We may redefine the corresponding $\dot{\Psi}$ similarly as the function that satisfies

$$\Psi'_{\theta_0} = \langle \dot{\Psi}, \theta - \theta_0 \rangle_{g_0},$$

and it immediately follows that $\dot{\Psi} = 1/g_0$. Moreover, under a strong positivity condition, namely $g_0(X) \geq \delta_g > 0$ a.s. for some δ_g, which is a typical condition in causal inference literature (van der Laan and Rose, 2018; Yang and Ding, 2018), then it is straightforward to show that $\delta_g \|\cdot\| \leq \|\cdot\|_{g_0} \leq \|\cdot\|$; that is, $\|\cdot\|_{g_0}$ is equivalent to $L^2(P_0)$-norm. Using this fact, it can be shown that all other conditions with respect to the $L^2(P_0)$-inner product are equivalent to the corresponding conditions with respect to $\langle \cdot, \cdot \rangle_{g_0}$.

Therefore, the data-adaptive series can be applied to estimation of the counterfactual mean outcome under our conditions for $L^2(P_0)$-inner product. If we use the targeted form in Remark 2, then we need a flexible estimator of g_0 and the procedure is almost identical to a TMLE (van der Laan and Rose, 2018). If we use the generalized data-adaptive series, we would require sufficient amount of smoothness for $g_0(\cdot)$. In the latter case, the change in norm when evaluating the conditions is a purely technical device and the estimation procedure is the same as would have been used if we had used the $L^2(P_0)$-norm. We also note that the same argument may be used to show that in Example 4, with $\ell(\theta) : v \mapsto a[z-\mu_1(x)]^2 + (1-a)[z-\mu_0(x)]^2$ being the usual squared-error loss, we may choose the alternative inner product $\langle \theta_1, \theta_2 \rangle_{g_0} := P_0\{\theta_1^\top \cdot \text{diag}(1-g_0, g_0) \cdot \theta_2\}$ and find that $\dot{\Psi} = (-2/(1-g_0)\cdot[(\mu_{01}-\mu_{00}) - P_0(\mu_{01}-\mu_{00})], 2/g_0 \cdot [(\mu_{01}-\mu_{00}) - P_0(\mu_{01}-\mu_{00})])^\top$, as we did in Section 2.5.3.

A.2 Additional conditions

Throughout the rest of this appendix, we use C to denote a general absolute positive constant that can vary line by line.

A.2.1 HAL

Condition B3 (Empirical processes conditions)**.** For any fixed $\vartheta \in \Theta_{\text{v},M}$ and some $\Delta > 0$, it holds that $\ell(\theta)$, $\ell_0'[\theta - \theta_0]$ and $\{r[\theta-\theta_0] - r[\theta + \delta(\vartheta-\theta)-\theta_0]\}/\delta$ are càdlàg for all $\theta \in \Theta_{\text{v},M}$ and all $\delta \in [0,\Delta]$. Moreover, the following terms are all finite:

$$\sup_{\theta\in\Theta_{\text{v},M}} \|\ell(\theta)\|_{\text{v}}, \sup_{\theta\in\Theta_{\text{v},M}} \|\ell_0'[\theta-\theta_0]\|_{\text{v}}, \sup_{\theta\in\Theta_{\text{v},M},\delta\in[0,\Delta]} \left\| \frac{r[\theta-\theta_0] - r[\theta-\theta_0+\delta(\vartheta-\theta)]}{\delta} \right\|_{\text{v}}.$$

In addition, $\|\ell_0'[\hat{\theta}_n - \theta_0]\|$ and $\sup_{\delta\in[0,\Delta]} \left\| \{r[\hat{\theta}_n - \theta_0] - r[\hat{\theta}_n - \theta_0 + \delta(\vartheta - \hat{\theta}_n)]\}/\delta \right\|$ converge to 0 in probability.

Condition B4 (Finite variance of influence function)**.** $\xi^2 := \text{Var}_{P_0}(\ell_0'[\dot{\Psi}](V))/\alpha_{0,\ell}^2 < \infty$.

A.2.2 Data-adaptive series

Condition C6 (Local Lipschitz continuity of $\Pi_{n,\theta_0}(\mathcal{I})$)**.** For sufficiently large n,

$$\|\Pi_{n,\theta_0}(\mathcal{I}) \circ \theta - \Pi_{n,\theta_0}(\mathcal{I}) \circ \theta_0\| \leq C\|\theta - \theta_0\|$$

for all $\theta \in \Theta$ with $\|\theta - \theta_0\| \leq n^{-1/4}$.

Condition C7 (Local Lipschitz continuity of $\dot{\psi}$ and $\Pi_{n,\theta_0}(\dot{\psi})$)**.** For sufficiently large n, for all $\theta \in \Theta$ with $\|\theta - \theta_0\| \leq n^{-1/4}$,

(a) $\|\dot{\psi} \circ \theta - \dot{\psi} \circ \theta_0\| \leq C\|\theta - \theta_0\|$;

(b) $\|\Pi_{n,\theta_0}(\dot{\psi}) \circ \theta - \Pi_{n,\theta_0}(\dot{\psi}) \circ \theta_0\| \leq C\|\theta - \theta_0\|$.

Condition C8 (Empirical process conditions)**.** There exists some constant $\Delta > 0$ such that

$$\sup_{\delta \in [0,\Delta]} \left| (P_n - P_0) \left\{ \frac{r[\theta_n^* - \theta_0] - r[\pi_n((1-\delta)\theta_n^* + \delta(\pm\dot{\Psi} + \theta_0)) - \theta_0]}{\delta} \right\} \right| = o_p(n^{-1/2}),$$

$$(P_n - P_0)\ell_0'[(\pm\dot{\Psi} + \theta_0) - \pi_n(\pm\dot{\Psi} + \theta_0)] = o_p(n^{-1/2}),$$

$$(P_n - P_0)\ell_0'[\theta_n^* - \theta_0] = o_p(n^{-1/2}).$$

Condition C9 (Finite variance of influence function)**.** $\xi^2 := \mathrm{Var}_{P_0}(\ell_0'[\dot{\Psi}](V))/\alpha_{0,\ell}^2 < \infty$.

A.2.3 Generalized data-adaptive series

Condition C6* (Local Lipschitz continuity of projected $\mathcal{I}$ for Θ_{n,θ_0})**.** For sufficiently large n, $\|\Pi_{n,\theta_0}(\mathcal{I}) \circ (\theta, \mathcal{I}_x) - \Pi_{n,\theta_0}(\mathcal{I}) \circ (\theta_0, \mathcal{I}_x)\| \leq C\|\theta - \theta_0\|$ for all $\|\theta - \theta_0\| \leq n^{-1/4}$.

Condition C7* (Local Lipschitz continuity of $\dot{\psi}$ and its projection for Θ_{n,θ_0})**.** For sufficiently large n, for all $\|\theta - \theta_0\| \leq n^{-1/4}$,

(a) $\|\dot{\psi} \circ (\theta, \mathcal{I}_x) - \dot{\psi} \circ (\theta_0, \mathcal{I}_x)\| \leq C\|\theta - \theta_0\|$;

(b) $\|\Pi_{n,\theta_0}(\dot{\psi}) \circ (\theta, \mathcal{I}_x) - \Pi_{n,\theta_0}(\dot{\psi}) \circ (\theta_0, \mathcal{I}_x)\| \leq C\|\theta - \theta_0\|$.

A.2.4 Conditions for efficiency of the plug-in estimator

Define a collection of submodels

$$\{\{P_{H,\delta} : \delta \in B_H \subseteq \mathbb{R}\} : H \in \mathcal{H}\}$$

for which: (i) $\mathcal{H}$ is a subset of $L_0^2(P_0)$ and the $L_0^2(P_0)$-closure of its linear span is $L_0^2(P_0)$; and (ii) each $\{P_{H,\delta} : \delta \in B_H \subseteq \mathbb{R}\}$ is a regular univariate parametric submodel that passes through P_0 and has score H for δ at $\delta = 0$. For each $H \in \mathcal{H}$ and $\delta \in B_H$, we define $\theta_{H,\delta} \in \text{argmin}_{\theta\in\Theta} P_{H,\delta}\ell(\theta)$. In this appendix, for all small o and big O notations, we let $\delta \to 0$ with H fixed.

Condition E1 (Sufficiently close risk minimizer)**.** For any given $H \in \mathcal{H}$, $\|\theta_{H,\delta} - \theta_0\| = o(\delta^{1/2})$.

Condition E2 (Quadratic behavior of loss function remainder near 0)**.** For any given $H \in \mathcal{H}$ and ϑ, there exists positive $\delta' = o(\delta)$ such that $(P_{H,\delta} - P_0)\{r[(1-\delta')(\theta_{H,\delta} - \theta_0) + \delta'\dot{\Psi}] - r[\theta_{H,\delta} - \theta_0]\}/\delta' = o(\delta)$.

A.3 Discussion of technical conditions for data-adaptive series and its generalization

A.3.1 Theorem 2.2

Condition C2 usually imposes an upper bound on the growth rate of K. To see this, we show that Condition C2 is equivalent to a term being $o_p(n^{-1/4})$, and an upper bound of this term is controlled by K. Let $\theta_n^\dagger \in \text{argmin}_{\theta\in\Theta_n} P_0\ell(\theta)$ be the true-risk minimizer in Θ_n. Under Conditions A2, C1, C3 and C6, by Lemma A.4, it follows that Condition C2 is equivalent to requiring that $\|\theta_n^* - \theta_n^\dagger\| = o_p(n^{-1/4})$. Note that θ_n^* minimizes the empirical risk in Θ_n, and M-estimation theory (van der Vaart and Wellner, 2000) can show that $\|\theta_n^* - \theta_n^\dagger\|$ can be upper bounded by an empirical process term, whose upper bound is related to the complexity of Θ_n, namely how fast K grows with sample size. To ensure this bound is $o_p(n^{-1/4})$, K must not grow too quickly.

Condition C3 assumes that the identity function can be well approximated by the series ϕ_k with the specified number of terms K in the $L^2(P_{\theta_0})$ sense. If $\text{Span}\{\phi_1, \ldots, \phi_K\}$ does not contain $\mathcal{I}$ for any K, then sufficiently many terms must be included to satisfy this condition; that is, this condition imposes a lower bound on the rate at which K should grow with n. Even if $\text{Span}\{\phi_1, \ldots, \phi_K\}$ does contain $\mathcal{I}$ for some finite K, this condition still requires that K is not too small.

Condition C4 is implied by the following condition in view of Lemma A.2:

Condition C4s. $\|[\dot{\psi} - \Pi_{n,\theta_0}(\dot{\psi})] \circ \theta_0\| = o(n^{-1/4})$.

This condition is similar to Condition C3. However, in general, we do not expect $\dot{\psi}$ to be contained in $\text{Span}\{\phi_1, \ldots, \phi_K\}$ for any K, and hence this condition generally imposes a lower bound on the rate of K. Note that Condition C4s is stronger than Condition C4, and there are interesting examples where C4 holds but C4s fails to hold. Indeed, if θ_n^* converges to θ_0 at a rate much faster than $n^{-1/4}$, then C4 can be satisfied even if $\|[\dot{\psi} - \Pi_{n,\theta_0}(\dot{\psi})] \circ \theta_0\|$ decays to zero in probability relatively slowly — that is, the convergence rate of θ_n^* can compensate for the approximation error of $\dot{\psi}$. This is one way in which we can benefit from using flexible ML algorithms to estimate θ_0: if θ_n^0 converges to θ_0 at a fast rate, then we can expect θ_n^* to also have a fast convergence rate.

Conditions C2, C3 and C4 are not stringent provided sufficient smoothness on derivatives of $\dot{\psi}$ and a reasonable series. For example, as noted in Chen (2007), when $\dot{\psi}$ has a bounded p-th order derivative and the polynomial, trigonometric series or spline with degree at least $p+1$ is used, then if $K^2/n \to 0$ ($K^3/n \to 0$ for polynomial series), the term in Condition C2 is $O_p(\sqrt{K/n})$; the terms in Condition C3 and the sufficient Condition C4s are $O(K^{-p/q})$. Therefore, we can select K to grow at a rate faster than $n^{q/(4p)}$ and slower than $n^{1/2}$ ($n^{1/3}$ for polynomial series). If p is large, then this allows for a wide range of rates for K. Typically $\dot{\Psi}$ (and hence $\dot{\psi}$) is only related to the summary of interest Ψ but not the true function θ_0. For example, for the summary $\Psi(\theta) = P_0(f \circ \theta)$ at the beginning of Section 2.4.1, $\dot{\psi} = f'$ is

variation independent of θ_0. It is often the case that Ψ is smooth and so is $\dot{\psi}$, so p is often sufficiently large for this window to be wide.

Condition C6 is usually easy to satisfy. Since $\Pi_{n,\theta_0}(\mathcal{I})$ is a linear combination of $\{\phi_k : k \in \{1,\ldots,K\}\}$ and is an approximation of a highly smooth function $\mathcal{I}$, if the series ϕ_k is smooth, then we can expect that $\Pi_{n,\theta_0}(\mathcal{I})$ will be Lipschitz uniformly over n, that is, that Condition C6 holds. For example, using polynomial series, cubic splines or trigonometric series imply that this condition holds.

Condition C7 imposes Lipschitz continuity conditions on $\dot{\psi}$ and $\Pi_{n,\theta_0}(\dot{\psi})$ uniformly over n. The Lipschitz continuity of $\dot{\psi}$ has been discussed above. As for $\Pi_{n,\theta_0}(\dot{\psi})$, similarly to Condition C6, as long as the series ϕ_k being used is smooth, $\Pi_{n,\theta_0}(\dot{\psi})$ would be Lipschitz continuous uniformly over n.

A.3.2 Theorem 2.5

The conditions are similar to those in Theorem 2.2. However, Condition C4* can be more stringent than Condition C4. For generalized data-adaptive series, the dimension of the argument of the series is larger. Hence, as noted in Chen (2007), C4* may require more smoothness of $\dot{\psi}$ in order that $\dot{\psi}$ can be well approximated by $\Pi_{n,\theta_0}(\dot{\psi})$. However, in general, we do not expect the smoothness of $\dot{\psi}$ to depend on Ψ alone but no components of P_0, so the amount of smoothness of $\dot{\psi}$ may be more limited in practice.

It is also worth noting that, similarly to Theorem 2.2, a sufficient condition for Condition C4* is the following:

Condition C4*s. $\|[\dot{\psi} - \Pi_{n,\theta_0}(\dot{\psi})] \circ (\theta_0, \mathcal{I}_x)\| = o(n^{-1/4})$.

A.4 Lemmas and technical proofs

A.4.1 Highly Adaptive Lasso (HAL)

Proof of Theorem 2.1. Under Conditions A2 and B1–B3, Lemma 1 and its corollary in van Der Laan (2017) show that $\|\hat{\theta}_n - \theta_0\| = o_p(n^{-1/4})$.

We show that the small perturbations of $\hat{\theta}_n$ in certain directions are contained in $\Theta_{\mathrm{v},M}$. Let $\vartheta_\delta = \hat{\theta}_n + \delta(\dot{\Psi} + \theta_0 - \hat{\theta}_n)$ be a path indexed by δ $(0 \le \delta < 1)$ that is a perturbation of $\hat{\theta}_n$. Note that for all δ, ϑ_δ is càdlàg by Condition B1 and we have that

$$\|\vartheta_\delta\|_{\mathrm{v}} = \|(1-\delta)\hat{\theta}_n + \delta(\dot{\Psi} + \theta_0)\|_{\mathrm{v}} \le (1-\delta)\|\hat{\theta}_n\|_{\mathrm{v}} + \delta(\|\dot{\Psi}\|_{\mathrm{v}} + \|\theta_0\|_{\mathrm{v}}) \le (1-\delta)M + \delta M = M$$

by Condition B2. Hence $\vartheta_\delta \in \Theta_{\mathrm{v},M}$. The same result holds for the path $\tilde{\vartheta}_\delta := \hat{\theta}_n + \delta(-\dot{\Psi} + \theta_0 - \hat{\theta}_n)$.

Combining this observation with the P_0-Donkser property of $\Theta_{\mathrm{v},M'}$ for any fixed $M' > 0$ (Gill et al., 1993) and Conditions A1–A2, B4, we have that all of the conditions of Theorem 1 in Shen (1997) are satisfied with all sieves being $\Theta_{\mathrm{v},M}$. The desired asymptotic linearity result follows. The efficiency result is shown in Appendix A.4.3. □

Proof of Lemma 2.1. Recall that $\mathcal{X} \subseteq \mathbb{R}^d$. Similar to $x^{(l)}$, let $x^{(u)} = \inf\{x : P_0(X \le x) = 1\}$ where inf and $\le$ are entrywise. To avoid clumsy notations, in this proof we drop the subscript in θ_0 and use θ instead. This should not introduce confusion because other functions (e.g., an estimator of θ_0) are not involved in the statement or proof. Using the results reviewed in Section 2.3.1,

$$\begin{aligned}\|\dot{\Psi}\|_{\mathrm{v}} &= |\dot{\Psi}(x^{(\ell)})| + \sum_{s\subseteq\{1,2,\ldots,d\},s\neq\emptyset} \int_{x_s^{(\ell)}}^{x_s^{(u)}} |\dot{\Psi}_s(du)| \\ &= |\dot{\Psi}(x^{(\ell)})| + \sum_{s\subseteq\{1,2,\ldots,d\},s\neq\emptyset} \int_{x_s^{(\ell)}}^{x_s^{(u)}} |\dot{\psi}'(z)|\Big|_{z=\theta_s(u)} |\theta_s(du)|.\end{aligned}$$

Since

$$|\theta(x)| = \left|\theta(x^{(\ell)}) + \sum_{s\subseteq\{1,2,\ldots,d\},s\neq\emptyset} \int_{x_s^{(\ell)}}^{x_s} \theta_s(du)\right|$$

$$\leq |\theta(x^{(\ell)})| + \sum_{s\subseteq\{1,2,\ldots,d\},s\neq\emptyset} \int_{x_s^{(\ell)}}^{x_s} |\theta_s(du)|$$
$$\leq |\theta(x^{(\ell)})| + \sum_{s\subseteq\{1,2,\ldots,d\},s\neq\emptyset} \int_{x_s^{(\ell)}}^{x_s^{(u)}} |\theta_s(du)| = \|\theta\|_{\mathrm{v}},$$

we have $|\dot{\psi}'(z)|\Big|_{z=\theta_s(u)} \leq \sup_{z':|z'|\leq\|\theta_0\|_{\mathrm{v}}} |\dot{\psi}'(z')| = B$ for all $x^{(\ell)} \leq u \leq x^{(u)}$, so

$$\|\dot{\Psi}\|_{\mathrm{v}} \leq |\dot{\Psi}(x^{(\ell)})| + \sum_{s\subseteq\{1,2,\ldots,d\},s\neq\emptyset} \int_{x_s^{(\ell)}}^{x_s^{(u)}} B|\theta_s(du)| \leq |\dot{\Psi}(x^{(\ell)})| + B\|\theta_0\|_{\mathrm{v}}.$$

□

Lemma A.1 (CV-selected bound not much smaller than the bound of the true function's variation norm)**.** *Suppose that Condition B1 holds, θ_0 is càdlàg, $\|\theta_0\|_{\mathrm{v}} < \infty$ and for any M, $\sup_{\theta\in\Theta_{\mathrm{v},M}} \|\ell(\theta)\| < \infty$. Let M_n be a (possibly random) sequence such that $P_0\{\ell(\hat{\theta}_{n,M_n}) - \ell(\theta_0)\} = o_p(1)$. Then for any $\epsilon > 0$, with probability tending to one, $M_n \geq \|\theta_0\|_{\mathrm{v}} - \epsilon$. Therefore, for any fixed $\epsilon > 0$, with probability tending to one, $M_n+\epsilon \geq (\|\theta_0\|_{\mathrm{v}}-\epsilon)+\epsilon = \|\theta_0\|_{\mathrm{v}}$.*

Proof of Lemma A.1. We prove by contradiction. Suppose the claim is not true, i.e. there exists $\epsilon, \delta > 0$ such that $P(M_n < \|\theta_0\|_{\mathrm{v}} - \epsilon) \geq \delta$ for all $n \in \mathcal{N}$, where $\mathcal{N}$ is an infinite set. Let $\theta_{0,M} \in \mathrm{argmin}_{\theta\in\Theta_{\mathrm{v},M}} P_0\ell(\theta)$. Then for all $n \in \mathcal{N}$, with probability at least δ,

$$\begin{aligned} P_0\{\ell(\hat{\theta}_{n,M_n}) - \ell(\theta_0)\} &= P_0\{\ell(\hat{\theta}_{n,M_n}) - \ell(\theta_{0,M_n})\} + P_0\{\ell(\theta_{0,M_n}) - \ell(\theta_0)\} \\ &\geq P_0\{\ell(\theta_{0,M_n}) - \ell(\theta_0)\} \\ &\geq P_0\{\ell(\theta_{0,\|\theta_0\|_{\mathrm{v}}-\epsilon}) - \ell(\theta_0)\}, \end{aligned}$$

which is a positive constant since the function class $\Theta_{\|\theta_0\|_{\mathrm{v}}-\epsilon}$ does not contain θ_0 and this term is non-negligible bias. This contradicts the assumption that $P_0\{\ell(\hat{\theta}_{n,M_n}) - \ell(\theta_0)\} = o_p(1)$ and hence the desired follows. □

Therefore, if $\|\dot{\Psi}\|_{\mathrm{v}} \leq F(\|\theta_0\|_{\mathrm{v}})$ for a known increasing function F, then with probability tending to one, $M_n + \epsilon + F(M_n + \epsilon)$ is a valid bound on $\|\hat{\theta}_n\|_{\mathrm{v}}$ that can be used to obtain

an efficient plug-in estimator. Moreover, if the bound is loose, i.e. $\|\dot{\Psi}\|_{\mathrm{v}} < F(\|\theta_0\|_{\mathrm{v}})$, and F is continuous, then there exists some $\epsilon > 0$ such that $\|\dot{\Psi}\|_{\mathrm{v}} \leq F(\|\theta_0\|_{\mathrm{v}} - \epsilon) - \epsilon$ and hence $\|\theta_0\|_{\mathrm{v}} + \|\dot{\Psi}\|_{\mathrm{v}} \leq M_n + F(M_n)$ with probability tending to one.

Note that this lemma only concerns learning a function-valued feature but not estimating $\Psi(\theta_0)$. There are examples where $\dot{\Psi}$ depends on components of P_0, say η_0, other than θ_0. However, if η_0 can be learned via HAL, then Lemma A.1 can be applied. Therefore, if it is known that $\|\dot{\Psi}\|_{\mathrm{v}} \leq F(\|\theta_0\|_{\mathrm{v}}, \|\eta_0\|_{\mathrm{v}})$ for a known increasing function F, then we can use a bound on $\|\hat{\theta}_n\|_{\mathrm{v}}$ obtained in a similar fashion as above from the sequence M_n to construct an efficient plug-in estimator $\Psi(\hat{\theta}_n)$.

Now consider obtaining M_n by k-fold CV from a set of candidate bounds. Then, under Conditions B1–B3, by (i) Lemma 1 and its corollary of van Der Laan (2017), and (ii) the oracle inequality for k-fold CV in van der laan and Dudoit (2003), $P_0\{\ell(\hat{\theta}_{n,M_n}) - \ell(\theta_0)\} = o_p(n^{-1/4})$ if (i) one candidate bound is no smaller than $\|\theta_0\|_{\mathrm{v}}$, and (ii) the number of candidate bounds is fixed. Therefore, the above results apply to this case.

A.4.2 Data-adaptive series estimation

We first present and prove two lemmas that lead to Theorems 2.2 and 2.5.

Lemma A.2 (Convergence rate of the sieve estimator)**.** *Under Conditions C1, C3 and C6,* $\|\pi_n(\theta_0) - \theta_0\| = o_p(n^{-1/4})$. *Under an additional condition C2,* $\|\theta_n^* - \theta_0\| = o_p(n^{-1/4})$.

Proof of Lemma A.2. By triangle inequality, $\|\pi_n(\theta_0) - \theta_0\| \leq \|\theta_0 - \theta_n^0\| + \|\theta_n^0 - \pi_n(\theta_n^0)\| + \|\pi_n(\theta_n^0) - \pi_n(\theta_0)\|$. We bound these three terms separately.

Term 1: By Condition C1, $\|\theta_0 - \theta_n^0\| = o_p(n^{-1/4})$.

Term 2: By the definition of projection operator,

$$\|\theta_n^0 - \pi_n(\theta_n^0)\| = \|\theta_n^0 - \Pi_{n,\theta_n^0}(\mathcal{I}) \circ \theta_n^0\| \leq \|\theta_n^0 - \Pi_{n,\theta_0}(\mathcal{I}) \circ \theta_n^0\|.$$

We bound the right-hand side by showing this term is close to $\|\theta_0 - \Pi_{n,\theta_0}(\mathcal{I}) \circ \theta_0\|$ up to an $o_p(n^{-1/4})$ term. By the reverse triangle inequality and the triangle inequality,

$$\left|\|\theta_n^0 - \Pi_{n,\theta_0}(\mathcal{I}) \circ \theta_n^0\| - \|\theta_0 - \Pi_{n,\theta_0}(\mathcal{I}) \circ \theta_0\|\right|$$

$$
\begin{aligned}
&\leq \|[\theta_n^0 - \Pi_{n,\theta_0}(\mathcal{I}) \circ \theta_n^0] - [\theta_0 - \Pi_{n,\theta_0}(\mathcal{I}) \circ \theta_0]\| \\
&= \|[\theta_n^0 - \theta_0] - [\Pi_{n,\theta_0}(\mathcal{I}) \circ \theta_n^0 - \Pi_{n,\theta_0}(\mathcal{I}) \circ \theta_0]\| \\
&\leq \|\theta_n^0 - \theta_0\| + \|\Pi_{n,\theta_0}(\mathcal{I}) \circ \theta_n^0 - \Pi_{n,\theta_0}(\mathcal{I}) \circ \theta_0\| \\
&\leq \|\theta_n^0 - \theta_0\| + C\|\theta_n^0 - \theta_0\|, \qquad \text{(Condition C6)}
\end{aligned}
$$

which is $o_p(n^{-1/4})$ by Condition C1. Therefore, by Condition C3,

$$\|\theta_n^0 - \pi_n(\theta_n^0)\| \leq \|\theta_n^0 - \Pi_{n,\theta_0}(\mathcal{I}) \circ \theta_n^0\| \leq \|\theta_0 - \Pi_{n,\theta_0}(\mathcal{I}) \circ \theta_0\| + o_p(n^{-1/4}) = o_p(n^{-1/4}).$$

Term 3: By the definition of projection and Condition C1, $\|\pi_n(\theta_n^0) - \pi_n(\theta_0)\| \leq \|\theta_n^0 - \theta_0\| = o_p(n^{-1/4})$.

Conclusion from the three bounds: $\|\pi_n(\theta_0) - \theta_0\| = o_p(n^{-1/4})$.

If, in addition, Condition C2 also holds, then $\|\theta_n^* - \theta_0\| \leq \|\pi_n(\theta_0) - \theta_0\| + \|\theta_n^* - \pi_n(\theta_0)\| = o_p(n^{-1/4})$. □

The same result holds for the generalized data-adaptive series under Conditions C1, C6*, C3* and C2 (if relevant). The proof is almost identical and is therefore omitted.

Lemma A.3 (Approximation error to $\dot{\psi}$)**.** *Under Condition C7,* $\|\dot{\psi} \circ \theta_0 - \pi_n(\dot{\psi} \circ \theta_0)\| \leq C\|\theta_n^0 - \theta_0\| + \|\dot{\psi} \circ \theta_0 - \Pi_{n,\theta_0}(\dot{\psi}) \circ \theta_0\|$*. Therefore, under Conditions C1–C4,* $\|\dot{\psi} \circ \theta_0 - \pi_n(\dot{\psi} \circ \theta_0)\| \cdot \|\theta_n^* - \theta_0\| = o_p(n^{-1/2})$.

Proof of Lemma A.3. By the definition of the projection operator and triangle inequality,

$$\|\dot{\psi} \circ \theta_0 - \pi_n(\dot{\psi} \circ \theta_0)\| \leq \|\dot{\psi} \circ \theta_0 - \pi_n(\dot{\psi} \circ \theta_n^0)\| \leq \|\dot{\psi} \circ \theta_0 - \dot{\psi} \circ \theta_n^0\| + \|\dot{\psi} \circ \theta_n^0 - \pi_n(\dot{\psi} \circ \theta_n^0)\|.$$

We bound the two terms on the right-hand side separately.

Term 1: By Condition C7, $\|\dot{\psi} \circ \theta_0 - \dot{\psi} \circ \theta_n^0\| \leq C\|\theta_0 - \theta_n^0\|$.

Term 2: This term can be bounded similarly as in Lemma A.2. By the reverse triangle inequality and the triangle inequality,

$$\left| \|\dot{\psi} \circ \theta_n^0 - \Pi_{n,\theta_0}(\dot{\psi}) \circ \theta_n^0\| - \|\dot{\psi} \circ \theta_0 - \Pi_{n,\theta_0}(\mathcal{I}) \circ \theta_0\| \right|$$

$$
\begin{aligned}
&\leq \|[\dot{\psi} \circ \theta_n^0 - \Pi_{n,\theta_0}(\dot{\psi}) \circ \theta_n^0] - [\dot{\psi} \circ \theta_0 - \Pi_{n,\theta_0}(\dot{\psi}) \circ \theta_0]\| \\
&= \|[\dot{\psi} \circ \theta_n^0 - \dot{\psi} \circ \theta_0] - [\Pi_{n,\theta_0}(\dot{\psi}) \circ \theta_n^0 - \Pi_{n,\theta_0}(\dot{\psi}) \circ \theta_0]\| \\
&\leq \|\dot{\psi} \circ \theta_n^0 - \dot{\psi} \circ \theta_0\| + \|\Pi_{n,\theta_0}(\dot{\psi}) \circ \theta_n^0 - \Pi_{n,\theta_0}(\dot{\psi}) \circ \theta_0\| \\
&\leq C\|\theta_n^0 - \theta_0\| + C\|\theta_n^0 - \theta_0\| \qquad \text{(Condition C7)} \\
&= C\|\theta_n^0 - \theta_0\|.
\end{aligned}
$$

Therefore, by the definition of the projection operator and Condition C7,

$$
\begin{aligned}
\|\dot{\psi} \circ \theta_n^0 - \pi_n(\dot{\psi} \circ \theta_n^0)\| &\leq \|\dot{\psi} \circ \theta_n^0 - \Pi_{n,\theta_0}(\dot{\psi}) \circ \theta_n^0\| \\
&\leq \|\dot{\psi} \circ \theta_0 - \Pi_{n,\theta_0}(\dot{\psi}) \circ \theta_0\| + C\|\theta_n^0 - \theta_0\|.
\end{aligned}
$$

Conclusion from the two bounds: $\|\dot{\psi}\circ\theta_0 - \pi_n(\dot{\psi}\circ\theta_0)\| \leq C\|\theta_n^0 - \theta_0\| + \|\dot{\psi}\circ\theta_0 - \Pi_{n,\theta_0}(\dot{\psi})\circ\theta_0\|$.

Under Conditions C1–C4, using Lemma A.2, it follows that $\|\dot{\psi}\circ\theta_0 - \pi_n(\dot{\psi}\circ\theta_0)\| \cdot \|\theta_n^* - \theta_0\| = o_p(n^{-1/2})$. □

Note that π_n is a linear operator. Lemma A.2 and A.3 along with other conditions essentially satisfy the assumptions in Corollary 2 in Shen (1997). We can prove the asymptotic linearity result of Theorem 2.2 similarly to this result as follows.

Proof of Theorem 2.2. We note that

$$
\begin{aligned}
P_n\ell(\theta_n^*) &= P_n\ell(\theta_0) + P_0[\ell(\theta_n^*) - \ell(\theta_0)] + (P_n - P_0)[\ell(\theta_n^*) - \ell(\theta_0)] \\
&= P_n\ell(\theta_0) + P_0[\ell(\theta_n^*) - \ell(\theta_0)] + (P_n - P_0)\ell_0'[\theta_n^* - \theta_0] \\
&\quad + (P_n - P_0)r[\theta_n^* - \theta_0].
\end{aligned}
$$

Let ϵ_n be an arbitrary sequence of positive real numbers that is $o(n^{-1/2})$. We may replace θ_n^* with $\pi_n((1-\epsilon_n)\theta_n^* + \epsilon_n(\theta_0 \pm \dot{\Psi}))$ in the above equation. We first consider $\pi_n((1-\epsilon_n)\theta_n^* +$

$\epsilon_n(\theta_0 + \dot{\Psi})$):

$$\begin{aligned} & P_n\ell\left(\pi_n((1-\epsilon_n)\theta_n^* + \epsilon_n(\theta_0+\dot{\Psi}))\right) \\ &= P_n\ell(\theta_0) + P_0[\ell(\pi_n((1-\epsilon_n)\theta_n^* + \epsilon_n(\theta_0+\dot{\Psi}))) - \ell(\theta_0)] \\ &\quad + (P_n - P_0)\ell_0'[\pi_n((1-\epsilon_n)\theta_n^* + \epsilon_n(\theta_0+\dot{\Psi})) - \theta_0] \\ &\quad + (P_n - P_0)r[\pi_n((1-\epsilon_n)\theta_n^* + \epsilon_n(\theta_0+\dot{\Psi})) - \theta_0]. \end{aligned} \tag{A.1}$$

Take the difference between the above two equations. By the linearity of ℓ_0' and π_n, we have that

$$\begin{aligned} & P_n\ell\left(\pi_n((1-\epsilon_n)\theta_n^* + \epsilon_n(\theta_0+\dot{\Psi}))\right) - P_n\ell(\theta_n^*) \\ &= P_0[\ell(\pi_n((1-\epsilon_n)\theta_n^* + \epsilon_n(\theta_0+\dot{\Psi}))) - \ell(\theta_0)] - P_0[\ell(\theta_n^*) - \ell(\theta_0)] \\ &\quad + (P_n - P_0)\ell_0'[\pi_n((1-\epsilon_n)\theta_n^* + \epsilon_n(\theta_0+\dot{\Psi})) - \theta_n^*] \\ &\quad + (P_n - P_0)\{r[\pi_n((1-\epsilon_n)\theta_n^* + \epsilon_n(\theta_0+\dot{\Psi})) - \theta_0] - r[\theta_n^* - \theta_0]\}. \end{aligned}$$

We next analyze the three lines on the right-hand side of the above equation separately.

Line 1: Under Condition A2,

$$\begin{aligned} & P_0[\ell(\pi_n((1-\epsilon_n)\theta_n^* + \epsilon_n(\theta_0+\dot{\Psi}))) - \ell(\theta_0)] - P_0[\ell(\theta_n^*) - \ell(\theta_0)] \\ &= \frac{\alpha_{0,\ell}}{2}\|\pi_n((1-\epsilon_n)\theta_n^* + \epsilon_n(\theta_0+\dot{\Psi})) - \theta_0\|^2 - \frac{\alpha_{0,\ell}}{2}\|\theta_n^* - \theta_0\|^2 \\ &\quad + o_p\left(\|\pi_n((1-\epsilon_n)\theta_n^* + \epsilon_n(\theta_0+\dot{\Psi})) - \theta_0\|^2 + \|\theta_n^* - \theta_0\|^2\right) \end{aligned}$$

We subtract and add $(1-\epsilon_n)\theta_n^* + \epsilon_n(\theta_0 + \dot{\Psi})$ in the first term. By the fact that π_n is linear and $\pi_n(\theta_n^*) = \theta_n^*$, the display continues as

$$\begin{aligned} &= \frac{\alpha_{0,\ell}}{2}\|\{\pi_n((1-\epsilon_n)\theta_n^* + \epsilon_n(\theta_0+\dot{\Psi})) - ((1-\epsilon_n)\theta_n^* + \epsilon_n(\theta_0+\dot{\Psi}))\} \\ &\quad + \{(1-\epsilon_n)\theta_n^* + \epsilon_n(\theta_0+\dot{\Psi}) - \theta_0\}\|^2 \\ &\quad - \frac{\alpha_{0,\ell}}{2}\|\theta_n^* - \theta_0\|^2 + o_p\left(\|\pi_n((1-\epsilon_n)\theta_n^* + \epsilon_n(\theta_0+\dot{\Psi})) - \theta_0\|^2 + \|\theta_n^* - \theta_0\|^2\right) \\ &= \frac{\alpha_{0,\ell}}{2}\|\epsilon_n\{\pi_n(\theta_0+\dot{\Psi}) - (\theta_0+\dot{\Psi})\} + (\theta_n^* - \theta_0) + \epsilon_n(\dot{\Psi} + \theta_0 - \theta_n^*)\|^2 - \frac{\alpha_{0,\ell}}{2}\|\theta_n^* - \theta_0\|^2 \end{aligned}$$

$$
\begin{aligned}
&+ o_p\left(\|\pi_n((1-\epsilon_n)\theta_n^* + \epsilon_n(\theta_0 + \dot{\Psi})) - \theta_0\|^2 + \|\theta_n^* - \theta_0\|^2\right) \\
&= \epsilon_n \alpha_{0,\ell}\langle \theta_n^* - \theta_0, \dot{\Psi}\rangle + \epsilon_n^2 \frac{\alpha_{0,\ell}}{2}\|\pi_n(\theta_0 + \dot{\Psi}) - \theta_n^*\|^2 \\
&+ \epsilon_n \alpha_{0,\ell}\langle \pi_n(\theta_0) - \theta_0, \theta_n^* - \theta_0\rangle + \epsilon_n \alpha_{0,\ell}\langle \pi_n(\dot{\Psi}) - \dot{\Psi}, \theta_n^* - \theta_0\rangle - \epsilon_n \alpha_{0,\ell}\|\theta_n^* - \theta_0\|^2 \\
&+ o_p\left(\|\pi_n((1-\epsilon_n)\theta_n^* + \epsilon_n(\theta_0 + \dot{\Psi})) - \theta_0\|^2 + \|\theta_n^* - \theta_0\|^2\right)
\end{aligned}
$$

By Cauchy-Schwards inequality, the display continues as

$$
\begin{aligned}
&\leq \epsilon_n \alpha_{0,\ell}\langle \theta_n^* - \theta_0, \dot{\Psi}\rangle + \epsilon_n^2 \frac{\alpha_{0,\ell}}{2}\|\pi_n(\theta_0 + \dot{\Psi}) - \theta_n^*\|^2 \\
&+ \epsilon_n \alpha_{0,\ell}\|\pi_n(\theta_0) - \theta_0\|\|\theta_n^* - \theta_0\| + \epsilon_n \alpha_{0,\ell}\|\pi_n(\dot{\Psi}) - \dot{\Psi}\|\|\theta_n^* - \theta_0\| - \epsilon_n \alpha_{0,\ell}\|\theta_n^* - \theta_0\|^2 \\
&+ o_p\left(\|\pi_n((1-\epsilon_n)\theta_n^* + \epsilon_n(\theta_0 + \dot{\Psi})) - \theta_0\|^2 + \|\theta_n^* - \theta_0\|^2\right)
\end{aligned}
$$

By Lemmas A.2–A.3 and the assumption that $\epsilon_n = o(n^{-1/2})$, the display continues as

$$
= \epsilon_n \alpha_{0,\ell}\langle \theta_n^* - \theta_0, \dot{\Psi}\rangle + \epsilon_n o_p(n^{-1/2}) + o_p\left(\|\pi_n((1-\epsilon_n)\theta_n^* + \epsilon_n(\theta_0 + \dot{\Psi})) - \theta_0\|^2 + \|\theta_n^* - \theta_0\|^2\right).
$$

Line 2: We subtract and add $(1-\epsilon_n)\theta_n^* + \epsilon_n(\theta_0 + \dot{\Psi})$. By linearity of ℓ_0', Condition C8, and the fact that $\pi_n(\theta_n^*) = \theta_n^*$, we have that

$$
\begin{aligned}
&(P_n - P_0)\ell_0'[\pi_n((1-\epsilon_n)\theta_n^* + \epsilon_n(\theta_0 + \dot{\Psi})) - \theta_n^*] \\
&= (P_n - P_0)\ell_0'[(1-\epsilon_n)\theta_n^* + \epsilon_n(\theta_0 + \dot{\Psi}) - \theta_n^*] \\
&\quad + \epsilon_n(P_n - P_0)\ell_0'[\pi_n(\theta_0 + \dot{\Psi}) - (\theta_0 + \dot{\Psi})] \\
&= \epsilon_n(P_n - P_0)\ell_0'[\dot{\Psi}] - \epsilon_n(P_n - P_0)\ell_0'[\theta_n^* - \theta_0] + \epsilon_n o_p(n^{-1/2}) \\
&= \epsilon_n(P_n - P_0)\ell_0'[\dot{\Psi}] + \epsilon_n o_p(n^{-1/2}).
\end{aligned}
$$

Line 3: By Condition C8, this term is $\epsilon_n o_p(n^{-1/2})$.

Conclusion of the three lines: It holds that

$$
\begin{aligned}
&P_n\ell\left(\pi_n((1-\epsilon_n)\theta_n^* + \epsilon_n(\theta_0 + \dot{\Psi}))\right) - P_n\ell(\theta_n^*) \\
&\leq \epsilon_n \alpha_{0,\ell}\langle \theta_n^* - \theta_0, \dot{\Psi}\rangle + \epsilon_n(P_n - P_0)\ell_0'[\dot{\Psi}]
\end{aligned}
$$

$$+ \epsilon_n o_p(n^{-1/2}) + o_p\left(\|\pi_n((1-\epsilon_n)\theta_n^* + \epsilon_n(\theta_0 + \dot{\Psi})) - \theta_0\|^2 + \|\theta_n^* - \theta_0\|^2\right).$$

Since θ_n^* is an empirical risk minimizer, the left-hand side is non-negative. Thus,

$$0 \leq \langle \theta_n^* - \theta_0, \dot{\Psi} \rangle + (P_n - P_0)\alpha_{0,\ell}^{-1}\ell_0'[\dot{\Psi}] + o_p(n^{-1/2}).$$

Similarly, by replacing $\pi_n((1-\epsilon_n)\theta_n^* + \epsilon_n(\theta_0 + \dot{\Psi}))$ with $\pi_n((1-\epsilon_n)\theta_n^* + \epsilon_n(\theta_0 - \dot{\Psi}))$ in (A.1), we derive that

$$0 \leq -\langle \theta_n^* - \theta_0, \dot{\Psi} \rangle - (P_n - P_0)\alpha_{0,\ell}^{-1}\ell_0'[\dot{\Psi}] + o_p(n^{-1/2}).$$

Therefore, $|\langle \theta_n^* - \theta_0, \dot{\Psi} \rangle + (P_n - P_0)\alpha_{0,\ell}^{-1}\ell_0'[\dot{\Psi}]| = o_p(n^{-1/2})$. By Conditions A3–A4 and Lemma A.2,

$$\begin{aligned}\Psi(\theta_n^*) - \Psi(\theta_0) &= \langle \theta_n^* - \theta_0, \dot{\Psi} \rangle + o_p(n^{-1/2}) \\ &= -(P_n - P_0)\alpha_{0,\ell}^{-1}\ell_0'[\dot{\Psi}] + o_p(n^{-1/2}).\end{aligned}$$

The asymptotic linearity of $\Psi(\theta_n^*)$ follows. We prove the efficiency in Appendix A.4.3. □

The proof of Theorem 2.5 is almost identical.

Nest we present and prove a lemma allows us to interpret Condition C2 as an upper bound on the rate of K.

Lemma A.4. *Under Conditions A2, C1, C3 (C3* resp.) and C6 (C6* resp.),* $\|\pi_n(\theta_0) - \theta_n^\dagger\| = o_p(n^{-1/4})$.

Proof of Lemma A.4. By definition of $\theta_n^\dagger$ and Condition A2, we have

$$\|\theta_n^\dagger - \theta_0\|^2 \leq CP_0\{\ell(\theta_n^\dagger) - \ell(\theta_0)\} \leq CP_0\{\ell(\pi_n(\theta_0)) - \ell(\theta_0)\} \leq C\|\pi_n(\theta_0) - \theta_0\|^2,$$

the right-hand side of which is $o_p(n^{-1/2})$ by Lemma A.2 (or its corresponding version under Conditions C6* and C3*). Therefore, $\|\theta_n^\dagger - \theta_0\| = o_p(n^{-1/4})$ and hence $\|\pi_n(\theta_0) - \theta_n^\dagger\| \leq \|\pi_n(\theta_0) - \theta_0\| + \|\theta_n^\dagger - \theta_0\| = o_p(n^{-1/4})$. □

We finally prove the efficiency of the data-adaptive series estimator with K selected by CV.

Proof of Theorem 2.4. By Lemma A.2 and Condition A2, for that existing deterministic K, $P_0\{\ell(\theta^*_K(\theta^0_n)) - \ell(\theta_0)\} \leq C\|\theta^*_K(\theta^0_n) - \theta_0\|^2 = o_p(n^{-1/2})$. By the oracle inequality for CV in van der laan and Dudoit (2003), $P_0\{\ell(\theta^\sharp_n) - \ell(\theta_0)\} = o_p(n^{-1/2})$. By Condition A2, $\|\theta^\sharp_n - \theta_0\|^2 \leq CP_0\{\ell(\theta^\sharp_n) - \ell(\theta_0)\} = o_p(n^{-1/2})$ and hence $\|\theta^\sharp_n - \theta_0\| = o_p(n^{-1/4})$. So with probability tending to one,

$$\begin{aligned}
\|\dot{\psi} \circ \theta^0_n - \pi_{K^*,\theta^0_n}(\dot{\psi} \circ \theta^0_n)\| &= \|\dot{\psi} \circ \theta^0_n - \Pi_{K^*,\theta^0_n}(\dot{\psi}) \circ \theta^0_n\| \\
&\leq C\|\theta^0_n - \Pi_{K^*,\theta^0_n}(\mathcal{I}) \circ \theta^0_n\| && \text{(Condition C5)} \\
&\leq C\|\theta^0_n - \theta^\sharp_n\| && \text{(definition of the projection operator)} \\
&\leq C(\|\theta^0_n - \theta_0\| + \|\theta^\sharp_n - \theta_0\|), && \text{(triangle inequality)}
\end{aligned}$$

which is $o_p(n^{-1/4})$ by Condition C1. Hence,

$$\begin{aligned}
\|\dot{\psi} \circ \theta_0 - \pi_{K^*,\theta^0_n}(\dot{\psi} \circ \theta_0)\| &\leq \|\dot{\psi} \circ \theta_0 - \pi_{K^*,\theta^0_n}(\dot{\psi} \circ \theta^0_n)\| \\
&\leq \|\dot{\psi} \circ \theta_0 - \dot{\psi} \circ \theta^0_n\| + \|\dot{\psi} \circ \theta^0_n - \pi_{K^*,\theta^0_n}(\dot{\psi} \circ \theta^0_n)\| \\
&\leq C\|\theta^0_n - \theta_0\| + o_p(n^{-1/4}), && \text{(Condition C7)}
\end{aligned}$$

which is $o_p(n^{-1/4})$ by Condition C1.

This bounds the approximation error $\|\dot{\psi} \circ \theta_0 - \pi_{K^*,\theta^0_n}(\dot{\psi} \circ \theta_0)\|$ for $\dot{\psi}$, a result that is similar to Lemma A.3 combined with Conditions C1 and C4*s. Similarly to Theorem 2.2, along with other conditions, the assumptions in Corollary 2 in Shen (1997) are essentially satisfied and hence an almost identical argument shows that $\Psi(\theta^\sharp_n)$ is an asymptotically linear estimator of $\Psi(\theta_0)$. We prove the efficiency in Appendix A.4.3. □

A.4.3 Efficiency

Proof of efficiency of the proposed estimators. It is sufficient to show that the influence function of our proposed estimators is the canonical gradient under a nonparametric model. Let

$H \in \mathcal{H}$ be fixed. In the rest of this proof, for all small o and big O notations, we let $\delta \to 0$. The proof is similar to the proof of asymptotic linearity in Shen (1997) except that the estimator of θ_0 and the empirical distribution P_n are replaced by $\theta_{H,\delta}$ and $P_{H,\delta}$ respectively.

Let δ' satisfy Condition E2. We note that

$$\begin{aligned} P_{H,\delta}\ell(\theta_{H,\delta}) &= P_{H,\delta}\ell(\theta_0) + P_0[\ell(\theta_{H,\delta}) - \ell(\theta_0)] + (P_{H,\delta} - P_0)[\ell(\theta_{H,\delta}) - \ell(\theta_0)] \\ &= P_{H,\delta}\ell(\theta_0) + P_0[\ell(\theta_{H,\delta}) - \ell(\theta_0)] + (P_{H,\delta} - P_0)\ell_0'[\theta_{H,\delta} - \theta_0] \\ &\quad + (P_{H,\delta} - P_0)r[\theta_{H,\delta} - \theta_0]. \end{aligned}$$

We also note that $(1-\delta')\theta_{H,\delta} + \delta'(\theta_0 \pm \dot{\Psi}) \in \Theta$ if $|\delta|$ is sufficiently small. Then, similarly, by replacing $\theta_{H,\delta}$ with $(1-\delta')\theta_{H,\delta} + \delta'(\theta_0 + \dot{\Psi})$ in the above equation, we have that

$$\begin{aligned} &P_{H,\delta}\ell((1-\delta')\theta_{H,\delta} + \delta'(\theta_0 + \dot{\Psi})) \\ &= P_{H,\delta}\ell(\theta_0) + P_0[\ell((1-\delta')\theta_{H,\delta} + \delta'(\theta_0 + \dot{\Psi})) - \ell(\theta_0)] \\ &\quad + (P_{H,\delta} - P_0)\ell_0'[(1-\delta')\theta_{H,\delta} + \delta'(\theta_0 + \dot{\Psi}) - \theta_0] \\ &\quad + (P_{H,\delta} - P_0)r[(1-\delta')(\theta_{H,\delta} - \theta_0) + \delta'\dot{\Psi}]. \end{aligned} \tag{A.2}$$

Take the difference between the above two equations. By the linearity of ℓ_0', we have that

$$\begin{aligned} &P_{H,\delta}\ell((1-\delta')\theta_{H,\delta} + \delta'(\theta_0 + \dot{\Psi})) - P_{H,\delta}\ell(\theta_{H,\delta}) \\ &= P_0[\ell((1-\delta')\theta_{H,\delta} + \delta'(\theta_0 + \dot{\Psi})) - \ell(\theta_0)] - P_0[\ell(\theta_{H,\delta}) - \ell(\theta_0)] \\ &\quad + \delta'(P_{H,\delta} - P_0)\ell_0'[\dot{\Psi} - \theta_{H,\delta} + \theta_0] \\ &\quad + (P_{H,\delta} - P_0)\{r[(1-\delta')(\theta_{H,\delta} - \theta_0) + \delta'\dot{\Psi}] - r[\theta_{H,\delta} - \theta_0]\} \\ &= \frac{\alpha_{0,\ell}}{2}\|(1-\delta')(\theta_{H,\delta} - \theta_0) + \delta'\dot{\Psi}\|^2 - \frac{\alpha_{0,\ell}}{2}\|\theta_{H,\delta} - \theta_0\|^2 \\ &\quad + o\left(\|\theta_{H,\delta} - \theta_0\|^2 + \|(1-\delta')(\theta_{H,\delta} - \theta_0) + \delta'\dot{\Psi}\|^2\right) && \text{(Condition A2)} \\ &\quad + \delta'(P_{H,\delta} - P_0)\ell_0'[\dot{\Psi}] - \delta'(P_{H,\delta} - P_0)\ell_0'[\theta_{H,\delta} - \theta_0] + \delta' o(\delta) && \text{(Condition E2)} \\ &= \delta'\alpha_{0,\ell}\langle\theta_{H,\delta} - \theta_0, \dot{\Psi}\rangle - \delta'\alpha_{0,\ell}\|\theta_{H,\delta} - \theta_0\|^2 + \delta'^2\frac{\alpha_{0,\ell}}{2}\|\theta_{H,\delta} - \theta_0 + \dot{\Psi}\|^2 \\ &\quad + o\left(\|\theta_{H,\delta} - \theta_0\|^2 + \|(1-\delta')(\theta_{H,\delta} - \theta_0) + \delta'\dot{\Psi}\|^2\right) \\ &\quad + \delta'(P_{H,\delta} - P_0)\ell_0'[\dot{\Psi}] + \delta' o(\delta) \end{aligned}$$

$$\leq \delta' \alpha_{0,\ell} \langle \theta_{H,\delta} - \theta_0, \dot{\Psi} \rangle + \delta'^2 \frac{\alpha_{0,\ell}}{2} \|\theta_{H,\delta} - \theta_0 + \dot{\Psi}\|^2$$
$$+ o\left(\|\theta_{H,\delta} - \theta_0\|^2 + \|(1-\delta')(\theta_{H,\delta} - \theta_0) + \delta' \dot{\Psi}\|^2 \right)$$
$$+ \delta'(P_{H,\delta} - P_0)\ell_0'[\dot{\Psi}] + \delta' o(\delta).$$

Since the left-hand side of the above display is nonnegative, by Condition E1, we have that

$$\begin{aligned} 0 &\leq \langle \theta_{H,\delta} - \theta_0, \dot{\Psi} \rangle + \alpha_{0,\ell}^{-1}(P_{H,\delta} - P_0)\ell_0'[\dot{\Psi}] + O(\delta') + o(\delta) \\ &= \langle \theta_{H,\delta} - \theta_0, \dot{\Psi} \rangle + \alpha_{0,\ell}^{-1}(P_{H,\delta} - P_0)\ell_0'[\dot{\Psi}] + o(\delta). \end{aligned}$$

Similarly, by replacing $(1-\delta')\theta_{H,\delta} + \delta'(\theta_0 + \dot{\Psi})$ with $(1-\delta')\theta_{H,\delta} + \delta'(\theta_0 - \dot{\Psi})$ in (A.2), we show that $0 \leq -\langle \theta_{H,\delta} - \theta_0, \dot{\Psi} \rangle - \alpha_{0,\ell}^{-1}(P_{H,\delta} - P_0)\ell_0'[\dot{\Psi}] + o(\delta)$. Therefore, $|\langle \theta_{H,\delta} + \theta_0, \dot{\Psi} \rangle + \alpha_{0,\ell}^{-1}(P_{H,\delta} - P_0)\ell_0'[\dot{\Psi}]| = o(\delta)$ and

$$\begin{aligned} \Psi(\theta_{H,\delta}) - \Psi(\theta_0) &= \langle \theta_{H,\delta} - \theta_0, \dot{\Psi} \rangle + O(\|\theta_{H,\delta} - \theta_0\|^2) \\ &= -\alpha_{0,\ell}^{-1}(P_{H,\delta} - P_0)\ell_0'[\dot{\Psi}] + o(\delta) + O(\|\theta_{H,\delta} - \theta_0\|^2) \\ &= -\alpha_{0,\ell}^{-1}(P_{H,\delta} - P_0)\ell_0'[\dot{\Psi}] + o(\delta). \qquad \text{(Condition E1)} \end{aligned}$$

Consequently, $\lim_{\delta \to 0}[\Psi(\theta_{H,\delta}) - \Psi(\theta_0)]/\delta = P_0\{-\alpha_{0,\ell}^{-1}\ell_0'[\dot{\Psi}] \cdot H\}$ and hence the canonical gradient of Ψ under a nonparametric model is $\alpha_{0,\ell}^{-1}\{-\ell_0'[\dot{\Psi}] + P_0\ell_0'[\dot{\Psi}]\}$. Since the influence functions of our asymptotically linear estimators are equal to this canonical gradient, our proposed estimators are efficient under a nonparametric model. □

A.5 Simulation setting details

In all simulations, since $\theta_0(x) = \mathrm{E}_{P_0}[Z|X = x]$ is the conditional mean function, the loss function was chosen to be the square loss $\ell(\theta) : v \mapsto (z - \theta(x))^2$.

A.5.1 HAL

In the simulation, we generate data from the distribution defined by

$$X \sim \mathrm{N}(0,1),\ \theta_0(x) = \exp\{-(-1 + 2x + 2x^2)/2\}, Z|X = x \sim \text{Exponential}(\text{rate} = 1/\theta_0(x)).$$

The sample sizes being considered are 500, 1000, 2000, 5000 and 10000. For each scenario we run 1000 replicates. We chose M.gcv+ to be 3.1 times M.cv.

A.5.2 Data-adaptive series

Demonstration of Theorem 2.4

In the simulation, we generate data from the distribution defined by $X \sim \text{Unif}(-1, 1)$, $Z|X = x \sim \text{N}(\theta_0(x), 0.25^2)$ where

$$\begin{aligned}\theta_0 : x \mapsto\ & I(-1 \le x < -3/4) + \pi I(-3/4 \le x < -1/2) + 10x^2 I(-1/4 \le x < 1/4) \\ & + \sqrt{2} I(1/4 \le x < 1/2) + \exp(-1) I(1/2 \le x < 3/4) + \sqrt[3]{3} I(3/4 \le x \le 1),\end{aligned}$$

When using the trigonometric series, we first shift and scale the initial function range to be $[-1/2, 1/2]$ and then use the basis for the interval $[-1, 1]$ (i.e. $\sin(j\pi z), \cos(j\pi z)$) in sieve estimation to avoid the poor behavior of trigonometric series near the boundary. We consider sample sizes 500, 1000, 2000, 5000, 10000 and 20000. For each sample size, we run 1000 simulations.

Violation of Condition C5

In the simulation, we generate data from the distribution defined by $X \sim \text{Unif}(-1, 1)$, $Z|X = x \sim \text{N}(\theta_0(x), 1)$ where $\theta_0 : x \mapsto \cos(10x)$. The estimand is $\Psi(\theta_0) = P_0(f \circ \theta_0)$ where

$$\begin{aligned}f : z \mapsto\ & \left[\frac{3}{10\pi}\cos(5\pi z) - \frac{3}{8}\right] I\left(z < -\frac{1}{2}\right) - \frac{3}{2} z^2 I\left(-\frac{1}{2} \le z < 0\right) \\ & + 3z^2 I\left(0 \le z < \frac{1}{2}\right) + \left[-\frac{3}{2}\exp(2 - 4z) - 3z + \frac{15}{4}\right] I\left(z \ge \frac{1}{2}\right).\end{aligned}$$

We consider sample sizes 500, 1000, 2000, 5000, 10000 and 20000; for each sample size, we run 1000 simulations. Our goal is to explore the behavior of the plug-in estimator when f, instead of θ_0, is rough, so we use kernel regression (Nadaraya, 1964) to estimate θ_0 for convenience.

A.5.3 Generalized data-adaptive series

Demonstration of Theorem 2.6

In the simulation, we generate data from the distribution defined by $X \sim \text{Unif}(-1, 1)$, $A|X = x \sim \text{Bern}(\text{expit}(-x))$, $Y|A = a, X = x \sim \text{N}(\mu_{0,a}(x), 0.25^2)$ where

$$\begin{aligned}\mu_{00} : x \mapsto\ & I(-1 \le x < -3/4) + \pi I(-3/4 \le x < -1/2) + 10x^2 I(-1/4 \le x < 1/4) \\ & + \sqrt{2} I(1/4 \le x < 1/2) + \exp(-1) I(1/2 \le x < 3/4) + \sqrt[3]{3} I(3/4 \le x \le 1), \\ \mu_{01} : x \mapsto\ & x^2 I(x < -1/3) + \exp(x) I(-1/3 \le x < 1/3) + I(x > 1/3)\end{aligned}$$

The series is the tensor product (Chen, 2007) of univariate trigonometric series in A.5.2. The sample sizes are the same as in A.5.2.

*Violation of Condition C5**

In the simulation, we generate data from the distribution defined by $X \sim \text{Unif}(-1, 1)$, $A|X = x \sim \text{Bern}(g_0(x))$, $Y|A = a, X = x \sim N(\mu_{0,a}(x), 0.25^2)$ where $\mu_{0a} : x \mapsto \exp(-x^2 + 0.8ax + 0.5a)$ $(a \in \{0, 1\})$ and

$$\begin{aligned}g_0 : x \mapsto \text{expit}\Bigg\{ & \left(-\frac{5}{3}x^3 - \frac{15}{4}x^2 - \frac{5}{3}x - \frac{25}{96}\right) I\left(x \le -\frac{1}{2}\right) + \left(\frac{5}{6}x^4 + \frac{5}{3}x^3\right) I\left(-\frac{1}{2} < x \le 0\right) \\ & + \frac{5}{3}x^3 I\left(0 < x \le \frac{1}{2}\right) + \left(5x^2 - \frac{15}{4}x + \frac{5}{6}\right) I\left(x > \frac{1}{2}\right) \Bigg\}.\end{aligned}$$

We consider sample sizes 500, 1000, 2000, 5000, 10000 and 20000; for each sample size, we run 1000 simulations. Our goal is to explore the behavior of the plug-in estimator when $\dot{\Psi}$, instead of θ_0, is rough, so we use kernel regression (Nadaraya, 1964) to estimate θ_0 for convenience.

Appendix B

SUPPORTING INFORMATION FOR CHAPTER 3

As in Chapter 3, when stating conditions/results on IER, we mainly focus on the first setting and often write the corresponding statement for the second setting in brackets.

B.1 Additional technical conditions for asymptotic linearity of proposed estimators

Here we list the additional technical conditions in Sections 3.4.1 and 3.4.2 that we omit in the main text.

B.1.1 Case I: optimal individualized treatment rules

Condition B6 (Sufficient rates for nuisance estimators)**.** $\|\mu_n^Z - \mu_0^Z\|_{2,P_0} = \mathrm{o}_p(1)$ and the following term is $\mathrm{o}_p(n^{-1/2})$:

$$\Big(\|\mu_n^Z - \mu_0^Z\|_{2,P_0} + \|\mu_n^A - \mu_0^A\|_{2,P_0} + \|\mu_n^Y - \mu_0^Y\|_{2,P_0} + \|\hat{\mu}_n^A - \mu_0^A\|_{2,P_0} + \|\hat{\mu}_n^Y - \mu_0^Y\|_{2,P_0}\Big) \times \Big(\|\mu_n^A - \mu_0^A\|_{2,P_0} + \|\mu_n^Y - \mu_0^Y\|_{2,P_0} + \|\hat{\mu}_n^A - \mu_0^A\|_{2,P_0} + \|\hat{\mu}_n^Y - \mu_0^Y\|_{2,P_0}\Big).$$

Condition B7 (Asymptotically uniformly bounded conditional ATE estimators)**.** There exists a constant $C_\Delta > 0$ such that with probability tending to one, $|\Delta_n(w)| \leq C_\Delta$ and $|\hat{\Delta}_n(w)| \leq C_\Delta$ for all $w \in \mathcal{W}$.

This condition would hold if (1) $Y \sim P_0$ has a bounded support, or (ii) Δ_0 is a bounded function and both of Δ_n and $\hat{\Delta}_n$ are consistent for Δ_0 in the $L^\infty(P_0)$ sense.

Let $D_{r,n}^T(o) := D^T(\hat{P}_n, \mu_n^Y, \mu_n^A, t_n, \tau_0^T)(o) - D^T(\hat{P}_n, \mu_n^Y, \mu_n^A, t_r, 0)(o)$.

Condition B8 (Consistency of estimated influence function)**.** $\|D_{r,n}^T - D_r^T(P_0)\|_{2,P_0} = o_P(1)$.

Condition B9 (Donsker condition)**.** There exists a fixed P_0-Donsker class $\mathcal{D}^T$ such that $D^T_{r,n}$ belongs to $\mathcal{D}^T$ with probability tending to 1.

Condition B10 (Glivenko-Cantelli condition)**.** $\|\Delta_{b,n} - \Delta_{b,0}\|_{1,P_0} = \mathrm{o}_p(1)$. Moreover, if, on the one hand, $\kappa < 1$, then, for any η sufficiently close to η_0^T, $v \mapsto I(\Delta_{b,n}(v) > \eta)$ belongs to a fixed P_0-Glivenko-Cantelli class with probability tending to 1. If, on the other hand, $\kappa = 1$, then, for any $\eta < 0$ with sufficiently large $|\eta|$, $v \mapsto I(\Delta_{b,n}(v) > \eta)$ belongs to a fixed P_0-Glivenko-Cantelli class with probability tending to 1.

B.1.2 Case II: optimal individualized encouragement rules

Let

$$
\begin{aligned}
D^E_{n,\mathrm{FR}}(o) &:= D^E(\hat{P}_n, e_n, \tau_0^E, \mu_n^A) - D^E(\hat{P}_n, e^{\mathrm{FR}}, 0, \mu_0^A),\\
\underline{D}^E_{n,\mathrm{FR}}(o) &:= \underline{D}^E(\underline{\hat{P}}_n, \underline{e}_n, \underline{\tau}_0^E, \mu_n^A) - D^E(\hat{P}_n, e^{\mathrm{FR}}, 0, \mu_0^A),\\
D^E_{n,\mathrm{RD}}(o) &:= D^E(\hat{P}_n, e_n, \tau_0^E, \mu_n^A) - D^E(\hat{P}_n, e_n^{\mathrm{RD}}, 0, \mu_0^A) + \frac{\Psi^E_{e_n^{\mathrm{RD}}}(\hat{P}_n)}{P_n\hat{\mu}_n^A(1,\cdot)} D(\hat{P}_n, \hat{\mu}_n^A),\\
\underline{D}^E_{n,\mathrm{RD}}(o) &:= \underline{D}^E(\underline{\hat{P}}_n, \underline{e}_n, \underline{\tau}_0^E, \mu_n^A)(o) - \underline{D}^E(\underline{\hat{P}}_n, \underline{e}_n^{\mathrm{RD}}, 0, \mu_0^A)(o)\\
&\qquad - \frac{\Psi^E_{\underline{e}_n^{\mathrm{RD}}}(\underline{\hat{P}}_n)}{\kappa - \varphi_n} D_1(\underline{\hat{P}}_n, \mu_n^A)(o) - \frac{\Psi^E_{\underline{e}_n^{\mathrm{RD}}}(\hat{P}_n)}{P_n\underline{\hat{\Delta}}_n^A} D_2(\underline{\hat{P}}_n, \underline{\hat{\mu}}_n^A)(o),\\
D^E_{n,\mathrm{TP}}(o) &:= D^E(\hat{P}_n, e_n, \tau_0^E, \mu_n^A) - G^E_{\mathrm{TP}}(\hat{P}_n)\\
\underline{D}^E_{n,\mathrm{TP}}(o) &:= \underline{D}^E(\underline{\hat{P}}_n, \underline{e}_n, \underline{\tau}_0^E, \mu_n^A) - \underline{G}^E_{\mathrm{TP}}(\underline{\hat{P}}_n).
\end{aligned}
$$

Condition C9 (Sufficient rates for nuisance estimators)**.** $\|\mu_n^Z - \mu_0^Z\|_{2,P_0} = \mathrm{o}_p(1)$ and

$$
\begin{aligned}
\|\mu_n^Z - \mu_0^Z\|_{2,P_0} \Big\{ &\|\mu_n^A(1,\cdot) - \mu_0^A(1,\cdot)\|_{2,P_0} + \|\mu_n^Y - \mu_0^Y\|_{2,P_0}\\
&+ \|\hat{\mu}_n^A(1,\cdot) - \mu_0^A(1,\cdot)\|_{2,P_0} + \|\hat{\mu}_n^Y - \mu_0^Y\|_{2,P_0} \Big\} = \mathrm{o}_p(n^{-1/2}).
\end{aligned}
$$

The next two conditions depend on the choice of reference rule, which is indexed by $\mathcal{R} \in \{\mathrm{FR}, \mathrm{RD}, \mathrm{TP}\}$.

Condition C10 (Consistency of estimated influence function)**.** In the first setting, $\|D(\hat{P}_n, \mu_n^A) - D(P_0, \mu_0^A)\|_{2,P_0} = \mathrm{o}_p(1)$ and $\|D_{n,\mathcal{R}}^E - D_{\mathcal{R}}^E(P_0)\|_{2,P_0} = \mathrm{o}_p(1)$; in the second setting, $\|D_1(\underline{\hat{P}}_n, \underline{\hat{\mu}}_n^A) - D_1(P_0, \mu_0^A)\|_{2,P_0} = \mathrm{o}_p(1)$, $\|D_2(\underline{\hat{P}}_n, \mu_n^A) - D_2(P_0, \mu_0^A)\|_{2,P_0} = \mathrm{o}_p(1)$, $\|[\underline{D}^E(\underline{\hat{P}}_n, \underline{e}_n, \underline{\tau}_0^E, \mu_n^A) - \underline{D}^E(\underline{\hat{P}}_n, \underline{e}_n^{\mathrm{RD}}, 0, \mu_0^A)] - [\underline{D}^E(P_0, \underline{e}_0, \underline{\tau}_0^E, \mu_0^A) - \underline{D}^E(P_0, \underline{e}_0^{\mathrm{RD}}, 0, \mu_0^A)]\|_{2,P_0} = \mathrm{o}_p(1)$ and $\|\underline{D}_{n,\mathcal{R}}^E - \underline{D}_{\mathcal{R}}^E(P_0)\|_{2,P_0} = \mathrm{o}_p(1)$.

Condition C11 (Donsker condition)**.** $\{o \mapsto d_{n,k}(v)D(\hat{P}_n, \mu_n^A)(o) : k \in [0,1]\}$ ($\{o \mapsto \underline{d}_{n,k}(v)D_2(\underline{\hat{P}}_n, \mu_n^A)(o) : k \in [0,1]\}$ resp.) is a subset of a fixed P_0-Donsker class with probability tending to 1. Additionally, $D_{n,\mathcal{R}}^E$ (each of $D_1(\underline{\hat{P}}_n, \mu_n^A)$ and $\underline{D}_{n,\mathcal{R}}^E$ resp.) belongs to a fixed P_0-Donsker class with probability tending to 1.

Condition C12 (Glivenko-Cantelli condition)**.** $\|\xi_n - \xi_0\|_{1,P_0} = \mathrm{o}_p(1)$ and $\|\mu_n^A(1,\cdot) - \mu_0^A(1,\cdot)\|_{1,P_0} = \mathrm{o}_p(1)$ ($\|\underline{\xi}_n - \underline{\xi}_0\|_{1,P_0} = \mathrm{o}_p(1)$ and $\|\Delta_n^A - \Delta_0^A\|_{1,P_0} = \mathrm{o}_p(1)$ resp.). Moreover, if, on the one hand, $\eta_0^E > -\infty$ ($\underline{\eta}_0^E > -\infty$ resp.), then, for any η sufficiently close to η_0^E ($\underline{\eta}_0^E$ resp.), $w \mapsto I(\xi_n(v) > \eta)\mu_n^A(1,w)$ ($w \mapsto I(\underline{\xi}_n(v) > \eta)\Delta_n^A(w)$ resp.) belongs to a P_0-Glivenko-Cantelli class with probability tending to 1. If, on the other hand, $\eta_0^E = -\infty$ ($\underline{\eta}_0^E = -\infty$ resp.), then, for any $\eta < 0$ with sufficiently large $|\eta|$, $w \mapsto I(\xi_n(v) > \eta)\mu_n^A(1,w)$ ($w \mapsto I(\underline{\xi}_n(v) > \eta)\Delta_n^A(w)$ resp.) belongs to a P_0-Glivenko-Cantelli class with probability tending to 1.

Remark 20. When $\kappa = 1$, then $w \mapsto \mu_0^A(1,w)$ and $\bar{\mu}_0^A$ (Δ_0^A and $\Delta_{b,0}^A$ resp.) do not need to be estimated. In this case, $\eta_0^E = -\infty$, $\tau_0^E = 0$ and we may take $\tau_n^E = 0$ without estimation. The needed conditions also change slightly. In particular, Conditions C4 and C7 can be dropped. Moreover, Condition C9 may be replaced by $\|\mu_n^Z - \mu_0^Z\|_{2,P_0}(\|\mu_n^Y - \mu_0^Y\|_{2,P_0} + \|\hat{\mu}_n^Y - \mu_0^Y\|_{2,P_0}) = \mathrm{o}_p(n^{-1/2})$. Finally, we note that, in this case, $e_0^{\mathrm{RD}} \equiv 1$ and Condition C5 does not hold, so our results regarding the reference rule e_0^{RD} do not hold.

B.2 Sufficient condition for fast convergence rate of estimated optimal rule

It may seem difficult to verify Conditions B5 and C8. Theorem B.1 below provides sufficient conditions for these conditions under Condition B2 or C2. These conditions are closely

related to the margin assumptions in Audibert and Tsybakov (2007), Qian and Murphy (2011), Luedtke and van der Laan (2016b) and Luedtke and van der Laan (2016a). Such margin assumptions essentially assume that the probability of an individual's conditional average treatment effect being close to the decision boundary is small. Throughout we use $\lesssim$ to denote $\leq$ up to a positive multiplicative constant that may depend on P_0.

In Theorem B.1 below, for ease of illustration, we consider the case where the estimated optimal individualized rule equals an indicator function P_0-almost surely. This would hold if $\Delta_{b,n}$ or $\Delta^Y_{b,n}$ is not a constant over any region in $\mathcal{V}$. We can readily generalize Theorem B.1 to the case where this region exists but shrinks at a certain rate, and when this rate is sufficiently fast, the resulting sufficient conditions for Conditions B5 and C8 remain the same.

Theorem B.1 (Fast rates for optimal rule estimation)**.**

1. *(Sufficient condition for Condition B5) Assume that, with probability tending to one,* $\int I\{t_n(v) = I(\Delta_{b,n}(v) > \tau_n^T)\}dP_0(v) = 1$ *and* $o \mapsto I(\Delta_{b,n}(v) > \eta_n^T)$ *belongs to a fixed* P_0*-Donsker class. Suppose that* $P_n I(\Delta_{b,n}(\cdot) > \eta_n^T) = \kappa \mathrm{O}_p(n^{-1/2})$ *and that the distribution of* $\Delta_{b,0}(V)$*,* $V \sim P_0$*, has nonzero finite continuous Lebesgue density in a neighborhood of* η_0^T *and a neighborhood of* τ_0^T*. The following implications are valid with probability tending to one:*

   - *if* $\|\Delta_{b,n} - \Delta_{b,0}\|_{q,P_0} = \mathrm{o}_p(1)$ *for some* $q \geq 1$*, then*

   $$|P_0\{(t_n - t_0)(\Delta_{b,0} - \tau_0^T)\}| \lesssim \|\Delta_{b,n} - \Delta_{b,0}\|_{q,P_0}^{2q/(q+1)} + \mathrm{O}_p(n^{-1}).$$

   - *if* $\|\Delta_{b,n} - \Delta_{b,0}\|_{\infty,P_0} = \mathrm{o}_p(1)$*, then*

   $$|P_0\{(t_n - t_0)(\Delta_{b,0} - \tau_0^T)\}| \lesssim \|\Delta_{b,n} - \Delta_{b,0}\|_{\infty,P_0}^2 + \mathrm{O}_p(n^{-1}).$$

2. *(Sufficient condition for Condition C8) Assume that, with probability tending to 1,* $\int I\{e_n(v) = I(\xi_n(v) > \tau_n^E)\}dP_0(v) = 1$ *and* $o \mapsto I(\xi_n(v) > \eta_n^E)$ *and* $o \mapsto I(\xi_n(v) >$

$\eta_n^E)\bar{\mu}_0^A(v)$ *belong to a fixed* P_0*-Donsker class. Suppose also that the distribution of* $\xi_0(V)$*,* $V \sim P_0$*, has nonzero finite continuous Lebesgue density in a neighborhood of* η_0^E *and a neighborhood of* τ_0^E*. The following implications are valid with probability tending to one:*

- *if* $\|\Delta_{b,n}^Y - \Delta_{b,0}^Y\|_{q,P_0} = \mathrm{o}_p(1)$ *for some* $q \geq 1$*, then*

$$|P_0\{(e_n - e_0)(\Delta_{b,0}^Y - \tau_0^E \bar{\mu}_0^A)\}| \lesssim \|\Delta_{b,n}^Y - \Delta_{b,0}^Y\|_{q,P_0}^{2q/(q+1)} + \mathrm{O}_p(n^{-1}).$$

- *if* $\|\Delta_{b,n}^Y - \Delta_{b,0}^Y\|_{\infty,P_0} = \mathrm{o}_p(1)$*, then*

$$|P_0\{(e_n - e_0)(\Delta_{b,0}^Y - \tau_0^E \bar{\mu}_0^A)\}| \lesssim \|\Delta_{b,n}^Y - \Delta_{b,0}^Y\|_{\infty,P_0}^2 + \mathrm{O}_p(n^{-1}).$$

Therefore, Condition B5 or C8 is satisfied if the corresponding distance between functions on the right hand side converges to 0 at $\mathrm{o}_p(n^{-3/8})$-rate in $L^2(P_0)$ sense or at $\mathrm{o}_p(n^{-1/4})$-rate in $L^\infty(P_0)$ sense. For example, Condition C8 is satisfied if $\|\xi_n - \xi_0\|_{2,P_0} = \mathrm{o}_p(n^{-3/8})$ or $\|\xi_n - \xi_0\|_{\infty,P_0} = \mathrm{o}_p(n^{-1/4})$.

The proof of this theorem can be found in Supplementary Appendix S9.

B.3 Modified procedure with sample splitting

As we mention in Remarks 16 and 19 in the main text, we can estimate an optimal individualized rule via sample splitting. This can reduce the positive bias of the average effect estimator in our original procedure. In this section, we describe these sample splitting procedures. In particular, we employ sample splitting in the step that estimates an optimal individualized rule and make corresponding modifications in the following steps. Note that, in this procedure, we only estimate the evaluation of an optimal individualized rule at all observations in the data, but do not estimate the rule as a function defined on $\mathcal{V}$. As we will show, we can still obtain a plug-in estimator, because with our construction of $\hat{P}_n$, for any $t, e : \mathcal{V} \to [0,1]$, $\Psi_t^T(\hat{P}_n)$ and $\Psi_e^E(\hat{P}_n)$ depend on t and e only through their values at all observations.

We use Λ to denote a user-specified fixed number of folds in this procedure. In practice, Λ may be taken as 5, 10 or 20.

B.3.1 Case I: optimal individualized treatment rules

1. Use the empirical distribution as an estimate $\hat{P}_{W,n}$ of the marginal distribution of W under P_0. Estimate μ_0^Y, μ_0^A and μ_0^Z with flexible methods. Denote the corresponding estimators by μ_n^Y, μ_n^A and μ_n^Z.

2. Let $\Delta_n^A : w \mapsto \mu_n^A(1,w) - \mu_n^A(0,w)$, $\Delta_n^Y : w \mapsto \mu_n^Y(1,w) - \mu_n^Y(0,w)$, and $\Delta_n : w \mapsto \Delta_n^Y(w)/\Delta_n^A(w)$.

3. Estimate the evaluation of an optimal ITR at each observation:

   (a) Create folds: divide the set of indices of observations $\{1, 2, \ldots, n\}$ into Λ mutually exclusive and exhaustive folds of (approximately) equal size. Denote these sets by S_λ, $\lambda = 1, 2, \ldots, \Lambda$. Let $S_{-\lambda} := \cup_{\lambda' \neq \lambda} S_{\lambda'}$. For each $i = 1, 2, \ldots, n$, let $\lambda(i)$ be the index of the fold that contains i, namely $\lambda(i)$ is the (unique) value of λ for which $i \in S_\lambda$.

   (b) Estimate the conditional ATE for each observation using sample splitting: for each $\lambda = 1, 2, \ldots, \Lambda$, let $\Delta_{b,n,S_{-\lambda}}$ be the (flexible) estimate of $\Delta_{b,0}$ based on the data $\{O_i : i \in S_{-\lambda}\}$. For each $i = 1, 2, \ldots, n$, let $\Delta_{b,n,i} := \Delta_{b,n,S_{-\lambda(i)}}(V_i)$.

   (c) Let $\eta_n^T := \inf\{\eta : \frac{1}{n}\sum_{i=1}^n I(\Delta_{b,n,i} > \eta) \leq \kappa\}$ and $\tau_n^T := \max\{\eta_n^T, 0\}$.

   (d) For each $i = 1, 2, \ldots, n$, estimate $t_0(V_i)$ by

$$t_{n,i} := \begin{cases} \frac{\kappa - \frac{1}{n}\sum_{i=1}^n I(\Delta_{b,n,i} > \tau_n^T)}{\frac{1}{n}\sum_{i=1}^n I(\Delta_{b,n,i} = \tau_n^T)}, & \text{if } \Delta_{b,n,i} = \tau_n^T,\ \tau_n^T > 0, \text{ and } \frac{1}{n}\sum_{i=1}^n I(\Delta_{b,n,i} = \tau_n^T) > 0, \\ I(\Delta_{b,n,i} > \tau_n^T), & \text{otherwise.} \end{cases}$$

4. Estimate the ATE relative to a reference rule t_r:

(a) Obtain a targeted estimate $\hat{\mu}_n^Y$ of μ_0^Y by running an ordinary least-squares linear regression using observations $i = 1, 2, \ldots, n$ with outcome Y_i, offset $\mu_n^Y(Z_i, W_i)$, no intercept and covariate $h(Z_i, W_i) := \frac{t_{n,i} - t_r(V_i)}{[Z_i + \mu_n^Z(W_i) - 1]\Delta_n^A(W_i)}$.

(b) Obtain a targeted estimate $\hat{\mu}_n^A$ of μ_0^A by running a logistic regression using observations $i = 1, 2, \ldots, n$ with outcome A_i, offset $\operatorname{logit} \mu_n^A(Z_i, W_i)$, no intercept and covariate $h(Z_i, W_i)\Delta_n(W_i)$.

(c) Letting $\hat{P}_n$ be a distribution with components $\hat{\mu}_n^Y$, $\hat{\mu}_n^A$ and $\hat{P}_{W,n}$, estimate the ATE of t_0 relative to that of t_r with $\frac{1}{n}\sum_{i=1}^n t_{n,i}\hat{\Delta}_n(W_i) - \Psi_{t_r}^T(\hat{P}_n)$, where recall that $\hat{\Delta}_n : w \mapsto \{\hat{\mu}_n^Y(1, w) - \hat{\mu}_n^Y(0, w)\}/\{\hat{\mu}_n^A(1, w) - \hat{\mu}_n^A(0, w)\}$.

B.3.2 Case II: optimal individualized encouragement rules

First setting:

1. Let $\hat{P}_{W,n}$ denote the empirical marginal distribution of W, and let μ_n^A, μ_n^Y and μ_n^Z denote (flexible) estimates of μ_0^A, μ_0^Y and μ_0^Z, respectively.

2. Estimate the evaluation of an optimal IER for each observation:

   (a) Create folds: perform Step 3a in Section B.3.1.

   (b) Estimate $\xi_0(V_i)$ using sample splitting: for each $\lambda = 1, 2, \ldots, \Lambda$, let $\Delta_{b,n,S_{-\lambda}}^Y$ and $\bar{\mu}_{n,S_{-\lambda}}^A$ denote the (flexible) estimates of $\Delta_{b,0}^Y$ and $\bar{\mu}_0^A$, respectively, based on data $\{O_i : i \in S_{-\lambda}\}$. For each $i = 1, 2, \ldots, n$, let $\xi_{n,i} := \Delta_{b,n,S_{-\lambda(i)}}^Y(V_i)/\bar{\mu}_{n,S_{-\lambda(i)}}^A(V_i)$.

   (c) Define $\Gamma_n : \tau \mapsto \frac{1}{n}\sum_{i:\xi_{n,i}>\tau} \mu_n^A(1, W_i)$, $\gamma_n : \tau \mapsto \frac{1}{n}\sum_{i:\xi_{n,i}=\tau} \mu_n^A(1, W_i)$. For any $k \in [0, 1]$, define $\eta_n^E(k) := \inf\{\tau : \Gamma_n(\tau) \leq k\}$, $\tau_n^E(k) := \max\{\eta_n^E(k), 0\}$, and, for $i = 1, 2, \ldots, n$,

$$d_{n,k,i} := \begin{cases} \frac{k - \Gamma_n(\eta_n^E(k))}{\gamma_n(\eta_n^E(k))}, & \text{if } \xi_{n,i} = \eta_n^E(k) \text{ and } \gamma_n(\eta_n^E(k)) > 0, \\ I\{\xi_{n,i} > \eta_n^E(k)\}, & \text{otherwise.} \end{cases}$$

(d) Compute k_n, which will be used to define an estimate of e_0 for which the plug-in estimator is asymptotically linear.

- If $\tau_n^E(\kappa) > 0$ *and* there is a solution in k to

$$\frac{1}{n}\sum_{i=1}^{n} d_{n,k,i}\left[\mu_n^A(1,W_i) + \frac{Z_i}{\mu_n^Z(W_i)}[A_i - \mu_n^A(1,W_i)]\right] = \kappa, \tag{B.1}$$

then let k_n be this solution.

- Else, let $k_n = \kappa$.

(e) For each $i = 1, 2, \ldots, n$, estimate $e_0(V_i)$ with

$$e_{n,i} := \begin{cases} \frac{k_n - \Gamma_n(\tau_n^E(k_n))}{\gamma_n(\tau_n^E(k_n))}, & \text{if } \xi_{n,i} = \tau_n^E(k_n), \text{ and } \gamma_n(\tau_n^E(k_n)) > 0, \\ I\{\xi_{n,i} > \tau_n^E(k_n)\}, & \text{otherwise.} \end{cases}$$

3. Estimate the reference rule $e_0^{\mathcal{R}}$ as follows:

- For $\mathcal{R} = \text{FR}$, let $e_n^{\mathcal{R}} : o \mapsto e^{\text{FR}}(v)$.
- For $\mathcal{R} = \text{RD}$,
  (a) Obtain a targeted estimate $\hat{\mu}_n^A(1,\cdot)$ of $\mu_0^A(1,\cdot)$ by running a logistic regression using observations $i = 1, 2, \ldots, n$ with outcome A_i, offset $\text{logit}\,\mu_n^A(1,W_i)$, no intercept and covariate $Z_i/\mu_n^Z(W_i)$. $\hat{\mu}_n^A$ is the fitted mean model.
  (b) Let $e_n^{\mathcal{R}} : o \mapsto \min\{1, \kappa/\hat{P}_{W,n}\hat{\mu}^A(1,\cdot)\}$, where we note that this function is constant in its input.
- For $\mathcal{R} = \text{TP}$, let $e_n^{\mathcal{R}} : o \mapsto z$.

4. Estimate AEE of e_0 compared with the reference rule $e_0^{\mathcal{R}}$:

(a) Obtain a targeted estimate $\hat{\mu}_n^Y$ of μ_0^Y by running an ordinary least-squares linear regression using observations $i = 1, 2, \ldots, n$ with outcome Y_i, offset $\mu_n^Y(Z_i, W_i)$, no intercept and covariate $[e_{n,i} - e_n^{\mathcal{R}}(O_i)]/[Z_i + \mu_n^Z(W_i) - 1]$. $\hat{\mu}_n^Y$ is the fitted mean function.

(b) Letting $\hat{P}_n$ be a distribution with components $\hat{\mu}_n^Y$ and $\hat{P}_{W,n}$, estimate the AEE of e_0 relative to $e_0^{\mathcal{R}}$ with $\frac{1}{n}\sum_{i=1}^n e_{n,i}\hat{\Delta}_n^Y(W_i) - \Psi_{e_n^{\mathcal{R}}}^E(\hat{P}_n)$ where recall that $\hat{\Delta}_n^Y : w \mapsto \hat{\mu}_n^Y(1,w) - \hat{\mu}_n^Y(0,w)$.

Second setting:

1. Same as the first setting.

2. Estimate the evaluation of an optimal IER for each observation:

   (a) Create folds: perform Step 3a in Section B.3.1.

   (b) Estimate $\underline{\xi}_0(V_i)$ using sample splitting: for each $\lambda = 1, 2, \ldots, \Lambda$, let $\Delta_{b,n,S_{-\lambda}}^Y$ and $\Delta_{b,n,S_{-\lambda}}^A$ denote the (flexible) estimates of $\Delta_{b,0}^Y$ and $\Delta_{b,0}^A$, respectively, based on data $\{O_i : i \in S_{-\lambda}\}$. For each $i = 1, 2, \ldots, n$, let $\underline{\xi}_{n,i} := \Delta_{b,n,S_{-\lambda(i)}}^Y(V_i)/\Delta_{b,n,S_{-\lambda(i)}}^A(V_i)$.

   (c) Define $\varphi_n := \frac{1}{n}\sum_{i=1}^n \mu_n^A(0,W_i) + \frac{1-Z_i}{1-\mu_n^Z(W_i)}[A_i - \mu_n^A(0,W_i)]$, $\underline{\Gamma}_n : \tau \mapsto \frac{1}{n}\sum_{i:\underline{\xi}_{n,i}>\tau} \Delta_n^A(W_i)$, $\underline{\gamma}_n : \tau \mapsto \frac{1}{n}\sum_{i:\underline{\xi}_{n,i}=\tau} \Delta_n^A(W_i)$. For any $k \in [0,1]$, define $\underline{\eta}_n^E(k) := \inf\{\tau : \underline{\Gamma}_n(\tau) \leq k - \varphi_n\}$, $\underline{\tau}_n^E(k) := \max\{\underline{\eta}_n^E(k), 0\}$, and, for $i = 1, 2, \ldots, n$,

$$d_{n,k,i} := \begin{cases} \frac{k-\varphi_n-\underline{\Gamma}_n(\underline{\eta}_n^E(k))}{\underline{\gamma}_n(\underline{\eta}_n^E(k))}, & \text{if } \underline{\xi}_{n,i} = \underline{\eta}_n^E(k) \text{ and } \underline{\gamma}_n(\underline{\eta}_n^E(k)) > 0, \\ I\{\underline{\xi}_{n,i} > \underline{\eta}_n^E(k)\}, & \text{otherwise.} \end{cases}$$

   (d) Compute $\underline{k}_n$, which will be used to define an estimate of e_0 for which the plug-in estimator is asymptotically linear.

   - If $\underline{\tau}_n^E(\kappa) > 0$ *and* there is a solution in k to

$$\frac{1}{n}\sum_{i=1}^n d_{n,k,i}\left[\Delta_n^A(W_i) + \frac{1}{Z_i + \mu_n^Z(W_i) - 1}[A_i - \mu_n^A(Z_i, W_i)]\right] = \kappa - \varphi_n, \quad \text{(B.2)}$$

     then let $\underline{k}_n$ be this solution.

   - Else, let $\underline{k}_n = \kappa$.

(e) For each $i = 1, 2, \ldots, n$, estimate $\underline{e}_0(V_i)$ with

$$\underline{e}_{n,i} := \begin{cases} \frac{\underline{k}_n - \varphi_n - \underline{\Gamma}_n(\underline{\tau}_n^E(\underline{k}_n))}{\underline{\gamma}_n(\underline{\tau}_n^E(\underline{k}_n))}, & \text{if } \underline{\xi}_{n,i} = \underline{\tau}_n^E(\underline{k}_n), \text{ and } \underline{\gamma}_n(\underline{\tau}_n^E(\underline{k}_n)) > 0, \\ I\{\underline{\xi}_{n,i} > \underline{\tau}_n^E(\underline{k}_n)\}, & \text{otherwise.} \end{cases}$$

3. Estimate the reference rule $\underline{e}_0^{\mathcal{R}}$ as follows:

   - For $\mathcal{R} = \text{FR}$ or TP, same as the first setting.
   - For $\mathcal{R} = \text{RD}$,
     (a) Obtain a targeted estimate $\underline{\hat{\mu}}_n^A$ of μ_0^A by running a logistic regression using observations $i = 1, 2, \ldots, n$ with outcome A_i, offset logit $\mu_n^A(Z_i, W_i)$, no intercept and covariate $1/(Z_i + \mu_n^Z(W_i) - 1)$. $\underline{\hat{\mu}}_n^A$ is the fitted mean model.
     (b) Let $\underline{e}_n^{\mathcal{R}} : o \mapsto \min\{1, (\kappa - \varphi_n)/\hat{P}_{W,n}\underline{\hat{\Delta}}_n^A\}$, where $\underline{\hat{\Delta}}_n^A : w \mapsto \underline{\hat{\mu}}_n^A(1, w) - \underline{\hat{\mu}}_n^A(0, w)$.

4. Same as the first setting with $(e_n, e_n^{\mathcal{R}})$ replaced by $(\underline{e}_n, \underline{e}_n^{\mathcal{R}})$.

B.4 Estimation of contrasts and V-specific means

Note that $\Delta_{b,0} : v \mapsto \mathrm{E}_0[\Delta_0(W) \mid V = v]$ is defined as the conditional mean of the ratio Δ_0 of two functions that can be readily estimated with existing supervised ML methods. A natural estimate Δ_n of this ratio is given by $w \mapsto (\mu_n^Y(1, w) - \mu_n^Y(0, w))/(\mu_n^A(1, w) - \mu_n^A(0, w))$, where here μ_n^Y and μ_n^A may be obtained from flexible ML methods, e.g. gradient boosting (Mason et al., 1999, 2000; Friedman, 2001, 2002). In the case that $V = W$, $\Delta_{b,0}$ is simply equal to the ratio Δ_0, and therefore Δ_n can also be used as an estimate $\Delta_{b,n}$ of $\Delta_{b,0}$. When $V \neq W$, $\Delta_{b,0}$ is instead equal to a conditional expectation of the ratio $\Delta_0(W)$. In this case, we therefore propose to estimate $\Delta_{b,0}$ by running a regression with outcome $\Delta_n(W)$ and covariate V. This regression can be performed using parametric or more flexible approaches.

To gain some insight into the performance of this estimation strategy, we will study a particular choice of regression algorithm for $\Delta_{b,n}$, namely an empirical risk minimizer (ERM)

over a class $\mathcal{F}$ (Vapnik, 1992). For simplicity, we will suppose that the class $\mathcal{F}$ contains $\Delta_{b,0}$. The ERM that we study is given by

$$\Delta_{b,n} \in \operatorname*{argmin}_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} [\Delta_n(W_i) - f(V_i)]^2.$$

We wish to establish a bound on the mean-squared error of this estimator of $\Delta_{b,0}$. To do this, it will be useful to define the true-risk minimizer, namely

$$\Delta_{b,n,0} \in \operatorname*{argmin}_{f \in \mathcal{F}} P_0(\Delta_n - f)^2. \tag{B.3}$$

By the triangle inequality, the root mean-squared error bounds as

$$\|\Delta_{b,n} - \Delta_{b,0}\|_{2,P_0} \leq \|\Delta_{b,n,0} - \Delta_{b,0}\|_{2,P_0} + \|\Delta_{b,n} - \Delta_{b,n,0}\|_{2,P_0}. \tag{B.4}$$

Hence, it suffices to study the *first-stage error* $\|\Delta_{b,n,0} - \Delta_{b,0}\|_{2,P_0}$ and the *second-stage error* $\|\Delta_{b,n} - \Delta_{b,n,0}\|_{2,P_0}$, where these errors are so named because they correspond to the errors induced by the first and second stages of our estimation procedure, namely the estimation of Δ_n and of the conditional expectation of $\Delta_n(W)$ given V, respectively. The following result bounds the first-stage error.

Theorem B.2 (First-stage error for ERM of $\Delta_{b,0}$)**.** *If* $\Delta_{b,0} \in \mathcal{F}$*, then* $\|\Delta_{b,n,0} - \Delta_{b,0}\|_{2,P_0} \leq 2\|\Delta_n - \Delta_0\|_{2,P_0}$.

Proof. The result is trivial if $\|\Delta_{b,n,0} - \Delta_{b,0}\|_{2,P_0} = 0$, so suppose that $\|\Delta_{b,n,0} - \Delta_{b,0}\|_{2,P_0} > 0$. For any $f \in \mathcal{F}$,

$$\begin{aligned} P_0(\Delta_0 - f)^2 =& P_0(\Delta_0 - \Delta_{b,0} + \Delta_{b,0} - f)^2 \\ =& P_0(\Delta_0 - \Delta_{b,0})^2 + 2P_0(\Delta_0 - \Delta_{b,0})(\Delta_{b,0} - f) + P_0(\Delta_{b,0} - f)^2 \\ =& P_0(\Delta_0 - \Delta_{b,0})^2 + P_0(\Delta_{b,0} - f)^2 \end{aligned}$$

since $\Delta_{b,0}(v) = \mathrm{E}_0[\Delta_0(W)|V = v]$. Therefore,

$$\|\Delta_{b,n,0} - \Delta_{b,0}\|_{2,P_0}^2 = P_0(\Delta_{b,n,0} - \Delta_{b,0})^2 = P_0(\Delta_0 - \Delta_{b,n,0})^2 - P_0(\Delta_0 - \Delta_{b,0})^2$$

$$\begin{aligned}
&\leq P_0(\Delta_0 - \Delta_{b,n,0})^2 - P_0(\Delta_0 - \Delta_{b,0})^2 - P_0(\Delta_n - \Delta_{b,n,0})^2 + P_0(\Delta_n - \Delta_{b,0})^2 \\
&= 2P_0(\Delta_n - \Delta_0)(\Delta_{b,n,0} - \Delta_{b,0}) \leq 2\|\Delta_n - \Delta_0\|_{2,P_0}\|\Delta_{b,n,0} - \Delta_{b,0}\|_{2,P_0},
\end{aligned}$$

where the first inequality holds by a combination of the fact that $\Delta_{b,0} \in \mathcal{F}$ and the minimization property of $\Delta_{b,n,0}$ given in (B.3), and the second inequality holds by Cauchy-Schwarz. Dividing both sides by $\|\Delta_{b,n,0} - \Delta_{b,0}\| > 0$ shows that $\|\Delta_{b,n,0} - \Delta_{b,0}\|_{2,P_0} \leq 2\|\Delta_n - \Delta_0\|_{2,P_0}$. □

Combining the preceding lemma with the decomposition in (B.4), we see that, provided $\Delta_{b,0} \in \mathcal{F}$, $\|\Delta_{b,n} - \Delta_{b,0}\|_{2,P_0} \leq 2\|\Delta_n - \Delta_0\|_{2,P_0} + \|\Delta_{b,n} - \Delta_{b,n,0}\|_{2,P_0}$. In other words, the root mean-squared error of the estimate $\Delta_{b,n}$ of $\Delta_{b,0}$ can be bounded by the sum of the $L^2(P_0)$ distances between (1) the estimated ratio Δ_n and the true ratio Δ_0, and (2) the empirical risk minimizer $\Delta_{b,n}$ and the true risk minimizer $\Delta_{b,n,0}$. We expect similar results to hold for more general regression methods.

When $V \neq W$, similar strategies can be used to estimate $\Delta_{b,0}^Y$ and $\bar{\mu}_0^A$ ($\Delta_{b,0}^A$ resp.). That is, for $\Delta_{b,0}^Y$ ($\bar{\mu}_0^A$ or $\Delta_{b,0}^A$, respectively), first estimate Δ_0^Y (μ_0^A or Δ_0^A, respectively) with Δ_n^Y (μ_n^A or Δ_n^A, respectively) and then run a regression with outcome $\Delta_n^Y(W)$ ($\mu_n^A(W)$ or $\Delta_n^A(W)$, respectively) and covariate V. We expect results similar to Theorem B.2 to hold for these estimators.

B.5 Identification of causal estimands

In this appendix, we prove identification results of causal estimands and the forms of optimal rules.

Theorem B.3 (Identification of conditional ATE and ATE)**.** *Under Conditions A1–A4, A5a and A5b with* $\delta^Z = \delta^A = 0$, $\mathbb{E}[Y(1) - Y(0) \mid W] = \Delta_0(W)$ *almost surely. In addition, for two ITRs* t *and* t_r, $\mathbb{E}[Y(t) - Y(t_r)] = \mathrm{E}_0[\{t(V) - t_r(V)\}\,\mathbb{E}[\Delta_0(W) \mid V]]$.

Proof. We first prove $\mathbb{E}[Y(1) - Y(0) \mid W] = \Delta_0(W)$ almost surely. We study the two cases where part (1) in A5b holds and part (2) in A5b holds separately.

First consider the case where part (1) in A5b holds. Our argument closely mirrors that given for the proof of Theorem 1 in Wang and Tchetgen Tchetgen (2018), but modifies the argument to account for the fact that we only assume an uncorrelated, rather than an independent, instrument. In the following display, we conditional on $W = w$ for w such that A1–A4, A5a and part (1) of A5b hold with $W = w$, and drop the conditioning in the notation. Note that the collection of such w has probability measure one. We have that

$$
\begin{aligned}
&\mathrm{E}_0[Y \mid Z=1] - \mathrm{E}_0[Y \mid Z=0] \\
&= \sum_{z=0,1} (2z-1)\, \mathrm{E}_0[Y \mid Z=z] \\
&= \sum_{z=0,1} (2z-1)\, \mathbb{E}\left[\mathbb{E}[Y \mid Z=z, U] \mid Z=z\right] \\
&= \sum_{z=0,1} (2z-1)\, \mathbb{E}\left[\mathbb{E}[YA \mid Z=z, U] + \mathbb{E}[Y\{1-A\} \mid Z=z, U] \mid Z=z\right]
\end{aligned}
$$

by Condition A5a, the display continues as

$$
\begin{aligned}
&= \sum_{z=0,1} (2z-1)\, \mathbb{E}\left[\mathbb{E}[Y(1)A \mid Z=z, U] + \mathbb{E}[Y(0)(1-A) \mid Z=z, U] \mid Z=z\right] \\
&= \sum_{z=0,1} (2z-1)\, \mathbb{E}\Bigg[\mathbb{E}[Y(1) \mid Z=z, U]\, \mathbb{E}[A \mid Z=z, U] \\
&\qquad\qquad + \mathbb{E}[Y(0) \mid Z=z, U](1 - \mathbb{E}[A \mid Z=z, U]) \;\Bigg|\; Z=z \Bigg] \\
&= \sum_{z=0,1} (2z-1)\, \mathbb{E}\left[\mathbb{E}[Y(1)-Y(0) \mid Z=z, U]\, \mathbb{E}[A \mid Z=z, U] + \mathbb{E}[Y(0) \mid Z=z, U] \mid Z=z\right]
\end{aligned}
$$

by Condition A5a and part (1)b of A5b, $\mathbb{E}[Y(1) - Y(0) \mid Z=z, U] = \mathbb{E}[Y(1) - Y(0)]$ and hence the display continues as

$$
\begin{aligned}
&= \sum_{z=0,1} (2z-1) \left\{\mathbb{E}[Y(1)-Y(0)]\, \mathbb{E}\left[\mathbb{E}[A \mid Z=z, U] \mid Z=z\right] + \mathbb{E}\left[\mathbb{E}[Y(0) \mid Z=z, U] \mid Z=z\right]\right\} \\
&= \mathbb{E}[Y(1)-Y(0)] \left\{\mathrm{E}_0[A \mid Z=1] - \mathrm{E}_0[A \mid Z=0]\right\} + \mathbb{E}[Y(0) \mid Z=1] - \mathbb{E}[Y(0) \mid Z=0].
\end{aligned}
$$

By part (1)a of A5b,

$$
0 = \mathbb{E}[Y(0)Z] - \mathbb{E}[Y(0)]\, \mathrm{E}_0[Z] = \mathbb{E}[Y(0) \mid Z=1]\, \mathrm{E}_0[Z] - \mathbb{E}[Y(0)]\, \mathrm{E}_0[Z].
$$

Under Condition A3, dividing $\mathrm{E}_0[Z]$ on both sides yields that $\mathbb{E}[Y(0) \mid Z = 1] = \mathbb{E}[Y(0)]$. Since $\mathbb{E}[Y(0)] = \mathbb{E}[Y(0) \mid Z = 1]\,\mathrm{E}_0[Z] + \mathbb{E}[Y(0) \mid Z = 0](1 - \mathrm{E}_0[Z])$, using Condition A3 again, it follows that $\mathbb{E}[Y(0) \mid Z = 1] = \mathbb{E}[Y(0) \mid Z = 0]$. Therefore,

$$\mathrm{E}_0[Y \mid Z = 1] - \mathrm{E}_0[Y \mid Z = 0] = \mathbb{E}[Y(1) - Y(0)]\left\{\mathrm{E}_0[A \mid Z = 1] - \mathrm{E}_0[A \mid Z = 0]\right\}.$$

Under Condition A2, $\mathrm{E}_0[A \mid Z = 1] - \mathrm{E}_0[A \mid Z = 0] \neq 0$ and hence the desired result follows by dividing this term on both sides.

We now consider our second case, namely that part (2) of A5b holds. Theorem 1 in Wang and Tchetgen Tchetgen (2018) shows that, in this case, $\mathbb{E}[Y(1) - Y(0) \mid W] = \Delta_0(W)$ almost surely.

We have thus proven that $\mathbb{E}[Y(1) - Y(0) \mid W] = \Delta_0(W)$ almost surely under the conditions of this theorem. Hence,

$$\begin{aligned}
\mathbb{E}[Y(t) - Y(t_r)] &= \mathbb{E}[\{t(V) - t_r(V)\}\{Y(1) - Y(0)\}] \\
&= \mathrm{E}_0[\{t(V) - t_r(V)\}\,\mathbb{E}[\mathbb{E}[Y(1) - Y(0) \mid W] \mid V]] \\
&= \mathrm{E}_0[\{t(V) - t_r(V)\}\,\mathrm{E}_0[\Delta_0(W) \mid V]],
\end{aligned}$$

where the second equality holds by the law of total expectation. □

Theorem B.4 (Optimality of t_0)**.** *t_0 is a solution to* (3.2)*.*

Proof. Let t be any other ITR that satisfies the constraint that $\mathrm{E}_0[t(V)] \leq \kappa$. We will show that $\mathrm{E}_0[t_0(V)\Delta_{b,0}(V)] \geq \mathrm{E}_0[t(V)\Delta_{b,0}(V)]$. As t was arbitrary, this will then show that t_0 is a solution to (3.2).

We start by noting that

$$\begin{aligned}
&\mathrm{E}_0[t_0(V)\Delta_{b,0}(V)] - \mathrm{E}_0[t(V)\Delta_{b,0}(V)] = \mathrm{E}_0[(t_0(V) - t(V))\Delta_{b,0}(V)] \\
&\quad = \mathrm{E}_0[(t_0(V) - t(V))\Delta_{b,0}(V)I(\Delta_{b,0}(V) > \tau_0^T)] + \mathrm{E}_0[(t_0(V) - t(V))\Delta_{b,0}(V)I(\Delta_{b,0}(V) < \tau_0^T)] \\
&\qquad + \mathrm{E}_0[(t_0(V) - t(V))\Delta_{b,0}(V)I(\Delta_{b,0}(V) = \tau_0^T)].
\end{aligned}$$

Note that $t_0(v) = 1 \geq t(v)$ if $\Delta_{b,0}(v) > \tau_0^T$, and $t_0(v) = 0 \leq t(v)$ if $\Delta_{b,0}(v) < \tau_0^T$. Combining this observation with the fact that $\tau_0^T \geq 0$, the above shows that

$$\begin{aligned}
&\mathrm{E}_0[t_0(V)\Delta_{b,0}(V)] - \mathrm{E}_0[t(V)\Delta_{b,0}(V)] \\
&\quad \geq \tau_0^T \,\mathrm{E}_0[(t_0(V) - t(V))I(\Delta_{b,0}(V) > \tau_0^T)] + \tau_0^T \,\mathrm{E}_0[(t_0(V) - t(V))I(\Delta_{b,0}(V) < \tau_0^T)] \\
&\qquad + \tau_0^T \,\mathrm{E}_0[(t_0(V) - t(V))I(\Delta_{b,0}(V) = \tau_0^T)] \\
&\quad = \tau_0^T \,\mathrm{E}_0[t_0(V) - t(V)].
\end{aligned}$$

We show that $\mathrm{E}_0[t_0(V)\Delta_{b,0}(V)] \geq \mathrm{E}_0[t(V)\Delta_{b,0}(V)]$ by considering two cases. First, suppose that $\tau_0^T = 0$. In this case, the above clearly gives the desired inequality. Second, suppose that $\tau_0^T \neq 0$. Since τ_0^T is nonnegative by construction, it must be that $\tau_0^T > 0$. Moreover, because t satisfies the constraint and t_0 saturates the constraint in the case that $t_0 > 0$, $\mathrm{E}_0[t(V)] \leq \kappa = \mathrm{E}_0[t_0(V)]$. Thus, the right-hand side above is the product of two nonnegative terms, and is therefore positive. It follows that $\mathrm{E}_0[t_0(V)\Delta_{b,0}(V)] \geq \mathrm{E}_0[t(V)\Delta_{b,0}(V)]$ in this second case as well. Thus, t_0 is a solution to (3.2). □

Theorem B.5 (Identification of AEE)**.** *Under Conditions A1 to A4 and A6a with* $\delta^Z = \delta^A = 0$*, it holds that* $\mathbb{E}[Y(A(1)) - Y(A(0)) \mid V] = \mathrm{E}_0[\Delta_0^Y(W) \mid V]$ *and* $\mathbb{E}[Y(A(e)) - Y(A(e_r))] = \mathrm{E}_0[\{e(V) - e_r(V)\}\,\mathrm{E}_0[\Delta_0^Y(W) \mid V]]$.

Proof. We start by noting that

$$\begin{aligned}
\mathbb{E}[Y(A(1)) \mid W] &= \mathbb{E}[Y(1)A(1) + Y(0)(1 - A(1)) \mid W] \\
&= \mathbb{E}[Y(1)A(1) \mid W] + \mathbb{E}[Y(0)(1 - A(1)) \mid W] \\
&= \mathbb{E}[Y(1)A(1) \mid Z = 1, W] + \mathbb{E}[Y(0)(1 - A(1)) \mid Z = 1, W] \\
&= \mathrm{E}_0[AY \mid Z = 1, W] + \mathrm{E}_0[Y(1 - A) \mid Z = 1, W] \\
&= \mathrm{E}_0[Y \mid Z = 1, W] = \mu_0^Y(1, W).
\end{aligned}$$

Similarly, $\mathbb{E}[Y(A(0)) \mid W] = \mathrm{E}_0[Y \mid Z = 0, W] = \mu_0^Y(0, W)$. Hence, $\mathbb{E}[Y(A(1)) - Y(A(0)) \mid W] = \Delta_0^Y(W)$. By the law of total expectation, this yields that $\mathbb{E}[Y(A(1)) - Y(A(0)) \mid V] =$

$\mathrm{E}_0[\Delta_0^Y(W) \mid V]$. This then yields that

$$\begin{aligned}\mathbb{E}[Y(A(e)) - Y(A(e_r))] &= \mathbb{E}[\{e(V) - e_r(V)\}\{Y(A(1)) - Y(A(0))\}] \\ &= \mathrm{E}_0[\{e(V) - e_r(V)\}\,\mathbb{E}[Y(A(1)) - Y(A(0)) \mid V]] \\ &= \mathrm{E}_0[\{e(V) - e_r(V)\}\,\mathrm{E}_0[\Delta_0^Y(W) \mid V]].\end{aligned}$$

□

Theorem B.6 (Optimality of e_0)**.** *e_0 is a solution to* (3.3).

Proof. The proof is similar to that of Theorem B.4. Let e be any other IER that satisfies the constraint that $\mathrm{E}_0[e(V)\bar{\mu}_0^A(V)] \leq \kappa$. We will show that $\mathrm{E}_0[e_0(V)\Delta_{b,0}^Y(V)] \geq \mathrm{E}_0[e(V)\Delta_{b,0}^Y(V)]$, implying that e_0 is a solution to (3.3).

Observe that

$$\begin{aligned}&\mathrm{E}_0[e_0(V)\Delta_{b,0}^Y(V)] - \mathrm{E}_0[e(V)\Delta_{b,0}^Y(V)] \\ =&\,\mathrm{E}_0[\{e_0(V) - e(V)\}\Delta_{b,0}^Y(V)] \\ =&\,\mathrm{E}_0[\{e_0(V) - e(V)\}\Delta_{b,0}^Y(V)I(\xi_0(V) > \tau_0^E)] + \mathrm{E}_0[\{e_0(V) - e(V)\}\Delta_{b,0}^Y(V)I(\xi_0(V) < \tau_0^E)] \\ &+ \mathrm{E}_0[\{e_0(V) - e(V)\}\Delta_{b,0}^Y(V)I(\xi_0(V) = \tau_0^E)]\end{aligned}$$

Note that $e_0(v) = 1 \geq e(v)$ if $\xi_0(v) > \tau_0^E$ and $e_0(v) = 0 \leq e(v)$ if $\xi_0(v) < \tau_0^E$. Combining this observation with the fact that $\tau_0^E \geq 0$, the above shows that

$$\begin{aligned}&\mathrm{E}_0[e_0(V)\Delta_{b,0}^Y(V)] - \mathrm{E}_0[e(V)\Delta_{b,0}^Y(V)] \\ \geq&\tau_0^E\,\mathrm{E}_0[\{e_0(V) - e(V)\}\bar{\mu}_0^A(V)I(\xi_0(V) > \tau_0^E)] + \tau_0^E\,\mathrm{E}_0[\{e_0(V) - e(V)\}\bar{\mu}_0^A(V)I(\xi_0(V) < \tau_0^E)] \\ &+ \tau_0^E\,\mathrm{E}_0[\{e_0(V) - e(V)\}\bar{\mu}_0^A(V)I(\xi_0(V) = \tau_0^E)] \\ =&\tau_0^E\,\mathrm{E}_0[\{e_0(V) - e(V)\}\bar{\mu}_0^A(V)]\end{aligned}$$

If $\tau_0^E = 0$, then clearly $\mathrm{E}_0[e_0(V)\Delta_{b,0}^Y(V)] - \mathrm{E}_0[e(V)\Delta_{b,0}^Y(V)] \geq 0$, as desired; otherwise, $\tau_0^E > 0$ and $\mathrm{E}_0[e(V)\bar{\mu}_0^A(V)] \leq \kappa = \mathrm{E}_0[e_0(V)\bar{\mu}_0^A(V)]$, and so we also have that $\mathrm{E}_0[e_0(V)\Delta_{b,0}^Y(V)] \geq \mathrm{E}_0[e(V)\Delta_{b,0}^Y(V)]$. Therefore, e_0 is an optimal IER. □

Theorem B.7 (Optimality of $\underline{e}_0$)**.** *$\underline{e}_0$ is a solution to* (3.4).

Proof. The proof is similar to that of Theorems B.4 and B.6. Let e be any other IER that satisfies the constraint that $\mathrm{E}_0[e(V)\Delta^A_{b,0}(V)] \leq \kappa - \varphi_0$. We will show that $\mathrm{E}_0[\underline{e}_0(V)\Delta^Y_{b,0}(V)] \geq \mathrm{E}_0[e(V)\Delta^Y_{b,0}(V)]$, implying that $\underline{e}_0$ is a solution to (3.4).

Observe that

$$\begin{aligned}
&\mathrm{E}_0[\underline{e}_0(V)\Delta^Y_{b,0}(V)] - \mathrm{E}_0[e(V)\Delta^Y_{b,0}(V)] \\
=&\,\mathrm{E}_0[\{\underline{e}_0(V) - e(V)\}\Delta^Y_{b,0}(V)] \\
=&\,\mathrm{E}_0[\{\underline{e}_0(V) - e(V)\}\Delta^Y_{b,0}(V)I(\underline{\xi}_0(V) > \underline{\tau}^E_0)] + \mathrm{E}_0[\{\underline{e}_0(V) - e(V)\}\Delta^Y_{b,0}(V)I(\underline{\xi}_0(V) < \underline{\tau}^E_0)] \\
&+ \mathrm{E}_0[\{\underline{e}_0(V) - e(V)\}\Delta^Y_{b,0}(V)I(\underline{\xi}_0(V) = \underline{\tau}^E_0)]
\end{aligned}$$

Note that $\underline{e}_0(v) = 1 \geq e(v)$ if $\underline{\xi}_0(v) > \underline{\tau}^E_0$ and $\underline{e}_0(v) = 0 \leq e(v)$ if $\underline{\xi}_0(v) < \underline{\tau}^E_0$. Combining this observation with the fact that $\underline{\tau}^E_0 \geq 0$, the above shows that

$$\begin{aligned}
&\mathrm{E}_0[\underline{e}_0(V)\Delta^Y_{b,0}(V)] - \mathrm{E}_0[e(V)\Delta^Y_{b,0}(V)] \\
\geq&\,\underline{\tau}^E_0\,\mathrm{E}_0[\{\underline{e}_0(V) - e(V)\}\Delta^A_{b,0}(V)I(\underline{\xi}_0(V) > \underline{\tau}^E_0)] + \underline{\tau}^E_0\,\mathrm{E}_0[\{\underline{e}_0(V) - e(V)\}\Delta^A_{b,0}(V)I(\underline{\xi}_0(V) < \underline{\tau}^E_0)] \\
&+ \underline{\tau}^E_0\,\mathrm{E}_0[\{\underline{e}_0(V) - e(V)\}\Delta^A_{b,0}(V)I(\underline{\xi}_0(V) = \underline{\tau}^E_0)] \\
=&\,\underline{\tau}^E_0\,\mathrm{E}_0[\{\underline{e}_0(V) - e(V)\}\Delta^A_{b,0}(V)]
\end{aligned}$$

If $\underline{\tau}^E_0 = 0$, then clearly $\mathrm{E}_0[\underline{e}_0(V)\Delta^Y_{b,0}(V)] - \mathrm{E}_0[\underline{e}(V)\Delta^Y_{b,0}(V)] \geq 0$, as desired; otherwise, $\underline{\tau}^E_0 > 0$ and $\mathrm{E}_0[e(V)\Delta^A_{b,0}(V)] \leq \kappa - \varphi_0 = \mathrm{E}_0[\underline{e}_0(V)\Delta^A_{b,0}(V)]$, and so we also have that $\mathrm{E}_0[\underline{e}_0(V)\Delta^Y_{b,0}(V)] \geq \mathrm{E}_0[e(V)\Delta^Y_{b,0}(V)]$. Therefore, $\underline{e}_0$ is an optimal IER. □

Theorem B.8 (Identification of LATE)**.** *Under Conditions A1 to A4, A6a and A6d with* $\delta^Z = \delta^A = 0$, $\mathbb{E}[Y(A(e)) - Y(A(e_r)) \mid A(1) > A(0)] = \frac{\mathrm{E}_0[\{e(V) - e_r(V)\}\,\mathrm{E}_0[\Delta^Y_0(W)|V]]}{\mathrm{E}_0[\Delta^A_0(W)]}$.

Proof. Arguments from Abadie (2003) show that both of the following equalities hold:

$$\mathrm{E}_0[(e(W) - e_r(W))Y(1) \mid A(1) > A(0)] = \frac{\mathrm{E}_0\left[\{e(V) - e_r(V)\}YA\frac{Z - g_0(W)}{g_0(W)(1 - g_0(W))}\right]}{P_0(A(1) > A(0))},$$

$$\mathrm{E}_0[(e(W) - e_r(W))Y(0) \mid A(1) > A(0)] = \frac{\mathrm{E}_0\left[\{e(V) - e_r(V)\}Y(1-A)\frac{-Z+g_0(W)}{g_0(W)(1-g_0(W))}\right]}{P_0(A(1) > A(0))}.$$

Therefore,

$$\begin{aligned}
&\mathbb{E}[Y(A(e)) - Y(A(e_r)) \mid A(1) > A(0)] \\
=&\mathbb{E}[\{e(V) - e_r(V)\}\{Y(A(1)) - Y(A(0))\} \mid A(1) > A(0)] \\
=&\mathbb{E}[\{e(V) - e_r(V)\}\{Y(1) - Y(0)\} \mid A(1) > A(0)] \\
=&\frac{\mathrm{E}_0\left[\{e(V) - e_r(V)\}Y\left\{A\frac{Z-g_0(W)}{g_0(W)(1-g_0(W))} - (1-A)\frac{-Z+g_0(W)}{g_0(W)(1-g_0(W))}\right\}\right]}{P_0(A(1) > A(0))} \\
=&\frac{\mathrm{E}_0\left[\{e(V) - e_r(V)\}Y\frac{Z-g_0(W)}{g_0(W)(1-g_0(W))}\right]}{P_0(A(1) > A(0))} \\
=&\frac{\mathrm{E}_0\left[\{e(V) - e_r(V)\}\,\mathrm{E}_0\left[Y\frac{Z-g_0(W)}{g_0(W)(1-g_0(W))} \mid V\right]\right]}{P_0(A(1) > A(0))},
\end{aligned}$$

where the final equality holds by the law of total expectation. Standard inverse probability weighting arguments show that $\mathrm{E}_0\left[Y\frac{Z-g_0(W)}{g_0(W)(1-g_0(W))} \mid V\right] = \mathrm{E}_0[\Delta_0^Y(W) \mid V]$, and so the numerator above is equal to $\mathrm{E}_0[\{e(V) - e_r(V)\}\,\mathrm{E}_0[\Delta_0^Y(W) \mid V]]$. By Abadie (2003), the denominator above is equal to $\mathrm{E}_0[\Delta_0^A(W)]$. This completes the proof. □

B.6 Overview of targeted minimum loss-based estimation

Our proposed estimators of average effects are targeted minimum loss-based estimators (TMLEs). In this appendix, we briefly introduce the intuition behind this general framework to construct root-n-consistent and asymptotically normal plug-in estimators.

Consider the general problem of estimating a real-valued summary $\Phi(P_0)$ of P_0. Under a suitable differentiability condition on the mapping $\Phi : \mathcal{M} \to \mathbb{R}$ (see, e.g., Chapter 1 of Pfanzagl, 1990), for each $P \in \mathcal{M}$, there exists a gradient $G(P) : \mathcal{O} \to \mathbb{R}$ of Φ at P such that the following first-order expansion holds:

$$\Phi(P) - \Phi(P_0) = -P_0 G(P) + R(P), \tag{B.5}$$

where $R(P)$ is a remainder term that is small if P is close to P_0 in an appropriate sense. We note that the dependencies of R on the parameter Φ, the distribution P_0, and the gradient $G(P)$ are suppressed in the notation.

Let $\tilde{P}_n$ be an estimator of P_0. Under a condition on the rate of convergence of $\tilde{P}_n$ to P_0, the remainder term $R(\tilde{P}_n)$ will be negligible, where throughout this section we refer to a term as "negligible" if it converges to zero in probability faster than $n^{-1/2}$. Under these conditions, (B.5) then shows that

$$\begin{aligned}\Phi(\tilde{P}_n) - \Phi(P_0) =& (P_n - P_0)G(\tilde{P}_n) - P_nG(\tilde{P}_n) + \mathrm{o}_p(n^{-1/2}) \\ =& (P_n - P_0)G(P_0) - P_nG(\tilde{P}_n) + T_n(\tilde{P}_n) + \mathrm{o}_p(n^{-1/2}),\end{aligned}$$

where $T_n(\tilde{P}_n) := (P_n - P_0)[G(\tilde{P}_n) - G(P_0)]$, where the dependence of T_n on P_n, P_0 and G are suppressed in the notation. The term $T_n(\tilde{P}_n)$ will be negligible under an empirical process condition. Consequently, under this condition, the above shows that $\Phi(\tilde{P}_n)$ would be an asymptotically linear estimator of $\Phi(P_0)$ if $P_nG(\tilde{P}_n)$ were negligible.

Unfortunately, when $\tilde{P}_n$ is obtained by the flexible estimation strategies that make it most likely that the remainder $R(\tilde{P}_n)$ will be negligible, it is often the case that $P_nG(\tilde{P}_n)$ is in turn *not* negligible. This motivates the development of TMLE: given an initial estimator $\tilde{P}_n$ of P_0, TMLE fluctuates $\tilde{P}_n$ according to a parametric submodel of $\mathcal{M}$ and obtains a distribution $\hat{P}_n$ that solves $P_nG(P) = 0$ in the distribution P belonging to this submodel, so that $P_nG(\hat{P}_n) = 0$. By standard Z-estimator theory (e.g., van der Vaart and Wellner, 2000), it is typically possible to show that $\hat{P}_n$ is close to $\tilde{P}_n$ in an appropriate sense, thereby ensuring that $R(\hat{P}_n)$ and the $T_n(\hat{P}_n)$ are negligible provided $R(\tilde{P}_n)$ and $T_n(\tilde{P}_n)$ are negligible. This then yields that

$$\Phi(\hat{P}_n) - \Phi(P_0) = (P_n - P_0)G(P_0) - P_nG(\hat{P}_n) + \mathrm{o}_p(n^{-1/2}), \tag{B.6}$$

and asymptotic linearity follows by the fact that $P_nG(\hat{P}_n) = 0$.

In our setting, it will turn out to be easier to construct a TMLE that solves an equation $P_nG_{\tilde{P}_n}(P) = 0$, rather than $P_nG(P) = 0$; here, each $G_{\tilde{P}_n}(P)$, $P \in \mathcal{M}$, is a pseudo-gradient

function that depends on the initial estimate $\tilde{P}_n$ of the distribution P_0 and allows for an expansion of the form given in (B.6): for $P \in \mathcal{M}$ that is close to P_0 in an appropriate sense,

$$\Phi(P) - \Phi(P_0) = (P_n - P_0)G(P_0) - P_n G_{\tilde{P}_n}(P) + R_{\tilde{P}_n}(P) + \mathrm{o}_p(n^{-1/2}), \tag{B.7}$$

where this remainder $R_{\tilde{P}_n}(P)$ equals the remainder term $R(P)$ from (B.5) up to a negligible term, and hence is itself negligible if the $R(P)$ is negligible. We will show that the fluctuation $\hat{P}_n$ defined via solving an equation based on $G_{\tilde{P}_n}(P)$, rather than on $G(P)$, also makes $\Phi(\hat{P}_n)$ an asymptotically linear estimator of $\Phi(P_0)$, by using the expansion in (B.7).

B.7 Derivation of canonical gradients

B.7.1 Preliminaries

In this appendix, we derive the canonical gradients of our estimands. Before doing so, we first briefly review the notion of pathwise differentiability. To do this, it will be useful to define the Hilbert space $L_0^2(P_0)$ of functions $f : \mathcal{O} \to \mathbb{R}$ satisfying $P_0 f^2 < \infty$ and $P_0 f = 0$, where this space has inner product $\langle f, g \rangle = P_0 fg$. Because the model $\mathcal{M}$ that we consider in this work is locally nonparametric at P_0, here we will only discuss pathwise differentiability relative to this locally nonparametric model. More general definitions can be found, for example, in Pfanzagl (1990).

Consider a general mapping $\Phi : \mathcal{M} \to \mathbb{R}$. Define a collection of submodels

$$\Big\{ \{P_{H,\epsilon} : \epsilon \in B_H \subseteq \mathbb{R}\} : H \in \mathcal{H} \Big\} \tag{B.8}$$

for which: (i) $\mathcal{H} \subseteq L_0^2(P_0)$ is such that the $L_0^2(P_0)$-closure of the linear span of $\mathcal{H}$ is $L_0^2(P_0)$; and (ii) each $\{P_{H,\epsilon} : \epsilon \in B_H \subseteq \mathbb{R}\} \subset \mathcal{M}$ is a regular, univariate parametric submodel that passes through P_0 at $\epsilon = 0$ and has score H for ϵ at $\epsilon = 0$. The parameter Φ is pathwise differentiable at P_0 if there exists a function $G(P_0) \in L_0^2(P_0)$, which does not depend on the choice of $H \in \mathcal{H}$, satisfying

$$\frac{d}{d\epsilon} \Phi(P_{H,\epsilon}) \bigg|_{\epsilon=0} = \mathrm{E}_{P_0}[G(P_0)(O)H(O)].$$

The function $G(P_0)$ is P_0-almost surely (a.s.) unique and is referred to as the canonical gradient of Φ at P_0. It can be shown that the pathwise differentiability property is invariant to the choice of the collection of submodels in (B.8) satisfying (i) and (ii): that is, if a parameter is pathwise differentiable with respect to a particular collection of submodels, then it is pathwise differentiable with respect to all such collections; moreover, the expression for the gradient $G(P_0)$ does not depend on the chosen collection.

Throughout this appendix, we will take $\mathcal{H}$ to be the collection of functions H whose range is contained in $[-1, 1]$. We note that the $L_0^2(P_0)$-closure of $\mathcal{H}$ is indeed $L_0^2(P_0)$. For a score $H \in \mathcal{H}$, we will let $H_{Z,W}$ denote $(z, w) \mapsto \mathrm{E}_0[H(O) \mid Z = z, W = w]$. For each $H \in \mathcal{H}$ and $\epsilon \in B_H := (1 - \sqrt{2}, \sqrt{2} - 1)$, we define the $P_{H,\epsilon}$ via its Radon-Nikodym derivative with respect to P_0:

$$\frac{dP_{H,\epsilon}}{dP_0} : o \mapsto [1 + \epsilon H(o) - \epsilon H_{Z,W}(z, w)] [1 + \epsilon H_{Z,W}(z, w)] . \tag{B.9}$$

We note that the interval B_H was chosen to ensure that the right-hand side is positive for all $o \in \mathcal{O}$, which can be shown to hold using the bound on the range of H. It is straightforward to verify that the score function for ϵ at $\epsilon = 0$ is indeed equal to H. In the remainder, we will omit the dependence of $P_{H,\epsilon}$ in the notation and will, in particular, write P_ϵ to refer to $P_{H,\epsilon}$.

We will see that, when applied to a distribution P_ϵ, each of the parameters that we consider writes cleanly as a mapping of some combination of the following marginal or conditional distributions of their input: the marginal distribution $P_{W,\epsilon}$ of W, the marginal distribution $P_{Z,W,\epsilon}$ of (Z, W), the conditional distribution $P_{A,\epsilon}$ of A given (Z, W), and the conditional distribution $P_{Y,\epsilon}$ of Y given (Z, W). As such, it is useful to have closed-form expressions for these features of P_ϵ. These expressions will rely on the functions $H_W : w \mapsto \mathrm{E}_0[H(O) \mid W = w]$, $H_A : (a \mid z, w) \mapsto \mathrm{E}_0[H(O) \mid A = a, Z = z, W = w] - H_{Z,W}(z, w)$, and $H_Y : (y \mid z, w) \mapsto \mathrm{E}_0[H(O) \mid Y = y, Z = z, W = w] - H_{Z,W}(z, w)$. For the distribution defined in (B.9), it can be shown that

$$\frac{dP_{W,\epsilon}}{dP_{W,0}} : w \mapsto 1 + \epsilon H_W(w) \quad \text{and} \quad \frac{dP_{Z,W,\epsilon}}{dP_{Z,W,0}} : (z, w) \mapsto 1 + \epsilon H_{Z,W}(z, w). \tag{B.10}$$

Moreover, for each (z, w), it holds that

$$\frac{dP_{A,\epsilon}}{dP_{A,0}}(\cdot \mid z, w) : a \mapsto 1 + \epsilon H_A(a \mid z, w) \quad \text{and} \quad \frac{dP_{Y,\epsilon}}{dP_{Y,0}}(\cdot \mid z, w) : y \mapsto 1 + \epsilon H_Y(y \mid z, w). \tag{B.11}$$

It will be convenient to note that $\mathrm{E}_0[H_W(W)] = 0$, $\mathrm{E}_0[H_{Z,W}(Z, W)] = 0$, $\mathrm{E}_0[H_A(A \mid Z, W) \mid Z, W] = 0$ P_0-a.s., and $\mathrm{E}_0[H_Y(Y \mid Z, W) \mid Z, W] = 0$ P_0-a.s.

We conclude by describing some notational conventions that we will use in this appendix. We use C to denote a generic positive constant that may vary line by line. We use S_0^E ($\underline{S}_0^E$ resp.) to denote the survival function of the distribution of $\xi_0(V)$ ($\underline{\xi}_0(V)$ resp.) when $V \sim P_0$, and S_0^T to denote the survival function of the distribution of $\Delta_{b,0}(V)$ when $V \sim P_0$. We also use the notation $\lesssim$, which was defined in Appendix B.2 as an inequality up to a positive multiplicative constant that may depend on P_0. For a generic function $f : \mathbb{R} \to \mathbb{R}$, we will use the big- and little-oh notations, namely $\mathrm{O}(f(\epsilon))$ and $\mathrm{o}(f(\epsilon))$, to denote the behavior of terms as $\epsilon \to 0$. Finally, for a general function or quantity f_P that depends on a distribution P, we use f_ϵ to denote f_{P_ϵ}. For example, $\mu_\epsilon^Y := \mu_{P_\epsilon}^Y$ and $\Delta_\epsilon := \Delta_{P_\epsilon}$. We will also write expectations under P_ϵ as E_ϵ.

B.7.2 Canonical gradient of fixed reference rule mean outcome (Theorem 3.3)

It is possible to show that $G_{\mathrm{FR}}^E(P_0)$ ($\underline{G}_{\mathrm{FR}}^E(P_0)$ resp.) is the canonical gradient of $P \mapsto \Psi_{e^{\mathrm{FR}}}^E(P)$ ($\Psi_{\underline{e}^{\mathrm{FR}}}^E(P)$ resp.) using nearly identical arguments to those given in Section 3.4 of Kennedy (2016); as such, these arguments are omitted.

B.7.3 Canonical gradient of true propensity reference rule mean outcome (Theorem 3.3)

Fix a score $H \in \mathcal{H}$. Note that, for all $P \in \mathcal{M}$, $\Psi_{e_P^{\mathrm{TP}}}^E(P) = \int z \Delta_P^Y(w) P_{Z,W}(dz, dw)$. Combining this (B.10), (B.11), and the chain rule yields that

$$\begin{aligned}
\frac{d}{d\epsilon} \Psi_{e_\epsilon^{\mathrm{TP}}}^E(P_\epsilon)\bigg|_{\epsilon=0} &= \int z \, \frac{d}{d\epsilon} \left[\Delta_\epsilon^Y(w) P_{Z,W,\epsilon}(dz, dw)\right]\bigg|_{\epsilon=0} \\
&= \int z \left(\frac{d}{d\epsilon} \Delta_\epsilon^Y(w)\bigg|_{\epsilon=0} \right) P_{Z,W,0}(dz, dw) + \int z \Delta_0^Y(w) \, \frac{d}{d\epsilon} P_{Z,W,\epsilon}(dz, dw)\bigg|_{\epsilon=0}
\end{aligned}$$

$$
\begin{aligned}
&= \int z \left(\frac{I(z=1)}{\mu_P^Z(w)} - \frac{I(z=0)}{1-\mu_P^Z(w)} \right) y H_Y(y \mid z, w) P_0(dy, dz, dw) \\
&\quad + \int z \Delta_0^Y(w) H_{Z,W}(z, w) P_{Z,W,0}(dz, dw) \\
&= \int G^E_{\mathrm{TP}}(P_0)(o) H(o) P_0(do),
\end{aligned}
$$

where the final equality uses that $\mathrm{E}_0[H_Y(Y \mid Z, W) \mid Z, W] = 0$ P_0-a.s. and $\mathrm{E}_0[H_{Z,W}(Z, W)] = 0$. As $H \in \mathcal{H}$ was arbitrary, the above shows that the canonical gradient of $P \mapsto \Psi^E_{e_P^{\mathrm{TP}}}(P)$ at P_0 is $G^E_{\mathrm{TP}}(P_0)$. Since $\underline{e}_P^{\mathrm{TP}} = e_P^{\mathrm{TP}}$, the same result holds for parameter $P \mapsto \Psi^E_{\underline{e}_P^{\mathrm{TP}}}(P)$.

B.7.4 Canonical gradient of randomly distributed reference rule mean outcome in the first setting (Theorem 3.3)

Fix a score $H \in \mathcal{H}$. We will establish that

$$
\left. \frac{d}{d\epsilon} \Psi^E_{e_\epsilon^{\mathrm{RD}}}(P_\epsilon) \right|_{\epsilon=0} = \int G^E_{\mathrm{RD}}(P_0)(o) H(o) P_0(do). \tag{B.12}
$$

As H was arbitrary, this will show that $P \mapsto \Psi^E_{e_P^{\mathrm{RD}}}(P)$ is pathwise differentiable with canonical gradient $G^E_{\mathrm{RD}}(P_0)$ at P_0.

We first note that, by similar arguments to those given in Section 3.4 of Kennedy (2016),

$$
\left. \frac{d}{d\epsilon} P_\epsilon \mu_\epsilon^A(1, \cdot) \right|_{\epsilon=0} = \int \left\{ \frac{z[a - \mu_0^A(1, w)]}{\mu_{P_0}^Z(w)} + \mu_0^A(1, w) - P_0 \mu_0^A(1, \cdot) \right\} H(o) P_0(do). \tag{B.13}
$$

Consequently, $P_\epsilon \mu_\epsilon^A(1, \cdot) = P_0 \mu_0^A(1, \cdot) + \mathrm{O}(\epsilon)$. It follows that, for all ϵ in a neighborhood of zero, Condition C5 implies that $\kappa / P_\epsilon \mu_\epsilon^A(1, \cdot) < 1$. Consequently, for each ϵ in this neighborhood, $\Psi^E_{e_\epsilon^{\mathrm{RD}}}(P_\epsilon) = \frac{\kappa \Psi^E_{v \mapsto 1}(P_\epsilon)}{P_\epsilon \mu_\epsilon^A(1, \cdot)}$, where we have used that $P_\epsilon \Delta_\epsilon^Y = \Psi^E_{v \mapsto 1}(P_\epsilon)$. It follows that the derivative on the right-hand side of (B.12) is the same as the derivative of $f : \epsilon \mapsto \frac{\kappa \Psi^E_{v \mapsto 1}(P_\epsilon)}{P_\epsilon \mu_\epsilon^A(1, \cdot)}$ at $\epsilon = 0$, provided this derivative exists. Noting that $v \mapsto 1$ is a particular instance of a fixed encouragement rule, the results in Section B.7.2 shows that

$$
\left. \frac{d}{d\epsilon} \Psi^E_{v \mapsto 1}(P_\epsilon) \right|_{\epsilon=0} = \int D^E(P_0, v \mapsto 1, 0, \mu_0^a)(o) H(o) P_0(do). \tag{B.14}
$$

As both the above derivative and the derivative in (B.13) exist, we can apply the chain rule to show that

$$\frac{d}{d\epsilon} f(\epsilon)\Big|_{\epsilon=0} = \frac{\kappa}{P_0\mu_0^A(1,\cdot)} \frac{d}{d\epsilon}\Psi^E_{v\mapsto 1}(P_\epsilon)\Big|_{\epsilon=0} - \frac{\kappa\Psi^E_{v\mapsto 1}(P_0)}{P_0\mu_0^A(1,\cdot)^2} \frac{d}{d\epsilon} P_\epsilon\mu_\epsilon^A(1,\cdot)\Big|_{\epsilon=0}.$$

Plugging (B.13) and (B.14) into the above and simplifying shows that the right-hand side of the above is equal to the right-hand side of (B.12). As $\frac{d}{d\epsilon} f(\epsilon)\big|_{\epsilon=0} = \frac{d}{d\epsilon}\Psi^E_{e^{\mathrm{RD}}_\epsilon}(P_\epsilon)\Big|_{\epsilon=0}$, we have shown that (B.12) holds, and thus that $P \mapsto \Psi^E_{e^{\mathrm{RD}}_P}(P)$ is pathwise differentiable at P_0 with canonical gradient $G^E_{\mathrm{RD}}(P_0)$.

B.7.5 Canonical gradient of randomly distributed reference rule mean outcome in the second setting (Theorem 3.3)

Fix a score $H \in \mathcal{H}$. We will establish that

$$\frac{d}{d\epsilon}\Psi^E_{\underline{e}^{\mathrm{RD}}_\epsilon}(P_\epsilon)\Big|_{\epsilon=0} = \int \underline{G}^E_{\mathrm{RD}}(P_0)(o)H(o)P_0(do). \tag{B.15}$$

As H was arbitrary, this will show that $P \mapsto \Psi^E_{\underline{e}^{\mathrm{RD}}_P}(P)$ is pathwise differentiable with canonical gradient $\underline{G}^E_{\mathrm{RD}}(P_0)$ at P_0.

We first note that, by similar arguments to those given in Section 3.4 of Kennedy (2016),

$$\frac{d}{d\epsilon} P_\epsilon\mu_\epsilon^A(0,\cdot)\Big|_{\epsilon=0} = \int \left\{ \frac{1-z}{1-\mu_0^Z(w)}[a-\mu_0^A(0,w)] + \mu_0^A(0,w) - P_0\mu_0^A(0,\cdot) \right\} H(o)P_0(do), \tag{B.16}$$

$$\frac{d}{d\epsilon} P_\epsilon\Delta_\epsilon^A\Big|_{\epsilon=0} = \int \left\{ \frac{1}{z+\mu_0^Z(w)-1}[a-\mu_0^A(z,w)] + \Delta_0^A(w) - P_0\Delta_0^A \right\} H(o)P_0(do). \tag{B.17}$$

Consequently, $P_\epsilon\mu_\epsilon^A(0,\cdot) = P_0\mu_0^A(0,\cdot) + \mathrm{O}(\epsilon)$ and $P_\epsilon\Delta_\epsilon^A = P_0\Delta_0^A + \mathrm{O}(\epsilon)$. It follows that, for all ϵ in a neighborhood of zero, Condition C5 implies that $(\kappa - P_\epsilon\mu_\epsilon^A(0,\cdot))/P_\epsilon\Delta_\epsilon^A < 1$. Consequently, for each ϵ in this neighborhood, $\Psi^E_{\underline{e}^{\mathrm{RD}}_\epsilon}(P_\epsilon) = \frac{\kappa - P_\epsilon\mu_\epsilon^A(0,\cdot)}{P_\epsilon\Delta_\epsilon^A}\Psi^E_{v\mapsto 1}(P_\epsilon)$, where we have used that $P_\epsilon\Delta_\epsilon^Y = \Psi^E_{v\mapsto 1}(P_\epsilon)$. It follows that the derivative on the right-hand side of (B.15) is the same as the derivative of $f : \epsilon \mapsto \frac{\kappa - P_\epsilon\mu_\epsilon^A(0,\cdot)}{P_\epsilon\Delta_\epsilon^A}\Psi^E_{v\mapsto 1}(P_\epsilon)$ at $\epsilon = 0$, provided this

derivative exists. Noting that $v \mapsto 1$ is a particular instance of a fixed encouragement rule, the results in Section B.7.2 shows that

$$\frac{d}{d\epsilon}\Psi^E_{v\mapsto 1}(P_\epsilon)\bigg|_{\epsilon=0} = \int D^E(P_0, v \mapsto 1, 0, \mu_0^a)(o)H(o)P_0(do). \tag{B.18}$$

As both the above derivative and the derivatives in (B.16) and (B.17) exist, we can apply the chain rule to show that

$$\begin{aligned}\frac{d}{d\epsilon}f(\epsilon)\bigg|_{\epsilon=0} = \frac{\kappa - P_0\mu_0^A(0,\cdot)}{P_0\Delta_0^A}\ \frac{d}{d\epsilon}\Psi^E_{v\mapsto 1}(P_\epsilon)\bigg|_{\epsilon=0} - \frac{(\kappa - P_0\mu_0^A(0,\cdot))\Psi^E_{v\mapsto 1}(P_0)}{(P_0\Delta_0^A)^2}\ \frac{d}{d\epsilon}P_\epsilon\Delta_\epsilon^A\bigg|_{\epsilon=0} \\ - \frac{\Psi^E_{v\mapsto 1}(P_0)}{P_0\Delta_0^A}\ \frac{d}{d\epsilon}P_\epsilon\mu_\epsilon^A(0,\cdot)\bigg|_{\epsilon=0}.\end{aligned}$$

Note that $\varphi_P := \Phi(P) = P\mu_P^A(0,\cdot)$. Plugging (B.16), (B.17) and (B.14) into the above and simplifying shows that the right-hand side of the above is equal to the right-hand side of (B.12). As $\frac{d}{d\epsilon}f(\epsilon)\big|_{\epsilon=0} = \frac{d}{d\epsilon}\Psi^E_{\underline{c}^{\mathrm{RD}}_\epsilon}(P_\epsilon)\Big|_{\epsilon=0}$, we have shown that (B.15) holds, and thus that $P \mapsto \Psi^E_{\underline{c}^{\mathrm{RD}}_P}(P)$ is pathwise differentiable at P_0 with canonical gradient $\underline{G}^E_{\mathrm{RD}}(P_0)$.

B.7.6 Canonical gradient of optimal IER mean outcome (Theorem 3.3)

Fix a score $H \in \mathcal{H}$. The argument that we use parallels that of Luedtke and van der Laan (2016a), but modifies the argument from that work to account for the fact that the resource constraint takes a different form in our setting.

Our proof will make use of the following inequalities, all of which hold for all ϵ sufficiently close to zero:

$$\sup_w |\mu_\epsilon^A(1,w) - \mu_0^A(1,w)| \lesssim |\epsilon|, \tag{B.19}$$

$$\sup_w |\Delta_\epsilon^A(w) - \Delta_0^A(w)| \lesssim |\epsilon|, \tag{B.20}$$

$$\sup_v |\Delta_{b,\epsilon}^Y(v) - \Delta_{b,0}^Y(v)| \lesssim |\epsilon|. \tag{B.21}$$

The derivations of these inequalities is straightforward and so are omitted. The first inequality above also readily yields that, for all ϵ close enough to zero,

$$\sup_v |\bar{\mu}_\epsilon^A(v) - \bar{\mu}_0^A(v)| \lesssim |\epsilon|, \tag{B.22}$$

$$\sup_v |\Delta^A_{b,\epsilon}(v) - \Delta^A_{b,0}(v)| \lesssim |\epsilon|. \tag{B.23}$$

Under Condition C4, the above and (B.21) also imply that

$$\sup_v |\xi_\epsilon(v) - \xi_0(v)| = \left| \frac{\Delta^Y_{b,\epsilon}(v)}{\bar{\mu}^A_\epsilon(v)} - \frac{\Delta^Y_{b,0}(v)}{\bar{\mu}^A_0(v)} \right| \lesssim |\epsilon|, \tag{B.24}$$

$$\sup_v |\underline{\xi}_\epsilon(v) - \underline{\xi}_0(v)| = \left| \frac{\Delta^Y_{b,\epsilon}(v)}{\Delta^A_{b,\epsilon}(v)} - \frac{\Delta^Y_{b,0}(v)}{\Delta^A_{b,0}(v)} \right| \lesssim |\epsilon|. \tag{B.25}$$

For $\epsilon \in B_H$, it will be useful to define the $[-\infty, \infty) \to \mathbb{R}$ functions

$$\Gamma_\epsilon : \eta \mapsto \mathrm{E}_\epsilon[I\{\xi_\epsilon(V) > \eta\}\bar{\mu}^A_\epsilon(V)],$$
$$\underline{\Gamma}_\epsilon : \eta \mapsto \mathrm{E}_\epsilon[I\{\underline{\xi}_\epsilon(V) > \eta\}\Delta^A_{b,\epsilon}(V)].$$

Also define $\Gamma'_\epsilon : \eta \mapsto \frac{d}{ds}\Gamma_\epsilon(s)|_{s=\eta}$ and $\underline{\Gamma}'_\epsilon : \eta \mapsto \frac{d}{ds}\underline{\Gamma}_\epsilon(s)|_{s=\eta}$.

We begin with two lemmas. Both of these lemmas refer depend on the H that was fixed at the beginning of this subsection and on the H-dependent submodel $\{P_\epsilon : \epsilon \in (1-\sqrt{2}, \sqrt{2}-1)\}$ introduced in (B.9). The first studies the convergence of η^E_ϵ ($\underline{\eta}^E_n$ reps/) to η^E_0 ($\underline{\eta}^E_0$). Because it may be the case that $\eta^E_0 = -\infty$ ($\underline{\eta}^E_0 = -\infty$ resp.), the convergence stated in this result should be understood as convergence in the extended real number line.

Lemma B.1. *If the conditions of Theorem 3.3 hold, then $\eta^E_\epsilon \to \eta^E_0$ ($\underline{\eta}^E_\epsilon \to \underline{\eta}^E_0$ resp.) as $\epsilon \to 0$.*

Proof of Lemma B.1. We separately consider the cases that $\eta^E_0 > -\infty$ ($\underline{\eta}^E_0 > -\infty$ resp.) and $\eta^E_0 = -\infty$ ($\underline{\eta}^E_0 = -\infty$ resp.).

Suppose that $\eta^E_0 > -\infty$ ($\underline{\eta}^E_0 > -\infty$ resp.). For any small enough $\delta > 0$, for sufficiently small $|\epsilon|$, (B.22) ((B.23) resp.), (B.24) ((B.25) resp.), and the fact that the range of H is contained in $[-1, 1]$ show that

$$\Gamma_\epsilon(\eta^E_0 + \delta) \le (1 + C|\epsilon|)\,\mathrm{E}_0[I\{\xi_\epsilon(V) > \eta^E_0 + \delta\}\bar{\mu}^A_0(V)] \le (1 + C|\epsilon|)\Gamma_0\left(\eta^E_0 + \delta - C|\epsilon|\right)$$

in the first setting, and

$$\underline{\Gamma}_\epsilon(\underline{\eta}^E_0 + \delta) + \varphi_\epsilon \le (1 + C|\epsilon|)\,\mathrm{E}_0[I\{\underline{\xi}_\epsilon(V) > \underline{\eta}^E_0 + \delta\}\Delta^A_{b,0}(V)] + \varphi_\epsilon \le (1 + C|\epsilon|)\underline{\Gamma}_0\left(\underline{\eta}^E_0 + \delta - C|\epsilon|\right) + \varphi_\epsilon$$

in the second setting. Condition C3 shows that, provided δ is small enough, the right-hand side converges to $\Gamma_0(\eta_0^E + \delta)$ ($\underline{\Gamma}_0(\underline{\eta}_0^E + \delta) + \varphi_0$ resp.) as $\epsilon \to 0$. Moreover, Conditions C3 and C4 (A6b rep.) can be combined to show that the derivative of Γ_0 ($\underline{\Gamma}_0$ resp.) is strictly negative for all $x \in [\eta_0^E, \eta_0^E + \delta]$ provided δ is small enough, and so $\Gamma_0(\eta_0^E) > \Gamma_0(\eta_0^E + \delta)$ ($\underline{\Gamma}_0(\underline{\eta}_0^E) > \underline{\Gamma}_0(\underline{\eta}_0^E + \delta)$ resp.). Because $\Gamma_0(\eta_0^E) = \kappa$ ($\underline{\Gamma}_0(\underline{\eta}_0^E) + \varphi_0 = \kappa$ resp.) by the definition of e_0^E ($\underline{e}_0^E$ resp.) under Condition C2, we have shown that, for all ϵ sufficiently close to zero, $\Gamma_\epsilon(\eta_0^E + \delta) < \kappa$ ($\underline{\Gamma}_\epsilon(\underline{\eta}_0^E + \delta) + \varphi_\epsilon < \kappa$ resp.). Recalling the definition $\eta_\epsilon^E := \inf\{\eta : \Gamma_\epsilon(\eta) \le \kappa\}$ ($\underline{\eta}_\epsilon^E := \inf\{\eta : \underline{\Gamma}_\epsilon(\eta) + \varphi_\epsilon \le \kappa\}$ resp.), we see that, for all ϵ sufficiently close to zero, $\eta_0^E + \delta \ge \eta_\epsilon^E$ ($\underline{\eta}_0^E + \delta \ge \underline{\eta}_\epsilon^E$ resp.), that is, that $\eta_\epsilon^E - \eta_0^E \le \delta$ ($\underline{\eta}_\epsilon^E - \underline{\eta}_0^E \le \delta$ resp.).

Similar arguments show that, for all ϵ sufficiently close to zero, $\eta_\epsilon^E - \eta_0^E \ge -\delta$ ($\underline{\eta}_\epsilon^E - \underline{\eta}_0^E \ge -\delta$ resp.). Indeed,

$$\Gamma_\epsilon(\eta_0^E - \delta) \ge (1 - C|\epsilon|)\Gamma_0(\eta_0^E - \delta + C|\epsilon|)$$

in the first setting, and

$$\underline{\Gamma}_\epsilon(\underline{\eta}_0^E - \delta) + \varphi_\epsilon \ge (1 - C|\epsilon|)\underline{\Gamma}_0(\underline{\eta}_0^E - \delta + C|\epsilon|) + \varphi_\epsilon.$$

in the second setting. The right-hand side converges to $\Gamma_0(\eta_0^E - \delta)$ ($\underline{\Gamma}_0(\underline{\eta}_0^E - \delta) + \varphi_0$ resp.) as $\epsilon \to 0$ provided δ is small enough. The derivative of Γ_0 ($\underline{\Gamma}_0$ resp.) is strictly negative on $[\eta_0^E - \delta, \eta_0^E]$ provided δ is small enough, and so $\Gamma_0(\eta_0^E - \delta) > \Gamma_0(\eta_0^E) = \kappa$ ($\underline{\Gamma}_0(\underline{\eta}_0^E - \delta) + \varphi_0 > \underline{\Gamma}_0(\underline{\eta}_0^E) + \varphi_0 = \kappa$ resp.) in this case. Hence, $\Gamma_\epsilon(\eta_0^E - \delta) > \kappa$ ($\underline{\Gamma}_\epsilon(\underline{\eta}_0^E - \delta) + \varphi_\epsilon > \kappa$ resp.). Recalling that $\eta_\epsilon^E := \inf\{\eta : \Gamma_\epsilon(\eta) \le \kappa\}$ ($\underline{\eta}_\epsilon^E := \inf\{\eta : \underline{\Gamma}_\epsilon(\eta) + \varphi_\epsilon \le \kappa\}$ resp.), we see that indeed $\eta_\epsilon^E - \eta_0^E \ge -\delta$ ($\underline{\eta}_\epsilon^E - \underline{\eta}_0^E \ge -\delta$ resp.). Combining this with the result from the preceding paragraph shows that, for all ϵ sufficiently close to zero, $|\eta_\epsilon^E - \eta_0^E| \le \delta$ ($|\underline{\eta}_\epsilon^E - \underline{\eta}_0^E| \le \delta$ resp.). Hence, $\limsup_{\epsilon\to 0} |\eta_\epsilon^E - \eta_0^E| \le \delta$ ($\limsup_{\epsilon\to 0} |\underline{\eta}_\epsilon^E - \underline{\eta}_0^E| \le \delta$ resp.). As δ is an arbitrary value that is sufficiently close to zero, it follows that $\limsup_{\epsilon\to 0} |\eta_\epsilon^E - \eta_0^E| = 0$ ($\limsup_{\epsilon\to 0} |\underline{\eta}_\epsilon^E - \underline{\eta}_0^E| = 0$ resp.). That is, $\eta_\epsilon^E \to \eta_0^E$ ($\underline{\eta}_\epsilon^E \to \underline{\eta}_0^E$ resp.) as $\epsilon \to 0$ in the case that $\eta_0^E > -\infty$ ($\underline{\eta}_0^E > -\infty$ resp.).

Now suppose that $\eta_0^E = -\infty$ ($\underline{\eta}_0^E = -\infty$ resp.). If $\kappa = 1$, then it is trivially true that $\eta_\epsilon^E = -\infty = \eta_0^E$ ($\underline{\eta}_\epsilon^E = -\infty = \underline{\eta}_0^E$ resp.) for all ϵ, and so the desired convergence holds.

Suppose now that $\kappa < 1$. Fix a small enough $\delta > 0$ so that the bound in (B.24) ((B.25) resp.) is valid for all $\epsilon \in [-\delta, \delta]$. Also fix $\epsilon \in [-\delta, \delta]$ and $\eta \in \mathbb{R}$. In the next paragraph, we will show that $\Gamma_\epsilon(\eta) < \kappa$ ($\underline{\Gamma}_\epsilon(\eta) + \varphi_\epsilon < \kappa$ resp.) for all ϵ sufficiently close to zero. This will then imply that $\eta_\epsilon^E \leq \eta$ ($\underline{\eta}_\epsilon^E \leq \eta$ resp.) for all ϵ sufficiently close to zero. The fact that $\eta \in \mathbb{R}$ was arbitrary will then show that $\eta_\epsilon^E \to \eta_0^E = -\infty$ ($\underline{\eta}_\epsilon^E \to \underline{\eta}_0^E = -\infty$), establishing the result.

By (B.24) ((B.25) resp.) and the bound on the range of H,

$$\Gamma_\epsilon(\eta) \leq (1 + C|\epsilon|)\,\mathrm{E}_0[I\{\xi_\epsilon(V) > \eta\}\bar{\mu}_0^A(V)] \leq (1 + C|\epsilon|)\Gamma_0(\eta - C|\epsilon|)$$

in the first setting, and

$$\underline{\Gamma}_\epsilon(\eta) + \varphi_\epsilon \leq (1 + C|\epsilon|)\,\mathrm{E}_0[I\{\underline{\xi}_\epsilon(V) > \eta\}\Delta_{0,b}^A(V)] + \varphi_\epsilon \leq (1 + C|\epsilon|)\underline{\Gamma}_0(\eta - C|\epsilon|) + \varphi_\epsilon.$$

in the second setting. Because Γ_0 ($\underline{\Gamma}_0$ resp.) is a nonnegative decreasing function, the right-hand side is upper bounded by $(1 + C|\epsilon|)\Gamma_0(\eta - C\delta)$ ($(1 + C|\epsilon|)\underline{\Gamma}_0(\eta - C\delta) + \varphi_\epsilon$ resp.). This upper bound tends to $\Gamma_0(\eta - C\delta)$ ($\underline{\Gamma}_0(\eta - C\delta) + \varphi_0$ resp.) as $\epsilon \to 0$. Hence, $\limsup_{\epsilon \to 0} \Gamma_\epsilon(\eta) \leq \Gamma_0(\eta - C\delta)$ ($\limsup_{\epsilon \to 0} \underline{\Gamma}_\epsilon(\eta) + \varphi_\epsilon \leq \underline{\Gamma}_0(\eta - C\delta) + \varphi_0$ resp.). By Condition C3 and the monotonicity of Γ_0 ($\underline{\Gamma}_0$ resp.), $\Gamma_0(\eta - C\delta) < \kappa$ ($\underline{\Gamma}_0(\eta - C\delta) + \varphi_0 < \kappa$ resp.), and so $\Gamma_\epsilon(\eta) < \kappa$ ($\underline{\Gamma}_\epsilon(\eta) + \varphi_\epsilon < \kappa$ resp.) for all ϵ sufficiently close to zero. By the definition of η_ϵ^E ($\underline{\eta}_\epsilon^E$ resp.), it follows that $\eta_\epsilon^E \leq \eta$ ($\underline{\eta}_\epsilon^E \leq \eta$ resp.) for all ϵ sufficiently close to zero, as desired. □

The next lemma establishes a rate of convergence of τ_ϵ^E ($\underline{\tau}_\epsilon^E$ resp.) to τ_0^E ($\underline{\tau}_0^E$ resp.) as $\epsilon \to 0$.

Lemma B.2. *If the conditions of Theorem 3.3 hold, then* $\tau_\epsilon^E = \tau_0^E + \mathrm{O}(\epsilon)$ ($\underline{\tau}_\epsilon^E = \underline{\tau}_0^E + \mathrm{O}(\epsilon)$ *resp.).*

Proof of Lemma B.2. We will separately consider the cases that $\eta_0^E < 0$ ($\underline{\eta}_0^E < 0$ resp.) and $\eta_0^E \geq 0$ ($\underline{\eta}_0^E \geq 0$ resp.). We start with the easier case. Suppose that $\eta_0^E < 0$ ($\underline{\eta}_0^E < 0$ resp.). In this case, Lemma B.1 shows that $\tau_\epsilon^E := \max\{\eta_\epsilon^E, 0\}$ ($\underline{\tau}_\epsilon^E := \max\{\underline{\eta}_\epsilon^E, 0\}$ resp.) is equal

to $\tau_0^E = 0$ ($\underline{\tau}_0^E = 0$ resp.) for all ϵ in a neighborhood of zero. Thus, $\tau_\epsilon^E - \tau_0^E = \mathrm{O}(|\epsilon|)$ ($\underline{\tau}_\epsilon^E - \underline{\tau}_0^E = \mathrm{O}(|\epsilon|)$ resp.).

Now suppose that $\eta_0^E \geq 0$ ($\underline{\eta}_0^E \geq 0$ resp.). The Lipschitz property of the function $x \mapsto \max\{x, 0\}$ can be used to show that $|\max\{\eta_\epsilon^E, 0\} - \max\{\eta_0^E, 0\}| \leq |\eta_\epsilon^E - \eta_0^E|$ ($|\max\{\underline{\eta}_\epsilon^E, 0\} - \max\{\underline{\eta}_0^E, 0\}| \leq |\underline{\eta}_\epsilon^E - \underline{\eta}_0^E|$ resp.). As a consequence, to show that $\tau_\epsilon^E - \tau_0^E = \mathrm{O}(\epsilon)$ ($\underline{\tau}_\epsilon^E - \underline{\tau}_0^E = \mathrm{O}(\epsilon)$ resp.), it suffices to show that $\eta_\epsilon^E - \eta_0^E = \mathrm{O}(\epsilon)$ ($\underline{\eta}_\epsilon^E - \underline{\eta}_0^E = \mathrm{O}(\epsilon)$ resp.). The remainder of this proof establishes this fact.

Fix ϵ in a neighborhood of zero. By the definition $\eta_\epsilon^E := \inf\{\eta : \Gamma_\epsilon(\eta) \leq \kappa\}$ ($\underline{\eta}_\epsilon^E := \inf\{\eta : \underline{\Gamma}_\epsilon(\eta) \leq \kappa - \varphi_0\}$), the bound on the range of H, and (B.24) ((B.25) resp.), it holds that $\kappa < \Gamma_\epsilon(\eta_\epsilon^E - |\epsilon|) \leq [1 + C|\epsilon|]\Gamma_0(\eta_\epsilon^E - [1+C]|\epsilon|)$ ($\kappa < \underline{\Gamma}_\epsilon(\underline{\eta}_\epsilon^E - |\epsilon|) + \varphi_\epsilon \leq [1 + C|\epsilon|]\underline{\Gamma}_0(\underline{\eta}_\epsilon^E - [1+C]|\epsilon|) + \varphi_\epsilon$ resp.). A Taylor expansion of Γ_0 about η_0^E, which is justified by Condition C3 provided $|\epsilon|$ is small enough, then shows that

$$\kappa < [1 + C|\epsilon|]\left[\Gamma_0(\eta_0^E) + \{\eta_\epsilon^E - \eta_0^E - (1 - C)|\epsilon|\}\{\Gamma_0'(\eta_0^E) + \mathrm{o}(1)\}\right]$$

in the first setting and

$$\kappa < [1 + C|\epsilon|]\left[\underline{\Gamma}_0(\underline{\eta}_0^E) + \{\underline{\eta}_\epsilon^E - \underline{\eta}_0^E - (1 - C)|\epsilon|\}\{\underline{\Gamma}_0'(\underline{\eta}_0^E) + \mathrm{o}(1)\}\right] + \varphi_0 + \mathrm{O}(\epsilon)$$

in the second setting. By Condition C3, $\Gamma_0(\eta_0^E) = \kappa$ ($\underline{\Gamma}_0(\underline{\eta}_0^E) + \varphi_0 = \kappa$). Plugging this into the above and rearranging shows that

$$0 < C\kappa|\epsilon| + [1 + C|\epsilon|]\left[\eta_\epsilon^E - \eta_0^E - (1 - C)|\epsilon|\right]\left[\Gamma_0'(\eta_0^E) + \mathrm{o}(1)\right]$$

in the first setting, and

$$0 < C\underline{\Gamma}_0(\underline{\eta}_0^E)|\epsilon| + [1 + C|\epsilon|]\left[\underline{\eta}_\epsilon^E - \underline{\eta}_0^E - (1 - C)|\epsilon|\right]\left[\underline{\Gamma}_0'(\underline{\eta}_0^E) + \mathrm{o}(1)\right] + \mathrm{O}(\epsilon)$$

in the second setting. Using that Condition C3 implies that $\Gamma_0'(\eta_0^E) \in (-\infty, 0)$ ($\underline{\Gamma}_0'(\underline{\eta}_0^E) \in (-\infty, 0)$ resp.), the above shows that, for all ϵ sufficiently close to zero, $0 < [\eta_\epsilon^E - \eta_0^E]\Gamma_0'(\eta_0^E) + C|\epsilon| + \mathrm{o}\left(\eta_\epsilon^E - \eta_0^E\right)$ ($0 < [\underline{\eta}_\epsilon^E - \underline{\eta}_0^E]\underline{\Gamma}_0'(\underline{\eta}_0^E) + C|\epsilon| + \mathrm{o}(\underline{\eta}_\epsilon^E - \underline{\eta}_0^E)$ resp.), which implies that there exists an $\mathrm{O}(\epsilon)$ sequence for which $\eta_\epsilon^E - \eta_0^E < \mathrm{O}(\epsilon)$ ($\underline{\eta}_\epsilon^E - \underline{\eta}_0^E < \mathrm{O}(\epsilon)$ resp.).

A similar argument, which is based on the observation that $\Gamma_\epsilon(\eta_\epsilon^E + |\epsilon|) \leq \kappa$ ($\underline{\Gamma}_\epsilon(\underline{\eta}_\epsilon^E + |\epsilon|) \leq \kappa$ resp.), can be used to show that there exists an $\mathrm{O}(\epsilon)$ sequence such that $\eta_\epsilon^E - \eta_0^E > \mathrm{O}(\epsilon)$ ($\underline{\eta}_\epsilon^E - \underline{\eta}_0^E > \mathrm{O}(\epsilon)$ resp.). Combining these two bounds shows that $\eta_\epsilon^E - \eta_0^E = \mathrm{O}(\epsilon)$ ($\underline{\eta}_\epsilon^E - \underline{\eta}_0^E = \mathrm{O}(\epsilon)$ resp.), as desired. This concludes the proof. □

Our derivation of the canonical gradient in the first setting will rely on the following decomposition:

$$
\begin{aligned}
&\Psi_{e_\epsilon}^E(P_\epsilon) - \Psi_{e_0}^E(P_0) \\
&= \Psi_{e_\epsilon}^E(P_\epsilon) - \Psi_{e_0}^E(P_\epsilon) + \Psi_{e_0}^E(P_\epsilon) - \Psi_{e_0}^E(P_0) \\
&= P_\epsilon\{[e_\epsilon - e_0]\Delta_{b,\epsilon}^Y\} + \Psi_{e_0}^E(P_\epsilon) - \Psi_{e_0}^E(P_0) \\
&= P_\epsilon\{[e_\epsilon - e_0](\Delta_{b,\epsilon}^Y - \tau_0^E\bar{\mu}_0^A)\} + \tau_0^E P_\epsilon\{(e_\epsilon - e_0)\bar{\mu}_0^A\} + \Psi_{e_0}^E(P_\epsilon) - \Psi_{e_0}^E(P_0) \\
&= P_\epsilon\{[e_\epsilon - e_0](\Delta_{b,\epsilon}^Y - \tau_0^E\bar{\mu}_0^A)\} + [\Psi_{e_0}^E(P_\epsilon) - \Psi_{e_0}^E(P_0)] + \tau_0^E\{P_\epsilon[\bar{\mu}_\epsilon^A e_\epsilon] - P_0[\bar{\mu}_0^A e_0]\} \\
&\quad - \tau_0^E P_\epsilon\{(\bar{\mu}_\epsilon^A - \bar{\mu}_0^A)e_0\} - \tau_0^E(P_\epsilon - P_0)\{\bar{\mu}_0^A e_0\}. \qquad \text{(B.26)}
\end{aligned}
$$

In what follows, we will separately study each of the five terms on the right-hand side, which we refer to as terms 1, 2, 3, 4, and 5.

Study of term 1 in (B.26)**:** We will show that this term is $\mathrm{o}(\epsilon)$. By Lemma B.2 and (B.21),

$$\sup_v |\Delta_{b,\epsilon}^Y(v) - \tau_\epsilon^E - \Delta_{b,0}^Y(v) + \tau_0^E| \leq \sup_v |\Delta_{b,\epsilon}^Y(v) - \Delta_{b,0}^Y(v)| + |\tau_\epsilon^E - \tau_0^E| \lesssim |\epsilon|.$$

Under Condition C2, which implies that $P_0\{\xi_0(V) = \tau_0^E\} = 0$, a similar argument as that used to prove Lemma 2 in van der Laan and Luedtke (2014) shows that

$$
\begin{aligned}
&\left|P_\epsilon\{[e_\epsilon - e_0](\Delta_{b,\epsilon}^Y - \tau_0^E\bar{\mu}_0^A)\}\right| \\
&\quad = \left|\int [e_\epsilon(v) - e_0(v)][\Delta_{b,\epsilon}^Y(v) - \tau_0^E\bar{\mu}_0^A(v)]P_{W,\epsilon}(dw)\right| \\
&\quad \leq \int |e_\epsilon(v) - e_0(v)|\left|\Delta_{b,\epsilon}^Y(v) - \tau_0^E\bar{\mu}_0^A(v)\right| P_{W,\epsilon}(dw).
\end{aligned}
$$

Because $e_\epsilon(v) \neq e_0(v)$ implies that either (i) $\xi_\epsilon(v) - \tau_\epsilon^E$ and $\xi_0(v) - \tau_0^E$ have different signs or (ii) only one of these quantities is zero, the display continues as

$$\leq \int I\{|\xi_0(v) - \tau_0^E| \leq |\xi_\epsilon(v) - \tau_\epsilon^E - \xi_0(v) + \tau_0^E|\}\left|\Delta_{b,\epsilon}^Y(v) - \tau_0^E\bar{\mu}_0^A(v)\right| P_{W,\epsilon}(dw)$$

$$\leq \int I\{|\xi_0(v) - \tau_0^E| \leq C|\epsilon|\} \left(\left|\Delta_{b,0}^Y(v) - \tau_0^E \bar{\mu}_0^A(v)\right| + C|\epsilon|\right) P_{W,\epsilon}(dw).$$

Using that $\inf_v \bar{\mu}_0^A(v) > 0$ by Condition C4, that $\sup_v \bar{\mu}_0^A(v) \leq 1$ since probabilities are no more than unity, and that $\xi_0(v) := \Delta_{b,0}^Y(v)/\bar{\mu}_0^A(v)$, the display continues as

$$\leq \int I\{|\xi_0(v) - \tau_0^E| \leq C|\epsilon|\} \left(\left|\xi_0(v) - \tau_0^E\right| + C|\epsilon|\right) P_{W,\epsilon}(dw)$$

Leveraging the bound on $|\xi_0(v) - \tau_0^E|$ that appears in the indicator function, we see that

$$\begin{aligned}
&\leq \int I\{|\xi_0(v) - \tau_0^E| \leq C|\epsilon|\}(C|\epsilon| + C|\epsilon|)P_{W,\epsilon}(dw) \\
&\lesssim |\epsilon| \int I\{|\xi_0(v) - \tau_0^E| \leq C|\epsilon|\}P_{W,0}(dw) \\
&= |\epsilon| \int I\{0 < |\xi_0(v) - \tau_0^E| \leq C|\epsilon|\}P_{W,0}(dw),
\end{aligned}$$

where the final equality holds by Condition C1. The integral in the final expression is o(1), and so this expression is o(ϵ).

Study of term 2 in (B.26)**:** By the result in Section B.7.2, the second term satisfies $\Psi_{e_0}^E(P_\epsilon) - \Psi_{e_0}^E(P_0) = \epsilon \int G_2^E(o)H(o)P(do) + \mathrm{o}(\epsilon)$, where $G_2^E \in L_0^2(P_0)$ is equal to $D^E(P_0, e_0, 0, \mu_0^A)$.

Study of term 3 in (B.26)**:** We will show that the third term is zero for all ϵ that are sufficiently close to zero. If $\tau_{P_0} = 0$, then this term is obviously zero, so suppose that $\tau_0^E = \eta_0^E > 0$. Lemma B.1 shows that, in this case, $\eta_\epsilon^E > 0$ for ϵ sufficiently close to zero. Hence, $\mathrm{E}_\epsilon[\bar{\mu}_\epsilon^A(V)e_\epsilon(V)] = \kappa = \mathrm{E}_0[\bar{\mu}_0^A(V)e_0(V)]$. Consequently, this third term is indeed equal to zero for all ϵ sufficiently close to zero.

Study of term 4 in (B.26)**:** We will show that this term writes as $\epsilon \int G_4^E(o)H(o)P_0(do)+\mathrm{o}(\epsilon)$ for an appropriately defined $G_4^E \in L_0^2(P_0)$ that does not depend on H. We start by noting that the perturbation implies that there exists a function $H_W : (w \mid v) \mapsto H_W(w \mid v)$ for which $\int H_W(w \mid v)P_{W,0}(dw \mid v) = 0$, $\sup_{w,v} |H_W(w \mid v)| < \infty$, and, for all v,

$$P_{W,\epsilon}(dw \mid v) = (1 + \epsilon H_W(w \mid v) + \mathrm{o}(\epsilon))P_{W,0}(dw \mid v).$$

The function H_W can be chosen so that the above $\mathrm{o}(\epsilon)$ term indicates little-oh behavior uniformly over w and v. Also recalling the definition of H_A from above (B.11), we see that

$$\begin{aligned}\bar{\mu}_\epsilon^A(v) - \bar{\mu}_0^A(v) =& \iint a\Big\{[1+\epsilon H_A(a \mid 1,w)][1+\epsilon H_W(w \mid v)+\mathrm{o}(\epsilon)] - 1\Big\}P_0(da \mid 1,w)P_0(dw \mid v) \\ =& \epsilon \iint a(H_A(a \mid 1,w) + H_W(w \mid v) + \mathrm{o}(1))P_0(da \mid 1,w)P_0(dw \mid v) + \mathrm{o}(\epsilon),\end{aligned}$$

where the little-oh terms are uniform over w and v. Hence,

$$\begin{aligned}&\frac{d}{d\epsilon}P_\epsilon\{[\bar{\mu}_\epsilon^A - \bar{\mu}_0^A]e_0\}\Big|_{\epsilon=0} \\ &= \iint e_0(v)a\{H_A(a \mid 1,w) + H_W(w \mid v)\}P_0(da \mid 1,w)P_0(dw) \\ &= \mathrm{E}_0\left[e_0(V)\left(\frac{Z}{\mu_0^Z(W)}\{A - \mu_0^A(1,W)\}H_A(A \mid 1,W) + \{\mu_0^A(1,W) - \bar{\mu}_0^A(V)\}H_W(W \mid V)\right)\right].\end{aligned}$$

Using that $\mathrm{E}_0[H_A(A \mid Z,W) \mid Z,W] = \mathrm{E}_0[H_W(W \mid V) \mid V] = 0$ P_0-a.s., the display continues as

$$= \mathrm{E}_0\left[e_0(V)\left(\frac{Z}{\mu_0^Z(W)}\{A - \mu_0^A(1,W)\} + \mu_0^A(1,W) - \bar{\mu}_0^A(V)\right)H(O)\right].$$

As a consequence of the above display, the fourth term satisfies

$$-\tau_0^E P_\epsilon\{[\bar{\mu}_\epsilon^A - \bar{\mu}_0^A]e_0\} = \epsilon \int G_4^E(o)H(o)P_0(do) + \mathrm{o}(\epsilon),$$

where

$$G_4^E : o \mapsto -\tau_0^E e_0(v)\left\{\frac{z}{\mu_0^Z(w)}[I(a=1) - \mu_0^A(1,w)] + \mu_0^A(1,w) - \bar{\mu}_{P_0}(v)\right\}.$$

Note that $P_0 G_4^E = 0$, and that $P_0[(G_4^E)^2] < \infty$ under Condition A3.

Study of term 5 in (B.26)**:** By (B.10) and the fact that $P_0\{\bar{\mu}_0^A(\cdot)e_0\} = \kappa$ whenever $\tau_0^E > 0$, we see that the fourth term satisfies $-\tau_0^E(P_\epsilon - P_0)\{\bar{\mu}_0^A e_0\} = \epsilon \int G_5^E(o)H_V(v)P_0(do)$, where $G_5^E \in L_0^2(P_0)$ is defined as $o \mapsto -\tau_0^E[\bar{\mu}_0^A(v)e_0(v) - \kappa]$. Using that H_V is defined as $v \mapsto \mathrm{E}_0[H(O) \mid V = v]$, we see that it also holds that $-\tau_0^E(P_\epsilon - P_0)\{\bar{\mu}_0^A e_0\} = \epsilon \int G_5^E(o)H(o)P_0(do)$.

Conclusion of the derivation of the canonical gradient of $P \mapsto \Psi^E_{e_P}(P)$: Combining our results regarding the five terms in (B.26), we see that

$$\Psi^E_{e_\epsilon}(P_\epsilon) - \Psi^E_{e_0}(P_0) = \epsilon \int \left[G^E_2(o) + G^E_4(o) + G^E_5(o)\right] H(o)P_0(do) + \mathrm{o}(\epsilon).$$

Dividing both sides by $\epsilon \neq 0$ and taking the limit as $\epsilon \to 0$, we see that $G^E_2 + G^E_4 + G^E_5$ is the canonical gradient of $P \mapsto \Psi^E_{e_P}(P)$ at P_0. The proof concludes by noting that $G^E_2 + G^E_4 + G^E_5$ is equal to $G^E(P_0)$.

Our derivation of the canonical gradient in the second setting is similar and hence arguments are slightly abbreviated. We consider the following decomposition:

$$\begin{aligned}
&\Psi^E_{\underline{e}_\epsilon}(P_\epsilon) - \Psi^E_{\underline{e}_0}(P_0) \\
&= P_\epsilon\{[\underline{e}_\epsilon - \underline{e}_0](\Delta^Y_{b,\epsilon} - \underline{\tau}^E_0 \Delta^A_{b,0})\} + [\Psi^E_{\underline{e}_0}(P_\epsilon) - \Psi^E_{\underline{e}_0}(P_0)] + \underline{\tau}^E_0\{P_\epsilon[\Delta^A_{b,\epsilon}\underline{e}_\epsilon] + \varphi_\epsilon - P_0[\Delta^A_{b,0}\underline{e}_0] - \varphi_0\} \\
&\quad - \underline{\tau}^E_0 P_\epsilon\{(\Delta^A_{b,\epsilon} - \Delta^A_{0,b})\underline{e}_0\} - \underline{\tau}^E_0(P_\epsilon - P_0)\{\Delta^A_{b,0}\underline{e}_0\} - \underline{\tau}^E_0\{\varphi_\epsilon - \varphi_0\}. \qquad \text{(B.27)}
\end{aligned}$$

By similar arguments as the first setting, term 1 is $\mathrm{o}(\epsilon)$; term 2 is $\epsilon \int \underline{G}^E_2(o)H(o)P(do) + \mathrm{o}(\epsilon)$, where $\underline{G}^E_2 \in L^2_0(P_0)$ is equal to $\underline{D}^E(P_0, e_0, 0, \mu^A_0)$; term 3 is zero; term 4 is $\epsilon \int \underline{G}^E_4(o)H(o)P_0(do) + \mathrm{o}(\epsilon)$ where

$$\underline{G}^E_4 : o \mapsto -\underline{\tau}^E_0 \underline{e}_0(v)\left\{\frac{1}{z + \mu^Z_0(w) - 1}[a - \mu^A_0(z,w)] + \Delta^A_0(w) - \Delta^A_{b,0}(v)\right\};$$

term 5 is $\epsilon \int \underline{G}^E_5(o)H(o)P_0(do)$ where $\underline{G}^E_5 : o \mapsto -\underline{\tau}^E_0[\Delta^A_{b,0}(v)\underline{e}_0(v) - \kappa + \varphi_0]$.

Study of term 6 in (B.27)**:** We have shown that

$$\varphi_\epsilon - \varphi_0 = \epsilon \int \left\{\frac{1-z}{1-\mu^Z_0(w)}[a - \mu^A_0(0,w)] + \mu^A_0(0,w) - \varphi_0\right\} H(o)P_0(do) + \mathrm{o}(\epsilon).$$

Therefore, $-\underline{\tau}^E_0(\varphi_\epsilon - \varphi_0) = \epsilon \int \underline{G}^E_6(o)H(o)P_0(do) + \mathrm{o}(\epsilon)$ where

$$\underline{G}^E_6 : o \mapsto -\underline{\tau}^E_0\left\{\frac{1-z}{1-\mu^Z_0(w)}[a - \mu^A_0(0,w)] + \mu^A_0(0,w) - \varphi_0\right\}.$$

Conclusion of the derivation of the canonical gradient of $P \mapsto \Psi^E_{\underline{e}_P}(P)$: Combining our results regarding the six terms in (B.27), we see that

$$\Psi^E_{\underline{e}_\epsilon}(P_\epsilon) - \Psi^E_{\underline{e}_0}(P_0) = \epsilon \int \left[\underline{G}^E_2(o) + \underline{G}^E_4(o) + \underline{G}^E_5(o) + \underline{G}^E_6(o)\right] H(o)P_0(do) + \mathrm{o}(\epsilon).$$

Thus, $\underline{G}_2^E + \underline{G}_4^E + \underline{G}_5^E + \underline{G}_6^E$ is the canonical gradient of $P \mapsto \Psi_{\underline{e}_P}^E(P)$ at P_0. The proof concludes by noting that $\underline{G}_2^E + \underline{G}_4^E + \underline{G}_5^E + \underline{G}_6^E$ is equal to $\underline{G}^E(P_0)$.

B.7.7 Canonical gradient of fixed treatment reference rule mean outcome (Theorem 3.1)

It is possible to show that $D^T(P_0, \mu_0^Y, \mu_0^A, t, 0)$ is the canonical gradient of $P \mapsto \Psi_t^T(P)$ at P_0 using nearly identical arguments to those given in Wang and Tchetgen Tchetgen (2018); as such, these arguments are omitted.

B.7.8 Canonical gradient of optimal ITR mean outcome (Theorem 3.1)

Fix a score $H \in \mathcal{H}$. The argument that we use parallels that of Luedtke and van der Laan (2016a), but modifies the argument from that work to account for the fact that our fixed-ITR parameter $P \mapsto \Psi_t^T(P)$ takes a different form from the fixed-ITR parameter considered in that earlier work. The argument is also similar to the one that we used in Appendix B.7.6. We therefore slightly abbreviated some arguments here.

It will be useful to note that the choice of submodel $\{P_\epsilon : \epsilon \in B_H\}$ ensures that, for all ϵ sufficiently close to zero,

$$\sup_v |\Delta_{b,\epsilon}(v) - \Delta_{b,0}(v)| \lesssim |\epsilon|. \tag{B.28}$$

We now present two lemmas, which are analogues of Lemmas B.1 and B.2.

Lemma B.3. *If the conditions of Theorem 3.1 hold, then $\eta_\epsilon^T \to \eta_0^T$ as $\epsilon \to 0$.*

Proof of Lemma B.3. If $\kappa = 1$, then, for all $\epsilon \in B_H$, $\eta_\epsilon^T = -\infty$, and so $\eta_\epsilon^T = \eta_0^T$ for all ϵ. Thus, the desired convergence clearly holds. Suppose now that $\kappa < 1$. In this case, it is necessarily true that, for all $\epsilon \in B_H$, the $(1-\kappa)$-quantile of $\Delta_{b,\epsilon}(V)$, $V \sim P_\epsilon$, is finite, that is, that η_ϵ^T is finite. The bound on the range of H, (B.10), and (B.28) show that, for any sufficiently small $\delta > 0$ and ϵ sufficiently close to zero,

$$S_\epsilon^T(\eta_0^T - \delta) \geq (1 - C|\epsilon|)P_0\{\Delta_{b,\epsilon}(V) > \eta_0^T - \delta\} \geq (1 - C|\epsilon|)S_0^T(\eta_0^T - \delta - C|\epsilon|).$$

Condition B2 implies that S_0^T is continuous in a neighborhood of η_0^T. Therefore, the right-hand side converges to $S_0^T(\eta_0^T - \delta)$ as ϵ converges to zero. Using that $S_0^T(\eta_0^T - \delta) > \kappa$ by Condition B2, we see that $S_\epsilon^T(\eta_0^T - \delta) > \kappa$ for all sufficiently small ϵ. Recalling that $\eta_\epsilon^T := \inf\{\eta : S_0^T(\eta) \leq \kappa\}$, this shows that $\eta_\epsilon^T - \eta_0^T \geq -\delta$ for all sufficiently small ϵ.

A similar argument, which makes use of the observation that $S_\epsilon^T(\eta_0^T + \delta) \leq (1 + C|\epsilon|)S_0^T(\eta_0^T + \delta + C|\epsilon|)$, can be used to show that $\eta_\epsilon^T - \eta_0^T \leq \delta$ for all sufficiently small ϵ. Combining these two results shows that $|\eta_\epsilon^T - \eta_0^T| \leq \delta$. As ϵ was an arbitrary number that was sufficiently close to zero, $\lim_{\epsilon \to 0} |\eta_\epsilon^T - \eta_0^T| \leq \delta$. As δ was an arbitrary small positive constant, this shows that $\lim_{\epsilon \to 0} |\eta_\epsilon^T - \eta_0^T| = 0$, as desired. □

Lemma B.4. *If the conditions of Theorem 3.1 hold, then* $\tau_\epsilon^T = \tau_0^T + \mathrm{O}(\epsilon)$.

Proof of Lemma B.4. If $\kappa = 1$, then $\eta_\epsilon^T = -\infty$ for all $\epsilon \in B_H$. Consequently, $\tau_\epsilon^T = \tau_0^T = 0$ for all ϵ, and so $\tau_\epsilon^T = \tau_0^T + \mathrm{O}(\epsilon)$ with much to spare.

Now suppose that $\kappa < 1$. In this case, $\eta_\epsilon^T > -\infty$ for all $\epsilon \in B_H$. Note that $|\max\{\eta_\epsilon^T, 0\} - \max\{\eta_0^T, 0\}| \leq |\eta_\epsilon^T - \eta_0^T|$. As a consequence, to show that $\tau_\epsilon^T - \tau_0^T = \mathrm{O}(\epsilon)$, it suffices to show that $\eta_\epsilon^T - \eta_0^T = \mathrm{O}(\epsilon)$. The remainder of this proof establishes this fact.

Fix ϵ in a neighborhood of zero. Let f_0^T denote the Lebesgue density of $\Delta_{b,0}(V)$, $V \sim P_0$, which exists and is continuous in a neighborhood of η_0^T by Condition B2. By the fact that $\eta_\epsilon^T := \inf\{\eta : S_\epsilon^T(\eta) \leq \kappa\}$, the bound on the range of H, and (B.28), it holds that $\kappa < S_\epsilon^T(\eta_\epsilon^T - |\epsilon|) \leq [1 + C|\epsilon|]S_0^T(\eta_\epsilon^T - [1 + C]|\epsilon|)$. A Taylor expansion of S_0^T about η_0^E shows that, for all ϵ sufficiently close to zero,

$$\kappa < [1 + C|\epsilon|]\left[S_0^T(\eta_0^T) + \{\eta_\epsilon^T - \eta_0^T - (1 - C)|\epsilon|\}\{-f_0^T(\eta_0^T) + \mathrm{o}(1)\}\right].$$

By Condition B2, $S_0^T(\eta_0^T) = \kappa$. Plugging this into the above and rearranging shows that

$$0 < C\kappa|\epsilon| + [1 + C|\epsilon|]\left[\eta_\epsilon^T - \eta_0^T - (1 - C)|\epsilon|\right]\left[-f_0^T(\eta_0^T) + \mathrm{o}(1)\right].$$

Using that Condition C3 implies that $f_0^T(\eta_0^T) \in (0, \infty)$, the above shows that, for all ϵ sufficiently close to zero, $0 < -[\eta_\epsilon^T - \eta_0^T]f_0^T(\eta_0^T) + C|\epsilon| + o\left(\eta_\epsilon^T - \eta_0^T\right)$, which implies that there exists an $\mathrm{O}(\epsilon)$ sequence for which $\eta_\epsilon^T - \eta_0^T < \mathrm{O}(\epsilon)$.

A similar argument can be used to show that there exists an $\mathrm{O}(\epsilon)$ sequence such that $\eta_\epsilon^T - \eta_0^T > \mathrm{O}(\epsilon)$. Combining these two bounds shows that $\eta_\epsilon^T - \eta_0^T = \mathrm{O}(\epsilon)$, as desired. $\square$

Our derivation of the canonical gradient relies on the following decomposition:

$$\begin{aligned}
\Psi_{t_\epsilon}^T(P_\epsilon) - \Psi_{t_0}^T(P_0) &= \Psi_{t_\epsilon}^T(P_\epsilon) - \Psi_{t_0}^T(P_\epsilon) + \Psi_{t_0}^T(P_\epsilon) - \Psi_{t_0}^T(P_0) \\
&= P_\epsilon\{(t_\epsilon - t_0)\Delta_{b,\epsilon}\} + [\Psi_{t_0}^T(P_\epsilon) - \Psi_{t_0}^T(P_0)] \\
&= P_\epsilon\{(t_\epsilon - t_0)(\Delta_{b,\epsilon} - \tau_0^T)\} + \tau_0^T(P_\epsilon t_\epsilon - P_0 t_0) \\
&\quad - \tau_0^T(P_\epsilon - P_0)t_0 + [\Psi_{t_0}^T(P_\epsilon) - \Psi_{t_0}^T(P_0)]. \qquad \text{(B.29)}
\end{aligned}$$

We analyze the four terms separately.

Study of term 1 in (B.29)**:** We will show that this term is $\mathrm{o}(\epsilon)$. Similar arguments to those used to study term 1 in (B.26), where the application of Lemma B.2 is replaced by an application of Lemma B.4, show that

$$\begin{aligned}
&|P_\epsilon\{(t_\epsilon - t_0)(\Delta_{b,\epsilon} - \tau_0^T)\}| \\
&= \left| \int \{t_\epsilon(v) - t_0(v)\}(\Delta_{b,\epsilon}(v) - \tau_0^T) P_{V,\epsilon}(dv) \right| \\
&\leq \int I\{|\Delta_{b,0}(v) - \tau_0^T| \leq |\Delta_{b,\epsilon}(v) - \tau_\epsilon^T - \Delta_{b,0}(v) + \tau_0^T|\}|\Delta_{b,\epsilon}(v) - \tau_0^T| P_{V,\epsilon}(dv) \\
&\leq \int I\{|\Delta_{b,0}(v) - \tau_0^T| \leq C|\epsilon|\}(|\Delta_{b,0}(v) - \tau_0^T| + C|\epsilon|) P_{V,\epsilon}(dv) \\
&\leq C|\epsilon| \int I\{|\Delta_{b,0}(v) - \tau_0^T| \leq C|\epsilon|\} P_{V,\epsilon}(dv) \\
&= C|\epsilon| \int I\{|\Delta_{b,0}(v) - \tau_0^T| \leq C|\epsilon|\}[1 + \epsilon H_v(v)] P_{V,0}(dv) \\
&\lesssim |\epsilon| \int I\{|\Delta_{b,0}(v) - \tau_0^T| \leq C|\epsilon|\} P_{V,0}(dv) \\
&\lesssim |\epsilon| \int I\{0 < |\Delta_{b,0}(v) - \tau_0^T| \leq C|\epsilon|\} P_{V,0}(dv),
\end{aligned}$$

where the final expression holds by Condition B1. The integral in this final expression is $\mathrm{o}(1)$, and therefore the expression is $\mathrm{o}(\epsilon)$.

Study of term 2 in (B.29)**:** We will show that this term is zero for all ϵ sufficiently close to zero. Consider two cases. First, if $\tau_0^T = 0$, then this term is zero for all ϵ. Second, if

$\tau_0^T > 0$, then $\tau_0^T = \eta_0^T$. Consequently $\eta_0^T > 0$, implying that $P_0 t_0 = \kappa$. By Lemma B.3, $\eta_\epsilon^T > 0$ for all ϵ sufficiently close to zero. Consequently, $P_\epsilon t_\epsilon = \kappa$ for all such ϵ. Thus, the for all ϵ sufficiently close to zero, $\tau_0^T(P_\epsilon t_\epsilon - P_0 t_0) = 0$, as desired.

Study of term 3 in (B.29)**:** It holds that $-\tau_0^T(P_\epsilon - P_0)t_0 = \epsilon P_0 G_3^T H$, where $G_3^T \in L_0^2(P_0)$ is the function $o \mapsto -\tau_0^T[t_0(v) - P_0 t_0]$. Because $P_0 t_0 = \kappa$ whenever $\tau_0^T \neq 0$, we see that G_3^T also writes as $o \mapsto -\tau_0^T[t_0(v) - \kappa]$.

Study of term 4 in (B.29)**:** The results from Appendix B.7.7 show that $\Psi_{t_0}^T(P_\epsilon) - \Psi_{t_0}^T(P_0) = \epsilon P_0 G_4^T H + \mathrm{o}(\epsilon)$, where $G_4^T = D^T(P_0, \mu_0^Y, \mu_0^A, t_0, 0)$.

Conclusion of the derivation of the canonical gradient of $P \mapsto \Psi_{t_P}^T(P)$: Combining our results regarding the four terms in (B.29), we see that

$$\Psi_{t_\epsilon}^T(P_\epsilon) - \Psi_{t_0}^T(P_0) = \epsilon \int \left[G_3^T(o) + G_4^T(o)\right] H(o) P_0(do) + \mathrm{o}(\epsilon).$$

Dividing both sides by $\epsilon \neq 0$ and taking the limit as $\epsilon \to 0$, we see that $G_3^T + G_4^T$ is the canonical gradient of $P \mapsto \Psi_{t_P}^T(P)$ at P_0. The proof concludes by noting that $G_3^T + G_4^T$ is equal to $D^T(P_0, \mu_0^Y, \mu_0^A, t_0, \tau_0^T)$.

B.8 Expansions based on gradients or pseudo-gradients

In this appendix, we present expansions of parameters that are analogous to (B.5) or (B.7). For some parameters, we consider expansions based on pseudo-gradients, rather than gradients, because the corresponding TMLE is easier to construct. Our proofs of the asymptotic linearity of our proposed estimators are based on these expansions.

B.8.1 Fixed reference rule mean outcome

For any given $e : \mathcal{W} \to [0, 1]$ and $P \in \mathcal{M}$, we define

$$\begin{aligned} R_e^E(P, P_0) :=& \Psi_e^E(P) - \Psi_e^E(P_0) + P_0 D^E(P, e, 0, \mu^A) \\ =& \mathrm{E}_0 \left[e(W) \left\{ \frac{\mu_P^Z(W) - \mu_0^Z(W)}{\mu_P^Z(W)} (\mu_P^Y(1, W) - \mu_0^Y(1, W)) \right. \right. \end{aligned}$$

$$+ \frac{\mu_P^Z(W) - \mu_0^Z(W)}{1 - \mu_P^Z(W)} (\mu_P^Y(0, W) - \mu_0^Y(0, W)) \Bigg\} \Bigg].$$

For any given $e : \mathcal{V} \to [0, 1]$, it will be convenient to let $R_e^E(P, P_0) := R_{w \mapsto e(V(w))}^E(P, P_0)$. The domain of e should be clear from the context in all subsequent uses of this notation.

For the mapping $P \mapsto \Psi_{e^{\mathrm{FR}}}^E(P)$, we consider the expansion based on the gradient $G_{\mathrm{FR}}(P)$. For $P \in \mathcal{M}$, the expansion is given by

$$\Psi_{e^{\mathrm{FR}}}^E(P) - \Psi_{e^{\mathrm{FR}}}^E(P_0) = -P_0 G_{\mathrm{FR}}(P) + R_{e^{\mathrm{FR}}}^E(P, P_0).$$

B.8.2 Randomly distributed reference rule mean outcome in the first setting

We consider the expansion based on the gradient, which, for $P \in \mathcal{M}$, is given by

$$\begin{aligned}
&\Psi_{e_P^{\mathrm{RD}}}^E(P) - \Psi_{e_{P_0}^{\mathrm{RD}}}^E(P_0) \\
&= -P_0 G_{\mathrm{RD}}^E(P) + R_{e_P^{\mathrm{RD}}}^E(P, P_0) + \kappa(P\Delta_P^Y - P_0\Delta_{P_0}^Y) \frac{P\bar{\mu}_P^A - P_0\bar{\mu}_0^A}{[P\bar{\mu}_P^A][P_0\bar{\mu}_0^A]} \\
&\quad - \frac{\kappa P\Delta_P^Y}{[P\bar{\mu}_P^A]^2} \mathrm{E}_0 \left[\frac{\mu_P^Z(W) - \mu_{P_0}^Z(W)}{\mu_P^Z(W)} \{\mu_P^A(1, W) - \mu_0^A(1, W)\} \right] - \kappa(P\Delta_P^Y) \frac{[P\bar{\mu}_P^A - P_0\bar{\mu}_0^A]^2}{[P\bar{\mu}_P^A]^2[P_0\bar{\mu}_0^A]}.
\end{aligned}$$

B.8.3 Randomly distributed reference rule mean outcome in the second setting

To facilitate analysis, we expand Ψ and $\underline{e}^{\mathrm{RD}}$ separately as follows:

$$\begin{aligned}
&\Psi_{\underline{e}_P^{\mathrm{RD}}}^E(P) - \Psi_{\underline{e}_{P_0}^{\mathrm{RD}}}^E(P_0) = P_0 \underline{D}^E(P, \underline{e}_P^{\mathrm{RD}}, 0, \mu_P^A) + R_{\underline{e}_P^{\mathrm{RD}}}^E(P, P_0) + (\underline{e}_P^{\mathrm{RD}} - \underline{e}_0^{\mathrm{RD}}) P_0 \Delta_0^Y, \\
&\underline{e}_P^{\mathrm{RD}} - \underline{e}_0^{\mathrm{RD}} = \frac{\kappa - \varphi_P}{P\Delta_P^A} - \frac{\kappa - \varphi_0}{P_0\Delta_0^A}, \\
&\kappa - \varphi_P = \kappa - \varphi_0 + P_0 D_1(P, \mu^A) + P_0 \left\{ (\mu^A(0, \cdot) - \mu^A(1, \cdot)) \frac{\mu_P^Z - \mu_0^Z}{1 - \mu_P^Z} \right\}, \\
&P\Delta_P^A = P_0\Delta_0^A - P_0 D_2(P, \mu_P^A) + P_0 \left\{ (\mu_P^A(1, \cdot) - \mu_0^A(1, \cdot)) \frac{\mu_P^Z - \mu_0^Z}{\mu_P^Z} + (\mu_P^A(0, \cdot) - \mu_0^A(0, \cdot)) \frac{\mu_P^Z - \mu_0^Z}{1 - \mu_P^Z} \right\}.
\end{aligned}$$

Here, the expansion of $\kappa - \varphi_P$ takes the form appearing in (B.7) with pseudo-gradient $D_1(P, \mu^A)$.

B.8.4 True propensity reference rule mean outcome

We consider the expansion based on the gradient, which, for $P \in \mathcal{M}$, is given by

$$\Psi^E_{e^{\mathrm{TP}}_P}(P) - \Psi^E_{e^{\mathrm{TP}}_{P_0}}(P_0) = -P_0 G^E_{\mathrm{TP}}(P) + R^E_{e^{\mathrm{TP}}_{P_0}}(P, P_0).$$

B.8.5 Optimal IER mean outcome

Let $P \in \mathcal{M}$, $e : \mathcal{V} \to [0,1]$, and $\mu^A : \{0,1\} \times \mathcal{W} \to \mathbb{R}$. We consider an expansion of the type appearing in (B.7) with pseudo-gradient $D^E(P, e, \tau^E_0, \mu^A)$ ($\underline{D}^E(P, e, \underline{\tau}^E_0, \mu^A)$ resp.). The remainder for this expansion is given by

$$\begin{aligned}
&\Psi^E_e(P) - \Psi^E_{e_0}(P_0) + P_0 D^E(P, e, \tau^E_0, \mu^A) \\
&\quad = \Psi^E_e(P) - \Psi^E_e(P_0) + P_0 D^E_e(P) + \Psi^E_e(P_0) - \Psi^E_{e_0}(P_0) \\
&\qquad - \tau^E_0 \left\{ \mathrm{E}_0[\mu^A(1,W)e(V)] - \kappa \right\} - \tau^E_0 \, \mathrm{E}_0 \left[e \frac{\mu^Z_0(W)}{\mu^Z_P(W)} [\mu^A_0(1,W) - \mu^A(1,W)] \right] \\
&\quad = R^E_e(P, P_0) + P_0\{(e - e_0)(\Delta^Y_{b,0} - \tau^E_0 \bar{\mu}^A_0)\} \\
&\qquad - \tau^E_0 \, \mathrm{E}_0 \left[e(V) \frac{\mu^Z_P(W) - \mu^Z_0(W)}{\mu^Z_P(W)} \{\mu^A(1,W) - \mu^A_0(1,W)\} \right]
\end{aligned}$$

in the first setting, and similarly,

$$\begin{aligned}
&\Psi^E_e(P) - \Psi^E_{\underline{e}_0}(P_0) + P_0 \underline{D}^E(P, e, \underline{\tau}^E_0, \mu^A) \\
&\quad = R^E_e(P, P_0) + P_0\{(e - \underline{e}_0)(\Delta^Y_{b,0} - \underline{\tau}^E_0 \Delta^A_{b,0})\} \\
&\qquad - \underline{\tau}^E_0 \, \mathrm{E}_0 \left[e(V) \frac{\mu^Z_P(W) - \mu^Z_0(W)}{\mu^Z_P(W)} \{\mu^A(1,W) - \mu^A_0(1,W)\} \right] \\
&\qquad + \underline{\tau}^E_0 \, \mathrm{E}_0 \left[(1 - e(V)) \frac{\mu^Z_P(W) - \mu^Z_0(W)}{1 - \mu^Z_P(W)} \{\mu^A(0,W) - \mu^A_0(0,W)\} \right]
\end{aligned}$$

in the second setting. Subtracting $P_0 D^E(P, e, \tau^E_0, \mu^A)$ ($P_0 \underline{D}^E(P, e, \underline{\tau}^E_0, \mu^A)$ resp.) from both sides yields the expansion.

B.8.6 Fixed treatment reference rule mean outcome

Let $P \in \mathcal{M}$, $\mu^Y : \{0,1\} \times \mathcal{W} \to \mathbb{R}$, $\mu^A : \{0,1\} \times \mathcal{W} \to \mathbb{R}$, and $t : \mathcal{V} \to [0,1]$. We consider an expansion of the type appearing in (B.7) with pseudo-gradient $D^T(P, \mu^Y, \mu^A, t, 0)$. We

suppose that μ^A is such that $\Delta^A : w \mapsto \mu^A(1,w) - \mu^A(0,w)$ satisfies the strong IV relevance assumption, in the sense that $|\Delta^A(W)| > \delta^A$ almost surely when $W \sim P_0$. We also define $\Delta : w \mapsto \frac{\mu^Y(1,w)-\mu^Y(0,w)}{\Delta^A(w)}$. The expansion is given by

$$\Psi_t^T(P) - \Psi_t^T(P_0) = -P_0 D^T(P, \mu^Y, \mu^A, t, 0) + R^T(P, \mu^Y, \mu^A, t, P_0),$$

where tedious calculations can be used to show that the remainder takes the form

$$\begin{aligned}
&R^T(P, \mu^Y, \mu^A, t, P_0) \\
&:= P_0 \left\{ t \frac{\Delta_P^A - \Delta_0^A}{\Delta_P^A} (\Delta_P - \Delta_0) \right\} - P_0 \left\{ \frac{t}{\Delta^A} \left[(\Delta - \Delta_P)(\Delta_0^A - \Delta_P^A) \right] \right\} \\
&\quad - P_0 \left\{ \frac{t}{\Delta^A \Delta_P^A} \left(\Delta_P^A - \Delta^A \right) \left[(\Delta_P^A - \Delta_0^Y) - (\Delta_P^A - \Delta_0^A)\Delta_P \right] \right\} \\
&\quad + \mathrm{E}_0 \left[\frac{t(V)}{\Delta^A(W)\mu_P^Z(W)} \{\mu_P^Z(W) - \mu_0^Z(W)\} \{\mu_P^Y(1,W) - \mu_0^Y(1,W)\} \right] \\
&\quad - \mathrm{E}_0 \left[\frac{t(V)\Delta(W)}{\Delta^A(W)\mu_P^Z(W)} \{\mu_P^Z(W) - \mu_0^Z(W)\} \{\mu_P^A(1,W) - \mu_0^A(1,W)\} \right] \\
&\quad + \mathrm{E}_0 \left[\frac{t(V)}{\Delta^A(W)\{1 - \mu_P^Z(W)\}} \{\mu_P^Z(W) - \mu_0^Z(W)\} \{\mu_P^Y(0,W) - \mu_0^Y(0,W)\} \right] \\
&\quad - \mathrm{E}_0 \left[\frac{t(V)\Delta(W)}{\Delta^A(W)\{1 - \mu_P^Z(W)\}} \{\mu_P^Z(W) - \mu_0^Z(W)\} \{\mu_P^A(0,W) - \mu_0^A(0,W)\} \right].
\end{aligned}$$

B.8.7 Optimal ITR mean outcome

Let $P \in \mathcal{M}$, $\mu^Y : \{0,1\} \times \mathcal{W} \to \mathbb{R}$, $\mu^A : \{0,1\} \times \mathcal{W} \to \mathbb{R}$, and $t : \mathcal{V} \to [0,1]$. We consider an expansion of the type appearing in (B.7) with pseudo-gradient $D^T(P, \mu^Y, \mu^A, t, \tau_0^T)$. The remainder for the expansion is given by

$$\begin{aligned}
&\Psi_t^T(P) - \Psi_{t_0}^T(P_0) + P_0 D_\kappa^T(\mu^Y, \mu^A, t, \tau_0^T, P) \\
&\quad = \Psi_t^T(P) - \Psi_t^T(P_0) + P_0 D_\kappa^T(\mu^Y, \mu^A, t, 0, P) + \Psi_t^T(P_0) - \Psi_\kappa^T(P_0) - \tau_0^T(P_0 t - \kappa) \\
&\quad = R^T(P, \mu^Y, \mu^A, t, P_0) + P_0\{(t - t_0)(\Delta_{b,0} - \tau_0^T)\}.
\end{aligned}$$

Subtracting $P_0 D_\kappa^T(\mu^Y, \mu^A, t, \tau_0^T, P)$ from both sides yields the expansion.

B.9 Proof of results about the proposed estimators

We use the $\lesssim$ to denote $\leq$ up to a positive constant that may depend on P_0 as in Section B.2 and survival functions that are defined in Appendix B.7.

B.9.1 Preliminary lemmas

Lemma B.5. *Let $\epsilon > 0$, $\eta \in \mathbb{R}$, $g : \mathcal{O} \to \mathbb{R}$ be bounded and functions $f_0 : \mathcal{O} \to \mathbb{R}$ and $f : \mathcal{O} \to \mathbb{R}$. Then*

$$\begin{aligned}|P_0([I(f > \eta) - I(f_0 > \eta)]g)| &\leq P_0|[I(f > \eta) - I(f_0 > \eta)]g| \\ &\lesssim P_0\{|f(O) - f_0(O)| > \epsilon\} + P_0\{|f_0(O) - \eta| \leq \epsilon\}.\end{aligned}$$

If g takes values in $[-1, 1]$, then $\lesssim$ can be replaced by $\leq$.

Proof of Lemma B.5. Note that

$$\begin{aligned}&|P_0([I(f > \eta) - I(f_0 > \eta)]g)| \\ &\quad \leq P_0|I(f > \eta) - I(f_0 > \eta)||g| \\ &\quad \lesssim P_0|I(f > \eta) - I(f_0 > \eta| \\ &\quad = P_0|I(f > \eta) - I(f_0 > \eta)|I(|f_0 - \eta| > \epsilon) + P_0|I(f > \eta) - I(f_0 > \eta)|I(|f_0 - \eta| \leq \epsilon).\end{aligned}$$

Because, for all $o \in \mathcal{O}$, $|I(f(o) > \eta) - I(f_0(o) > \eta)| \leq 1$ and is nonzero if and only if (i) $f(o) - \eta$ and $f_0(o) - \eta$ take difference signs or (ii) exactly one of these quantities is zero, the display continues as

$$\begin{aligned}&\leq P_0|I(f > \eta) - I(f_0 > \eta)|I(|f - f_0| > \epsilon) + P_0|I(f > \eta) - I(f_0 > \eta)|I(|f_0 - \eta| \leq \epsilon) \\ &\leq P_0 I(|f - f_0| > \epsilon) + P_0 I(|f_0 - \eta| \leq \epsilon).\end{aligned}$$

If g takes values in $[-1, 1]$, the $\lesssim$ above can be replaced by $\leq$. □

Lemma B.6 (Consistency of the sample quantile)**.** *Let $f_0 : \mathcal{O} \to \mathbb{R}$ and f_n be an estimator of f_0 based on observations $O_1, \ldots, O_n$ that satisfies $\|f_n - f_0\|_{1,P_0} = o_p(1)$. Let $S_0 : \tau \mapsto$*

$P\{f_0(O) > \tau\}$ be the survival function of $f_0(O)$, $O \sim P_0$ and let $S_n : \tau \mapsto \frac{1}{n}\sum_{i=1}^{n} I\{f_n(O_i) > \tau\}$ be an estimator thereof. Let $\eta_0 = \inf\{\tau : S_0(\tau) \leq \kappa\}$ be the $(1-\kappa)$-quantile of $f_0(O)$, $O \sim P_0$ and let $\eta_n = \inf\{\tau : S_n(\tau) \leq \kappa\}$ be an estimator thereof. If $\eta_0 > -\infty$, then also suppose that, for any η in a neighborhood of η_0, S_0 is continuous at η and $o \mapsto I\{f_n(o) > \eta\}$ belongs to a fixed P_0-Glivenko-Cantelli class with probability tending to one. Under these conditions, $\eta_n \xrightarrow{p} \eta_0$.

Proof of Lemma B.6. We separately consider the cases that $\eta_0 = -\infty$ and $\eta_0 > -\infty$. In the simpler case where $\eta_0 = -\infty$, it must hold that $\kappa = 1$ and so, by definition, $\eta_n = -\infty = \eta_0$.

Now suppose that $\eta_0 > -\infty$, in which case $\kappa < 1$. For any η sufficiently close to η_0,

$$\begin{aligned} |S_n(\eta) - S_0(\eta)| &= |P_n I(f_n > \eta) - P_0 I(f_0 > \eta)| \\ &\leq |P_0[I(f_n > \eta) - I(f_0 > \eta)]| + |(P_n - P_0)I(f_n > \eta)|. \end{aligned} \tag{B.30}$$

The second term is $o_p(1)$ because $I(f_n > \eta)$ belongs to a P_0-Glivenko-Cantelli class with probability tending to one. We now provide an argument showing that the first term is also $o_p(1)$ provided η is sufficiently close to η_0. Fix $\epsilon > 0$ and $\delta > 0$. We will show that there exists an N such that $P_0\{|P_0[I(f_n > \eta) - I(f_0 > \eta)]| > \epsilon\} \leq \delta$ for all $n \geq N$, establishing this result. We start by noting the following bound, which, by Lemma B.5, holds for any $\epsilon' > 0$:

$$\begin{aligned} |P_0[I(f_n > \eta) - I(f_0 > \eta)]| &\leq P_0 I(|f_n - f_0| > \epsilon') + P_0 I(|f_0 - \eta| \leq \epsilon') \\ &\leq \frac{P_0|f_n - f_0|}{\epsilon'} + P_0 I(|f_0 - \eta| \leq \epsilon') \qquad \text{(Markov's inequality)} \\ &= \frac{\|f_n - f_0\|_{1,P_0}}{\epsilon'} + P_0 I(|f_0 - \eta| \leq \epsilon') \end{aligned}$$

If η is sufficiently close to η_0 and $\epsilon' > 0$ is sufficiently small, then our conditions ensure that S_0 is continuous in $[\eta - \epsilon', \eta + \epsilon']$. Consequently, there exists an $\epsilon' > 0$ such that $P_0 I(|f_0 - \eta| \leq \epsilon') \leq \epsilon/2$. Combining this observation with the above display shows that

$$P_0\left\{|P_0[I(f_n > \eta) - I(f_0 > \eta)]| > \epsilon\right\} \leq P_0\left\{\frac{\|f_n - f_0\|_{1,P_0}}{\epsilon'} + P_0 I(|f_0 - \eta| \leq \epsilon') > \epsilon\right\}$$

$$\leq P_0 \left\{ \frac{\|f_n - f_0\|_{1,P_0}}{\epsilon'} > \epsilon/2 \right\}.$$

Now, because $\|f_n - f_0\|_{1,P_0} = \mathrm{o}_p(1)$ by assumption, there also exists an N such that $P\{\|f_n - f_0\|_{1,P_0} > \epsilon\epsilon'/2\} \leq \delta$ for all $n \geq N$. Hence, for all $n \geq N$, the left-hand side above is no larger than δ. Recalling, (B.30), we have shown that $S_n(\eta) - S_0(\eta) = o_P(1)$, for all η sufficiently close to η_0.

Fix an arbitrary $\epsilon > 0$ that is small enough such that S_0 is continuous in $[\eta_0 - \epsilon, \eta_0 + \epsilon]$. The above result implies that $S_n(\eta_0 - \epsilon) = S_0(\eta_0 - \epsilon) + \mathrm{o}_p(1)$ and $S_n(\eta_0 + \epsilon) = S_0(\eta_0 + \epsilon) + \mathrm{o}_p(1)$. Because $S_0(\eta_0 - \epsilon) > \kappa$ and $S_0(\eta_0 + \epsilon) < \kappa$, this shows that, with probability tending to one, it holds that $S_n(\eta_0 - \epsilon) > \kappa$ and $S_n(\eta_0 + \epsilon) < \kappa$. Hence, $\eta_0 - \epsilon \leq \eta_n \leq \eta_0 + \epsilon$ with probability tending to one. Because ϵ was arbitrary, we have shown that $\eta_n - \eta_0 = \mathrm{o}_p(1)$. □

Lemma B.7 (Convergence rate of sample quantile)**.** *Let κ, f_n, f_0, S_n, S_0, η_n and η_0 be as defined in Lemma B.6. Suppose that $\kappa < 1$ and that $o \mapsto I\{f_n(o) > \eta_n\}$ belongs to a P_0-Donsker class with probability tending to one. Further suppose that the distribution of $f_0(O)$, $O \sim P_0$, has nonzero finite continuous Lebesgue density in a neighborhood of η_0. Finally, suppose that $S_n(\eta_n) = \kappa + \mathrm{O}_p(n^{-1/2})$. Under these conditions, the following implications are valid with probability tending to one:*

i) if $\|f_n - f_0\|_{q,P_0} = \mathrm{o}_p(1)$ for some $0 < q < \infty$, then $|\eta_n - \eta_0| \lesssim \|f_n - f_0\|_{q,P_0}^{q/q+1} + \mathrm{O}_p(n^{-1/2})$;

ii) if $\|f_n - f_0\|_{\infty,P_0} = \mathrm{o}_p(1)$, then $|\eta_n - \eta_0| \lesssim \|f_n - f_0\|_{\infty,P_0} + \mathrm{O}_p(n^{-1/2})$.

We note that, in the setting of the above lemma, the condition that $S_n(\eta_n) = \kappa + \mathrm{O}_p(n^{-1/2})$ is satisfied if the distribution of $f_n(O)$ has Lebesgue density near η_0, in which case $S_n(\eta_n) = \kappa + \mathrm{O}_p(n^{-1})$, so that the stated condition holds with much to spare.

Proof of Lemma B.7. Because $\kappa < 1$, it holds that $\eta_0 > -\infty$. Because $f_0(O)$ is continuous when $O \sim P_0$, $S_0(\eta_0) = \kappa$. Therefore, the fact that $S_n(\eta_n) = \kappa + \mathrm{O}_p(n^{-1/2})$ yields that

$$\mathrm{O}_p(n^{-1/2}) = S_n(\eta_n) - S_0(\eta_0) = [S_n(\eta_n) - S_0(\eta_n)] + [S_0(\eta_n) - S_0(\eta_0)].$$

A Taylor expansion shows that the second term on the right is equal to $S_0'(\eta_0)(\eta_n - \eta_0) + o_p(\eta_n - \eta_0)$ whenever η_n is sufficiently close to η_0. Because Lemma B.6 implies that $\eta_n \xrightarrow{p} \eta_0$, this expansion is valid with probability tending to one. Therefore, with probability tending to one,

$$O_p(n^{-1/2}) = [S_n(\eta_n) - S_0(\eta_n)] + S_0'(\eta_0)(\eta_n - \eta_0) + o_p(\eta_n - \eta_0).$$

Rearranging and taking an absolute value, and subsequently applying the triangle inequality shows that

$$\begin{aligned} |S_0'(\eta_0)(\eta_n - \eta_0) + o_p(\eta_n - \eta_0)| &= |S_n(\eta_n) - S_0(\eta_n) + O_p(n^{-1/2})| \\ &\leq |S_n(\eta_n) - S_0(\eta_n)| + O_p(n^{-1/2}). \end{aligned} \tag{B.31}$$

For the first term above, we note that, with probability tending to one,

$$\begin{aligned} S_n(\eta_n) - S_0(\eta_n) =& P_n I(f_n > \eta_n) - P_0 I(f_0 > \eta_n) \\ =& P_0[I(f_n > \eta_n) - I(f_0 > \eta_n)] + (P_n - P_0) I(f_n > \eta_n) \\ =& P_0[I(f_n > \eta_n) - I(f_0 > \eta_n)] + O_p(n^{-1/2}), \end{aligned}$$

where the last step follows from Donkser assumption on $o \mapsto I\{f_n(o) > \eta_n\}$. Plugging this into (B.31) then shows that

$$|S_0'(\eta_0)(\eta_n - \eta_0) + o_p(\eta_n - \eta_0)| \leq |P_0[I(f_n > \eta_n) - I(f_0 > \eta_n)]| + O_p(n^{-1/2}). \tag{B.32}$$

We now study the first term on the right. By Lemma B.5, the following holds for any $\epsilon > 0$:

$$\begin{aligned} |P_0[I(f_n > \eta_n) - I(f_0 > \eta_n)]| &\leq P_0 I(|f_n - f_0| > \epsilon) + P_0 I(|f_0 - \eta_n| \leq \epsilon) \\ &= P_0 I(|f_n - f_0| > \epsilon) + P_0[I(f_0 \geq \eta_n - \epsilon) - I(f_0 > \eta_n + \epsilon)]. \end{aligned} \tag{B.33}$$

Fix a positive sequence $\{\epsilon_n\}_{n=1}^{\infty}$, where each ϵ_n may be random through observations $O_1, \ldots, O_n$, such that $\epsilon_n \xrightarrow{p} 0$ as $n \to \infty$. By a Taylor expansion of S_0 around η_0, the following display holds with probability tending to one:

$$P_0[I(f_0 \geq \eta_n - \epsilon_n) - I(f_0 > \eta_n + \epsilon_n)]$$

$$= S_0(\eta_n - \epsilon_n) - S_0(\eta_n + \epsilon_n)$$
$$= [S_0(\eta_0) + S_0'(\eta_0)(\eta_n - \eta_0 - \epsilon_n)] - [S_0(\eta_0) + S_0'(\eta_0)(\eta_n - \eta_0 + \epsilon_n)] + \mathrm{o}_p(|\eta_n - \eta_0| + \epsilon_n).$$

The right-hand side is $\mathrm{O}_p(\epsilon_n)+\mathrm{o}_p(|\eta_n-\eta_0|)$ because $S_0'(\eta_0)$ is finite by assumption. Combining this with (B.33) and (B.32) shows that

$$|S_0'(\eta_0)(\eta_n - \eta_0) + \mathrm{o}_p(\eta_n - \eta_0)| \leq P_0 I(|f_n - f_0| > \epsilon_n) + \mathrm{o}_p(|\eta_n - \eta_0|) + \mathrm{O}_p(\epsilon_n) + \mathrm{O}_p(n^{-1/2}).$$

Using that $S_0'(\eta_0) \neq 0$, this shows that

$$|\eta_n - \eta_0| \leq P_0 I(|f_n - f_0| > \epsilon_n) + \mathrm{O}_p(\epsilon_n) + \mathrm{O}_p(n^{-1/2}).$$

By Markov's inequality, for any $0 < q < \infty$, $P_0 I(|f_n - f_0| > \epsilon_n) = P_0 I(|f_n - f_0|^q > \epsilon_n^q) \leq \|f_n - f_0\|_{q,P_0}^q / \epsilon_n^q$. If $\|f_n - f_0\|_{q,P_0} = \mathrm{o}_p(1)$, taking $\epsilon_n = \|f_n - f_0\|_{q,P_0}^{q/q+1}$ shows that $|\eta_n - \eta_0| \lesssim \|f_n - f_0\|_{q,P_0}^{q/q+1} + \mathrm{O}_p(n^{-1/2})$. If $\|f_n - f_0\|_{\infty,P_0} = \mathrm{o}_p(1)$, taking $\epsilon_n = \|f_n - f_0\|_{\infty,P_0}$ yields that $P_0 I(|f_n - f_0| > \epsilon_n) = 0$, and therefore shows that $|\eta_n - \eta_0| \lesssim \|f_n - f_0\|_{\infty,P_0} + \mathrm{O}_p(n^{-1/2})$. □

Lemma B.8. *Fix functions $\mu^A : \{0,1\} \times \mathcal{W} \to [0,1]$ and $\mu^Y : \{0,1\} \times \mathcal{W} \to \mathbb{R}$, and suppose that $P_0\mu^Y(0,\cdot)^2 < \infty$ and $P_0\mu^Y(1,\cdot)^2 < \infty$. If Condition A3 holds, then*

$$\|\mu^Y(1,\cdot) - \mu_0^Y(1,\cdot)\|_{2,P_0} + \|\mu^Y(0,\cdot) - \mu_0^Y(0,\cdot)\|_{2,P_0} \simeq \|\mu^Y - \mu_0^Y\|_{2,P_0},$$
$$\|\mu^A(1,\cdot) - \mu_0^A(1,\cdot)\|_{2,P_0} + \|\mu^A(0,\cdot) - \mu_0^A(0,\cdot)\|_{2,P_0} \simeq \|\mu^A - \mu_0^A\|_{2,P_0},$$

where $a \simeq b$ is defined as $a \lesssim b$ and $b \lesssim a$.

Proof of Lemma B.8. Observe that

$$\begin{aligned}\|\mu^Y - \mu_0^Y\|_{2,P_0}^2 &= \mathrm{E}_0\left[\{\mu^Y(Z,W) - \mu_0^Y(Z,W)\}^2\right] \\ &= \mathrm{E}_0\left[\left\{Z[\mu^Y(1,W) - \mu_0^Y(1,W)] + (1-Z)[\mu^Y(0,W) - \mu_0^Y(0,W)]\right\}^2\right] \\ &= \mathrm{E}_0\left[\mu_0^Z(W)\{\mu^Y(1,W) - \mu_0^Y(1,W)\}^2 + \{1 - \mu_0^Z(W)\}\{\mu^Y(0,W) - \mu_0^Y(0,W)\}^2\right].\end{aligned}$$

Because $\mu_0^Z(W) \in (\delta^Z, 1-\delta^Z)$ P_0-a.s. for $\delta^Z > 0$, we have that

$$\|\mu^Y(1,\cdot) - \mu_0^Y(1,\cdot)\|_{2,P_0}^2 + \|\mu^Y(0,\cdot) - \mu_0^Y(0,\cdot)\|_{2,P_0}^2 \simeq \|\mu^Y - \mu_0^Y\|_{2,P_0}^2.$$

Noting that $(a+b)/\sqrt{2} \leq \sqrt{a^2+b^2} \leq a+b$ for any real numbers $a, b \geq 0$, it follows that

$$\|\mu^Y(1,\cdot) - \mu_0^Y(1,\cdot)\|_{2,P_0} + \|\mu^Y(0,\cdot) - \mu_0^Y(0,\cdot)\|_{2,P_0} \simeq \|\mu^Y - \mu_0^Y\|_{2,P_0}.$$

The same proof applies to μ^A and μ_0^A. □

B.9.2 Individualized encouragement rules

Lemma B.9 (Asymptotic linearity of φ_n and $P_n\underline{\hat{\Delta}}_n^A$)**.** *Under the conditions of Theorem 3.4,*

$$\varphi_n - \varphi_0 = (P_n - P_0)D_1(P_0, \mu_0^A) + \mathrm{o}_p(n^{-1/2}) = \mathrm{O}_p(n^{-1/2}),$$
$$P_n\underline{\hat{\Delta}}_n^A - P_0\Delta_0^A = (P_n - P_0)D_2(P_0, \mu_0^A) + \mathrm{o}_p(n^{-1/2}) = \mathrm{O}_p(n^{-1/2}).$$

This result follows from the facts that (i) φ_n is a one-step correction estimator of φ_0 (e.g., Pfanzagl, 1982), and (ii) $P_n\underline{\hat{\Delta}}_n^A$ is a TMLE for $P_0\Delta_0^A$ (e.g., van Der Laan, 2017; van der Laan and Rose, 2018). Therefore the proof is omitted.

Recall the definitions $\Gamma_0 : \eta \mapsto \mathrm{E}_0[I\{\xi_0(V) > \eta\}\mu_0^A(1,W)]$, $\underline{\Gamma}_0 : \eta \mapsto \mathrm{E}_0[I\{\underline{\xi}_0(V) > \eta\}\Delta_0^A(W)]$, $\Gamma_n : \eta \mapsto n^{-1}\sum_{i=1}^n I\{\xi_n(V_i) > \eta\}\mu_n^A(1,W_i)$ and $\underline{\Gamma}_n : \eta \mapsto n^{-1}\sum_{i=1}^n I\{\underline{\xi}_n(V_i) > \eta\}\Delta_n^A(W_i)$. We also define $\eta_n^E := \eta_n^E(k_n)$, $\tau_n^E := \tau_n^E(k_n)$, $\underline{\eta}_n^E := \underline{\eta}_n^E(\underline{k}_n)$ and $\underline{\tau}_n^E := \underline{\tau}_n^E(\underline{k}_n)$.

Lemma B.10 (Consistency of $\eta_n^E(\kappa)$ and $\underline{\eta}_n^E(\kappa)$)**.** *Under Conditions C2, C3 and C12,* $\eta_n^E(\kappa) \xrightarrow{p} \eta_0^E$ *(*$\underline{\eta}_n^E(\kappa) \xrightarrow{p} \underline{\eta}_0^E$ *resp.).*

This lemma is a stochastic variant of the deterministic result in Lemma B.1 and, as such, has a similar proof. The proof is also similar to that of the stochastic result in Lemma B.6. Due to the similarity of this proof to our earlier results, the arguments are slightly abbreviated here.

Proof of Lemma B.10. We separately consider the cases that $\eta_0^E > -\infty$ ($\underline{\eta}_0^E > -\infty$ resp.) and $\eta_0^E = -\infty$ ($\underline{\eta}_0^E = -\infty$ resp.).

First consider the case where $\eta_0^E > -\infty$ ($\underline{\eta}_0^E > -\infty$ resp.). We start by showing that, for any η sufficiently close to η_0^E ($\underline{\eta}_0^E$ resp.), it holds that $\Gamma_n(\eta) - \Gamma_0(\eta) = \mathrm{o}_p(1)$ ($\underline{\Gamma}_n(\eta) - \underline{\Gamma}_0(\eta) =$

$o_p(1)$ resp.). Fix an η in a neighborhood of η_0^E ($\underline{\eta}_0^E$ resp.). By the triangle inequality,

$$\begin{aligned}|\Gamma_n(\eta) - \Gamma_0(\eta)| \leq\ & |P_0[I\{\xi_n(V(\cdot)) > \eta\} - I(\xi_0(V(\cdot)) > \eta)]\mu_0^A(1,\cdot)| \\ & + |P_0 I\{\xi_n(V(\cdot)) > \eta\}[\mu_n^A(1,\cdot) - \mu_0^A(1,\cdot)]| \\ & + |(P_n - P_0) I\{\xi_n(V(\cdot)) > \eta\}\mu_n^A(1,\cdot)| \end{aligned} \tag{B.34}$$

in the first setting, and

$$\begin{aligned}|\underline{\Gamma}_n(\eta) - \underline{\Gamma}_0(\eta)| \leq\ & |P_0[I\{\underline{\xi}_n(V(\cdot)) > \eta\} - I(\underline{\xi}_0(V(\cdot)) > \eta)]\Delta_0^A| \\ & + |P_0 I\{\underline{\xi}_n(V(\cdot)) > \eta\}[\Delta_n^A - \Delta_0^A]| \\ & + |(P_n - P_0) I\{\underline{\xi}_n(V(\cdot)) > \eta\}\Delta_n^A| \end{aligned} \tag{B.35}$$

in the second setting. We will show that the right-hand side is $o_p(1)$. By Condition C12, the third term on the right is $o_p(1)$ provided η is sufficiently close to η_0^E ($\underline{\eta}_0^E$ resp.). Moreover, because the second term bounds as $|P_0 I\{\xi_n(V(\cdot)) > \eta\}[\mu_n^A(1,\cdot) - \mu_0^A(1,\cdot)]| \leq \|\mu_n^A(1,\cdot) - \mu_0^A(1,\cdot)\|_{1,P_0}$ ($|P_0 I\{\underline{\xi}_n(V(\cdot)) > \eta\}[\Delta_n^A - \Delta_0^A]| \leq \|\Delta_n^A - \Delta_0^A\|_{1,P_0}$ resp.), Condition C12 also implies that this second term is also $o_p(1)$. We will now argue that the first term is $o_p(1)$. By Lemma B.5, for any $\epsilon > 0$,

$$\begin{aligned}|P_0[I(\xi_n(V(\cdot)) > \eta) - I(\xi_0(V(\cdot)) > \eta)]\mu_0^A(1,\cdot)| &\leq P_0|I(\xi_n > \eta) - I(\xi_0 > \eta)| \\ &\leq P_0 I(|\xi_n - \xi_0| > \epsilon) + P_0 I(|\xi_0 - \eta| \leq \epsilon) \\ &\leq \frac{\|\xi_n - \xi_0\|_{1,P_0}}{\epsilon} + P_0 I(|\xi_0 - \eta| \leq \epsilon)\end{aligned}$$

in the first setting, where the final relation follows from Markov's inequality, and similarly,

$$|P_0[I(\underline{\xi}_n(V(\cdot)) > \eta) - I(\underline{\xi}_0(V(\cdot)) > \eta)]\Delta_0^A| \leq \frac{\|\underline{\xi}_n - \underline{\xi}_0\|_{1,P_0}}{\epsilon} + P_0 I(|\underline{\xi}_0 - \eta| \leq \epsilon),$$

in the second setting. Similarly to as was done in the proof of Lemma B.6, we can then show that the last line is $o_p(1)$, and hence so is the first term. Recalling (B.34), we have shown that $\Gamma_n(\eta) - \Gamma_0(\eta) = o_p(1)$ ($\underline{\Gamma}_n(\eta) - \underline{\Gamma}_0(\eta) = o_p(1)$ resp.) for any η that is sufficiently close to η_0^E ($\underline{\eta}_0^E$ resp.).

Fix $\epsilon > 0$. Provided ϵ is sufficiently small, the above result (along with Lemma B.9 in the second setting) implies that $\Gamma_n(\eta_0^E - \epsilon) = \Gamma_0(\eta_0^E - \epsilon) + \mathrm{o}_p(1)$ ($\underline{\Gamma}_n(\underline{\eta}_0^E - \epsilon) + \varphi_n = \underline{\Gamma}_0(\underline{\eta}_0^E - \epsilon) + \varphi_0 + \mathrm{o}_p(1)$ resp.) and $\Gamma_n(\eta_0^E + \epsilon) = \Gamma_0(\eta_0^E + \epsilon) + \mathrm{o}_p(1)$ ($\underline{\Gamma}_n(\underline{\eta}_0^E + \epsilon) + \varphi_n = \underline{\Gamma}_0(\underline{\eta}_0^E + \epsilon) + \varphi_0 + \mathrm{o}_p(1)$ resp.). By Condition C3, $\Gamma_0(\eta_0^E - \epsilon) > \kappa > \Gamma_0(\eta_0^E + \epsilon)$ ($\underline{\Gamma}_0(\underline{\eta}_0^E - \epsilon) + \varphi_0 > \kappa > \underline{\Gamma}_0(\underline{\eta}_0^E + \epsilon) + \varphi_0$ resp.) provided ϵ is sufficiently small. It follows that, with probability tending to one, $\Gamma_n(\eta_0^E - \epsilon) > \kappa > \Gamma_n(\eta_0^E + \epsilon)$ ($\underline{\Gamma}_n(\underline{\eta}_0^E - \epsilon) + \varphi_n > \kappa > \underline{\Gamma}_n(\underline{\eta}_0^E + \epsilon) + \varphi_n$ resp.), and hence $\eta_0^E - \epsilon \le \eta_n^E(\kappa) \le \eta_0^E + \epsilon$ ($\underline{\eta}_0^E - \epsilon \le \underline{\eta}_n^E(\kappa) \le \underline{\eta}_0^E + \epsilon$ resp.). Because ϵ was arbitrary, it follows that $\eta_n^E(\kappa) \xrightarrow{p} \eta_0^E$ ($\underline{\eta}_n^E(\kappa) \xrightarrow{p} \underline{\eta}_0^E$ resp.).

The case where $\eta_0^E = -\infty$ ($\underline{\eta}_0^E = -\infty$ resp.) can be proved similarly. If $\kappa = 1$, then it trivially holds that $\eta_n^E(\kappa) = -\infty = \eta_0^E$ ($\underline{\eta}_n^E(\kappa) = -\infty = \underline{\eta}_0^E$ resp.) for all n, which easily implies the desired convergence. Otherwise, for any $\eta < 0$ for which $|\eta|$ is sufficiently large, a nearly identical argument to that used above shows that $\Gamma_n(\eta) - \Gamma_0(\eta) = \mathrm{o}_p(1)$ ($\underline{\Gamma}_n(\eta) + \varphi_n - \underline{\Gamma}_0(\eta) - \varphi_0 = \mathrm{o}_p(1)$ resp.). By Condition C3 and monotonicity of Γ_0 ($\underline{\Gamma}_0$ resp.), it follows that $\Gamma_0(\eta) < \kappa$ ($\underline{\Gamma}_0(\eta) + \varphi_0 < \kappa$ resp.), and so, with probability tending to one, $\Gamma_n(\eta) < \kappa$ ($\underline{\Gamma}_n(\eta) + \varphi_n < \kappa$ resp.) and hence $\eta_n^E(\kappa) \le \eta$ ($\underline{\eta}_n^E(\kappa) \le \eta$ resp.). Because η was arbitrary, $\eta_n^E(\kappa) \xrightarrow{p} -\infty = \eta_0^E$ ($\underline{\eta}_n^E(\kappa) \xrightarrow{p} -\infty = \underline{\eta}_0^E$ resp.). □

Lemma B.11 (Consistency of τ_n^E ($\underline{\tau}_n^E$ resp.) and existence of solution to Eq. 3.5 (Eq. 3.6 resp.) when $\eta_0^E > -\infty$ ($\underline{\eta}_0^E > -\infty$ resp.))**.** *Assume that the conditions for Theorem 3.4 hold. The following statements are valid:*

i) if $\eta_0^E > -\infty$ ($\underline{\eta}_0^E > -\infty$ resp.), then, with probability tending to one, a solution $k_n' \in [0,1]$ ($\underline{k}_n' \in [0,1]$ resp.) to (3.5) ((3.6) resp.) exists. For convenience, we let $k_n = k_n'$ ($\underline{k}_n = \underline{k}_n'$ resp.) when $\eta_n^E(\kappa) > 0$ ($\underline{\eta}_n^E(\kappa) > 0$ resp.) and k_n' ($\underline{k}_n'$ resp.) exists. Hence, if $\eta_0^E > 0$ ($\underline{\eta}_0^E > 0$ resp.), $\eta_n^E = \eta_n^E(k_n) = \eta_n^E(k_n')$ ($\underline{\eta}_n^E = \underline{\eta}_n^E(\underline{k}_n) = \underline{\eta}_n^E(\underline{k}_n')$ resp.);

ii) if a solution k_n' ($\underline{k}_n'$ resp.) to (3.5) ((3.6) resp.) exists, then with probability tending to one, $P_0\{d_{n,k_n'}\bar{\mu}_n^A\} = \kappa + \mathrm{O}_p(n^{-1/2})$ ($P_0\{\underline{d}_{n,\underline{k}_n'}\Delta_n^A\} + \varphi_0 = \kappa + \mathrm{O}_p(n^{-1/2})$ resp.);

iii) $\tau_n^E - \tau_0^E = \mathrm{o}_p(1)$ ($\underline{\tau}_n^E - \underline{\tau}_0^E = \mathrm{o}_p(1)$ resp.).

We separately prove i), ii), and iii) in the case that $\eta_0^E > 0$ ($\underline{\eta}_0^E > 0$ resp.), and then we separately prove iii) in the cases that $\eta_0^E = 0$ and $\eta_0^E < 0$.

Proof of i) from Lemma B.11. Our strategy for showing the existence of a solution to (3.5) ((3.6) resp.) is as follows. First, we show that the left-hand side of (3.5) ((3.6) resp.) consistently estimates the treatment resource being used uniformly over rules $\{d_{n,k}$ ($\underline{d}_{n,k}$ resp.) : $k \in [0,1]\}$. Next, we show that the left-hand side of (3.5) ((3.6) resp.) is a continuous function in k that takes different signs at $k = 0$ and $k = 1$ with probability tending to one.

Define the function $f_{n,k} : o \mapsto d_{n,k}(v)\left[\mu_n^A(1,w) + \frac{z}{\mu_n^Z(w)}[a - \mu_n^A(1,w)]\right]$ ($\underline{f}_{n,k} : o \mapsto \underline{d}_{n,k}(v)\left[\Delta_n^A(w) + \frac{1}{z+\mu_n^Z(w)-1}[a - \mu_n^A(z,w)]\right]$ resp.). We first show that

$$\sup_{k\in[0,1]} |P_n f_{n,k} - P_0\{d_{n,k}\bar{\mu}_0^A\}| = \mathrm{O}_p(n^{-1/2}) \tag{B.36}$$

in the first setting and

$$\sup_{k\in[0,1]} |P_n \underline{f}_{n,k} - P_0\{\underline{d}_{n,k}\Delta_0^A\}| = \mathrm{O}_p(n^{-1/2}). \tag{B.37}$$

in the second setting. The arguments that we use to establish the above rely on the fact that, for fixed $d_{n,k}$ ($\underline{d}_{n,k}$ resp.), $P_n f_{n,k}$ ($P_n \underline{f}_{n,k}$ resp.) is a one-step estimator of $P_0\{d_{n,k}\bar{\mu}_0^A\}$ ($P_0\{\underline{d}_{n,k}\Delta_0^A\}$ resp.). In the first setting, observe that

$$\begin{aligned}\sup_{k\in[0,1]} \left|P_n f_{n,k} - P_0\{d_{n,k}\bar{\mu}_0^A\}\right| \leq{}& \sup_{k\in[0,1]} \left|(P_n - P_0)d_{n,k}D(\hat{P}_n, \mu_n^A)\right| \\ &+ \sup_{k\in[0,1]} \left|P_0\left\{d_{n,k}(V(\cdot))\frac{\mu_n^Z(\cdot) - \mu_0^Z(\cdot)}{\mu_n^Z(\cdot)}[\mu_n^A(1,\cdot) - \mu_0^A(1,\cdot)]\right\}\right|.\end{aligned}$$

Conditions C10 and C11 imply that the first term on the right-hand side is $\mathrm{O}_p(n^{-1/2})$. For the second term, we note that the boundedness of the range of each $d_{n,k}$ and the Cauchy-Schwarz inequality imply that

$$\begin{aligned}&\sup_{k\in[0,1]} \left|P_0\left\{d_{n,k}(V(\cdot))\frac{\mu_n^Z(\cdot) - \mu_0^Z(\cdot)}{\mu_n^Z(\cdot)}[\mu_n^A(1,\cdot) - \mu_0^A(1,\cdot)]\right\}\right| \\ &\leq P_0\left|\frac{\mu_n^Z(\cdot) - \mu_0^Z(\cdot)}{\mu_n^Z(\cdot)}[\mu_n^A(1,\cdot) - \mu_0^A(1,\cdot)]\right| \lesssim \|\mu_n^Z - \mu_0^Z\|_{2,P_0}\|\mu_n^A(1,\cdot) - \mu_0^A(1,\cdot)\|_{2,P_0}.\end{aligned}$$

The right-hand side is $o_p(n^{-1/2})$ by Condition C9. Combining the previous two displays shows that (B.36) holds. Similarly, in the second setting,

$$\begin{aligned}\sup_{k\in[0,1]}\left|P_n\underline{f}_{n,k}-P_0\{\underline{d}_{n,k}\Delta_0^A\}\right| \leq & \sup_{k\in[0,1]}\left|(P_n-P_0)\underline{d}_{n,k}D_2(\hat{P}_n,\mu_n^A)\right| \\ & + \sup_{k\in[0,1]}\left|P_0\left\{\underline{d}_{n,k}(V(\cdot))\frac{\mu_n^Z(\cdot)-\mu_0^Z(\cdot)}{\mu_n^Z(\cdot)}[\mu_n^A(1,\cdot)-\mu_0^A(1,\cdot)]\right\}\right| \\ & + \sup_{k\in[0,1]}\left|P_0\left\{\underline{d}_{n,k}(V(\cdot))\frac{\mu_n^Z(\cdot)-\mu_0^Z(\cdot)}{1-\mu_n^Z(\cdot)}[\mu_n^A(0,\cdot)-\mu_0^A(0,\cdot)]\right\}\right|.\end{aligned}$$

By similar arguments, we can show that (B.37) holds.

Apply (B.36) ((B.37) resp.) at $k=0$ (along with Lemma B.9 in the second setting) shows that $P_nf_{n,0}=P_0\{d_{n,0}\bar{\mu}_0^A\}+O_p(n^{-1/2})=O_p(n^{-1/2})$ ($P_n\underline{f}_{n,0}+\varphi_n=P_0\{\underline{d}_{n,0}\Delta_0^A\}+\varphi_0+O_p(n^{-1/2})=\varphi_0+O_p(n^{-1/2})$ resp.), and therefore that $P_nf_{n,0}<\kappa$ ($P_n\underline{f}_{n,0}+\varphi_n<\kappa$ resp.) with probability tending to one. Applying this result at $k=1$ shows that $P_nf_{n,1}=P_0\{d_{n,1}\bar{\mu}_0^A\}+O_p(n^{-1/2})=P_0\bar{\mu}_0^A+O_p(n^{-1/2})$ ($P_n\underline{f}_{n,1}+\varphi_n=P_0\{\underline{d}_{n,1}\Delta_0^A\}+\varphi_0+O_p(n^{-1/2})=P_0\Delta_0^A+\varphi_0+O_p(n^{-1/2})$ resp.), and, combining this fact with the fact that $P_0\bar{\mu}_0^A>\kappa$ ($P_0\Delta_0^A+\varphi_0>\kappa$ resp.) whenever $\eta_0^E>-\infty$ ($\underline{\eta}_0^E>-\infty$ resp.) shows that $P_nf_{n,1}>\kappa$ ($P_n\underline{f}_{n,1}+\varphi_n>\kappa$ resp.) with probability tending to one. Combining these results at $k=0$ and $k=1$ with the fact that $k\mapsto P_nf_{n,k}$ ($k\mapsto P_n\underline{f}_{n,k}$ resp.) is a continuous function shows that, with probability tending to one, there exists a $k_n'\in[0,1]$ ($\underline{k}_n'\in[0,1]$ resp.) such that $P_nf_{n,k_n'}=\kappa$ ($P_n\underline{f}_{n,\underline{k}_n'}+\varphi_n=\kappa$ resp.). Lemma B.10 then implies $\eta_n^E=\eta_n^E(k_n)$ ($\underline{\eta}_n^E=\underline{\eta}_n^E(k_n)$ resp.) with probability tending to 1. □

Proof of ii) from Lemma B.11. By (B.36) and the fact that i) has been shown to hold, $P_0\{d_{n,k_n}\bar{\mu}_0^A\}=P_nf_{n,k_n}+O_p(n^{-1/2})=\kappa+O_p(n^{-1/2})$ ($P_0\{\underline{d}_{n,\underline{k}_n}\Delta_0^A\}+\varphi_0=P_n\underline{f}_{n,\underline{k}_n}+\varphi_n+O_p(n^{-1/2})=\kappa+O_p(n^{-1/2})$ resp.), as desired. □

Proof of iii) from Lemma B.11 when $\eta_0^E>0$ ($\underline{\eta}_0^E>0$ resp.). In this proof, we use P_0^n to denote a probability statement over the draws of $O_1,\ldots,O_n$. Fix $\epsilon>0$. We will argue by contradiction to show that $P_0^n\left\{\eta_n^E\geq\eta_0^E+\epsilon\right\}\to 0$ ($P_0^n\left\{\underline{\eta}_n^E\geq\underline{\eta}_0^E+\epsilon\right\}\to 0$ resp.) and $P_0^n\left\{\eta_n^E\leq\eta_0^E-\epsilon\right\}\to 0$ ($P_0^n\left\{\underline{\eta}_n^E\leq\underline{\eta}_0^E-\epsilon\right\}\to 0$ resp.) as $n\to\infty$, implying the consistency

of η_n^E ($\underline{\eta}_n^E$ resp.) and, consequently, τ_n^E ($\underline{\tau}_n^E$ resp.). We study these two cases separately. First, we suppose that

$$\limsup_n P_0^n \left\{\eta_n^E \geq \eta_0^E + \epsilon\right\} > 0 \qquad (\limsup_n P_0^n \left\{\underline{\eta}_n^E \geq \underline{\eta}_0^E + \epsilon\right\} > 0 \text{ resp.}). \tag{B.38}$$

In this case, there exists a $\delta > 0$ such that, for all n in an infinite sequence $N \subseteq \mathbb{N}$, the probability $P_0^n \left\{\eta_n^E \geq \eta_0^E + \epsilon\right\}$ ($P_0^n \left\{\underline{\eta}_n^E \geq \underline{\eta}_0^E + \epsilon\right\}$ resp.) is at least δ. Now, in the first setting, for any $n \in N$, the following holds with probability at least δ:

$$\begin{aligned} P_0\{d_{n,k_n}\bar{\mu}_0^A\} - \kappa &\leq P_0\{I(\xi_n > \eta_0^E + \epsilon/2)\bar{\mu}_0^A\} - \kappa \\ &= P_0\{[I(\xi_n > \eta_0^E + \epsilon/2) - I(\xi_0 > \eta_0^E + \epsilon/2)]\bar{\mu}_0^A\} + \Gamma_0(\eta_0^E + \epsilon/2) - \kappa. \end{aligned} \tag{B.39}$$

We now argue that the first term is $\mathrm{o}_p(1)$. For any $x > 0$ and $n \in \mathbb{N}$, Lemma B.5 shows that

$$\begin{aligned} |P_0\{[I(\xi_n > \eta_0^E + \epsilon/2) - I(\xi_0 > \eta_0^E + \epsilon/2)]\bar{\mu}_0^A\}| &\leq P_0 I(|\xi_n - \xi_0| > x) + P_0 I(|\xi_0 - \eta_n^E + \epsilon/2| \leq x) \\ &\leq \frac{\|\xi_n - \xi_0\|_{1,P_0}}{x} + P_0 I(|\xi_0 - \eta_n^E + \epsilon/2| \leq x). \end{aligned}$$

Similarly to the proof of Lemma B.6, the fact that $\|\xi_n - \xi_0\|_{1,P_0} = \mathrm{o}_p(1)$ (Condition C12) ensures that we can substitute in a sequence x_n for x to show that $P_0\{[I(\xi_n > \eta_0^E + \epsilon/2) - I(\xi_0 > \eta_0^E + \epsilon/2)]\bar{\mu}_0^A = \mathrm{o}_p(1)$. By Condition C3, $\Gamma_0(\eta_0^E + \epsilon/2) - \Gamma_0(\eta_0^E)$ is a negative constant. Because (B.39) holds with probability at least $\delta > 0$ for infinitely many n, this shows that $P_0\{d_{n,k_n}\bar{\mu}_0^A\} - \kappa$ *is not* $\mathrm{o}_p(1)$. This contradicts our result from the already-proven part ii) of this lemma. Therefore, (B.38) is false, that is, $\limsup_n P_0^n \left\{\eta_n^E \geq \eta_0^E + \epsilon\right\} = 0$. Almost identical arguments show that (B.38) is false in the second setting.

The second part of the contradiction argument involves assuming that, for some $\epsilon > 0$, $\limsup_n P_0^n \left\{\eta_n^E \leq \eta_0^E - \epsilon\right\} > 0$ ($\limsup_n P_0^n \left\{\underline{\eta}_n^E \leq \underline{\eta}_0^E - \epsilon\right\} > 0$ resp.). In this case, there exists a $\delta > 0$ such that, for all n in an infinite sequence $N \subseteq \mathbb{N}$, $P_0^n \left\{\eta_n^E \leq \eta_0^E - \epsilon\right\} \geq \delta$ ($P_0^n \left\{\underline{\eta}_n^E \leq \underline{\eta}_0^E - \epsilon\right\} \geq \delta$ resp.). Now, in the first setting, for any $n \in N$, the following holds with probability at least δ:

$$P_0\{d_{n,k_n}\bar{\mu}_0^A\} - \kappa \geq P_0\{I(\xi_n > \eta_0^E - \epsilon)\bar{\mu}_0^A\} - \kappa$$

$$= P_0\{[I(\xi_n > \eta_0^E - \epsilon) - I(\xi_0 > \eta_0^E - \epsilon)]\bar{\mu}_0^A\} + \Gamma_0(\eta_0^E - \epsilon) - \kappa.$$

The rest of the argument is extremely similar to the contradiction argument in the previous case, and is therefore omitted.

Since ϵ is arbitrary, combining the results of these two contradiction arguments shows that $\tau_n^E - \tau_0^E = \eta_n^E - \eta_0^E + \mathrm{o}_p(1) = \mathrm{o}_p(1)$ ($\underline{\tau}_n^E - \underline{\tau}_0^E = \underline{\eta}_n^E - \underline{\eta}_0^E + \mathrm{o}_p(1) = \mathrm{o}_p(1)$ resp.), as desired. □

Proof of iii) from Lemma B.11 when $\eta_0^E = 0$ ($\underline{\eta}_0^E = 0$ resp.). If $\eta_0^E = 0$ ($\underline{\eta}_0^E = 0$ resp.), then the construction of η_n^E ($\underline{\eta}_n^E$ resp.) implies that η_n^E ($\underline{\eta}_n^E$ resp.) takes values from two sequences: $\eta_n^E(\kappa)$ ($\underline{\eta}_n^E(\kappa)$ resp.) and $\eta_n^E(k_n)$ where k_n is a solution to (3.5) ($\underline{\eta}_n^E(\underline{k}_n)$ where $\underline{k}_n$ is a solution to (3.6), resp.). By Lemma B.10, $\eta_n^E(\kappa)$ ($\underline{\eta}_n^E(\kappa)$ resp.) is consistent for η_0^E ($\underline{\eta}_0^E$ resp.). When a solution to (3.5) ((3.5) resp.) exists and equals k_n ($\underline{k}_n$ resp.), iii) from Lemma B.11 shows that $\eta_n^E(k_n)$ ($\underline{\eta}_n^E(\underline{k}_n)$ resp.) is consistent for η_0^E ($\underline{\eta}_0^E$ resp.). Because $|\tau_n^E - \tau_0^E| \leq |\eta_n^E - \eta_0^E|$ ($|\underline{\tau}_n^E - \underline{\tau}_0^E| \leq |\underline{\eta}_n^E - \underline{\eta}_0^E|$ resp.), it follows that $\tau_n^E - \tau_0^E = \mathrm{o}_p(1)$ ($\underline{\tau}_n^E - \underline{\tau}_0^E = \mathrm{o}_p(1)$ resp.). □

Proof of iii) from Lemma B.11 when $\eta_0^E < 0$ ($\underline{\eta}_0^E < 0$ resp.). If $\eta_0^E < 0$ ($\underline{\eta}_0^E < 0$ resp.), then by Lemma B.10, $\eta_n^E(\kappa) \leq 0$ ($\underline{\eta}_n^E(\kappa) \leq 0$ resp.) with probability tending to one. Hence, with probability tending to one, $\tau_n^E = 0 = \tau_0^E$ ($\underline{\tau}_n^E = 0 = \underline{\tau}_0^E$ resp.). Therefore, iii) holds in the case that $\eta_0^E < 0$ ($\underline{\eta}_0^E < 0$ resp.). □

Lemma B.12. *Under Conditions A3, C6 and C9,*

$$\sup_{e:\mathcal{W}\to[0,1]} \left| R_e^E(\hat{P}_n, P_0) \right| = \mathrm{o}_p(n^{-1/2}).$$

Note that, because V is a function of W, the formulation of e in the above lemma includes the special case where e is a function of V.

Proof of Lemma B.12. By the boundedness of the range of e,

$$\sup_{e:\mathcal{W}\to[0,1]} \left| R_e^E(\hat{P}_n, P_0) \right|$$

$$\leq P_0 \left| e(\cdot) \left[\frac{\mu_n^Z(\cdot) - \mu_0^Z(\cdot)}{\mu_n^Z(\cdot)} \{\hat{\mu}_n^Y(1,\cdot) - \mu_0^Y(1,\cdot)\} + \frac{\mu_n^Z(\cdot) - \mu_0^Z(\cdot)}{1 - \mu_n^Z} \{\hat{\mu}_n^Y(0,\cdot) - \mu_0^Y(0,\cdot)\} \right] \right|$$
$$\leq P_0 \left| \left[\frac{\mu_n^Z(\cdot) - \mu_0^Z(\cdot)}{\mu_n^Z(\cdot)} \{\hat{\mu}_n^Y(1,\cdot) - \mu_0^Y(1,\cdot)\} + \frac{\mu_n^Z(\cdot) - \mu_0^Z(\cdot)}{1 - \mu_n^Z(\cdot)} \{\hat{\mu}_n^Y(0,\cdot) - \mu_0^Y(0,\cdot)\} \right] \right|.$$

Using Condition C6 and Lemma B.8, the display continues as

$$\lesssim P_0 \left| (\mu_n^Z(\cdot) - \mu_0^Z(\cdot))[\hat{\mu}_n^Y(1,\cdot) - \mu_0^Y(1,\cdot)] \right| + P_0 \left| (\mu_n^Z(\cdot) - \mu_0^Z(\cdot))[\hat{\mu}_n^Y(0,\cdot) - \mu_0^Y(0,\cdot)] \right|$$
$$\leq \|\mu_n^Z - \mu_0^Z\|_{2,P_0} \|\hat{\mu}_n^Y(1,\cdot) - \mu_0^Y(1,\cdot)\|_{2,P_0} + \|\mu_n^Z - \mu_0^Z\|_{2,P_0} \|\hat{\mu}_n^Y(0,\cdot) - \mu_0^Y(0,\cdot)\|_{2,P_0}$$
$$\lesssim \|\mu_n^Z - \mu_0^Z\|_{2,P_0} \|\hat{\mu}_n^Y - \mu_0^Y\|_{2,P_0}.$$

The right-hand side is $o_p(n^{-1/2})$ by Condition C9. □

Lemma B.13. *Under the conditions of Theorem 3.4,*

$$P_n \hat{\mu}_n^A - P_0 \mu_0^A = (P_n - P_0) D(P_0, \mu_0^A) + o_p(n^{-1/2}) = O_p(n^{-1/2}).$$

This result follows from the fact that $P_n \hat{\mu}_n^A$ is a TMLE for $P_0 \mu_0^A$ (e.g., van Der Laan, 2017; van der Laan and Rose, 2018), and therefore the proof is omitted.

We now prove Theorem 3.4.

Proof of Theorem 3.4. By the expansion of $P \mapsto \Psi_{e_P}^E(P)$ and $P \mapsto \Psi_{\underline{e}_P}^E(P)$ presented in Section B.8.5,

$$\begin{aligned}
&\Psi_{e_n}^E(\hat{P}_n) - \Psi_{e_0}^E(P_0) \\
&\quad = P_0 D^E(\hat{P}_n, e_n, \tau_0^E, \mu_n^A) + R_{e_n}^E(\hat{P}_n, P_0) \\
&\qquad - \tau_0^E P_0 \left\{ e_n \frac{\mu_n^Z(\cdot) - \mu_0^Z(\cdot)}{\mu_n^Z(\cdot)} [\mu_n^A(1,\cdot) - \mu_0^A(1,\cdot)] \right\} + P_0\{(e_n - e_0)(\Delta_{b,0}^Y - \tau_0^E \bar{\mu}_0^A)\} \\
&\quad = (P_n - P_0) D^E(P_0, e_0, \tau_0^E, \mu_0^A) - P_n D^E(\hat{P}_n, e_n, \tau_0^E, \mu_n^A) \\
&\qquad + (P_n - P_0) \left[D^E(\hat{P}_n, e_n, \tau_0^E, \mu_n^A) - D^E(P_0, e_0, \tau_0^E, \mu_0^A) \right] \\
&\qquad + R_{e_n}^E(\hat{P}_n, P_0) + P_0\{(e_n - e_0)(\Delta_{b,0}^Y - \tau_0^E \bar{\mu}_0^A)\} \\
&\qquad - \tau_0^E P_0 \left\{ e_n(\cdot) \frac{\mu_n^Z(\cdot) - \mu_0^Z(\cdot)}{\mu_n^Z(\cdot)} [\mu_n^A(1,\cdot) - \mu_0^A(1,\cdot)] \right\},
\end{aligned}$$

$$
\begin{aligned}
&\Psi^E_{\underline{e}_n}(\hat{\underline{P}}_n) - \Psi^E_{\underline{e}_0}(P_0) \\
&\quad = P_0 \underline{D}^E(\hat{\underline{P}}_n, \underline{e}_n, \underline{\tau}^E_0, \mu^A_n) + R^E_{\underline{e}_n}(\hat{\underline{P}}_n, P_0) + P_0\{(\underline{e}_n - \underline{e}_0)(\Delta^Y_{b,0} - \underline{\tau}^E_0 \Delta^A_{b,0})\} \\
&\qquad - \underline{\tau}^E_0 P_0 \left\{ \underline{e}_n(\cdot) \frac{\mu^Z_n(\cdot) - \mu^Z_0(\cdot)}{\mu^Z_n(\cdot)} [\mu^A_n(1,\cdot) - \mu^A_0(1,\cdot)] \right\} \\
&\qquad + \underline{\tau}^E_0 P_0 \left\{ (1 - \underline{e}_n(\cdot)) \frac{\mu^Z_n(\cdot) - \mu^Z_0(\cdot)}{1 - \mu^Z_n(\cdot)} [\mu^A_n(0,\cdot) - \mu^A_0(0,\cdot)] \right\} \\
&\quad = (P_n - P_0) \underline{D}^E(P_0, \underline{e}_0, \underline{\tau}^E_0, \mu^A_0) - P_n \underline{D}^E(\hat{\underline{P}}_n, \underline{e}_n, \underline{\tau}^E_0, \mu^A_n) \\
&\qquad + (P_n - P_0) \left[\underline{D}^E(\hat{\underline{P}}_n, \underline{e}_n, \underline{\tau}^E_0, \mu^A_n) - \underline{D}^E(P_0, \underline{e}_0, \underline{\tau}^E_0, \mu^A_0) \right] \\
&\qquad + R^E_{\underline{e}_n}(\hat{\underline{P}}_n, P_0) + P_0\{(\underline{e}_n - \underline{e}_0)(\Delta^Y_{b,0} - \underline{\tau}^E_0 \Delta^A_{b,0})\} \\
&\qquad - \underline{\tau}^E_0 P_0 \left\{ \underline{e}_n(\cdot) \frac{\mu^Z_n(\cdot) - \mu^Z_0(\cdot)}{\mu^Z_n(\cdot)} [\mu^A_n(1,\cdot) - \mu^A_0(1,\cdot)] \right\} \\
&\qquad + \underline{\tau}^E_0 P_0 \left\{ (1 - \underline{e}_n(\cdot)) \frac{\mu^Z_n(\cdot) - \mu^Z_0(\cdot)}{1 - \mu^Z_n(\cdot)} [\mu^A_n(0,\cdot) - \mu^A_0(0,\cdot)] \right\}.
\end{aligned}
$$

Similarly, by the expansion presented in Section B.8.1,

$$
\begin{aligned}
\Psi^E_{e^{\mathrm{FR}}}(\hat{P}_n) - \Psi^E_{e^{\mathrm{FR}}}(P_0) &= (P_n - P_0) D^E(P_0, e^{\mathrm{FR}}, 0, \mu^A_0) - P_n D^E(\hat{P}_n, e^{\mathrm{FR}}, 0, \mu^A_0) \\
&\quad + (P_n - P_0) \left[D^E(\hat{P}_n, e^{\mathrm{FR}}, 0, \mu^A_0) - D^E(P_0, e^{\mathrm{FR}}, 0, \mu^A_0) \right] + R^E_{e^{\mathrm{FR}}}(\hat{P}_n, P_0);
\end{aligned}
$$

by the expansion presented in Section B.8.2

$$
\begin{aligned}
&\Psi^E_{e^{\mathrm{RD}}_n}(\hat{P}_n) - \Psi^E_{e^{\mathrm{RD}}_0}(P_0) \\
&= (P_n - P_0) \left[D^E(P_0, e^{\mathrm{RD}}_0, 0, \mu^A_0) - \frac{\Psi^E_{e^{\mathrm{RD}}_0}(P_0)}{P_0 \mu^A_0(1,\cdot)} D(P_0, \mu^A_0) \right] \\
&\quad - P_n \left[D^E(\hat{P}_n, e^{\mathrm{RD}}_n, 0, \mu^A_0) - \frac{\Psi^E_{e^{\mathrm{RD}}_n}(\hat{P}_n)}{P_n \mu^A_n(1,\cdot)} D(\hat{P}_n, \hat{\mu}^A_n) \right] \\
&\quad + (P_n - P_0) \Bigg[D^E(\hat{P}_n, e^{\mathrm{RD}}_n, 0, \mu^A_0) - \frac{\Psi^E_{e^{\mathrm{RD}}_n}(\hat{P}_n)}{P_n \mu^A_n(1,\cdot)} D(\hat{P}_n, \hat{\mu}^A_n) \\
&\qquad\qquad - D^E(P_0, e^{\mathrm{RD}}_0, 0, \mu^A_0) + \frac{\Psi^E_{e^{\mathrm{RD}}_0}(P_0)}{P_0 \mu^A_0(1,\cdot)} D(P_0, \mu^A_0) \Bigg] \\
&\quad + R^E_{e^{\mathrm{RD}}_n}(\hat{P}_n, P_0) + \kappa(P_n \hat{\Delta}^Y_n - P_0 \Delta^Y_0) \frac{P_n \hat{\mu}^A_n(1,\cdot) - P_0 \mu^A_0(1,\cdot)}{[P_n \hat{\mu}^A_n(1,\cdot)][P_0 \mu^A_0(1,\cdot)]}
\end{aligned}
$$

$$- \kappa \frac{P_n \hat{\Delta}_n^Y}{[P_n \hat{\mu}_n^A(1,\cdot)]^2} P_0 \left\{ \frac{\mu_n^Z(\cdot) - \mu_0^Z(\cdot)}{\mu_n^Z(\cdot)} [\hat{\mu}_n^A(1,\cdot) - \mu_0^A(1,\cdot)] \right\} - \kappa (P_n \hat{\Delta}_n^Y) \frac{[P_n \hat{\mu}_n^A(1,\cdot) - P_0 \mu_0^A(1,\cdot)]^2}{[P_n \hat{\mu}_n^A(1,\cdot)]^2 [P_0 \mu_0^A(1,\cdot)]};$$

by the expansion presented in Section B.8.3,

$$\begin{aligned}
&\Psi^E_{\underline{e}_n^{\mathrm{RD}}}(\underline{\hat{P}}_n) - \Psi^E_{\underline{e}_0^{\mathrm{RD}}}(P_0) \\
&= (P_n - P_0) \underline{D}^E(P_0, \underline{e}_0^{\mathrm{RD}}, 0, \mu_0^A) - P_n \underline{D}^E(\underline{\hat{P}}_n, \underline{e}_n^{\mathrm{RD}}, 0, \mu_0^A) \\
&\quad + (P_n - P_0) \left[\underline{D}^E(\underline{\hat{P}}_n, \underline{e}_n^{\mathrm{RD}}, 0, \mu_0^A) - \underline{D}^E(P_0, \underline{e}_0^{\mathrm{RD}}, 0, \mu_0^A) \right] \\
&\quad + R^E_{e_n^{\mathrm{RD}}}(\hat{P}_n, P_0) + (\underline{e}_n^{\mathrm{RD}} - \underline{e}_0^{\mathrm{RD}}) P_0 \Delta_0^Y;
\end{aligned}$$

by the expansion presented in Section B.8.4,

$$\begin{aligned}
&\Psi^E_{e_n^{\mathrm{TP}}}(\hat{P}_n) - \Psi^E_{e_0^{\mathrm{TP}}}(P_0) \\
&\quad = (P_n - P_0) G^E_{\mathrm{TP}}(P_0) - P_n G^E_{\mathrm{TP}}(\hat{P}_n) + (P_n - P_0) \left[G^E_{\mathrm{TP}}(\hat{P}_n) - G^E_{\mathrm{TP}}(P_0) \right] + R^E_{e_0^{\mathrm{TP}}}(\hat{P}_n, P_0).
\end{aligned}$$

Momentarily, we will separately consider the following three cases: $\mathcal{R} = \mathrm{FR}$, $\mathcal{R} = \mathrm{RD}$, and $\mathcal{R} = \mathrm{TP}$. Before doing this, we note the following facts, which will be sufficient to ensure that the remainders and empirical process terms in all of the first-order expansions given above are $\mathrm{o}_p(n^{-1/2})$. By Condition C9, Lemma B.12, and Lemma B.13, using the Cauchy-Schwarz inequality and boundedness of the range of an IER, the following terms are all $\mathrm{o}_p(n^{-1/2})$:

$$\begin{aligned}
&R^E_e(\hat{P}_n, P_0) \text{ for } e = e_n, e^{\mathrm{FR}}, e_n^{\mathrm{RD}}, e_0^{\mathrm{TP}}, \underline{e}_n, \underline{e}^{\mathrm{FR}}, \underline{e}_n^{\mathrm{RD}}, \underline{e}_0^{\mathrm{TP}}, \\
&\tau_0^E P_0 \left\{ e_n \frac{\mu_n^Z(\cdot) - \mu_0^Z(\cdot)}{\mu_n^Z(\cdot)} [\mu_n^A(1,\cdot) - \mu_0^A(1,\cdot)] \right\}, \\
&\tau_0^E P_0 \left\{ \underline{e}_n(\cdot) \frac{\mu_n^Z(\cdot) - \mu_0^Z(\cdot)}{\mu_n^Z(\cdot)} [\mu_n^A(1,\cdot) - \mu_0^A(1,\cdot)] \right\}, \\
&\tau_0^E P_0 \left\{ (1 - \underline{e}_n(\cdot)) \frac{\mu_n^Z(\cdot) - \mu_0^Z(\cdot)}{1 - \mu_n^Z(\cdot)} [\mu_n^A(0,\cdot) - \mu_0^A(0,\cdot)] \right\}, \\
&\kappa(P_n \hat{\Delta}_n^Y) \frac{[P_n \hat{\mu}_n^A(1,\cdot) - P_0 \mu_0^A(1,\cdot)]^2}{[P_n \hat{\mu}_n^A(1,\cdot)]^2 [P_0 \mu_0^A(1,\cdot)]}, \\
&\kappa(P_n \hat{\Delta}_n^Y - P_0 \Delta_0^Y) \frac{P_n \hat{\mu}_n^A(1,\cdot) - P_0 \mu_0^A(1,\cdot)}{[P_n \hat{\mu}_n^A(1,\cdot)][P_0 \mu_0^A(1,\cdot)]},
\end{aligned}$$

$$\kappa \frac{P_n \hat{\Delta}_n^Y}{[P_n \hat{\mu}_P^A(1,\cdot)]^2} P_0 \left\{ \frac{\mu_n^Z(\cdot) - \mu_0^Z(\cdot)}{\mu_n^Z(\cdot)} [\hat{\mu}_n^A(1,\cdot) - \mu_0^A(1,\cdot)] \right\}.$$

By Condition C8, $P_0\{(e_n - e_0)(\Delta_{b,0}^Y - \tau_0^E \bar{\mu}_0^A)\} = \mathrm{o}_p(n^{-1/2})$ ($P_0\{(\underline{e}_n - \underline{e}_0)(\Delta_{b,0}^Y - \underline{\tau}_0^E \Delta_{b,0}^A)\} = \mathrm{o}_p(n^{-1/2})$ resp.). By Conditions C10 and C11, $(P_n - P_0)\left[D_{n,\mathcal{R}}^E - D_{\mathcal{R}}^E(P_0)\right] = \mathrm{o}_p(n^{-1/2})$ and $(P_n - P_0)\left[\underline{D}_{n,\mathcal{R}}^E - \underline{D}_{\mathcal{R}}^E(P_0)\right] = \mathrm{o}_p(n^{-1/2})$ for all $\mathcal{R} \in \{\mathrm{FR}, \mathrm{RD}, \mathrm{TP}\}$; $(P_n - P_0)\left[\underline{D}^E(\hat{\underline{P}}_n, \underline{e}_n^{\mathrm{RD}}, 0, \mu_0^A) - \underline{D}^E(P_0, \underline{e}_0^{\mathrm{RD}}, 0, \mu_0^A)\right] = \mathrm{o}_p(n^{-1/2})$. Therefore, all relevant remainders and empirical process terms are $\mathrm{o}_p(n^{-1/2})$.

Case I: $\mathcal{R} = FR$. In this case, for the first setting,

$$\begin{aligned}
\psi_n^E - \psi_0^E &= (P_n - P_0) D_{\mathrm{FR}}^E(P_0) - P_n D_{n,\mathrm{FR}}^E + \mathrm{o}_p(n^{-1/2}) \\
&= (P_n - P_0) D_{\mathrm{FR}}^E(P_0) \\
&\quad - \tau_0^E \left\{ \frac{1}{n} \sum_{i=1}^n e_n(V_i) \left[\mu_n^A(1, W_i) + \frac{Z_i}{\mu_n^Z(W_i)} [A_i - \mu_n^A(1, W_i)] \right] - \kappa \right\} + \mathrm{o}_p(n^{-1/2}),
\end{aligned}$$

where the last step follows from the TMLE construction of $\hat{P}_n$ (Step 4a of our estimator), which implies that

$$\frac{1}{n} \sum_{i=1}^n \left\{ \frac{e_n(V_i) - e^{\mathrm{FR}}(V)}{Z_i + \mu_n^Z(W_i) - 1} [Y_i - \hat{\mu}_n^Y(Z_i, W_i)] \right\} = 0.$$

If $\tau_0^E = 0$, then this term is zero. Otherwise, $\tau_0^E = \eta_0^E > 0$, so, by Lemma B.11, the following holds with probability tending to one:

$$\frac{1}{n} \sum_{i=1}^n e_n(V_i) \left[\mu_n^A(1, W_i) + \frac{Z_i}{\mu_n^Z(W_i)} [A_i - \mu_n^A(1, W_i)] \right] = \kappa.$$

Therefore, $\psi_n^E - \psi_0^E = (P_n - P_0) D_{\mathrm{FR}}^E(P_0) + \mathrm{o}_p(n^{-1/2})$.

Similarly, for the second setting, the TMLE construction of $\hat{\underline{P}}_n$ implies that

$$\frac{1}{n} \sum_{i=1}^n \left\{ \frac{\underline{e}_n(V_i) - \underline{e}^{\mathrm{FR}}(V)}{Z_i + \mu_n^Z(W_i) - 1} [Y_i - \hat{\underline{\mu}}_n^Y(Z_i, W_i)] \right\} = 0.$$

In addition, either $\underline{\tau}_0^E = 0$ or

$$\frac{1}{n} \sum_{i=1}^n \underline{e}_n(V_i) \left\{ \Delta_n^A(W_i) + \frac{1}{Z_i + \mu_n^Z(W_i) - 1} \left[A_i - \mu_n^A(Z_i, W_i) \right] \right\}$$

$$+ \mu_n^A(0, W_i) + \frac{1 - Z_i}{1 - \mu_n^Z(W_i)}[A_i - \mu_n^A(0, W_i)] = \kappa$$

with probability tending to 1 by Lemma B.11 and definition of φ_n. Almost identical arguments show that $\underline{\psi}_n^E - \underline{\psi}_0^E = (P_n - P_0)\underline{D}_{\mathrm{FR}}^E(P_0) + \mathrm{o}_p(n^{-1/2})$.

Case II: $\mathcal{R} = RD$. In this case, in the first setting,

$$\psi_n^E - \psi_0^E = (P_n - P_0)D_{\mathrm{RD}}^E(P_0) - P_n D_{n,\mathrm{RD}}^E + \mathrm{o}_p(n^{-1/2}).$$

The TMLE construction of $\hat{P}_n$ (Step 3a and 4a of our estimator) implies that

$$\frac{1}{n}\sum_{i=1}^n \frac{e_n(V_i) - e_n^{\mathrm{RD}}(V_i)}{Z_i + \mu_n^Z(W_i) - 1}[Y_i - \hat{\mu}_n^Y(Z_i, W_i)] = 0,$$
$$P_n D(\hat{P}_n, \hat{\mu}_n^A) = \frac{1}{n}\sum_{i=1}^n \frac{Z_i}{\mu_n^Z(W_i)}[A_i - \mu^A(1, W_i)] = 0,$$

and hence

$$P_n D_{n,\mathrm{RD}}^E = \tau_0^E \left\{ \frac{1}{n}\sum_{i=1}^n e_n(V_i)\left[\mu_n^A(1, W_i) + \frac{Z_i}{\mu_n^Z(W_i)}[A_i - \mu_n^A(1, W_i)]\right] - \kappa \right\},$$

which is zero with probability tending to one as proved above. Therefore,

$$\psi_n^E - \psi_0^E = (P_n - P_0)D_{\mathrm{RD}}^E(P_0) + \mathrm{o}_p(n^{-1/2}).$$

The proof for the second setting is similar in view of the fact that $\underline{e}_n^{\mathrm{RD}}$ is an asymptotically linear estimator of $\underline{e}_0^{\mathrm{RD}}$ by Lemma B.9.

Case 3: $\mathcal{R} = TP$. In this case, for the first setting

$$\psi_n^E - \psi_0^E = (P_n - P_0)D_{\mathrm{TP}}^E(P_0) - P_n D_{n,\mathrm{TP}}^E + \mathrm{o}_p(n^{-1/2}).$$

The TMLE construction of $\hat{P}_n$ (Step 4a of our estimator) implies that

$$\frac{1}{n}\sum_{i=1}^n \left\{ \frac{e_n(V_i) - Z_i}{Z_i + \mu_n^Z(W_i) - 1}[Y_i - \hat{\mu}_n^Y(Z_i, W_i)] \right\} = 0,$$

so

$$P_n D_{n,\mathrm{TP}}^E = \tau_0^E \left\{ \frac{1}{n}\sum_{i=1}^n e_n(V_i)\left[\mu_n^A(1, W_i) + \frac{Z_i}{\mu_n^Z(W_i)}[A_i - \mu_n^A(1, W_i)]\right] - \kappa \right\},$$

which is zero with probability tending to one as proved above. Therefore,

$$\psi_n^E - \psi_0^E = (P_n - P_0)D_{\mathrm{TP}}^E(P_0) + \mathrm{o}_p(n^{-1/2}).$$

The proof for the second setting is almost identical.

Conclusion: asymptotic normality results. For any $\mathcal{R} \in \{\mathrm{FR}, \mathrm{RD}, \mathrm{TP}\}$, by the central limit theorem and Slutsky's theorem, the asymptotic linearity of ψ_n^E ($\underline{\psi}_n^E$ resp.) imply that $\sqrt{n}(\psi_n^E - \psi_0^E) \xrightarrow{d} \mathrm{N}(0, P_0 D_{\mathcal{R}}^E(P_0)^2)$ ($\sqrt{n}\left(\underline{\psi}_n^E - \underline{\psi}_0^E\right) \xrightarrow{d} \mathrm{N}\left(0, P_0 \underline{D}_{\mathcal{R}}^E(P_0)^2\right)$ resp.). □

Proof of Corollary 3.4.1. The desired result follows from straightforward application of the delta method for influence functions. Indeed, for the first setting,

$$\begin{aligned}
\frac{\psi_n^E}{\phi_n} =& \frac{\psi_0^E + (P_n - P_0)D_{\mathcal{R}}^E(P_0) + \mathrm{o}_p(n^{-1/2})}{\phi_0 + (P_n - P_0)D_A(P_0) + \mathrm{o}_p(n^{-1/2})} \\
=& \{\psi_0^E + (P_n - P_0)D_{\mathcal{R}}^E(P_0) + \mathrm{o}_p(n^{-1/2})\} \left\{ \frac{1}{\phi_0} - (P_n - P_0)\frac{D_A(P_0)}{\phi_0^2} + \mathrm{o}_p(n^{-1/2}) \right\} \\
=& \frac{\psi_0^E}{\phi_0} + (P_n - P_0)\left\{ \frac{D_{\mathcal{R}}^E(P_0)}{\phi_0} - \frac{\psi_0^E}{\phi_0^2} D_A(P_0) \right\} + \mathrm{o}_p(n^{-1/2}).
\end{aligned}$$

Subtracting $\frac{\psi_0^E}{\phi_0}$ from both sides yields the desired result. The derivation for the second setting is almost identical. The asymptotic normality result follows the central limit theorem and Slutsky's theorem. □

B.9.3 Individualized treatment rules

Lemma B.14. *Under Condition A2, A3, B3, B4, B6 and B7,*

$$\sup_{t:\mathcal{V}\to[0,1]} \left| R^T(\hat{P}_n, \mu_n^Y, \mu_n^A, t, P_0) \right| = \mathrm{o}_p(n^{-1/2}).$$

Proof of Lemma B.14. The remainder $R^T(\hat{P}_n, \mu_n^Y, \mu_n^A, t, P_0)$ is a sum of five terms and we deal with them separately. By the boundedness of the range of t,

$$\left| P_0 \left\{ t \frac{\hat{\Delta}_n^A - \Delta_0^A}{\hat{\Delta}_n^A} (\hat{\Delta}_n - \Delta_0) \right\} \right| \leq P_0 \left| t \frac{\hat{\Delta}_n^A - \Delta_0^A}{\hat{\Delta}_n^A} (\hat{\Delta}_n - \Delta_0) \right| \leq P_0 \left| \frac{\hat{\Delta}_n^A - \Delta_0^A}{\hat{\Delta}_n^A} (\hat{\Delta}_n - \Delta_0) \right|.$$

By Condition B4, Cauchy-Schwarz inequality and Lemma B.8, the display continues as

$$\lesssim P_0|(\hat{\Delta}_n^A - \Delta_0^A)(\hat{\Delta}_n - \Delta_0)| \leq \|\hat{\Delta}_n^A - \Delta_0^A\|_{2,P_0}\|\hat{\Delta}_n - \Delta_0\|_{2,P_0}$$
$$\lesssim \|\hat{\mu}_n^A - \mu_0^A\|_{2,P_0}(\|\hat{\mu}_n^A - \mu_0^A\|_{2,P_0} + \|\hat{\mu}_n^Y - \mu_0^Y\|_{2,P_0}).$$

The right-hand side is $o_p(n^{-1/2})$ by Condition B6. We study the remaining four terms similarly.

$$\begin{aligned}
&\left|P_0\left\{t\left(\frac{1}{\Delta_n^A} - \frac{1}{\hat{\Delta}_n^A}\right)(\Delta_0^Y - \hat{\Delta}_n^Y)\right\}\right| \\
&\lesssim P_0|(\Delta_n^A - \hat{\Delta}_n^A)(\Delta_0^Y - \hat{\Delta}_n^Y)| && \text{(boundedness of } t\text{, Condition B4)} \\
&\leq \|\Delta_n^A - \hat{\Delta}_n^A\|_{2,P_0}\|\Delta_0^Y - \hat{\Delta}_n^Y\|_{2,P_0} && \text{(Cauchy-Schwarz)} \\
&\lesssim \|\hat{\mu}_n^A - \mu_n^A\|_{2,P_0}\|\hat{\mu}_n^Y - \mu_0^Y\|_{2,P_0}, && \text{(Lemma B.8)}
\end{aligned}$$

which is $o_p(n^{-1/2})$ by Condition B6.

$$\begin{aligned}
&\left|P_0\left\{t\left(\frac{1}{\Delta_n^A} - \frac{1}{\hat{\Delta}_n^A}\right)\hat{\Delta}_n \cdot (\Delta_0^A - \hat{\Delta}_n^A)\right\}\right| \\
&\leq P_0\left|t\left(\frac{1}{\Delta_n^A} - \frac{1}{\hat{\Delta}_n^A}\right)\hat{\Delta}_n \cdot (\Delta_0^A - \hat{\Delta}_n^A)\right| \\
&\lesssim P_0|(\Delta_n^A - \hat{\Delta}_n^A)(\Delta_0^A - \hat{\Delta}_n^A)| && \text{(boundedness of } t\text{, Condition B4)} \\
&\leq \|\Delta_n^A - \hat{\Delta}_n^A\|_{2,P_0}\|\Delta_0^A - \hat{\Delta}_n^A\|_{2,P_0}, && \text{(Cauchy-Schwarz)}
\end{aligned}$$

which is $o_p(n^{-1/2})$ by Condition B6.

$$\begin{aligned}
&\left|P_0\left\{\frac{t}{\Delta_n^A}(\Delta_n - \hat{\Delta}_n)(\Delta_0^A - \hat{\Delta}_n^A)\right\}\right| \\
&\leq P_0\left|\frac{t}{\Delta_n^A}(\Delta_n - \hat{\Delta}_n)(\Delta_0^A - \hat{\Delta}_n^A)\right| \\
&\lesssim P_0|(\Delta_n - \hat{\Delta}_n)(\Delta_0^A - \hat{\Delta}_n^A)| && \text{(boundedness of } t\text{, Condition B4)} \\
&\leq \|\Delta_n - \hat{\Delta}_n\|_{2,P_0}\|\Delta_0^A - \hat{\Delta}_n^A\|_{2,P_0} && \text{(Cauchy-Schwarz)} \\
&\lesssim (\|\mu_n^Y - \hat{\mu}_n^Y\|_{2,P_0} + \|\mu_n^A - \hat{\mu}_n^A\|_{2,P_0})\|\mu_0^A - \hat{\mu}_n^A\|_{2,P_0}, && \text{(Condition B4, Lemma B.8)}
\end{aligned}$$

which is $o_p(n^{-1/2})$ by Condition B6. By the boundedness of the range of t, Conditions B3 and B4,

$$\begin{aligned}
&\left| P_0 \left\{ \frac{t(V(\cdot))}{\Delta_n^A(\cdot)} \left[\frac{\mu_0^Z(\cdot) - \mu_n^Z(\cdot)}{\mu_n^Z(\cdot)} [(\mu_0^Y(1,\cdot) - \hat{\mu}_n^Y(1,\cdot)) - \Delta_n(\cdot)(\mu_0^A(1,\cdot) - \hat{\mu}_n^A(1,\cdot))] \right.\right.\right. \\
&\quad \left.\left.\left. - \frac{\mu_n^Z(\cdot) - \mu_0^Z(\cdot)}{1 - \mu_n^Z(\cdot)} [(\mu_0^Y(0,\cdot) - \hat{\mu}_n^Y(0,\cdot)) - \Delta_n(\cdot)(\mu_0^A(0,\cdot) - \hat{\mu}_n^A(0,\cdot))] \right] \right\} \right| \\
&\lesssim P_0 \left| (\mu_0^Z(\cdot) - \mu_n^Z(\cdot))(\mu_0^Y(1,\cdot) - \hat{\mu}_n^Y(1,\cdot)) \right| + P_0 \left| (\mu_0^Z(\cdot) - \mu_n^Z(\cdot))(\mu_0^Y(0,\cdot) - \hat{\mu}_n^Y(0,\cdot)) \right| \\
&\quad + P_0 \left| (\mu_0^Z(\cdot) - \mu_n^Z(\cdot))(\mu_0^A(1,\cdot) - \hat{\mu}_n^A(1,\cdot)) \right| + P_0 \left| (\mu_0^Z(\cdot) - \mu_n^Z(\cdot))(\mu_0^A(0,\cdot) - \hat{\mu}_n^A(0,\cdot)) \right|.
\end{aligned}$$

By Cauchy-Schwarz, the display continues as

$$\begin{aligned}
&\leq \|\mu_0^Z - \mu_n^Z\|_{2,P_0} \Big(\|\mu_0^Y(1,\cdot) - \hat{\mu}_n^Y(1,\cdot)\|_{2,P_0} + \|\mu_0^Y(0,\cdot) - \hat{\mu}_n^Y(0,\cdot)\|_{2,P_0} \\
&\quad + \|\mu_0^A(1,\cdot) - \hat{\mu}_n^A(1,\cdot)\|_{2,P_0} + \|\mu_0^A(0,\cdot) - \hat{\mu}_n^A(0,\cdot)\|_{2,P_0} \Big).
\end{aligned}$$

By Lemma B.8, the display continues as

$$\lesssim \|\mu_n^Z - \mu_0^Z\|_{2,P_0} (\|\hat{\mu}_n^Y - \mu_0^Y\|_{2,P_0} + \|\hat{\mu}_n^A - \mu_0^A\|_{2,P_0}),$$

which is $o_p(n^{-1/2})$ by Condition B6.

Note that all above bounds hold uniformly over $t : \mathcal{V} \to \mathbb{R}$. Combining the above results shows that $\sup_{t:\mathcal{V}\to[0,1]} \left| R^T(\hat{P}_n, \mu_n^Y, \mu_n^A, t, P_0) \right| = o_p(n^{-1/2})$, as desired. □

Proof of Theorem 3.2. By the expansion of $\Psi_{t_P}^T(P)$ given in Section B.8.7,

$$\begin{aligned}
&\Psi_{t_n}^T(\hat{P}_n) - \Psi_{t_0}^T(P_0) \\
&\quad = -P_0 D^T(\hat{P}_n, \mu_n^Y, \mu_n^A, t_n, \tau_0^T) + R^T(\hat{P}_n, \mu_n^Y, \mu_n^A, t_n, P_0) + P_0\{(t_n - t_0)(\Delta_{b,0} - \tau_0^T)\} \\
&\quad = (P_n - P_0) D^T(\mu_0^Y, \mu_0^A, t_0, \tau_0^T, P_0) - P_n D^T(\mu_n^Y, \mu_n^A, t_n, \tau_0^T, \hat{P}_n) \\
&\qquad + (P_n - P_0) \left[D^T(\mu_n^Y, \mu_n^A, t_n, \tau_0^T, \hat{P}_n) - D^T(\mu_0^Y, \mu_0^A, t_0, \tau_0^T, P_0) \right] + R^T(\hat{P}_n, \mu_n^Y, \mu_n^A, t_n, P_0) \\
&\qquad + P_0\{(t_n - t_0)(\Delta_{b,0} - \tau_0^T)\}.
\end{aligned}$$

Similarly, by the expansion given in Section B.8.6,

$$\begin{aligned}
&\Psi^T_{t_r}(\hat{P}_n) - \Psi^T_{t_r}(P_0) \\
&\quad = (P_n - P_0)D^T(P_0, \mu^Y_0, \mu^A_0, t_r, 0) - P_n D^T(\hat{P}_n, \mu^Y_n, \mu^A_n, t_r, 0) \\
&\qquad + (P_n - P_0)\left[D^T(\hat{P}_n, \mu^Y_n, \mu^A_n, t_r, 0) - D^T(P_0, \mu^Y_0, \mu^A_0, t_r, 0)\right] + R^T(\hat{P}_n, \mu^Y_n, \mu^A_n, t_r, P_0).
\end{aligned}$$

By Lemma B.14, $R^T(\hat{P}_n, \mu^Y_n, \mu^A_n, t, P_0) = \mathrm{o}_p(n^{-1/2})$ uniformly over $t : \mathcal{V} \to \mathbb{R}$, and therefore $R^T(\hat{P}_n, \mu^Y_n, \mu^A_n, t_n, P_0)$ and $R^T(\hat{P}_n, \mu^Y_n, \mu^A_n, t_r, P_0)$ are both $\mathrm{o}_p(n^{-1/2})$. By Condition B5, $P_0\{(t_n - t_0)(\Delta_{b,0} - \tau^T_0)\} = \mathrm{o}_p(n^{-1/2})$. By Conditions B8 and B9, $(P_n - P_0)[D^T_{r,n} - D^T_r(P_0)] = \mathrm{o}_p(n^{-1/2})$. Therefore, the remainders and the empirical process terms in the displays above are $\mathrm{o}_p(n^{-1/2})$ and

$$\begin{aligned}
\psi^T_n - \Psi^T_r(P_0) &= (P_n - P_0)D^T_r(P_0) - P_n D^T_{r,n} + \mathrm{o}_p(n^{-1/2}) \\
&= (P_n - P_0)D^T_r(P_0) + \tau^T_0(P_n t_n - \kappa) + \mathrm{o}_p(n^{-1/2}),
\end{aligned}$$

where the last step follows from the TMLE construction of $\hat{P}_n$ (Step 4a and 4b of our estimator), which yields that

$$\frac{1}{n}\sum_{i=1}^{n} \frac{t_n(V_i) - t_r(V_i)}{[Z_i + \mu^Z_n(W_i) - 1]\Delta^A_n(W_i)} \left\{(Y_i - \hat{\mu}^Y_n(Z_i, W_i)) + \Delta_n(W_i)(A_i - \hat{\mu}^A_n(Z_i, W_i))\right\} = 0.$$

When $\tau^T_0 = 0$, $\tau^T_0(P_n t_n - \kappa) = 0$; otherwise, by Lemma B.6, with probability tending to one, $\tau^T_n > 0$ and, hence $P_n t_n = \kappa$. Therefore, $\psi^T_n - \Psi^T_r(P_0) = (P_n - P_0)D^T_r(P_0) + \mathrm{o}_p(n^{-1/2})$.

Conclusion: asymptotic normality results. By the central limit theorem and Slutsky's theorem, the asymptotic linearity of ψ^T_n imply that $\sqrt{n}[\psi^T_n - \Psi^T_r(P_0)] \xrightarrow{d} \mathrm{N}(0, P_0 D^T_r(P_0)^2)$. □

B.10 Proof of Theorem B.1

Recall the definitions $\eta^E_n := \eta^E_n(k_n)$, $\underline{\eta}^E_n := \underline{\eta}^E_n(k_n)$, $\tau^E_n := \tau^E_n(k_n)$ and $\underline{\tau}^E_n := \underline{\tau}^E_n(k_n)$.

Lemma B.15 (Convergence rate of τ^E_n ($\underline{\tau}^E_n$ resp.) if $\eta^E_0 > -\infty$ ($\underline{\eta}^E_0 > -\infty$ resp.))**.** *Assume that the conditions for Theorem 3.4 hold. Suppose that $\eta^E_0 > -\infty$ ($\underline{\eta}^E_0 > -\infty$ resp.), that the Lebesgue density of the distribution of $\xi_0(V)$ ($\underline{\xi}_0(V)$ resp.) under $V \sim P_0$ is well-defined,*

nonzero and finite in a neighborhood of η_0^E ($\underline{\eta}_0^E$ *resp.)and that* $P_0 I(\xi_n = \eta_n^E) = \mathrm{O}_p(n^{-1/2})$ ($P_0 I(\underline{\xi}_n = \underline{\eta}_n^E) = \mathrm{O}_p(n^{-1/2})$ *resp.). Under these conditions, the following implications hold with probability tending to one:*

- *If* $\|\xi_n - \xi_0\|_{q,P_0} = \mathrm{o}_p(1)$ ($\|\underline{\xi}_n - \xi_0\|_{q,P_0} = \mathrm{o}_p(1)$ *resp.) for some* $0 < q < \infty$*, then* $|\tau_n^E - \tau_0^E| \lesssim \|\xi_n - \xi_0\|_{q,P_0}^{q/q+1} + \mathrm{O}_p(n^{-1/2})$ ($|\underline{\tau}_n^E - \tau_0^E| \lesssim \|\underline{\xi}_n - \xi_0\|_{q,P_0}^{q/q+1} + \mathrm{O}_p(n^{-1/2})$ *resp.).*

- *If* $\|\xi_n - \xi_0\|_{\infty,P_0} = \mathrm{o}_p(1)$ ($\|\underline{\xi}_n - \xi_0\|_{\infty,P_0} = \mathrm{o}_p(1)$ *resp.), then* $|\tau_n^E - \tau_0^E| \lesssim \|\xi_n - \xi_0\|_{\infty,P_0} + \mathrm{O}_p(n^{-1/2})$ ($|\underline{\tau}_n^E - \tau_0^E| \lesssim \|\underline{\xi}_n - \xi_0\|_{\infty,P_0} + \mathrm{O}_p(n^{-1/2})$ *resp.).*

The condition that $P_0 I(\xi_n = \eta_n^E) = \mathrm{O}_p(n^{-1/2})$ ($P_0 I(\underline{\xi}_n = \underline{\eta}_n^E) = \mathrm{O}_p(n^{-1/2})$ resp.) is reasonable if $\xi_n(V)$ ($\underline{\xi}_0(V)$ resp.) has a continuous distribution when $V \sim P_0$, in which case $P_0 I(\xi_n = \eta_n^E) = 0$ ($P_0 I(\underline{\xi}_n = \underline{\eta}_n^E) = 0$ resp.).

Proof of Lemma B.15. The proof is similar to that of Lemma B.7. We study the three cases where $\eta_0^E > 0$, $\eta_0^E < 0$ and $\eta_0^E = 0$ ($\underline{\eta}_0^E > 0$, $\underline{\eta}_0^E < 0$ and $\underline{\eta}_0^E = 0$ resp.) separately.

First consider the case where $\eta_0^E > 0$ ($\underline{\eta}_0^E > 0$ resp.). By Lemma B.11, with probability tending to one, $\eta_n^E = \eta_n^E(k_n)$ ($\underline{\eta}_n^E = \underline{\eta}_n^E(\underline{k}_n)$ resp.), where k_n ($\underline{k}_n$ resp.) is a solution to (3.5) ((3.6) resp.), and in the first setting,

$$P_0\{[I(\xi_n > \eta_n^E) - I(\xi_0 > \eta_0^E)]\bar{\mu}_0^A\} = P_0\{d_{n,k_n}\bar{\mu}_0^A\} - \kappa + \mathrm{O}_p(n^{-1/2}) = \mathrm{O}_p(n^{-1/2}),$$

while in the second setting,

$$P_0\{[I(\underline{\xi}_n > \underline{\eta}_n^E) - I(\underline{\xi}_0 > \underline{\eta}_0^E)]\Delta_0^A\} = P_0\{\underline{d}_{n,\underline{k}_n}\Delta_0^A\} - (\kappa - \varphi_0) + \mathrm{O}_p(n^{-1/2}) = \mathrm{O}_p(n^{-1/2}).$$

We argue conditionally on the event that k_n ($\underline{k}_n$ resp.) is a solution to (3.5) ((3.5) resp.). Adding $\Gamma_0(\eta_n^E) - P_0\{I(\xi_n > \eta_n^E)\bar{\mu}_0^A\}$ ($\underline{\Gamma}_0(\underline{\eta}_n^E) - P_0\{I(\underline{\xi}_n > \underline{\eta}_n^E)\Delta_0^A\}$ resp.) to both sides shows that $\Gamma_0(\eta_n^E) - \Gamma_0(\eta_0^E) = -P_0\{[I(\xi_n > \eta_n^E) - I(\xi_0 > \eta_n^E)]\bar{\mu}_0^A\} + \mathrm{O}_p(n^{-1/2})$ ($\underline{\Gamma}_0(\underline{\eta}_n^E) - \underline{\Gamma}_0(\underline{\eta}_0^E) = -P_0\{[I(\underline{\xi}_n > \underline{\eta}_n^E) - I(\underline{\xi}_0 > \underline{\eta}_n^E)]\Delta_0^A\} + \mathrm{O}_p(n^{-1/2})$ resp.). By a Taylor expansion of Γ_0 ($\underline{\Gamma}_0$) under Conditions C2, C3 and A2 (A6b resp.), the left-hand side is equal to $-C(\eta_n^E - \eta_0^E) + \mathrm{o}_p(\eta_n^E - \eta_0^E)$ ($-C(\underline{\eta}_n^E - \underline{\eta}_0^E) + \mathrm{o}_p(\underline{\eta}_n^E - \underline{\eta}_0^E)$ resp.) for some $C > 0$. In the first setting, this yields that

$$[C + o_P(1)][\eta_n^E - \eta_0^E] = P_0\{[I(\xi_n > \eta_n^E) - I(\xi_0 > \eta_n^E)]\bar{\mu}_0^A\} + \mathrm{O}_p(n^{-1/2}),$$

which immediately implies that

$$\eta_n^E - \eta_0^E = \mathrm{O}_p\Big(P_0\{[I(\xi_n > \eta_n^E) - I(\xi_0 > \eta_n^E)]\bar{\mu}_0^A\}\Big) + \mathrm{O}_p(n^{-1/2}), \tag{B.40}$$

where we recall that the expression $P_0\{[I(\xi_n > \eta_n^E) - I(\xi_0 > \eta_n^E)]\bar{\mu}_0^A\}$ is random through the random draws $O_1, \ldots, O_n \sim P_0$ used to define ξ_n and η_n^E. In the second setting, similarly, we can show that

$$\underline{\eta}_n^E - \underline{\eta}_0^E = \mathrm{O}_p\Big(P_0\{[I(\underline{\xi}_n > \underline{\eta}_n^E) - I(\underline{\xi}_0 > \underline{\eta}_n^E)]\Delta_0^A\}\Big) + \mathrm{O}_p(n^{-1/2}). \tag{B.41}$$

By Lemma B.5, for any $\epsilon > 0$ it holds that

$$|P_0\{[I(\xi_n > \eta_n^E) - I(\xi_0 > \eta_n^E)]\bar{\mu}_0^A\}| \leq P_0 I(|\xi_n - \xi_0| > \epsilon) + P_0 I(|\xi_0 - \eta_n^E| \leq \epsilon)$$

in the first setting, and

$$|P_0\{[I(\underline{\xi}_n > \underline{\eta}_n^E) - I(\underline{\xi}_0 > \underline{\eta}_n^E)]\Delta_0^A\}| \leq P_0 I(|\underline{\xi}_n - \underline{\xi}_0| > \epsilon) + P_0 I(|\underline{\xi}_0 - \underline{\eta}_n^E| \leq \epsilon)$$

in the second setting. Fix a positive sequence $\{\epsilon_n\}_{n=1}^{\infty}$, where each ϵ_n may be random through observations $O_1, \ldots, O_n$, such that $\epsilon_n \xrightarrow{p} 0$ as $n \to \infty$. By a Taylor expansion of S_0^E ($\underline{S}_0^E$ resp.) around η_0^E ($\underline{\eta}_0^E$ resp.), which is valid under Condition C2 provided ϵ_n is sufficiently small, it follows that

$$|P_0\{[I(\xi_n > \eta_n^E) - I(\xi_0 > \eta_n^E)]\bar{\mu}_0^A\}| \leq P_0 I(|\xi_n - \xi_0| > \epsilon_n) - 2(S_0^E)'(\eta_0^E)\epsilon_n + \mathrm{o}_p(\epsilon_n)$$

in the first setting, and

$$|P_0\{[I(\underline{\xi}_n > \underline{\eta}_n^E) - I(\underline{\xi}_0 > \underline{\eta}_n^E)]\Delta_0^A\}| \leq P_0 I(|\underline{\xi}_n - \underline{\xi}_0| > \epsilon_n) - 2(\underline{S}_0^E)'(\underline{\eta}_0^E)\epsilon_n + \mathrm{o}_p(\epsilon_n).$$

in the second setting. Here we recall that $(S_0^E)'(\eta_0^E)$ ($(\underline{S}_0^E)'(\underline{\eta}_0^E)$ resp.) is finite by Condition C2. Returning to (B.40),

$$\eta_n^E - \eta_0^E = \mathrm{O}_p\Big(P_0 I(|\xi_n - \xi_0| > \epsilon_n)\Big) - [2(S_0^E)'(\eta_0^E) + \mathrm{o}_p(1)]\epsilon_n + \mathrm{O}_p(n^{-1/2}),$$

and for (B.41),

$$\underline{\eta}_n^E - \underline{\eta}_0^E = \mathrm{O}_p\Big(P_0 I(|\underline{\xi}_n - \underline{\xi}_0| > \epsilon_n)\Big) - [2(\underline{S}_0^E)'(\underline{\eta}_0^E) + \mathrm{o}_p(1)]\epsilon_n + \mathrm{O}_p(n^{-1/2}).$$

Now consider the first setting. If $\|\xi_n - \xi_0\|_{q,P_0} = \mathrm{o}_p(1)$ for some $0 < q < \infty$, by Markov's inequality, $P_0 I(|\xi_n - \xi_0| > \epsilon_n) \leq \|\xi_n - \xi_0\|_{q,P_0}^q / \epsilon_n^q$. In this case, taking $\epsilon_n = \|\xi_n - \xi_0\|_{q,P_0}^{q/(q+1)}$ yields that $|\eta_n^E - \eta_0^E| \lesssim \|\xi_n - \xi_0\|_{q,P_0}^{q/(q+1)} + \mathrm{O}_p(n^{-1/2})$ with probability tending to one. If $\|\xi_n - \xi_0\|_{\infty,P_0} = \mathrm{o}_p(1)$, then taking $\epsilon_n = \|\xi_n - \xi_0\|_{\infty,P_0}$ yields that $P_0 I(|\xi_n - \xi_0| > \epsilon_n) = 0$, and hence that $|\eta_n^E - \eta_0^E| \lesssim \|\xi_n - \xi_0\|_{\infty,P_0}^2 + \mathrm{O}_p(n^{-1/2})$ with probability tending to one. The desired result follows by noting that $\tau_0^E = \eta_0^E$ and in both cases, $\tau_n^E = \eta_n^E(k_n)$ with probability tending to one. For the second setting, note the identical structure in the above equation for the two settings, and hence almost identical argument leads to the desired result.

We now study the case where $\eta^E < 0$ ($\underline{\eta}^E < 0$ resp.). By Lemma B.10, with probability tending to one, $\eta_n^E < 0$ ($\underline{\eta}_n^E < 0$ resp.) and hence $\tau_n^E = 0 = \tau_0^E$ ($\underline{\tau}_n^E = 0 = \underline{\tau}_0^E$ resp.), as desired.

We finally study the case where $\eta_0^E = 0$ ($\underline{\eta}_0^E = 0$ resp.). We argue conditional on the event that a solution k_n' ($\underline{k}_n'$ resp.) to (3.5) ((3.6) resp.) exists, which happens with probability tending to one by Lemma B.11. Recall that for convenience we let $k_n = k_n'$ ($\underline{k}_n = \underline{k}_n'$ resp.) when $\eta_n^E(\kappa) > 0$ ($\underline{\eta}_n^E(\kappa) > 0$ resp.). Then, exactly one of the following two events happen: (i) $\eta_n^E(\kappa) \leq 0$ ($\underline{\eta}_n^E(\kappa) \leq 0$ resp.) or $\eta_n^E(k_n') \leq 0$ ($\underline{\eta}_n^E(\underline{k}_n') \leq 0$ resp.), in which case $\tau_n^E = 0 = \tau_0^E$ ($\underline{\tau}_n^E = 0 = \underline{\tau}_0^E$ resp.); (2) $\eta_n^E(\kappa) > 0$ ($\underline{\eta}_n^E(\kappa) > 0$ resp.) and $\eta_n^E(k_n') > 0$ ($\underline{\eta}_n^E(\underline{k}_n') > 0$ resp.), in which case a similar argument as the above proof for the case where $\eta_0^E > 0$ ($\underline{\eta}_0^E > 0$ resp.) shows that the distance between $\tau_n^E = \eta_n^E(k_n')$ ($\underline{\tau}_n^E = \underline{\eta}_n^E(\underline{k}_n')$ resp.) and τ_0^E ($\underline{\tau}_0^E$ resp.) has the desired bound. The desired result holds conditional on either event, so it holds unconditional on either event. □

We conclude by proving Theorem B.1.

Proof of Theorem B.1. This proof is motivated by the proofs in Luedtke and van der Laan (2016b) and Luedtke and van der Laan (2016a), and mirrors arguments given earlier in

Audibert and Tsybakov (2007). Throughout the proof, we let $\{\epsilon_n\}_{n=1}^{\infty}$ be a positive sequence, where each ϵ_n is random through the observations $O_1, \ldots, O_n$, such that $\epsilon_n \xrightarrow{p} 0$ as $n \to \infty$.

We first present the proof for the sufficient conditions for Condition C8. Observe that

$$\begin{aligned}
|P_0\{(e_n - e_0)(\Delta_{b,0}^Y - \tau_0^E \bar{\mu}_0^A)\}| &\leq P_0|\{I(\xi_n > \tau_n^E) - I(\xi_0 > \tau_0^E)\}(\xi_0 - \tau_0^E)\bar{\mu}_0^A| \\
&\leq P_0|\{I(\xi_n > \tau_n^E) - I(\xi_0 > \tau_0^E)\}(\xi_0 - \tau_0^E)| \\
&\leq P_0|\{I(\xi_n > \tau_n^E) - I(\xi_0 > \tau_n^E)\}(\xi_0 - \tau_n^E)| \\
&\quad + P_0|\{I(\xi_0 > \tau_n^E) - I(\xi_0 > \tau_0^E)\}(\xi_0 - \tau_0^E)| \\
&\quad + |\tau_n^E - \tau_0^E| P_0|I(\xi_n > \tau_n^E) - I(\xi_0 > \tau_n^E)|
\end{aligned} \tag{B.42}$$

in the first setting, and similarly,

$$\begin{aligned}
|P_0\{(\underline{e}_n - \underline{e}_0)(\Delta_{b,0}^Y - \underline{\tau}_0^E \Delta_0^A)\}| &\leq P_0|\{I(\underline{\xi}_n > \underline{\tau}_n^E) - I(\underline{\xi}_0 > \underline{\tau}_n^E)\}(\underline{\xi}_0 - \underline{\tau}_n^E)| \\
&\quad + P_0|\{I(\underline{\xi}_0 > \underline{\tau}_n^E) - I(\underline{\xi}_0 > \underline{\tau}_0^E)\}(\underline{\xi}_0 - \underline{\tau}_0^E)| \\
&\quad + |\underline{\tau}_n^E - \underline{\tau}_0^E| P_0|I(\underline{\xi}_n > \underline{\tau}_n^E) - I(\underline{\xi}_0 > \underline{\tau}_n^E)|.
\end{aligned} \tag{B.43}$$

in the second setting. Our proof is based on bounds on the right hand side of (B.42) and (B.43). Since the structures are identical, we only prove the first setting and note that the proof for the second setting is almost identical. We denote the three terms on the right-hand side of (B.42) by terms 1, 2, and 3, and study these terms separately. It is useful to note that $\tau_n^E - \tau_0^E = \mathrm{o}_p(1)$, so the Lebesgue density of the distribution of $\xi_0(V)$, $V \sim P_0$, is finite in a neighborhood of τ_n^E with probability tending to one.

Study of term 1 in (B.42)**:** Observe that

$$\begin{aligned}
&P_0|\{I(\xi_n > \tau_n^E) - I(\xi_0 > \tau_n^E)\}(\xi_0 - \tau_n^E)| \\
&\quad = P_0|\{I(\xi_n > \tau_n^E) - I(\xi_0 > \tau_n^E)\}(\xi_0 - \tau_n^E)| I(0 < |\xi_0 - \tau_n^E|).
\end{aligned}$$

First consider the bound with the $L^q(P_0)$-distance. Because $I(\xi_n(v) > \tau_n^E) \neq I(\xi_0(v) > \tau_n^E)$ if and only if (i) $\xi_n(v) - \tau_n^E$ and $\xi_0(v) - \tau_n^E$ take different signs or (ii) only one of them is zero, this event implies $|\xi_0(v) - \tau_n^E| \leq |\xi_n(v) - \xi_0(v)|$, and so this term is upper bounded by

$$P_0|\{I(\xi_n > \tau_n^E) - I(\xi_0 > \tau_n^E)\}(\xi_0 - \tau_n^E)| I(0 < |\xi_0 - \tau_n^E| \leq \epsilon_n)$$

$$
\begin{aligned}
&+ P_0|\{I(\xi_n > \tau_n^E) - I(\xi_0 > \tau_n^E)\}(\xi_0 - \tau_n^E)|I(|\xi_0 - \tau_n^E| > \epsilon_n) \\
&\leq P_0|\xi_n - \xi_0|I(0 < |\xi_0 - \tau_n^E| \leq \epsilon_n) + P_0|\xi_n - \xi_0|I(|\xi_n - \xi_0| > \epsilon_n) \\
&\leq \|\xi_n - \xi_0\|_{q,P_0} \left\{P_0(0 < |\xi_0(V) - \tau_n^E| \leq \epsilon_n)\right\}^{(q-1)/q} + \frac{P_0|\xi_n - \xi_0|^q}{\epsilon_n^{q-1}} \\
&\lesssim \|\xi_n - \xi_0\|_{q,P_0} \cdot \epsilon_n^{(q-1)/q} + \frac{\|\xi_n - \xi_0\|_{q,P_0}^q}{\epsilon_n^{q-1}},
\end{aligned}
$$

where second to last relation holds by Hölder's inequality and Markov's inequality, and the last relation holds with probability tending to one by the assumption that the distribution of $\xi_0(V)$, $V \sim P_0$, has a continuous finite Lebesgue density in a neighborhood of τ_0^E and Lemma B.11. Taking $\epsilon_n = \|\xi_n - \xi_0\|_{q,P_0}^{q/(q+1)}$ yields that $|P_0\{I(\xi_n > \tau_n^E) - I(\xi_0 > \tau_n^E)\}(\xi_0 - \tau_n^E)| \lesssim \|\xi_n - \xi_0\|_{q,P_0}^{2q/(q+1)}$.

Next consider the bound with the $L^\infty(P_0)$-distance. We have that

$$
\begin{aligned}
P_0|\{I(\xi_n > \tau_n^E) - I(\xi_0 > \tau_n^E)\}(\xi_0 - \tau_n^E)| &\leq P_0 I(|\xi_0 - \tau_n^E| \leq |\xi_n - \xi_0|)|\xi_0 - \tau_n^E| \\
&= P_0 I(0 < |\xi_0 - \tau_n^E| \leq |\xi_n - \xi_0|)|\xi_0 - \tau_n^E| \\
&\leq P_0 I(0 < |\xi_0 - \tau_n^E| \leq \|\xi_n - \xi_0\|_{\infty,P_0})|\xi_0 - \tau_n^E| \\
&\leq \|\xi_n - \xi_0\|_{\infty,P_0} P_0(0 < |\xi_0(V) - \tau_n^E| \leq \|\xi_n - \xi_0\|_{\infty,P_0}) \\
&\lesssim \|\xi_n - \xi_0\|_{\infty,P_0}^2.
\end{aligned}
$$

Therefore, the first term is upper bounded by both $\|\xi_n - \xi_0\|_{q,P_0}^{2q/(q+1)}$ and $\|\xi_n - \xi_0\|_{\infty,P_0}^2$, up to an absolute constant.

Study of term 2 in (B.42)**:** Because $I(\xi_0(v) > \tau_n^E) \neq I(\xi_0(v) > \tau_0^E)$ if and only if the two indicators take different signs or only one of them is zero, these indicators only take different values if $|\xi_0(v) - \tau_0^E| \leq |\tau_n^E - \tau_0^E|$. Therefore, term 2 bounds as

$$
\begin{aligned}
P_0|\{I(\xi_0 > \tau_n^E) - I(\xi_0 > \tau_0^E)\}(\xi_0 - \tau_0^E)| &\leq P_0 I(|\xi_0 - \tau_0^E| \leq |\tau_n^E - \tau_0^E|)|\xi_0 - \tau_0^E| \\
&\leq |\tau_n^E - \tau_0^E| P_0 I(|\xi_0 - \tau_0^E| \leq |\tau_n^E - \tau_0^E|) \\
&\lesssim |\tau_n^E - \tau_0^E|^2,
\end{aligned}
$$

where the last step holds for with probability tending to one by the assumption that the distribution of $\xi_0(V)$, $V \sim P_0$, has a continuous finite Lebesgue density in a neighborhood of τ_0^E and Lemma B.11. If $\eta_0^E > -\infty$, by Lemma B.15, with probability tending to one,

$$P_0|I(\xi_0 > \tau_n^E) - I(\xi_0 > \tau_0^E)||\xi_0 - \tau_0^E| \lesssim \begin{cases} \|\xi_n - \xi_0\|_{q,P_0}^{2q/(q+1)} + \mathrm{O}_p(n^{-1}), & \text{if } \|\xi_n - \xi_0\|_{q,P_0} = \mathrm{o}_p(1) \\ \|\xi_n - \xi_0\|_{\infty,P_0}^2 + \mathrm{O}_p(n^{-1}), & \text{if } \|\xi_n - \xi_0\|_{\infty,P_0} = \mathrm{o}_p(1) \end{cases}.$$

Otherwise, by Lemma B.10, with probability tending to one, $\tau_n^E = 0 = \tau_0^E$ and the above result still holds.

Study of term 3 in (B.42)**:** By Lemma B.5,

$$P_0|I(\xi_n > \tau_n^E) - I(\xi_0 > \tau_n^E)| \leq P_0 I(|\xi_n - \xi_0| > \epsilon_n) + P_0 I(|\xi_0 - \tau_n^E| \leq \epsilon_n).$$

By a Taylor expansion of S_0^E around τ_0^E, similarly to the proof of Lemma B.7, with probability tending to one,

$$P_0 I(|\xi_0 - \tau_n^E| \leq \epsilon_n) = -2(S_0^E)'(\tau_0^E)\epsilon_n + \mathrm{o}_p(|\tau_n^E - \tau_0^E| + \epsilon_n),$$

where $|(S_0^E)'(\tau_0^E)| < \infty$. If $\|\xi_n - \xi_0\|_{q,P_0} = \mathrm{o}_p(1)$ for some $1 < q < \infty$, then $P_0 I(|\xi_n - \xi_0| > \epsilon_n) \leq \|\xi_n - \xi_0\|_{q,P_0}^q / \epsilon_n^q$. Taking $\epsilon_n = \|\xi_n - \xi_0\|_{q,P_0}^{q/(q+1)}$ yields that $|P_0\{I(\xi_n > \tau_n^E) - I(\xi_0 > \tau_n^E)\}| \lesssim \|\xi_n - \xi_0\|_{q,P_0}^{q/(q+1)}$. If $\|\xi_n - \xi_0\|_{\infty,P_0} = \mathrm{o}_p(1)$, then taking $\epsilon_n = \|\xi_n - \xi_0\|_{\infty,P_0}$ yields that $|P_0\{I(\xi_n > \tau_n^E) - I(\xi_0 > \tau_n^E)\}| \lesssim \|\xi_n - \xi_0\|_{\infty,P_0}$ with probability tending to one. Also note that, by Lemma B.15, if $\eta_0^E > -\infty$, then, with probability tending to one,

$$|\tau_n^E - \tau_0^E| \leq |\eta_n^E - \eta_0^E| \lesssim \begin{cases} \|\xi_n - \xi_0\|_{q,P_0}^{q/(q+1)} + \mathrm{O}_p(n^{-1/2}), & \text{if } \|\xi_n - \xi_0\|_{q,P_0} = \mathrm{o}_p(1), \\ \|\xi_n - \xi_0\|_{\infty,P_0} + \mathrm{O}_p(n^{-1/2}), & \text{if } \|\xi_n - \xi_0\|_{\infty,P_0} = \mathrm{o}_p(1). \end{cases}$$

The same holds when $\eta_0^E = -\infty$ since then $|\tau_n^E - \tau_0^E| = 0$ with probability tending to one.

Therefore, with probability tending to one,

$$|\tau_n^E - \tau_0^E| P_0|I(\xi_n > \tau_n^E) - I(\xi_0 > \tau_n^E)|$$

$$\lesssim \begin{cases} \|\xi_n - \xi_0\|_{q,P_0}^{2q/(q+1)} + \|\xi_n - \xi_0\|_{q,P_0}^{q/(q+1)} \mathrm{O}_p(n^{-1/2}) & \text{if } \|\xi_n - \xi_0\|_{q,P_0} = \mathrm{o}_p(1), \\ \|\xi_n - \xi_0\|_{\infty,P_0}^2 + \|\xi_n - \xi_0\|_{\infty,P_0} \mathrm{O}_p(n^{-1/2}) & \text{if } \|\xi_n - \xi_0\|_{\infty,P_0} = \mathrm{o}_p(1). \end{cases}$$

Conclusion of the bound in (B.42)**:** We finally combine the bounds for all three terms. Note that $a_n \mathrm{O}_p(b_n) \lesssim a_n^2 + \mathrm{O}_p(b_n^2)$ for any sequence of non-negative random variables a_n and sequence of constants b_n. It follows that, with probability tending to one,

$$|P_0\{(e_n - e_0)(\Delta_{b,0}^Y - \tau_0^E \bar{\mu}_0^A)\}| \lesssim \begin{cases} \|\xi_n - \xi_0\|_{q,P_0}^{2q/(q+1)} + \mathrm{O}_p(n^{-1}), & \text{if } \|\xi_n - \xi_0\|_{q,P_0} = \mathrm{o}_p(1), \\ \|\xi_n - \xi_0\|_{\infty,P_0}^2 + \mathrm{O}_p(n^{-1}), & \text{if } \|\xi_n - \xi_0\|_{\infty,P_0} = \mathrm{o}_p(1). \end{cases}$$

We now prove the counterpart for Condition B5. Because the arguments are similar, they are abbreviated.

$$\begin{aligned} |P_0\{(t_n - t_0)(\Delta_{b,0} - \tau_0^T)\}| &\leq P_0|\{I(\Delta_{b,n} > \tau_n^T) - I(\Delta_{b,0} > \tau_0^T)\}(\Delta_{b,0} - \tau_0^T)| \\ &\leq P_0|\{I(\Delta_{b,n} > \tau_n^T) - I(\Delta_{b,0} > \tau_n^T)\}(\Delta_{b,0} - \tau_n^T)| \\ &\quad + P_0|\{I(\Delta_{b,0} > \tau_n^T) - I(\Delta_{b,0} > \tau_0^T)\}(\Delta_{b,0} - \tau_0^T)| \\ &\quad + |\tau_n^T - \tau_0^T| P_0|I(\Delta_{b,n} > \tau_n^T) - I(\Delta_{b,0} > \tau_n^T)|. \end{aligned} \tag{B.44}$$

Study of term 1 in (B.44)**:** Observe that

$$\begin{aligned} &P_0|\{I(\Delta_{b,n} > \tau_n^T) - I(\Delta_{b,0} > \tau_n^T)\}(\Delta_{b,0} - \tau_n^T)| \\ &\quad = P_0|\{I(\Delta_{b,n} > \tau_n^T) - I(\Delta_{b,0} > \tau_n^T)\}(\Delta_{b,0} - \tau_n^T)| I(0 < |\Delta_{b,0} - \tau_n^T|). \end{aligned}$$

For the $L^q(P_0)$-distance, this term is upper bounded by

$$\begin{aligned} &P_0|\{I(\Delta_{b,n} > \tau_n^T) - I(\Delta_{b,0} > \tau_n^T)\}(\Delta_{b,0} - \tau_n^T)| I(0 < |\Delta_{b,0} - \tau_n^T| \leq \epsilon_n) \\ &\quad + P_0|\{I(\Delta_{b,n} > \tau_n^T) - I(\Delta_{b,0} > \tau_n^T)\}(\Delta_{b,0} - \tau_n^T)| I(|\Delta_{b,0} - \tau_n^T| > \epsilon_n) \\ &\leq P_0|\Delta_{b,n} - \Delta_{b,0}| I(0 < |\Delta_{b,0} - \tau_n^T| \leq \epsilon_n) + P_0|\Delta_{b,n} - \Delta_{b,0}| I(|\Delta_{b,n} - \Delta_{b,0}| > \epsilon_n) \\ &\leq \|\Delta_{b,n} - \Delta_{b,0}\|_{q,P_0} \left\{P_0(0 < |\Delta_{b,0}(V) - \tau_n^T| \leq \epsilon_n)\right\}^{(q-1)/q} + \frac{P_0|\Delta_{b,n} - \Delta_{b,0}|^q}{\epsilon_n^{q-1}} \\ &\lesssim \|\Delta_{b,n} - \Delta_{b,0}\|_{q,P_0} \cdot \epsilon_n^{(q-1)/q} + \frac{\|\Delta_{b,n} - \Delta_{b,0}\|_{q,P_0}^q}{\epsilon_n^{q-1}}, \end{aligned}$$

where the last relation holds with probability tending to one by the assumption that $\Delta_{b,0}(V)$, $V \sim P_0$, has a continuous finite Lebesgue density in a neighborhood of τ_0^T and Lemma B.6. Taking $\epsilon_n = \|\Delta_{b,n} - \Delta_{b,0}\|_{q,P_0}^{q/(q+1)}$ shows that $|P_0\{I(\Delta_{b,n} > \tau_n^T) - I(\Delta_{b,0} > \tau_n^T)\}(\Delta_{b,0} - \tau_n^T)| \lesssim \|\Delta_{b,n} - \Delta_{b,0}\|_{q,P_0}^{2q/(q+1)}$ with probability tending to one.

For the $L^\infty(P_0)$-distance,

$$\begin{aligned}
&P_0|\{I(\Delta_{b,n} > \tau_n^T) - I(\Delta_{b,0} > \tau_n^T)\}(\Delta_{b,0} - \tau_n^T)| \\
&\quad \le P_0 I(|\Delta_{b,0} - \tau_n^T| \le |\Delta_{b,n} - \Delta_{b,0}|)|\Delta_{b,0} - \tau_n^T| \\
&\quad \le \|\Delta_{b,n} - \Delta_{b,0}\|_{\infty,P_0} P_0(0 < |\Delta_{b,0} - \tau_n^T| \le \|\Delta_{b,n} - \Delta_{b,0}\|_{\infty,P_0}) \\
&\quad \lesssim \|\Delta_{b,n} - \Delta_{b,0}\|_{\infty,P_0}^2
\end{aligned}$$

with probability tending to one.

Study of term 2 in (B.44)**:** Observe that

$$\begin{aligned}
&P_0|\{I(\Delta_{b,0} > \tau_n^T) - I(\Delta_{b,0} > \tau_0^T)\}(\Delta_{b,0} - \tau_0^T)| \\
&\quad \le P_0 I(|\Delta_{b,0} - \tau_0^T| \le |\tau_n^T - \tau_0^T|)|\Delta_{b,0} - \tau_0^T| \\
&\quad \le |\tau_n^T - \tau_0^T| P_0 I(|\Delta_{b,0} - \tau_0^T| \le |\tau_n^T - \tau_0^T|) \\
&\quad \lesssim |\tau_n^T - \tau_0^T|^2 \begin{cases} \le |\eta_n^T - \eta_0^T|^2, & \text{if } \eta_0^T > -\infty, \\ = 0 \text{ with probability tending to one}, & \text{otherwise.} \end{cases}
\end{aligned}$$

By Lemma B.7, it follows that, with probability tending to one,

$$\begin{aligned}
&P_0|\{I(\Delta_{b,0} > \tau_n^T) - I(\Delta_{b,0} > \tau_0^T)\}(\Delta_{b,0} - \tau_0^T)| \\
&\quad \lesssim \begin{cases} \|\Delta_{b,n} - \Delta_{b,0}\|_{q,P_0}^{2q/(q+1)} + \mathrm{O}_p(n^{-1}), & \text{if } \|\Delta_{b,n} - \Delta_{b,0}\|_{q,P_0} = \mathrm{o}_p(1), \\ \|\Delta_{b,n} - \Delta_{b,0}\|_{\infty,P_0}^2 + \mathrm{O}_p(n^{-1}), & \text{if } \|\Delta_{b,n} - \Delta_{b,0}\|_{\infty,P_0} = \mathrm{o}_p(1). \end{cases}
\end{aligned}$$

Study of term 3 in (B.44)**:** By Lemma B.5,

$$P_0|\{I(\Delta_{b,n} > \tau_n^T) - I(\Delta_{b,0} > \tau_n^T)\}| \le P_0 I(|\Delta_{b,n} - \Delta_{b,0}| > \epsilon_n) + P_0 I(|\Delta_{b,0} - \tau_n^T| \le \epsilon_n).$$

By a Taylor expansion of S_0^T around τ_0^T, the following holds with probability tending to one,

$$P_0 I(|\Delta_{b,0} - \tau_n^T| \leq \epsilon_n) = -2(S_0^T)'(\tau_0^T)\epsilon_n + \mathrm{o}_p(|\tau_n^T - \tau_0^T| + \epsilon_n).$$

If $\|\Delta_{b,n} - \Delta_{b,0}\|_{q,P_0} = \mathrm{o}_p(1)$, then $P_0 I(|\Delta_{b,n} - \Delta_{b,0}| > \epsilon_n) \leq \|\Delta_{b,n} - \Delta_{b,0}\|_{q,P_0}^q / \epsilon_n^q$. Taking $\epsilon_n = \|\Delta_{b,n} - \Delta_{b,0}\|_{q,P_0}^{q/(q+1)}$ shows that $|P_0\{I(\Delta_{b,n} > \tau_n^T) - I(\Delta_{b,0} > \tau_n^T)\}| \lesssim \|\Delta_{b,n} - \Delta_{b,0}\|_{q,P_0}^{q/(q+1)}$ with probability tending to one. If $\|\Delta_{b,n} - \Delta_{b,0}\|_{\infty,P_0} = \mathrm{o}_p(1)$, then taking $\epsilon_n = \|\Delta_{b,n} - \Delta_{b,0}\|_{\infty,P_0}$ shows that $|P_0\{I(\Delta_{b,n} > \tau_n^T) - I(\Delta_{b,0} > \tau_n^T)\}| \lesssim \|\Delta_{b,n} - \Delta_{b,0}\|_{\infty,P_0}$ with probability tending to one. By Lemma B.7, if $\eta_0^T > -\infty$, then

$$|\tau_n^T - \tau_0^T| \leq |\eta_n^T - \eta_0^T| \lesssim \begin{cases} \|\Delta_{b,n} - \Delta_{b,0}\|_{q,P_0}^{q/(q+1)} + \mathrm{O}_p(n^{-1/2}), & \text{if } \|\Delta_{b,n} - \Delta_{b,0}\|_{q,P_0} = \mathrm{o}_p(1), \\ \|\Delta_{b,n} - \Delta_{b,0}\|_{\infty,P_0} + \mathrm{O}_p(n^{-1/2}), & \text{if } \|\Delta_{b,n} - \Delta_{b,0}\|_{\infty,P_0} = \mathrm{o}_p(1). \end{cases}$$

If $\eta_0^T = -\infty$, then $|\tau_n^T - \tau_0^T| = 0$ with probability tending to one and the same result holds.

Therefore, with probability tending to one,

$$\begin{aligned} &|\tau_n^T - \tau_0^T P_0 I(\Delta_{b,n} > \tau_n^T) - I(\Delta_{b,0} > \tau_n^T)| \\ &\lesssim \begin{cases} \|\Delta_{b,n} - \Delta_{b,0}\|_{q,P_0}^{2q/(q+1)} + \|\Delta_{b,n} - \Delta_{b,0}\|_{q,P_0}^{q/(q+1)} \mathrm{O}_p(n^{-1/2}), & \text{if } \|\Delta_{b,n} - \Delta_{b,0}\|_{q,P_0} = \mathrm{o}_p(1), \\ \|\Delta_{b,n} - \Delta_{b,0}\|_{\infty,P_0}^2 + \|\Delta_{b,n} - \Delta_{b,0}\|_{\infty,P_0} \mathrm{O}_p(n^{-1/2}), & \text{if } \|\Delta_{b,n} - \Delta_{b,0}\|_{\infty,P_0} = \mathrm{o}_p(1). \end{cases} \end{aligned}$$

Conclusion of the bound in (B.44)**:** Combining the above results shows that, with probability tending to one,

$$|P_0\{(t_n - t_0)(\Delta_{b,0} - \tau_0^T)\}| \lesssim \begin{cases} \|\Delta_{b,n} - \Delta_{b,0}\|_{q,P_0}^{2q/(q+1)} + \mathrm{O}_p(n^{-1}), & \text{if } \|\Delta_{b,n} - \Delta_{b,0}\|_{q,P_0} = \mathrm{o}_p(1), \\ \|\Delta_{b,n} - \Delta_{b,0}\|_{\infty,P_0}^2 + \mathrm{O}_p(n^{-1}), & \text{if } \|\Delta_{b,n} - \Delta_{b,0}\|_{\infty,P_0} = \mathrm{o}_p(1). \end{cases}$$

□

Appendix C

SUPPORTING INFORMATION FOR CHAPTER 4

C.1 Proofs

C.1.1 Proof of Theorem 4.1 and Corollary 4.1.1

Lemma C.1. *If $\{\Omega_\ell\}_{\ell=1}^{\infty}$ is an increasing sequence of subsets of $\mathcal{M}$ such that $\bigcup_{\ell=1}^{\infty}\Omega_\ell = \mathcal{M}$, then, for any $d \in \mathcal{D}$, $r_{\sup}(d, \tilde{\Gamma}_\ell) \nearrow r_{\sup}(d, \Gamma)$ ($\ell \to \infty$).*

Proof of Lemma C.1. Since $\tilde{\Gamma}_\ell \subseteq \tilde{\Gamma}_{\ell+1} \subseteq \Gamma$, it holds that $r_{\sup}(d, \tilde{\Gamma}_\ell) \leq r_{\sup}(d, \tilde{\Gamma}_{\ell+1}) \leq r_{\sup}(d, \Gamma)$, and so we only need to lower bound $r_{\sup}(d, \tilde{\Gamma}_\ell)$. Fix $\epsilon > 0$. By Corollary 5 of Pinelis (2016), $r_{\sup}(d, \Gamma)$ can be approximated by $r(d, \nu)$ arbitrarily well for priors $\nu \in \Gamma$ with a finite support; that is, there exists $\nu \in \Gamma$ with finite support such that $r(d, \nu) \geq r_{\sup}(d, \Gamma) - \epsilon$. For sufficiently large ℓ, Ω_ℓ contains all support points of ν and hence $r_{\sup}(d, \tilde{\Gamma}_\ell) \geq r(d, \nu) \geq r_{\sup}(d, \Gamma) - \epsilon$. The desired result follows. □

Lemma C.2. *Under Condition 3, $d \mapsto r(d, \pi)$ is Lipschitz continuous with Lipschitz constant L; moreover, $d \mapsto r_{\sup}(d, \Gamma')$ is Lipschitz continuous with Lipschitz constant L for any $\Gamma' \subseteq \Gamma$.*

Proof of Lemma C.2. By Condition 3, $|R(d_1, P) - R(d_2, P)| \leq L\varrho(d_1, d_2)$ for any $d_1, d_2 \in \mathcal{D}$ and any $P \in \mathcal{M}$. Then, for any $\pi \in \Gamma$ and any $d_1, d_2 \in \mathcal{D}$,

$$\begin{aligned} |r(d_1, \pi) - r(d_2, \pi)| &= \left| \int [R(d_1, P) - R(d_2, P)] \pi(\mathrm{d}P) \right| \\ &\leq \int |R(d_1, P) - R(d_2, P)| \, \pi(\mathrm{d}P) \\ &\leq L\varrho(d_1, d_2). \end{aligned}$$

This proves that $d \mapsto r(d, \pi)$ is Lipschitz continuous with a universal Lipschitz constant L. We now prove that $d \mapsto r_{\sup}(d, \Gamma)$ is Lipschitz continuous with Lipschitz constant L. Let

$\epsilon > 0$. For any $d_1 \in \mathcal{D}$, there exists $\pi_1 \in \Gamma'$ such that $r_{\sup}(d_1, \Gamma') \leq r(d_1, \pi_1) + \epsilon$. Then, for any $d_2 \in \mathcal{D}$,

$$r_{\sup}(d_1, \Gamma') - r_{\sup}(d_2, \Gamma') \leq r(d_1, \pi_1) + \epsilon - r(d_2, \pi_1) \leq L\varrho(d_1, d_2) + \epsilon.$$

Since ϵ is arbitrary, we have that $r_{\sup}(d_1, \Gamma') - r_{\sup}(d_2, \Gamma') \leq L\varrho(d_1, d_2)$. Reversing the role of d_1 and d_2, we derive that $r_{\sup}(d_2, \Gamma') - r_{\sup}(d_1, \Gamma') \leq L\varrho(d_1, d_2)$. Therefore, $|r_{\sup}(d_1, \Gamma') - r_{\sup}(d_2, \Gamma')| \leq L\varrho(d_1, d_2)$ for any $d_1, d_2 \in \mathcal{D}$. □

Proof of Theorem 4.1. Let $\epsilon > 0$. There exists $d' \in \mathcal{D}$ such that

$$r_{\sup}(d', \Gamma) \leq \inf_{d \in \mathcal{D}} r_{\sup}(d, \Gamma) + \epsilon.$$

Moreover, there exists $\pi_\ell \in \Gamma_\ell$ such that

$$r_{\sup}(d', \Gamma_\ell) \leq r(d', \pi_\ell) + \epsilon.$$

Using the fact that d^*_ℓ is Γ_ℓ-minimax and the definition of $r_{\sup}$, it holds that

$$\begin{aligned} r_{\sup}(d^*_\ell, \Gamma_\ell) &\leq r_{\sup}(d', \Gamma_\ell) \leq r(d', \pi_\ell) + \epsilon \\ &\leq r_{\sup}(d', \Gamma) + \epsilon \leq \inf_{d \in \mathcal{D}} r_{\sup}(d, \Gamma) + 2\epsilon. \end{aligned}$$

Since this inequality holds for any $\epsilon > 0$, we have that $r_{\sup}(d^*_\ell, \Gamma_\ell) \leq \inf_{d \in \mathcal{D}} r_{\sup}(d, \Gamma)$. An almost identical argument shows that the sequence $\{r_{\sup}(d^*_\ell, \Gamma_\ell)\}_{\ell=1}^{\infty}$ is nondecreasing. Therefore, this sequence converges to some limit $\mathcal{R} \leq \inf_{d \in \mathcal{D}} r_{\sup}(d, \Gamma) \leq r_{\sup}(d^*, \Gamma)$.

We next prove that $r_{\sup}(d^*, \Gamma) \leq \mathcal{R}$. Let $\epsilon > 0$. Without loss of generality, we may assume that $\mathcal{M}_\ell \subseteq \Omega_\ell$ for all $\ell = 1, 2, \ldots$ in Condition 4. (Otherwise, we may instead consider the sequence $\{\Omega_{\tilde{\ell}}\}_{\tilde{\ell}=1}^{\infty}$ where $\Omega_{\tilde{\ell}} = \bigcap_{\ell' : \Omega_{\ell'} \supseteq \mathcal{M}_\ell} \Omega_{\ell'}$. Note that Condition 4 also holds for $\{\Omega_{\tilde{\ell}}\}_{\tilde{\ell}=1}^{\infty}$.) By Lemma C.1, there exists ℓ_0 such that $r_{\sup}(d^*, \tilde{\Gamma}_{\ell_0}) \geq r_{\sup}(d^*, \Gamma) - \epsilon/3$. By Condition 4, there exists i_1 such that $r_{\sup}(d^*, \Gamma_{i_1|\ell_0}) \geq r_{\sup}(d^*, \tilde{\Gamma}_{\ell_0}) - \epsilon/3$. Without loss of generality, suppose that $d^*_\ell \to d^*$ (otherwise, take a convergent subsequence to this accumulation point). This then implies that there exists $i_2 > i_1$ such that $\varrho(d^*_{i_2}, d^*) \leq \epsilon/(3L)$.

By Lemma C.2, $r_{\sup}(d^*_{i_2}, \Gamma_{i_1|\ell_0}) \geq r_{\sup}(d^*, \Gamma_{i_1|\ell_0}) - \epsilon/3$. Moreover, since $\Gamma_{i_1|\ell_0} \subseteq \Gamma_{i_1} \subseteq \Gamma_{i_2}$, it holds that $r_{\sup}(d^*_{i_2}, \Gamma_{i_2}) \geq r_{\sup}(d^*_{i_2}, \Gamma_{i_1|\ell_0})$. Therefore, $r_{\sup}(d^*_{i_2}, \Gamma_{i_2}) \geq r_{\sup}(d^*, \Gamma) - \epsilon$. Since the sequence $\{r_{\sup}(d^*_\ell, \Gamma_\ell)\}_{\ell=1}^\infty$ is nondecreasing, it holds that $r_{\sup}(d^*_\ell, \Gamma_\ell) \geq r_{\sup}(d^*, \Gamma) - \epsilon$ for all $\ell \geq i_2$. Therefore, $\liminf_{\ell\to\infty} r_{\sup}(d^*_\ell, \Gamma_\ell) \geq r_{\sup}(d^*, \Gamma)$, and hence $\mathcal{R} \geq r_{\sup}(d^*, \Gamma)$.

Combining the results from the preceding two paragraphs, $\mathcal{R} = \inf_{d\in\mathcal{D}} r_{\sup}(d, \Gamma) = r_{\sup}(d^*, \Gamma)$. Consequently, d^* is Γ-minimax. Moreover, as $\{r_{\sup}(d^*_\ell, \Gamma_\ell)\}_{\ell=1}^\infty$ increases to $\mathcal{R}$, this sequence also increases to $r_{\sup}(d^*, \Gamma)$. This concludes the proof. □

Proof of Corollary 4.1.1. We first establish the strict convexity of $d \mapsto r(d, \pi)$ for any $\pi \in \Gamma$. We then establish the strict convexity of $d \mapsto r_{\sup}(d, \Gamma)$. We then establish that there is a unique minimizer of $d \mapsto r_{\sup}(d, \Gamma)$ and show that the desired result follows from Theorem 4.1.

Let $d_1, d_2 \in \mathcal{D}$ and $c \in (0, 1)$ be arbitrary, then by the convexity of $\mathcal{D}$ and the strict convexity of $d \mapsto R(d, P)$ for each $P \in \mathcal{M}$,

$$\begin{aligned} r(cd_1 + (1-c)d_2, \pi) &= \int R(cd_1 + (1-c)d_2, P)\pi(\mathrm{d}P) \\ &< \int \{cR(d_1, P) + (1-c)R(d_2, P)\}\pi(\mathrm{d}P) \\ &= cr(d_1, \pi) + (1-c)r(d_2, \pi). \end{aligned}$$

Therefore, $d \mapsto r(d, \pi)$ is strictly convex for any $\pi \in \Gamma$.

Let $d_1, d_2 \in \mathcal{D}$ and $c \in (0, 1)$ be arbitrary. Since $r_{\sup}(d, \Gamma)$ is attainable for any $d \in \mathcal{D}$, there exists $\tilde{\pi} \in \Gamma$ such that

$$\begin{aligned} r_{\sup}(cd_1 + (1-c)d_2, \Gamma) &= r(cd_1 + (1-c)d_2, \tilde{\pi}) \\ &< cr(d_1, \tilde{\pi}) + (1-c)r(d_2, \tilde{\pi}) \\ &\leq cr_{\sup}(d_1, \Gamma) + (1-c)r_{\sup}(d_2, \Gamma). \end{aligned}$$

Thus, $d \mapsto r_{\sup}(d, \Gamma)$ is strictly convex.

As $d \mapsto r_{\sup}(d, \Gamma)$ is continuous by Condition 3 and $\mathcal{D}$ is compact by Condition 2, $d \mapsto r_{\sup}(d, \Gamma)$ achieves at least one minimum on $\mathcal{D}$. As $d \mapsto r_{\sup}(d, \Gamma)$ is strictly convex and $\mathcal{D}$ is convex, this function achieves exactly one minimum on $\mathcal{D}$. By Theorem 4.1, any

accumulation point d^* of $\{d^*_\ell\}_{\ell=1}^\infty$ is a minimizer of $d \mapsto r_{\sup}(d, \Gamma)$, and so the sequence has a limit point, which is also the unique Γ-minimax estimator. □

C.1.2 Proof of Theorem 4.2

We prove Theorem 4.2 by checking that Assumptions 3.1 and 3.6 in Lin et al. (2019) are satisfied and using Theorem E.3 and E.4 in Lin et al. (2019), respectively. Since Assumption 3.1 is satisfied by our construction of $\hat{R}$, we focus on Assumption 3.6 for the rest of this section.

Let $\mathcal{M}_\ell = \{P_1, P_2, \ldots, P_\Lambda\} \subseteq \mathcal{M}$. For any $\pi \in \Gamma_\ell$, let π_λ denote the probability weight of π on P_λ ($\lambda = 1, \ldots, \Lambda$). For the rest of this section, we also use π to denote the vector $(\pi_1, \ldots, \pi_\Lambda)$. We also use $\lesssim$ to denote less than equal to up to a universal positive constant that may depend on ℓ. Then, straightforward calculations imply that $\nabla_\beta r(\beta, \pi) = \sum_{\lambda=1}^\Lambda \pi_\lambda \nabla_\beta R(\beta, P_\lambda)$ and $\nabla_\pi r(\beta, \pi) = (R(\beta, P_1), \ldots, R(\beta, P_\Lambda))^\top$.

For each $\ell = 1, 2, \ldots$, for any $\beta^1, \beta^2 \in \mathcal{H}$ and $\pi^1, \pi^2 \in \Gamma_\ell$, by Conditions 5 and 6,

$$
\begin{aligned}
&\left\| \nabla_\beta r(\beta, \pi)|_{\beta=\beta^1, \pi=\pi^1} - \nabla_\beta r(\beta, \pi)|_{\beta=\beta^2, \pi=\pi^2} \right\| \\
&= \left\| \sum_{\lambda=1}^\Lambda \left\{ \pi^1_\lambda \, \nabla_\beta R(\beta, P_\lambda)|_{\beta=\beta_1} - \pi^2_\lambda \, \nabla_\beta R(\beta, P_\lambda)|_{\beta=\beta_2} \right\} \right\| \\
&\leq \sum_{\lambda=1}^\Lambda \pi^1_\lambda \left\| \nabla_\beta R(\beta, P_\lambda)|_{\beta=\beta_1} - \nabla_\beta R(\beta, P_\lambda)|_{\beta=\beta_2} \right\| + \left\| \sum_{\lambda=1}^\Lambda (\pi^1_\lambda - \pi^2_\lambda) \, \nabla_\beta R(\beta, P_\lambda)|_{\beta=\beta_2} \right\| \\
&\lesssim \|\beta^1 - \beta^2\| + \|\pi^1 - \pi^2\| \\
&\lesssim \|(\beta^1, \pi^1) - (\beta^2, \pi^2)\|,
\end{aligned}
$$

and similarly for $\nabla_\pi r(\beta, \pi)$,

$$
\begin{aligned}
&\left\| \nabla_\pi r(\beta, \pi)|_{\beta=\beta^1, \pi=\pi^1} - \nabla_\pi r(\beta, \pi)|_{\beta=\beta^2, \pi=\pi^2} \right\| \\
&= \left\| \big(R(\beta^1, P_1) - R(\beta^2, P_1), R(\beta^1, P_2) - R(\beta^2, P_2), \ldots, R(\beta^1, P_\Lambda) - R(\beta^2, P_\Lambda) \big)^\top \right\| \\
&\lesssim \|\beta^1 - \beta^2\| \leq \|(\beta^1, \pi^1) - (\beta^2, \pi^2)\|.
\end{aligned}
$$

This implies that for each ℓ, the gradient of $r(\beta, \pi)$ ($\beta \in \mathcal{H}$, $\pi \in \Gamma_\ell$) is Lipschitz continuous.

For each $\ell = 1, 2, \ldots$, for any $\beta^1, \beta^2 \in \mathcal{H}$ and $\pi \in \Gamma_\ell$, Condition 5 implies that

$$\begin{aligned}\left|r(\beta^1, \pi) - r(\beta^2, \pi)\right| &= \left|\sum_{\lambda=1}^{\Lambda} \pi_\lambda \left[R(\beta^1, P_\lambda) - R(\beta^2, P_\lambda)\right]\right| \\ &\leq \sum_{\lambda=1}^{\Lambda} \pi_\lambda \left|R(\beta^1, P_\lambda) - R(\beta^2, P_\lambda)\right| \lesssim \|\beta^1 - \beta^2\|.\end{aligned}$$

Therefore, $\beta \mapsto r(\beta, \pi)$ is Lipschitz continuous with a universal Lipschitz constant independent of $\pi \in \Gamma_\ell$.

Finally, it is straightforward to check that (i) $\pi \mapsto r(\beta, \pi)$ is concave for any $\beta \in \mathcal{H}$, and (ii) Γ_ℓ is parameterized by a convex subset of a simplex in a Euclidean space, which is a convex and bounded set. These results show that Assumption 3.6 in Lin et al. (2019) is satisfied for Algorithm 2 and 3.

C.1.3 Proof of Theorem 4.3

Proof of Theorem 4.3. Let $\pi_{(t),0}$ denote a maximizer of $\pi \mapsto r(\beta_{(t-1)}, \pi)$. It holds that

$$\begin{aligned}0 &\leq r(\beta_{(t-1)}, \pi_{(t),0}) - r(\beta_{(t-1)}, \pi_{(t)}) \\ &\leq \frac{1}{J'} \sum_{j=1}^{J'} \hat{r}(\beta_{(t-1)}, \pi_{(t)}, \xi_j) - \frac{1}{J'} \sum_{j=1}^{J'} \hat{r}(\beta_{(t-1)}, \pi_{(t),0}, \xi_j) \\ &\quad + r(\beta_{(t-1)}, \pi_{(t),0}) - r(\beta_{(t-1)}, \pi_{(t)}) \\ &= \frac{1}{J'} \sum_{j=1}^{J'} \Bigg\{ \left[\hat{r}(\beta_{(t-1)}, \pi_{(t)}, \xi_j) - \hat{r}(\beta_{(t-1)}, \pi_{(t),0}, \xi_j)\right] \\ &\quad - \mathrm{E}\left[\hat{r}(\beta_{(t-1)}, \pi_{(t)}, \xi) - \hat{r}(\beta_{(t-1)}, \pi_{(t),0}, \xi)\right] \Bigg\} \\ &\leq \sup_{\beta \in \mathbb{R}^D, \pi_1, \pi_2 \in \Gamma_\ell} \Bigg| \frac{1}{J'} \sum_{j=1}^{J'} \Bigg\{ \left[\hat{r}(\beta, \pi_1, \xi_j) - \hat{r}(\beta, \pi_2, \xi_j)\right] \\ &\quad - \mathrm{E}\left[\hat{r}(\beta, \pi_1, \xi) - \hat{r}(\beta, \pi_2, \xi)\right] \Bigg\} \Bigg|.\end{aligned}$$

Note that the right hand side does not depend on t. Therefore,

$$\begin{aligned}
0 &\leq \sup_t \left\{ r(\beta_{(t-1)}, \pi_{(t),0}) - \mathrm{E}[r(\beta_{(t-1)}, \pi_{(t)})] \right\} \\
&\leq \mathrm{E}^* \sup_{\beta \in \mathbb{R}^D, \pi_1, \pi_2 \in \Gamma_\ell} \left| \frac{1}{J'} \sum_{j=1}^{J'} \left\{ [\hat{r}(\beta, \pi_1, \xi_j) - \hat{r}(\beta, \pi_2, \xi_j)] \right.\right. \\
&\qquad \left.\left. - \mathrm{E}\left[\hat{r}(\beta, \pi_1, \xi) - \hat{r}(\beta, \pi_2, \xi)\right] \right\} \right|,
\end{aligned}$$

where E^* stands for outer expectation. We may apply Corollary 9.27 in Kosorok (2008) to $\mathcal{F} := \{\xi \mapsto \hat{r}(\beta, \pi, \xi) : \beta \in \mathbb{R}^D, \pi \in \Gamma_\ell\}$ and show that $\mathcal{F} - \mathcal{F} := \{f_1 - f_2 : f_1, f_2 \in \mathcal{F}\} \supseteq \{\xi \mapsto \hat{r}(\beta, \pi_1, \xi) - \hat{r}(\beta, \pi_2, \xi) : \beta \in \mathbb{R}^D, \pi_1, \pi_2 \in \Gamma_\ell\}$ is a Ξ-Glivenko-Cantelli class. Therefore,

$$\begin{aligned}
&\left\{ \sup_{\beta \in \mathbb{R}^D, \pi_1, \pi_2 \in \Gamma_\ell} \left| \frac{1}{J'} \sum_{j=1}^{J'} \left\{ [\hat{r}(\beta, \pi_1, \xi_j) - \hat{r}(\beta, \pi_2, \xi_j)] \right.\right.\right. \\
&\qquad \left.\left.\left. - \mathrm{E}\left[\hat{r}(\beta, \pi_1, \xi) - \hat{r}(\beta, \pi_2, \xi)\right] \right\} \right| \right\}^* \\
&\leq \left\{ \sup_{f \in \mathcal{F} - \mathcal{F}} \left| \frac{1}{J'} \sum_{j=1}^{J'} \{f(\xi_j) - \mathrm{E}[f(\xi)]\} \right| \right\}^* \xrightarrow{a.s.} 0,
\end{aligned}$$

as $J' \to \infty$. Here, X^* stands for the minimal measurable majorant with respect to Ξ of a (possibly non-measurable) mapping X (van der Vaart and Wellner, 2000).

By Problem 1 of Section 2.4 in van der Vaart and Wellner (2000), there exists a random variable F such that $F \geq \sup_{f \in \mathcal{F} - \mathcal{F}} |f(\xi) - \mathrm{E}[f(\xi')]|$ Ξ-almost surely and $\mathrm{E}[F] < \infty$. Then,

$$\sup_{f \in \mathcal{F} - \mathcal{F}} \left| \frac{1}{J'} \sum_{j=1}^{J'} \{f(\xi_j) - \mathrm{E}[f(\xi_j)]\} \right| \leq F$$

Ξ-almost surely. By dominated convergence theorem,

$$\begin{aligned}
&\mathrm{E}^* \sup_{\beta \in \mathbb{R}^D, \pi_1, \pi_2 \in \Gamma_\ell} \left| \frac{1}{J'} \sum_{j=1}^{J'} \left\{ [\hat{r}(\beta, \pi_1, \xi_j) - \hat{r}(\beta, \pi_2, \xi_j)] \right.\right. \\
&\qquad \left.\left. - \mathrm{E}\left[\hat{r}(\beta, \pi_1, \xi_j) - \hat{r}(\beta, \pi_2, \xi_j)\right] \right\} \right| \to 0
\end{aligned}$$

as $J' \to \infty$, and so does $\sup_t \left\{ r(\beta_{(t-1)}, \pi_{(t),0}) - \mathrm{E}[r(\beta_{(t-1)}, \pi_{(t)})] \right\}$. Thus, for any $\zeta > 0$, there exists a sufficiently large J' such that $\mathrm{E}[r(\beta_{(t-1)}, \pi_{(t)})] \geq r(\beta_{(t-1)}, \pi_{(t),0}) - \zeta$ for all t. $\square$

C.1.4 Proof of Theorem 4.4

Our proof of Theorem 4.4 builds on that of Robinson (1951). Major modifications are needed to allow for more general definitions that can accommodate for potentially infinite spaces of pure strategies and a more careful control on a bound on $r(\overline{d}(\varpi_{(t-1)}), \pi^\dagger_{(t)}) - r(d^\dagger_{(t)}, \pi_{(t-1)})$ towards the end of the proof.

We first introduce the notion of cumulative Bayes risk functions. Under Algorithm 5, we let $U_0 : \mathcal{D} \to \mathbb{R}$ and $V_0 : \Gamma_\ell \to \mathbb{R}$ be any two continuous functions such that

$$\min_{d \in \mathcal{D}} U_0(d) = \max_{\pi \in \Gamma_\ell} V_0(\pi) \tag{C.1}$$

and recursively define

$$U_{t+1}(d) := U_t(d) + r(d, \pi^\dagger_{(t)}), \quad V_{t+1}(\pi) := V_t(\pi) + r(d^\dagger_{(t)}, \pi) \tag{C.2}$$

for $d \in \mathcal{D}$ and $\pi \in \Gamma_\ell$. Here, we let $\pi^\dagger_{(t)} \in \operatorname{argmax}_{\pi \in \Gamma_\ell} V_{t-1}(\pi)$ and $d^\dagger_{(t)} \in \operatorname{argmin}_{d \in \mathcal{D}} U_{t-1}(d)$. Note that the choices of $\pi^\dagger_{(t)}$ and $d_{(t)}$ in Algorithm 5 corresponds to setting $U_0 \equiv 0$ and $V_0 \equiv 0$, in which case $U_t(d) = t \cdot r(d, \pi_{(t)})$ and $V_t(\pi) = t \cdot r(\overline{d}(\varpi_{(t)}), \pi)$. In general,

$$U_t(d) = U_0(d) + t \cdot r(d, \pi_{(t)}), \quad V_t(\pi) = V_0(\pi) + t \cdot r(\overline{d}(\varpi_{(t)}), \pi) \tag{C.3}$$

for some $\pi_{(t)} \in \Gamma$ and $\overline{d}(\varpi_{(t)}) \in \overline{\mathcal{D}}$. Later in this section, we will also make use of U_t and V_t with other initializations U_0 and V_0.

To make notations concise, we define $\min_{d \in \mathcal{D}'} U_t := \min_{d \in \mathcal{D}'} U_t(d)$ for any $\mathcal{D}' \subseteq \mathcal{D}$, and define $\max_{\mathcal{D}'} U_t$, $\min_{\Pi'} V_t$ and $\max_{\Pi'} V_t$ ($\Pi' \subseteq \Gamma_\ell$) similarly. We also drop the subscript denoting the set when the set is the whole space we consider, i.e., $\mathcal{D}$ or Γ_ℓ. Note that for any $t_1, t_2 = 1, 2, \ldots$, under the setting of Algorithm 5 and (4.1), it holds that

$$\begin{aligned}
&\min U_{t_1}/t_1 = \min_{\overline{d} \in \overline{\mathcal{D}}} r(\overline{d}, \pi_{(t_1)}) \\
\le &\max_{\pi \in \Gamma_\ell} \min_{\overline{d} \in \overline{\mathcal{D}}} r(\overline{d}, \pi) = r(\overline{d}(\varpi^*_\ell), \pi^*_\ell) = \min_{\overline{d} \in \overline{\mathcal{D}}} \max_{\pi \in \Gamma_\ell} r(\overline{d}, \pi) \\
\le &\max_{\pi \in \Gamma_\ell} r(\overline{d}(\varpi_{(t_2)}), \pi) = \max V_{t_2}/t_2
\end{aligned}$$

Therefore, to prove the first result in Theorem 4.4, it suffices to show that $\limsup_{t\to\infty}(\max V_t - \min U_t)/t \leq 0$.

We next introduce additional definitions related to iterations. We say that $\pi \in \Gamma_\ell$ is eligible in the interval $[t_1, t_2]$ if there exists $t \in [t_1, t_2]$ such that $V_t(\pi) = \max V_t$; we say that $d \in \mathcal{D}$ is eligible in the interval $[t_1, t_2]$ if there exists $t \in [t_1, t_2]$ such that $U_t(d) = \min U_t$. We also define eligibility for sets. We say that $\Pi' \subseteq \Gamma_\ell$ is eligible in the interval $[t_1, t_2]$ if there exists $\pi \in \Pi'$ that is eligible in that interval; we say that $\mathcal{D}' \subseteq \mathcal{D}$ is eligible in the interval $[t_1, t_2]$ if there exists $d \in \mathcal{D}'$ that is eligible in the interval $[t_1, t_2]$. In addition, for any $\mathcal{D}' \subseteq \mathcal{D}$, we define maximum variation $\mathrm{MV}_t(\mathcal{D}') := \sup_{d\in\mathcal{D}'} U_t(d) - \inf_{d\in\mathcal{D}'} U_t(d)$ and $\mathrm{MV}_t(\Pi')$ similarly for any $\Pi' \subset \Gamma_\ell$. By Condition 3, there exists $B \in (0, \infty)$ such that $R \in [-B, B]$. Note that by Condition 2 and Lemma C.2, given an arbitrary desired approximation accuracy $\epsilon > 0$, $\mathcal{D}$ can be covered by finitely many compact subsets with the maximum variation of each subset bounded by ϵt for all t; by Condition 3, since Γ_ℓ is parameterized by a compact subset of a simplex in a Euclidean space, Γ_ℓ can also be covered by finitely many compact subsets with the maximum variation of each subset bounded by ϵt for all t. These covers can be viewed as discrete finite approximations to $\mathcal{D}$ and Γ_ℓ, respectively.

All of the above definitions are associated with the space of estimators $\mathcal{D}$ and the set of priors Γ_ℓ. We call $\{(U_t, V_t)\}_t$ a pair of cumulative Bayes risk functions constructed from the pair $(\mathcal{D}, \Gamma_\ell)$ of the space of estimators and the set of priors, and will consider pairs of cumulative Bayes risk functions constructed from other pairs $(\mathcal{D}', \Pi')$ of the space of estimators and the set of priors in the subsequent proof. We can define the above quantities similarly for such cases.

The following lemma gives an upper bound on the maximum variation of U_{s+t} and V_{s+t} over the corresponding entire space from which they are constructed after t iterations from s when essentially all parts of these spaces are eligible in $[s, s + t]$.

Lemma C.3. *Suppose that $\{(U_t, V_t)\}_t$ is a pair of cumulative Bayes risk functions con-*

structed from $(\mathcal{D}', \Pi')$. *Suppose that* $\mathcal{D}' = \bigcup_{i=1}^{I} \mathcal{D}_i$ *and* $\Pi' = \bigcup_{j=1}^{J} \Pi_j$ *where*

$$\sup_{i,t} \mathrm{MV}_t(\mathcal{D}_i)/t \leq A, \quad \sup_{j,t} \mathrm{MV}_t(\Pi_j)/t \leq A$$

for $A < \infty$. *If all* $\mathcal{D}_i$ *and* Π_j *are eligible in* $[s, s+t]$, *then* $\max_{\mathcal{D}'} U_{s+t} - \min_{\mathcal{D}'} U_{s+t} \leq (2B+A)t$ *and* $\max_{\Pi'} V_{s+t} - \min_{\Pi'} V_{s+t} \leq (2B + A)t$.

Proof of Lemma C.3. Without loss of generality, assume that $\tilde{d} \in (\mathrm{argmax}_{d \in \mathcal{D}'} U_{s+t}) \bigcap \mathcal{D}_1$. Since $\mathcal{D}_1$ is eligible in $[s, t]$, there exists $\tilde{t} \in [s, s + t]$ such that $(\mathrm{argmin}_{d \in \mathcal{D}'} U_{\tilde{t}}) \bigcap \mathcal{D}_1 \neq \emptyset$. By the recursive definition of the sequence $\{U_t\}_t$ in (C.2), the bound on the risk, and the assumption that $\sup_{i,t} \mathrm{MV}_t(\mathcal{D}_i)/t \leq A$, we have that $\max_{\mathcal{D}'} U_{s+t} = U_{s+t}(\tilde{d}) \leq U_{\tilde{t}}(\tilde{d}) + B(s + t - \tilde{t}) \leq \min_{\mathcal{D}'} U_{\tilde{t}} + At + B(s + t - \tilde{t}) \leq \min_{\mathcal{D}'} U_{\tilde{t}} + (A + B)t$. Letting $\tilde{d}' \in \mathrm{argmin}_{d \in \mathcal{D}'} U_{s+t}$, by the bound on the risk, we can derive that $\min_{\mathcal{D}'} U_{s+t} = U_{s+t}(\tilde{d}') \geq U_{\tilde{t}}(\tilde{d}') - B(s + t - \tilde{t}) \geq \min_{\mathcal{D}'} U_{\tilde{t}} - Bt$. Combine these two inequalities and we have that $\max_{\mathcal{D}'} U_{s+t} - \min_{\mathcal{D}'} U_{s+t} \leq (2B + A)t$. An identical argument applied to the sequence $\{V_t\}_t$ shows that $\max_{\Pi'} V_{s+t} - \min_{\Pi'} V_{s+t} \leq (2B + A)t$. □

The next lemma builds on the previous lemma and provides an upper bound on $\max V_{s+t} - \min U_{s+t}$ under the same conditions.

Lemma C.4. *Under the same setup and conditions as in Lemma C.3,* $\max_{\Pi'} V_{s+t} - \min_{\mathcal{D}'} U_{s+t} \leq (4B + 2A)t$.

Proof of Lemma C.4. Summing the two inequalities in Lemma C.3 and rearranging the terms, we have that $\max_{\Pi'} V_{s+t} - \min_{\mathcal{D}'} U_{s+t} \leq (4B + 2A)t + \min_{\Pi'} V_{s+t} - \max_{\mathcal{D}'} U_{s+t}$. It therefore suffices to show that $\min_{\Pi'} V_{s+t} \leq \max_{\mathcal{D}'} U_{s+t}$.

Let $\tau := s + t$. There exists $\pi' \in \Pi'$ and a stochastic strategy $\overline{d}' \in \mathcal{D}'$ such that $U_\tau(d) = U_0(d) + \tau \cdot r(d, \pi')$ and $V_\tau(\pi) = V_0(\pi) + \tau \cdot r(\overline{d}', \pi)$ for all $d \in \mathcal{D}'$ and all $\pi \in \Pi'$. Therefore, for this choice of π' and $\overline{d}'$, using (C.1), $\min_{\Pi'} V_\tau \leq V_\tau(\pi') = V_0(\pi') + \tau \cdot r(\overline{d}', \pi') \leq \max_{\Pi'} V_0 + \tau \cdot r(\overline{d}', \pi') = \min_{\mathcal{D}'} U_0 + \tau \cdot r(\overline{d}', \pi') \leq U_0(\overline{d}') + \tau \cdot r(\overline{d}', \pi') = U_\tau(\overline{d}') \leq \max_{\mathcal{D}'} U_\tau$. □

Proof of Theorem 4.4. It suffices to show that $\limsup_{t\to\infty}(\max V_t - \min U_t)/t \le 0$ by letting $U_0 \equiv 0$ and $V_0 \equiv 0$, which corresponds to Algorithm 5. Let $\epsilon > 0$. Note that r is Lipschitz continuous by Lemma C.2 and the fact that $r(d,\pi)$ is an average of bounded risks with weights π. Furthermore, $\mathcal{D}$ and Γ_ℓ are both compact. In addition, U_0 and V_0 are both continuous. Therefore, there exist covers $\mathcal{D} = \bigcup_{i=1}^{I} \mathcal{D}_i$ and $\Gamma_\ell = \bigcup_{j=1}^{J} \Pi_j$ such that (i) $\mathcal{D}_i$ and Π_j are all compact, and (ii) $\sup_{i,t} \mathrm{MV}_t(\mathcal{D}_i)/t \le \epsilon$, $\sup_{j,t} \mathrm{MV}_t(\Pi_j)/t \le \epsilon$. (Note that I and J may depend on ϵ.) For index sets $\mathcal{I} \subseteq \{1,2,\dots,I\}$ and $\mathcal{J} \subseteq \{1,2,\dots,J\}$, define $\mathcal{D}_\mathcal{I} := \bigcup_{i\in\mathcal{I}} \mathcal{D}_i$ and $\Pi_\mathcal{J} := \bigcup_{j\in\mathcal{J}} \Pi_j$. We show that $\max V_t - \min U_t \le C\epsilon t$ for an absolute constant C and all sufficiently large t via induction on the sizes of $\mathcal{I}$ and $\mathcal{J}$.

Let $\{(U_t, V_t)\}_t$ be a pair of cumulative Bayes risk functions constructed from $(\mathcal{D}_\mathcal{I}, \Pi_\mathcal{J})$ where $|\mathcal{I}| = |\mathcal{J}| = 1$. By (C.3) and the fact that $\mathrm{MV}_t(\mathcal{D}_\mathcal{I}) \le \epsilon t$ and $\mathrm{MV}_t(\Pi_\mathcal{J}) \le \epsilon t$, we have that

$$\begin{aligned}
\min_{\mathcal{D}_\mathcal{I}} U_t &= \min_{d\in\mathcal{D}_\mathcal{I}} [U_0(d) + t\cdot r(d,\pi_{(t)})] \ge \mathrm{E}_{d\sim\varpi_{(t)}}[U_0(d)] + t\cdot r(\overline{d}(\varpi_{(t)}),\pi_{(t)}) - \epsilon t \\
&\ge \min_{d\in\mathcal{D}_\mathcal{I}} U_0(d) + t\cdot r(\overline{d}(\varpi_{(t)}),\pi_{(t)}) - \epsilon t \\
&= \max_{\pi\in\Pi_\mathcal{J}} V_0(\pi) + t\cdot r(\overline{d}(\varpi_{(t)}),\pi_{(t)}) - \epsilon t \\
&\ge V_0(\pi_{(t)}) + t\cdot r(\overline{d}(\varpi_{(t)}),\pi_{(t)}) - \epsilon t \\
&\ge \max_{\pi\in\Pi_\mathcal{J}} [V_0(\pi) + t\cdot r(\overline{d}(\varpi_{(t)}),\pi)] - 2\epsilon t = \max_{\Pi_\mathcal{J}} V_t - 2\epsilon t.
\end{aligned}$$

Therefore, $\max_{\Pi_\mathcal{J}} V_t - \min_{\mathcal{D}_\mathcal{I}} U_t \le 2\epsilon t$.

Let $\epsilon' > 0$ be arbitrary. Suppose that there exists t_0 such that, for any $\mathcal{I}' \subseteq \mathcal{I}$ and $\mathcal{J}' \subseteq \mathcal{J}$ such that $\mathcal{I}' \ne \mathcal{I}$ or $\mathcal{J}' \ne \mathcal{J}$, for any pair of cumulative Bayes risk functions $\{(U_t, V_t)\}_t$ constructed from $(\mathcal{D}_{\mathcal{I}'}, \Pi_{\mathcal{J}'})$, it holds that $\max_{\Pi_{\mathcal{J}'}} V_t - \min_{\mathcal{D}_{\mathcal{I}'}} U_t \le \epsilon' t$ for all $t \ge t_0$. We next obtain a slightly greater bound on $\max_{\Pi_\mathcal{J}} V_t - \min_{\mathcal{D}_\mathcal{I}} U_t$ for all sufficiently large t.

We first prove that if, for a given pair of cumulative Bayes risk functions $\{(U_t, V_t)\}_t$ constructed from $(\mathcal{D}_\mathcal{I}, \Pi_\mathcal{J})$, there exists $i' \in \mathcal{I}$ or $j' \in \mathcal{J}$ such that $\mathcal{D}_{i'}$ or $\Pi_{j'}$ is not eligible

in an interval $[s, s+t_0]$, then

$$\max_{\Pi_{\mathcal{J}}} V_{s+t_0} - \min_{\mathcal{D}_{\mathcal{I}}} U_{s+t_0} \leq \max_{\Pi_{\mathcal{J}}} V_s - \min_{\mathcal{D}_{\mathcal{I}}} U_s + \epsilon' t_0. \tag{C.4}$$

Suppose that $\mathcal{D}_{i'}$ is not eligible in $[s, s+t_0]$, then define $U'_t := U_{s+t}$ and $V'_t := V_{s+t} - \max_{\Pi_{\mathcal{J}}} V_s + \min_{\mathcal{D}_{\mathcal{I}}} U_s$ for all $t \geq 0$. It is straightforward to check that $\{(U'_t, V'_t)\}_{t=0}^{t_0}$ satisfies the recursive definition of a pair of cumulative Bayes risk functions constructed from $(\mathcal{D}_{\mathcal{I}\setminus\{i'\}}, \Pi_{\mathcal{J}})$. By the induction hypothesis, $\max_{\Pi_{\mathcal{J}}} V'_{t_0} - \min_{\mathcal{D}_{\mathcal{I}\setminus\{i'\}}} U'_{t_0} \leq \epsilon' t_0$. Therefore, $\max_{\Pi_{\mathcal{J}}} V_{s+t_0} - \min_{\mathcal{D}_{\mathcal{I}}} U_{s+t_0} = \max_{\Pi_{\mathcal{J}}} V'_{t_0} - \min_{\mathcal{D}_{\mathcal{I}\setminus\{i'\}}} U'_{t_0} + \max_{\Pi_{\mathcal{J}}} V_s - \min_{\mathcal{D}_{\mathcal{I}}} U_s \leq \max_{\Pi_{\mathcal{J}}} V_s - \min_{\mathcal{D}_{\mathcal{I}}} U_s + \epsilon' t_0$. Similar argument can be applied if $\Pi_{j'}$ is not eligible in $[s, s+t_0]$.

Now we obtain a bound on $\max_{\Pi_{\mathcal{J}}} V_t - \min_{\mathcal{D}_{\mathcal{I}}} U_t$. Let $t > t_0$, $\mathcal{Q} := \lfloor t/t_0 \rfloor \geq 1$ and $\mathcal{R} := t/t_0 - \mathcal{Q} \in [0, 1)$. There are two cases.

Case 1: There exists $s_0 \leq \mathcal{Q}$ such that $\mathcal{D}_i$ and Π_j are eligible in $[(s_0 - 1 + \mathcal{R})t_0, (s_0 + \mathcal{R})t_0]$ for all $i \in \mathcal{I}$ and $j \in \mathcal{J}$. Take s_0 to be the largest such integer. Then, repeatedly apply (C.4) to intervals $[(s_0+\mathcal{R})t_0, (s_0+1+\mathcal{R})t_0], [(s_0+1+\mathcal{R})t_0, (s_0+2+\mathcal{R})t_0], \ldots, [(\mathcal{Q}-1+\mathcal{R})t_0, (\mathcal{Q}+\mathcal{R})t_0] = [t - t_0, t]$ and we derive that

$$\max_{\Pi_{\mathcal{J}}} V_t - \min_{\mathcal{D}_{\mathcal{I}}} U_t \leq \max_{\Pi_{\mathcal{J}}} V_{(s_0+\mathcal{R})t_0} - \min_{\mathcal{D}_{\mathcal{I}}} U_{(s_0+\mathcal{R})t_0} + \epsilon'(\mathcal{Q} - s_0)t_0.$$

By Lemma C.4, $\max_{\Pi_{\mathcal{J}}} V_{(s_0+\mathcal{R})t_0} - \min_{\mathcal{D}_{\mathcal{I}}} U_{(s_0+\mathcal{R})t_0} \leq (4B + \epsilon)t_0$. Therefore,

$$\max_{\Pi_{\mathcal{J}}} V_t - \min_{\mathcal{D}_{\mathcal{I}}} U_t \leq (4B + \epsilon)t_0 + \epsilon'(\mathcal{Q} - s_0)t_0 \leq (4B + \epsilon)t_0 + \epsilon' t.$$

Case 2: There is no integer s_0 satisfying the condition in Case 1. Then, repeatedly apply (C.4) to intervals $[\mathcal{R}t_0, (1+\mathcal{R})t_0], [(1+\mathcal{R})t_0, (2+\mathcal{R})t_0], \ldots, [(\mathcal{Q}-1+\mathcal{R})t_0, (\mathcal{Q}+\mathcal{R})t_0] = [t-t_0, t]$, we derive that

$$\max_{\Pi_{\mathcal{J}}} V_t - \min_{\mathcal{D}_{\mathcal{I}}} U_t \leq \max_{\Pi_{\mathcal{J}}} V_{\mathcal{R}t_0} - \min_{\mathcal{D}_{\mathcal{I}}} U_{\mathcal{R}t_0} + \epsilon' \mathcal{Q} t_0.$$

By the bound on the risk, $\max_{\Pi_{\mathcal{J}}} V_{\mathcal{R}t_0} \leq B\mathcal{R}t_0$ and $\min_{\mathcal{D}_{\mathcal{I}}} U_{\mathcal{R}t_0} \geq -B\mathcal{R}t_0$. Hence,

$$\max_{\Pi_{\mathcal{J}}} V_t - \min_{\mathcal{D}_{\mathcal{I}}} U_t \leq 2B\mathcal{R}t_0 + \epsilon' \mathcal{Q} t_0 \leq (4B + \epsilon)t_0 + \epsilon' t.$$

Thus, in both cases, it holds that $\max_{\Pi_{\mathcal{J}}} V_t - \min_{\mathcal{D}_{\mathcal{I}}} U_t \leq (4B+\epsilon)t_0 + \epsilon' t$ for $t > t_0$. Let $C > 0$ be any constant (which may depend on ϵ, the approximation error of the covers, that is, the bound on MV_t/t). The following holds for any sufficiently large t,

$$\max_{\Pi_{\mathcal{J}}} V_t - \min_{\mathcal{D}_{\mathcal{I}}} U_t \leq (4B+\epsilon)t_0 + \epsilon' t \leq (1+C)\epsilon' t. \tag{C.5}$$

In other words, we show that after increasing the size of either index set by 1, for all sufficiently large t, we obtain a bound on $\max_{\Pi_{\mathcal{J}}} V_t - \min_{\mathcal{D}_{\mathcal{I}}} U_t$ that grows by a multiplicative factor of $(1+C)$ relative to the original bound.

It takes finitely many, say N, steps to induct from the initial case where the sizes of both index sets are one to the case of interest with index sets $\{1,\ldots,I\}$ and $\{1,\ldots,J\}$. (Note that N may also depend on ϵ through its dependence on I and J.) Take $C = 1/N$ in (C.5) and we derive that, for all sufficiently large t,

$$\max V_t - \min U_t = \max_{\Pi_{\{1,\ldots,J\}}} V_t - \min_{\mathcal{D}_{\{1,\ldots,I\}}} U_t \leq (1+1/N)^N \cdot 2\epsilon t \leq 2e\epsilon t$$

where e is the base of natural logarithm. Since ϵ is arbitrary, we show that $\limsup_{t\to\infty}(\max V_t - \min U_t)/t \leq 0$. □

C.1.5 Derivation of Γ*-minimax estimator of the mean in Section 4.5.1*

In this section, we show that, for the problem of estimating the mean in Section 4.5.1, one Γ-minimax estimator lies in $\mathcal{D}_{\text{linear}}$. This is formally presented below.

Proposition C.1. *Let $\mathcal{M}$ consist of all probability distributions defined on the Borel σ-algebra on $[0,1]$. Let $X_1,\ldots,X_n \overset{\text{iid}}{\sim} P_0 \in \mathcal{M}$ and $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ be the observed data. Let $\Psi : P \mapsto \int xP(\mathrm{d}x)$ denote the mean parameter and $\Gamma = \{\pi \in \Pi : \int \Psi(P)\pi(\mathrm{d}P) = \mu\}$ be the set of priors that represent prior information. Let $\mathcal{D}$ denote the space of estimators that are square-integrable with respect to all $P \in \mathcal{M}$. Consider the risk in Example 5, $R : (d,P) \mapsto \mathrm{E}_P[(d(\mathbf{X}) - \Psi(P))^2]$. Define $\bar{X} = \sum_{i=1}^n X_i/n$ and $d_0 : \mathbf{X} \mapsto (\mu + \sqrt{n}\bar{X})/(1+\sqrt{n})$. Then $d_0 \in \mathcal{D}_{\text{linear}}$ is Γ-minimax over $\mathcal{D}$.*

We first present a theorem on a criterion of Γ-minimaxity.

Theorem C.1. *Suppose that $d_0 \in \mathcal{D}$ is a Bayes estimator for $\pi_0 \in \Gamma$ and $r(d_0, \pi_0) = r_{\sup}(d_0, \Gamma)$. Then d_0 is a Γ-minimax estimator in $\mathcal{D}$.*

Proof of Theorem C.1. Clearly $r_{\sup}(d_0, \Gamma) \geq \inf_{d \in \mathcal{D}} r_{\sup}(d, \Gamma)$. Fix $d' \in \mathcal{D}$. Then, $r_{\sup}(d', \Gamma) \geq r(d', \pi_0) \geq r(d_0, \pi_0) = r_{\sup}(d_0, \Gamma)$. Since d' is arbitrary, this shows that $\inf_{d \in \mathcal{D}} r_{\sup}(d, \Gamma) \geq r_{\sup}(d_0, \Gamma)$. Thus, $r_{\sup}(d_0, \Gamma) = \inf_{d \in \mathcal{D}} r_{\sup}(d, \Gamma)$ and d_0 is Γ-minimax. □

We now present a lemma that is used to prove Proposition C.1.

Lemma C.5. *Let $a < b$ and suppose that $\mathcal{M}$ denotes the model space that consists of all probability distributions defined on the Borel σ-algebra on $[a, b] \subseteq \mathbb{R}$ with mean $\mu \in [a, b]$. Let X denote a generic random variable generated from some $P \in \mathcal{M}$. Then $\max_{P \in \mathcal{M}} \mathrm{Var}_P(X) = \mathrm{Var}_{P^*}(X) = (b - \mu)(\mu - a)$, where P^* is defined by $P^*(X = a) = (b - \mu)/(b - a)$ and $P^*(X = b) = (\mu - a)/(b - a)$.*

Proof of Lemma C.5. Without loss of generality, we may assume that $a = -1$ and $b = 1$. Note that for any $P \in \mathcal{M}$, it holds that $\mathrm{Var}_P(X) = \mathrm{E}_P[X^2] - \mathrm{E}_P[X]^2 = \mathrm{E}_P[X^2] - \mu^2 \leq 1 - \mu^2$, where the equality is attained if $P(X \in \{-1, 1\}) = 1$. Therefore, the maximum variance is achieved at the distribution with the specified mean μ and support being $\{a, b\}$, that is, at the distribution P^* defined in the lemma statement. Straightforward calculations show that $\mathrm{Var}_{P^*}(X) = (b - \mu)(\mu - a)$. □

Proof of Proposition C.1. Let $\mathcal{M}' := \{\text{Bernoulli}(\theta) : \theta \in (0, 1)\} \subseteq \mathcal{M}$ and let π_0 be a prior distribution over $\mathcal{M}'$ such that the prior distribution on the success probability θ is $\text{Beta}(\mu\sqrt{n}, (1 - \mu)\sqrt{n})$. By Theorem 1.1 in Chapter 4 of Lehmann and Casella (1998), a Bayes estimator for π_0 minimizes the risk under the posterior distribution, whose minimizer over $\mathcal{D}$ is the posterior mean d_0 for our choice of risk. That is, d_0 is a Bayes estimator in $\mathcal{D}$ for π_0.

We next show that $r(d_0, \pi_0) = \sup_{\pi \in \Gamma} r(d_0, \pi)$. Let $\pi \in \Gamma$ be arbitrary. Since $\mathrm{E}_P[\bar{X}] = \Psi(P)$ and $\mathrm{Var}_P(\bar{X}) = \mathrm{Var}_P(X_1)/n$, we can derive that

$$\begin{aligned} r(d_0, \pi) &= \int \mathrm{E}_P \left[\left\{ \frac{\mu + \sqrt{n}\bar{X}}{1 + \sqrt{n}} - \Psi(P) \right\}^2 \right] \pi(\mathrm{d}P) \\ &= \int \mathrm{E}_P \left[\left\{ \frac{\sqrt{n}}{1 + \sqrt{n}} \left(\bar{X} - \Psi(P) \right) + \frac{\mu - \Psi(P)}{1 + \sqrt{n}} \right\}^2 \right] \pi(\mathrm{d}P) \\ &= \int \left\{ \frac{1}{(1 + \sqrt{n})^2} \mathrm{Var}_P(X_1) + \frac{(\mu - \Psi(P))^2}{(1 + \sqrt{n})^2} \right\} \pi(\mathrm{d}P) \end{aligned}$$

Apply Lemma C.5 to $\mathrm{Var}_P(X_1)$ and the display continues as

$$\begin{aligned} &\leq \int \left\{ \frac{1}{(1 + \sqrt{n})^2} \Psi(P)(1 - \Psi(P)) + \frac{(\mu - \Psi(P))^2}{(1 + \sqrt{n})^2} \right\} \pi(\mathrm{d}P) \\ &= \int \frac{1}{(1 + \sqrt{n})^2} \left\{ \mu^2 + (1 - 2\mu)\Psi(P) \right\} \pi(\mathrm{d}P) = \frac{\mu(1 - \mu)}{(1 + \sqrt{n})^2}. \end{aligned}$$

This upper bound can be attained by any π with support contained in $\mathcal{M}'$, for example, π_0. Therefore, $r_{\sup}(d_0, \Gamma) = r(d_0, \pi_0)$. By Theorem C.1, d_0 is Γ-minimax over $\mathcal{D}$. □