To the Directors of the Program on Computer Science:

I certify that I have read the thesis of Karthik Seetharaman

in its final form for submission and have found it to be satisfactory for the degree of

Bachelor of Science with Honors.

May 4, 2026

Vasileios Syrgkanis (reader's signature)

_____ (name of reader)

MS&E _____ (name of reader's department)

To the Directors of the Program on Computer Science:

I certify that I have read the thesis of Karthik Seetharaman

in its final form for submission and have found it to be satisfactory for the degree of

Bachelor of Science with Honors.

May 4, 2026

_____ (reader's signature)

Jiafeng Chen _____ (name of reader)

Economics _____ (name of reader's department)

INFLUENCE FUNCTIONS AND TARGETED LEARNING

Karthik Seetharaman

Stanford University
June 2026

An honors thesis submitted to the department of
Computer Science
in partial fulfillment of the requirements for the undergraduate
honors program

Advisor: Vasilis Syrgkanis

Abstract

The typical problem in causal inference involves estimating a finite-dimensional functional of an unknown data-generating distribution. This estimation often requires learning infinite-dimensional nuisance components such as conditional means, densities, or propensity scores. This makes statistically accurate estimation of the parameter of interest difficult, as naive plug-in estimators can inherit first-order bias from these nuisance estimates, preventing $O(n^{-\frac{1}{2}})$ convergence and asymptotic linearity and asymptotic normality even when the target parameter itself is low-dimensional. These properties are desirable as they allow for the construction of statistically valid confidence intervals. The *influence function* of an estimator characterizes its first-order asymptotic behavior, making it central to devising methods to remove first-order bias in estimators.

This thesis synthesizes the literature of influence functions in semiparametric efficiency theory and targeted learning. We begin with nonparametric models, where the influence function is unique. We then extend the theory to semiparametric models and develop a broader geometry of influence functions, allowing us to characterize the *efficient influence function* with the lowest possible asymptotic variance. After describing methods for deriving influence functions in practice, we turn to Targeted Maximum Likelihood Estimation (TMLE), a debiasing framework that updates an initial estimate of the data-generating distribution along a fluctuation submodel chosen using the efficient influence function. This targeting step yields a plug-in estimator whose first-order behavior is governed by the efficient influence function, allowing for asymptotic linearity and valid inference under appropriate conditions.

The final chapters study several extensions of the basic TMLE construction. One-Step TMLE uses least favorable submodels to achieve targeting in a single update rather than the multi-step iteration of basic TMLE. Cross-Validated TMLE uses sample splitting, allowing us to verify asymptotic linearity without appealing to Donsker (empirical process) conditions that can be difficult to verify in practice. Collaborative TMLE targets nuisance parameters to increase the accuracy of estimation of the target parameter, allowing for further bias reduction. Higher-Order TMLE extends targeting beyond first-order bias correction, leveraging higher-order derivatives of the target parameter (or approximations of these derivatives) so that asymptotic linearity depends only on controlling higher-order remainder terms. This thesis provides a self-contained exposition of targeted learning and its underlying semiparametric efficiency theory, unifying a relatively scattered literature.

Acknowledgements

I am deeply indebted to Vasilis Syrgkanis for his supervision and support throughout this project. I reached out to Vasilis in my sophomore year and we started having meetings to discuss interesting papers. This quickly turned into a research idea on preference optimization which, two years later, has been published. I was introduced to causal inference and causal machine learning through his class in junior fall, where I completed a project on proving the asymptotic linearity of TMLE in a special case. That project naturally extended into this thesis. I am constantly in awe of Vasilis's subject matter knowledge, sharpness, and attention to detail, and I am forever grateful for how giving of his time he was, even in weeks where I struggled. I hope to embody some of these qualities in my career going forward.

I am also deeply grateful to Jiafeng Chen for being a second reader on this thesis and for providing me with new perspectives on the topic. A special thank you must go to Ravi Jagadeesan for advising me on economics research while at Stanford, and for being a fantastic mentor as I begin to navigate the world of academia. I would also like to extend my sincere thank you to Noah Rosenberg and Lily Agranat-Tamir for working with me on interesting projects in discrete mathematics and mathematical biology. I would be remiss without thanking Neil Band for his research mentorship and for being the source of many interesting conversations during my senior year. Finally, I would like to thank my family for their unwavering support, and my friends for making Stanford such a fun place to be these past four years.

Contents

Chapter 1

Introduction

The usual formulation of a causal inference problem is to estimate some functional of interest $\psi_0 := \Psi(P_0)$, a function of an unknown distribution $P_0 \in \mathcal{M}$ (here, $\mathcal{M}$ is a *model*, or a set of distributions), given observations $O_1, O_2, \ldots, O_n \sim P_0$. This can be rather difficult as, most of the time, we are not working in a fully parametric setting, but rather P_0 can contain functions like conditional expectations or propensity scores which are infinite-dimensional. Hence, we have a finite-dimensional parameter of interest but infinite-dimensional "nuisances" that are part of our data-generating distribution. To accurately estimate the finite-dimensional ψ, we often must also estimate these infinite-dimensional nuisances accurately - this is the *semiparametric setting*.

The most natural approach is to construct some estimate $\hat{P_n}$ of the actual data-generating distribution P_0 and then construct the plug-in estimate $\hat{\psi_n} = \Psi(\hat{P_n})$. However, this naive approach is not ideal as it fails to have desirable statistical properties. In particular, we would like our estimator to achieve $O(\frac{1}{\sqrt{n}})$ convergence rates and asymptotic normality, so that we can construct confidence intervals. The estimation of infinite-dimensional nuisances makes this difficult - first-order bias from these estimates propagates to the naive plug-in estimator, destroying these statistical properties. Hence, it is necessary to incorporate a *debiasing* procedure to ensure first-order bias from nuisance estimation does not propagate to our estimation of the functional, allowing us to maintain desirable statistical properties.

The most desirable estimators are *asymptotically linear* estimators, or estimators $\hat{\psi_n}$ that satisfy an expansion of the form

$$\hat{\psi_n} - \psi_0 = (P_n - P_0)\phi(O; P_0) + o_P\left(\frac{1}{\sqrt{n}}\right),$$

where P_n is the empirical distribution of the observations O, and ϕ is a mean-zero, finite-variance function. Note that, in this case, asymptotic linearity automatically implies asymptotic normality and $\sqrt{n}$-consistency, allowing for the construction of confidence intervals. The function ϕ is known as an *influence function*, and is the focus of the first half of this thesis. Note that, by the Central Limit Theorem, the asymptotic variance of $\hat{\psi_n}$ is determined by ϕ.

It turns out that, even in the semiparametric and fully nonparametric cases, there is still a lower bound on the asymptotic variance of an estimator $\hat{\psi_n}$, presupposing that it is asymptotically linear and regular (a condition that will be elucidated in Chapter 3). This lower bound is achieved by the *efficient influence function*. It is unique in the fully nonparametric case, and in the semiparametric case, it can be characterized

by considering the underlying geometry of the *tangent space* of the set of distributions to be considered.

The second half of the thesis turns to *targeted learning*, a debiasing framework that uses the efficient influence function for the construction of asymptotically linear estimators that achieve the semiparametric efficiency lower bound. The idea is to, beginning with an initial estimate of P_0, update ("target") this estimate in a particular direction such that we get closer and closer to a linear expansion in the efficient influence function. After targeting the distribution appropriately, applying a plug-in estimate on this new distribution (the *targeted maximum likelihood estimator*) will remove first-order bias, thereby achieving the desirable statistical properties mentioned earlier.

The aim of this thesis is to develop the theoretical foundations underlying these ideas. We begin, in Chapter 2, by introducing influence functions in nonparametric models, where they are unique. In Chapter 3, we extend this to the semiparametric setting, developing the theory of tangent spaces. Chapter 4 provides a quick primer on methods to actually derive these influence functions in practice, allowing us to use the theory developed in the previous chapters with explicit objects.

After developing the theory of influence functions, we move to developing targeted learning. In general, we focus on developing asymptotic linearity results in various settings. Chapter 5 introduces the basic TMLE and proves asymptotic linearity under mild conditions. The last four chapters of the thesis introduce four different variants of the TMLE. Chapter 6 describes a version of TMLE that can converge in only one step, as opposed to the iterative process of the standard TMLE. Chapter 8 describes a TMLE where the nuisance estimates are also targeted to improve the estimation of the target parameter. Chapter 7 describes a TMLE where cross-validation is applied to every step; asymptotic linearity is shown to still hold here, making this useful in practice. Finally, Chapter 9 describes an extension to the TMLE where we control bias beyond the first-order.

Chapter 2

Influence Functions in Nonparametric Models

We begin our discussion in the simplest case of a *nonparametric model*. Here, the parameter of interest is a functional of a completely unrestricted distribution, but, amazingly, we can still develop a notion of optimality for estimators (the *nonparametric efficiency bound*) even without underlying structure.

2.1 The Nonparametric Efficiency Bound

Definition 2.1.1 *A **model** $\mathcal{P}$ is a set of distributions, and such a model is **nonparametric** if we make no assumptions on the structure of the underlying distributions.*

We now consider the following setup, akin to (Kennedy, 2024). Suppose we observe i.i.d. $Z_1, \ldots, Z_n \sim \mathbb{P}$, where $\mathbb{P} \in \mathcal{P}$ and $\mathcal{P}$ is a nonparametric model. Our goal is then to estimate some functional of interest $\psi : \mathcal{P} \to \mathbb{R}^q$.

If our estimate is $\hat{\psi}$, we want to have some idea of optimality for how good our estimate $\hat{\psi}$ is for ψ. There are two usual levers we can use for defining optimality, which are the *bias* and *variance* of $\hat{\psi}$. Ideally, $\hat{\psi}$ should be unbiased for ψ (so $\hat{\psi}$ is a *U-estimator*), so restricting to U-estimators of ψ, the estimator $\hat{\psi}$ that we choose should be the one with the lowest variance.

Remark 2.1.2 *In this chapter, we motivate the nonparametric efficiency bound for unbiased estimators. In Chapter 3, we present the analogous semiparametric efficiency bound for regular and asymptotically linear estimators, which are the types of estimators we tend to care about in practice.*

How exactly to determine this lowest variance estimator in a fully nonparametric model is unclear, so we first turn to the fully parametric case, where $\mathcal{P} = \{P_\theta : \theta \in \mathbb{R}\}$ is parameterized by θ. Our parametric model could be specified by higher-dimensional parameters, but for now, we focus on the scalar case.

In the parametric case, the well-known Cramér-Rao bound gives us a lower bound on the variance of any U-estimator $\hat{\psi}$:

Theorem 2.1.3 (Cramér-Rao) *Let ψ be a smooth functional of a smooth parametric model $\mathcal{P} = \{P_\theta : \theta \in \mathbb{R}\}$, so that P_θ and $\psi(\theta)$ are both differentiable in θ. Then, the variance of any U-estimator $\hat{\psi}$ satisfies*

$$Var_\theta(\hat{\psi}) \geq \frac{\psi'(\theta_0)^2}{Var_\theta(s_\theta(Z))},$$

where $s_\theta(z) = \frac{\partial}{\partial\theta} \log p_\theta(z)$ is the score function (derivative of the log-density).

To use Cramér-Rao style bounds in nonparametric settings, we want to be able to generalize the role of the score. We can no longer differentiate the whole density with respect to θ, so instead we differentiate along parametric paths that go through the true distribution. These paths are *parametric submodels*, and defining such submodels allows us to still use parametric ideas in the nonparametric setting.

Definition 2.1.4 *A **parametric submodel** $\mathcal{P}_\varepsilon$ of a nonparametric model $\mathcal{P}$ is a smooth parametric model $\mathcal{P}_\varepsilon = \{P_\varepsilon : \varepsilon \in \mathbb{R}\}$ such that $\mathcal{P}_\varepsilon \subseteq \mathcal{P}$ and $P_0 = \mathbb{P}$.*

Say we are to estimate ψ within a parametric submodel $\mathcal{P}_\varepsilon \subseteq \mathcal{P}$. Cramér-Rao gives us a lower bound on the variance of U-estimators $\hat{\psi}$ that are estimated over $\mathcal{P}_\varepsilon$. Since the parametric submodel is a subset of the entire nonparametric model, the Cramér-Rao lower bound for $\mathcal{P}_\varepsilon$ is, in turn, also a lower bound for the variance of a U-estimator $\hat{\psi}$ estimated over the entire nonparametric model $\mathcal{P}$. This gives us some direction in lower bounding our variance over $\mathcal{P}$, by taking a supremum over parametric submodels.

Now, consider the parametric submodel

$$p_\varepsilon(z) = p(z)\{1 + \varepsilon h(z)\}$$

for mean-zero h, where, to make the densities valid, we require $||h||_\infty \leq M < \infty$ and $\varepsilon < \frac{1}{M}$. We can calculate the score function

$$s_\varepsilon(z) = \frac{\partial}{\partial\varepsilon} \log p_\varepsilon(z)\bigg|_{\varepsilon=0} = \frac{\partial}{\partial\varepsilon} \log(1 + \varepsilon h(z))\bigg|_{\varepsilon=0} = h(z).$$

Thus, the Cramér-Rao bound for this submodel is (noting that the true parameter is $\varepsilon = 0$):

$$\mathrm{Var}_{P_\varepsilon}(\hat{\psi}) \geq \frac{\left(\frac{\partial}{\partial\varepsilon}\psi(P_\varepsilon)|_{\varepsilon=0}\right)^2}{\mathrm{Var}_{P_\varepsilon}(s_\varepsilon(z))} = \frac{\left(\frac{\partial}{\partial\varepsilon}\psi(P_\varepsilon)|_{\varepsilon=0}\right)^2}{\mathbb{E}_{P_\varepsilon}[h(Z)^2]}.$$

The above holds for $\mathcal{P}_\varepsilon$ defined by any mean-zero function h, so to find the best such lower bound, we have to optimize this over all P_ε in a parametric submodel. To do this generally, we define the notion of pathwise differentiability. For simplicity, we state the definition in the case of a scalar parameter $\psi : \mathcal{P} \to \mathbb{R}$.

Definition 2.1.5 *We say $\psi : \mathcal{P} \to \mathbb{R}$ is **pathwise differentiable** if there exists some mean-zero, finite-variance function $\varphi(z; P)$ (so $\int \varphi(z; P)dP(z) = 0$ and $\int \varphi(z; P)^2 dP(z) < \infty$) such that*

$$\frac{\partial}{\partial\varepsilon}\psi(P_\varepsilon)\bigg|_{\varepsilon=0} = \int \varphi(z; \mathbb{P}) s_\varepsilon(z) d\mathbb{P}(z)$$

*for every smooth submodel P_ε. We call φ the **influence function** (also called the **pathwise derivative**).*

Remark 2.1.6 *For now, we ignore the question of the uniqueness of φ, which we will return to in the next section. For nonparametric models $\mathcal{P}$, the choice of φ is unique.*

In a sense, φ acts as the "derivative" term in the definition of pathwise differentiability. For more motivation on why this is the case, we can note that the condition of pathwise differentiability is implied by the *von Mises expansion*:

Definition 2.1.7 *The **von Mises expansion** writes, for arbitrary distributions $\overline{P}$ and P,*

$$\psi(\overline{P}) - \psi(P) = \int \varphi(z; \overline{P}) d(\overline{P} - P)(z) + R_2(\overline{P}, P)$$

for a second-order remainder term R_2 (so R_2 only depends on products and squares).

Note that this expansion is essentially a distributional first-order Taylor expansion with φ acting as the derivative term.

Lemma 2.1.8 *If a functional ψ satisfies the von Mises expansion, it also satisfies pathwise differentiability (assuming regularity conditions such that the integral and derivative are interchangeable).*

Proof: Taking $\overline{P} = P$ and $P = P_\varepsilon$ in the von Mises expansion of ψ gives

$$\psi(P) - \psi(P_\varepsilon) = \int \varphi(z; P) d(P - P_\varepsilon)(z) + R_2(P, P_\varepsilon).$$

Differentiating both sides and ignoring terms that are constant with respect to ε, we can write

$$\frac{\partial}{\partial \varepsilon} \psi(P_\varepsilon) \bigg|_{\varepsilon=0} = \frac{\partial}{\partial \varepsilon} \left(\int \varphi(z; P) dP_\varepsilon(z) \right) \bigg|_{\varepsilon=0} - \frac{\partial}{\partial \varepsilon} R_2(P, P_\varepsilon) \bigg|_{\varepsilon=0}.$$

Proceeding term-by-term, under regularity conditions that allow us to swap the integral and derivative, we can write

$$\frac{\partial}{\partial \varepsilon} \left(\int \varphi(z; P) dP_\varepsilon(z) \right) \bigg|_{\varepsilon=0} = \int \varphi(z; P) \frac{\partial}{\partial \varepsilon} dP_\varepsilon(z) \bigg|_{\varepsilon=0}.$$

Note that

$$s_\varepsilon(z) = \frac{\partial}{\partial \varepsilon} \log dP_\varepsilon(z) \bigg|_{\varepsilon=0} = \frac{\frac{\partial}{\partial \varepsilon} dP_\varepsilon(z) \Big|_{\varepsilon=0}}{dP(z)},$$

so we can write

$$\int \varphi(z; P) \frac{\partial}{\partial \varepsilon} dP_\varepsilon(z) \bigg|_{\varepsilon=0} = \int \varphi(z; P) s_\varepsilon(z) dP(z).$$

On the other hand, since R_2 is second-order, we have

$$\frac{\partial}{\partial \varepsilon} R_2(P, P_\varepsilon) \bigg|_{\varepsilon=0} = 0,$$

so we get the pathwise differentiability condition. □

Remark 2.1.9 *Pathwise differentiability is also roughly equivalent to Neyman orthogonality, which states that, given a functional $\psi(W;\theta,\eta)$ where θ is the parameter of interest and η is a nuisance with true values θ_0, η_0,*

$$D_0[\eta - \eta_0] = 0$$

for all estimators $\hat{\eta_0}$ that occur with high probability, where we define the pathwise (Gâteaux) derivative

$$D_r[\eta - \eta_0] = \frac{\partial}{\partial r}\mathbb{E}_P[\psi(W;\theta_0;\eta_0 + r(\eta - \eta_0))]$$

(Chernozhukov et al., 2018; Foster & Syrgkanis, 2023).

This relationship is elaborated on in (Y. Chen, Kennedy, & Balakrishnan, 2026), and has to do with the introduction of the extra condition of ***local product structure****, which essentially is the condition that the parameter of interest and nuisance parameter can vary independently of each other within the model.*

Theorem 2.1.10 (Nonparametric Efficiency Bound) *Let $\psi : \mathcal{P} \to \mathbb{R}$ be pathwise differentiable with influence function $\varphi(Z)$. Then, the variance of any U-estimator $\hat{\psi}$ is at least the variance of $\varphi(Z)$.*

Proof: We return to the parametric submodel $p_\varepsilon(z) = d\mathbb{P}(z)\{1 + \varepsilon h(z)\}$, which has score function $h(z)$. By pathwise differentiability of ψ, we can write

$$\frac{\partial}{\partial \varepsilon}\psi(P_\varepsilon)\bigg|_{\varepsilon=0} = \int \varphi(z;\mathbb{P})h(z)d\mathbb{P}(z)$$

since $s_\varepsilon(z) = h(z)$. As derived earlier, the Cramér-Rao bound over any parametric submodel $\mathcal{P}_\varepsilon$ is given by

$$\text{Var}(\hat{\psi}) \geq \frac{(\frac{\partial}{\partial \varepsilon}\psi(P_\varepsilon)|_{\varepsilon=0})^2}{\mathbb{E}[h(Z)^2]} = \frac{\mathbb{E}[\varphi(Z;\mathbb{P})h(Z)]^2}{\mathbb{E}[h(Z)^2]}.$$

Hence, taking a supremum over Cramér-Rao bounds of submodels $\mathcal{P}_\varepsilon$ with generic element P_ε, we can write

$$\sup_{P_\varepsilon} \frac{(\frac{\partial}{\partial \varepsilon}\psi(P_\varepsilon)|_{\varepsilon=0})^2}{\text{Var}(s_\varepsilon(Z))} = \sup_h \frac{\mathbb{E}[\varphi(Z;\mathbb{P})h(Z)]^2}{\mathbb{E}[h(Z)^2]} \leq \mathbb{E}[\varphi(Z;\mathbb{P})^2] = \text{Var}(\varphi(Z)),$$

where the inequality is by Cauchy-Schwarz.

This bound, which is the *nonparametric efficiency bound*, gives an analogue of the Cramér-Rao bound for nonparametric models. Of course, for it to be useful, it should be sharp; in particular, we would like it to actually be achieved by some function h. Recall that the equality case of Cauchy-Schwarz is when h is a multiple of φ. Our other requirement is that h is a valid score function.

In the case of nonparametric models, $h = \varphi(z;\mathbb{P})$ is a valid choice (in that $\varphi(z;\mathbb{P})$ is a valid score function for some distribution), so for nonparametric models, $\text{Var}(\varphi(z;\mathbb{P}))$ is the nonparametric efficiency bound, and we can set $\varphi(z;\mathbb{P})$ to be the *efficient influence function*. In particular, this means that if we have an estimator $\hat{\psi}$ which satisfies

$$\sqrt{n}(\hat{\psi} - \psi) \xrightarrow{d} \mathcal{N}(0, \text{Var}(\varphi(z))),$$

this estimator attains the nonparametric efficiency bound and is optimal in the sense of achieving minimal variance. □

Remark 2.1.11 *In light of Remark 2.1.6, the distinction between an influence function and efficient influence function seems somewhat pointless, as φ is unique. However, in the case of semiparametric models, as in the next chapter, this distinction becomes much more meaningful, as there are many possible influence functions.*

Chapter 3

Semiparametric Efficiency Theory

We now move from the fully nonparametric setting to the semiparametric setting, which is the setting most commonly encountered when dealing with causal estimation problems. We first develop the theory in the parametric case before extending our results to the semiparametric case. The main goal here is to extend the efficiency bound from last chapter into a *semiparametric efficiency bound*, and in turn, characterize what influence functions are possible in the semiparametric setting. This will be done by examining the underlying geometry of these functions through the *tangent spaces*. Working in the parametric case allows us to explicitly characterize the tangent space objects in a finite-dimensional case first, making the later extension to infinite-dimensional nuisance parameters in the semiparametric case much easier.

Many of the results discussed in this chapter are taken from Chapters 3 and 4 of (Tsiatis, 2006). Additional sources for this information are (Y.-C. Chen, 2024) and (Sen, 2018).

3.1 RAL Estimators

We work in the setting of observations $Z_1, \ldots, Z_n$ being i.i.d. random vectors drawn from a density $p_Z(z;\theta)$ with dominating measure ν_Z. We have a parameter of interest $\theta = (\beta^T, \eta^T)^T$, where β is a q-dimensional parameter of interest, and η is an r-dimensional nuisance parameter. Later, we will extend our results to the case where the nuisance parameter is infinite-dimensional.

Definition 3.1.1 *An estimator $\hat{\beta}_n$ of β is **asymptotically linear** if there exists some mean-zero random vector $\varphi(Z)$ such that $\mathbb{E}[\varphi\varphi^T]$ is finite and non-singular, and*

$$n^{\frac{1}{2}}(\hat{\beta}_n - \beta_0) = n^{-\frac{1}{2}} \sum_{i=1}^{n} \varphi(Z_i) + o_P(1).$$

*Then, φ is the **influence function** of the estimator $\hat{\beta}_n$.*

Remark 3.1.2 *Given an asymptotically linear estimator $\hat{\beta}_n$, by the central limit theorem, we can write*

$$n^{\frac{1}{2}}(\hat{\beta}_n - \beta_0) \xrightarrow{d} \mathcal{N}(0, \mathbb{E}[\varphi\varphi^T]).$$

In particular, this means that the asymptotic properties of the estimator are determined by the influence function.

Theorem 3.1.3 *An asymptotically linear estimator has an almost surely unique influence function.*

Proof: If not, then there exist two influence functions φ^*, φ for $\hat{\beta}_n$ such that

$$n^{\frac{1}{2}}(\hat{\beta}_n - \beta_0) = n^{-\frac{1}{2}} \sum_{i=1}^{n} \varphi(Z_i) + o_P(1)$$

and similarly for φ^*. Subtracting the two equations, we get

$$n^{-\frac{1}{2}} \sum_{i=1}^{n} (\varphi(Z_i) - \varphi^*(Z_i)) = o_P(1).$$

By the Central Limit Theorem,

$$n^{-\frac{1}{2}} \sum_{i=1}^{n} (\varphi(Z_i) - \varphi^*(Z_i)) \xrightarrow{d} \mathcal{N}(0, \mathbb{E}[(\varphi - \varphi^*)(\varphi - \varphi^*)^T]).$$

We need this limiting distribution to be $o_P(1)$, which implies the covariance matrix must be 0. Hence, we have $\mathbb{E}[(\varphi - \varphi^*)(\varphi - \varphi^*)^T] = 0$, which implies the conclusion. □

We restrict our attention to regular estimators, defined as below.

Definition 3.1.4 *A local data generating process (LDGP) is a set of variables such that, for each n, the data $Z_{1n}, Z_{2n}, \ldots, Z_{nn}$ are distributed according to θ_n (i.i.d. $p(z, \theta_n)$), where $n^{\frac{1}{2}}(\theta_n - \theta^*)$ converges to a constant for some fixed parameter θ^*.*

Definition 3.1.5 *Given an LDGP, an estimator $\hat{\beta}_n = \hat{\beta}_n(Z_{1n}, \ldots, Z_{nn})$ is regular if, for all θ^*, $n^{\frac{1}{2}}(\hat{\beta}_n - \beta_n)$ has a limiting distribution independent of the LDGP.*

To see why this is useful, we construct the following pathological estimator, due to Hodges:

Example 3.1.6 *Let $Z_1, \ldots, Z_n \sim \mathcal{N}(\mu, 1)$. To estimate μ, the sample mean $\overline{Z_n}$ is the MLE, so $n^{\frac{1}{2}}(\overline{Z_n} - \mu) \xrightarrow{d} \mathcal{N}(0, 1)$. Hodges constructs the estimator $\hat{\mu}_n = \begin{cases} \overline{Z_n} & |\overline{Z_n}| > n^{-\frac{1}{4}} \\ 0 & |\overline{Z_n}| \leq n^{-\frac{1}{4}} \end{cases}$. For $\mu \neq 0$, $\overline{Z_n}$ moves away from 0 with increasing probability, so*

$$n^{\frac{1}{2}}(\overline{Z_n} - \mu) = n^{\frac{1}{2}}(\hat{\mu}_n - \mu) + o_P(1),$$

which implies $n^{\frac{1}{2}}(\hat{\mu}_n - \mu) \xrightarrow{d} \mathcal{N}(0, 1)$. However, if $\mu = 0$, then $\overline{Z_n}$ will be within $\pm n^{-\frac{1}{4}}$ of the origin with increasing probability (since $\overline{Z_n} \sim \mathcal{N}(0, \frac{1}{n})$, so $P(|\overline{Z_n}| < n^{-\frac{1}{4}}) = P(|\sqrt{n}\overline{Z_n}| < n^{\frac{1}{4}})$, where $\sqrt{n}\overline{Z_n}$ is a standard normal). This implies that $P(\hat{\mu}_n = 0) \to 1$ if $\mu = 0$. Thus, the asymptotic variance of $n^{\frac{1}{2}}(\hat{\mu}_n - \mu)$ is 1 for $\mu \neq 0$ and 0 for $\mu = 0$, implying Hodges' estimator is super-efficient.

However, Hodges' estimator possesses local properties that are not favorable. If we consider the sequence $\mu_n = n^{-\frac{1}{3}}$, then $\overline{Z_n}$ concentrates in an $O(n^{-\frac{1}{2}})$ neighborhood about μ_n (by a similar argument as before), and since μ_n is contained in $[-n^{-\frac{1}{4}}, n^{-\frac{1}{4}}]$ with probability 1 as n increases, we have, as n increases,

$$P_{\mu_n}\left[n^{\frac{1}{2}}(\hat{\mu}_n - \mu_n) = -n^{\frac{1}{2}}\mu_n\right] \to 1,$$

since $P_{\mu_n}(\hat{\mu}_n = 0) \to 1$. *However, for* $\mu_n = n^{-\frac{1}{3}}$, $-n^{\frac{1}{2}}\mu_n \to -\infty$, *which implies* $n^{\frac{1}{2}}(\hat{\mu}_n - \mu_n) \to -\infty$, *which is certainly undesirable.*

Regularity rules out estimators like Hodges' estimator. In particular, we can show that Hodges' estimator is not regular by considering $\mu_n = \frac{t}{\sqrt{n}}$ *(for fixed* t*) and* $\mu^* = 0$. *Then,* $n^{\frac{1}{2}}(\mu_n - \mu^*) = t$. *However,* $n^{\frac{1}{2}}(\hat{\mu}_n - \mu_n) \xrightarrow{d} -t$ *by a similar argument as before, which is clearly dependent on the LDGP. Hence, Hodges' estimator is not regular.*

Remark 3.1.7 *It turns out that for the theory we will develop, if we restrict ourselves to regular estimators, we can restrict ourselves to regular and asymptotically linear (RAL) estimators without losing anything. A regular estimator is not necessarily asymptotically linear, but the Hájek-Le Cam convolution theorem (Theorem 2.3.1 of (Bickel et al., 1993)) states the following:*

Let T_n *be a regular estimator of a parameter* $q(\theta) : \Theta \to \mathbb{R}^m$, *where we work in a parametric model. Assume that* q *is differentiable in* θ, *and let* $I^{-1}_{q(\theta)} = \dot{q(\theta)} I(\theta)^{-1} \dot{q(\theta)}^T$ *be the information bound for* q *and* $\psi_{q(\theta)}$ *be the EIF for* q. *Then:*

1. *There exist independent random vectors* $Z_\theta \sim \mathcal{N}(0, I^{-1}_{q(\theta)}), \Delta_\theta$ *such that*

$$\begin{pmatrix} \sqrt{n}(T_n - q(\theta)) - \frac{1}{\sqrt{n}} \sum_{i=1}^n \psi_{q(\theta)}(x_i) \\ \frac{1}{\sqrt{n}} \sum_{i=1}^n \psi_{q(\theta)}(x_i) \end{pmatrix} \xrightarrow{d} \begin{pmatrix} \Delta_\theta \\ Z_\theta \end{pmatrix}$$

2. *If* $\theta \mapsto \dot{q_\theta}$ *is continuous, then the previous convergence holds uniformly on compact subsets of* Θ. *Also, we have* $\Delta_\theta = 0$ *for all* θ *if and only if* T_n *is uniformly asymptotically linear with influence function* $\psi_{q(\theta)}$.

It is clearly a consequence of this theorem that the most efficient regular estimator is asymptotically linear, so we can restrict our attention to RAL estimators.

In what follows, define the score vectors

$$S_\theta(z, \theta_0) = \left.\frac{\partial \log p_Z(z,\theta)}{\partial \theta}\right|_{\theta=\theta_0}, S_\beta(z, \theta_0) = \left.\frac{\partial \log p_Z(z,\theta)}{\partial \beta}\right|_{\theta=\theta_0}, S_\eta(z, \theta_0) = \left.\frac{\partial \log p_Z(z,\theta)}{\partial \eta}\right|_{\theta=\theta_0},$$

so we can write $S_\theta(Z, \theta_0) = (S_\beta(Z, \theta_0), S_\eta(Z, \theta_0))^T$.

Theorem 3.1.8 *Let* $\beta(\theta)$ *be a* q*-dimensional function of the* p*-dimensional parameter* θ *with* $q < p$, *and let* $\Gamma(\theta) = \frac{\partial \beta(\theta)}{\partial \theta^T}$ *have rank* q *and be continuous around* θ_0. *Let* $\hat{\beta}_n$ *be an asymptotically linear estimator with influence function* $\varphi(Z)$ *such that* $\mathbb{E}_\theta[\varphi^T \varphi]$ *exists and is continuous around* θ_0. *If* $\hat{\beta}_n$ *is regular, we then have*

$$\mathbb{E}[\varphi(Z) S_\theta^T(Z, \theta_0)] = \Gamma(\theta_0).$$

If $\theta = (\beta^T, \eta^T)^T$ *as in parametric models, we then have*

$$\mathbb{E}[\varphi(Z) S_\beta^T(Z, \theta_0)] = I, \mathbb{E}[\varphi(Z) S_\eta^T(Z, \theta_0)] = 0.$$

Proof: Let $p_{0n}(v_n) = \prod p(z_{in}, \theta_0)$, and consider the LDGP $p_{1n}(v_n) = \prod p(z_{in}, \theta_n)$ with $n^{\frac{1}{2}}(\theta_n - \theta_0) \to \tau$, for some constant vector τ. Since $\hat{\beta}_n$ is asymptotically linear with influence function $\varphi(Z)$, we can write

$$n^{\frac{1}{2}}(\hat{\beta}_n - \beta(\theta_0)) = n^{-\frac{1}{2}} \sum_{i=1}^{n} \varphi(Z_{in}) + o_{P_{0n}}(1).$$

Lemma 3.1.9 *Under nice smoothness and regularity conditions, the sequence P_{1n} is contiguous to the sequence P_{0n}, meaning that, for any sequence of events A_n, $P_{0n}(A_n) \xrightarrow{n\to\infty} 0 \implies P_{1n}(A_n) \xrightarrow{n\to\infty} 0$.*

Proof: We recall Le Cam's third lemma, which states that if

$$\log\left(\frac{p_{1n}(V_n)}{p_{0n}(V_n)}\right) \xrightarrow{\mathcal{D}(P_{0n})} \mathcal{N}\left(-\frac{\sigma^2}{2}, \sigma^2\right),$$

then P_{1n} is contiguous to P_{0n}. Hence, it suffices to show convergence of the log-likelihood ratio. To this end, let

$$L_n(V_n) = \frac{p_{1n}(V_n)}{p_{0n}(V_n)} = \prod_{i=1}^{n} \frac{p(Z_{in}, \theta_n)}{p(Z_{in}, \theta_0)}.$$

Taking logs, we get

$$\log(L_n(V_n)) = \sum_{i=1}^{n} (\log p(Z_{in}, \theta_n) - \log p(Z_{in}, \theta_0)).$$

Then, performing a second-order Taylor series expansion around θ_0, for some intermediate value θ_n^*, we have

$$\log(L_n(V_n)) = (\theta_n - \theta_0)^T \sum_{i=1}^{n} S_\theta(Z_{in}, \theta_0) + \frac{1}{2}(\theta_n - \theta_0)^T \left(\sum_{i=1}^{n} S_{\theta\theta}(Z_{in}, \theta_n^*)\right)(\theta_n - \theta_0).$$

Inserting factors of $n^{\frac{1}{2}}$ in the appropriate places to help us analyze convergence rates, we can write

$$\log(L_n(V_n)) = n^{\frac{1}{2}}(\theta_n - \theta_0)^T \left(n^{-\frac{1}{2}} \sum_{i=1}^{n} S_\theta(Z_{in}, \theta_0)\right) + \frac{1}{2}(n^{\frac{1}{2}}(\theta_n - \theta_0)^T) \left(n^{-1} \sum_{i=1}^{n} S_{\theta\theta}(Z_{in}, \theta_n^*)\right)(n^{\frac{1}{2}}(\theta_n - \theta_0)).$$

We can now analyze convergence of each of these terms. By the CLT,

$$n^{-\frac{1}{2}} \sum_{i=1}^{n} S_\theta(Z_{in}, \theta_0) \xrightarrow{\mathcal{D}(P_{0n})} \mathcal{N}(0, I(\theta_0)),$$

where $I(\theta_0)$ is the Fisher information. Furthermore, $\theta_n^* \to \theta_0$, and $S_{\theta\theta}(Z_{in}, \theta_0)$ for $i = 1, \ldots, n$ are i.i.d. with mean $-I(\theta_0)$, we can write

$$n^{-1} \sum_{i=1}^{n} S_{\theta\theta}(Z_{in}, \theta_n^*) \xrightarrow{p} -I(\theta_0).$$

Finally, since $n^{\frac{1}{2}}(\theta_n - \theta_0) \to \tau$, we can use Slutsky's theorem to get that

$$\log(L_n(V_n)) \xrightarrow{\mathcal{D}(P_{0n})} \mathcal{N}\left(-\frac{\tau^T I(\theta_0)\tau}{2}, \tau^T I(\theta_0)\tau\right),$$

so we can apply Le Cam's third lemma to finish. □

By the lemma, since P_{1n} is contiguous to P_{0n}, $o_{P_{0n}}(1) \implies o_{P_{1n}}(1)$. Hence, we can write

$$n^{\frac{1}{2}}(\hat{\beta}_n - \beta(\theta_0)) = n^{-\frac{1}{2}} \sum_{i=1}^{n} \varphi(Z_{in}) + o_{P_{1n}}(1).$$

Hence, we can write

$$n^{\frac{1}{2}}(\hat{\beta}_n - \beta(\theta_n)) = n^{-\frac{1}{2}} \sum_{i=1}^{n} [\varphi(Z_{in}) - \mathbb{E}_{\theta_n}[\varphi(Z)]] + n^{\frac{1}{2}} \mathbb{E}_{\theta_n}[\varphi(Z)] - n^{\frac{1}{2}}(\beta(\theta_n) - \beta(\theta_0)) + o_{P_{1n}}(1).$$

By regularity, since the limiting distribution of $n^{\frac{1}{2}}(\hat{\beta}_n - \beta(\theta_n))$ is independent of the LDGP, we can write

$$n^{\frac{1}{2}}(\hat{\beta}_n - \beta(\theta_n)) \xrightarrow{\mathcal{D}(P_{1n})} \mathcal{N}(0, \mathbb{E}_{\theta_0}[\varphi\varphi^T])$$

(since $n^{\frac{1}{2}}(\hat{\beta}_n - \beta_0) \xrightarrow{\mathcal{D}(P_{0n})} \mathcal{N}(0, \mathbb{E}_{\theta_0}[\varphi\varphi^T])$).

Under P_{1n}, $\varphi(Z_{in}) - \mathbb{E}_{\theta_n}[\varphi(Z)]$ are mean-zero with variance $\mathbb{E}_{\theta_n}[\varphi\varphi^T] - \mathbb{E}_{\theta_n}[\varphi]\mathbb{E}_{\theta_n}[\varphi^T]$. By smoothness, this converges to $\mathbb{E}_{\theta_0}[\varphi\varphi^T]$, so by the CLT, we can write

$$n^{-\frac{1}{2}} \sum_{i=1}^{n} [\varphi(Z_{in}) - \mathbb{E}_{\theta_n}[\varphi(Z)]] \xrightarrow{\mathcal{D}(P_{1n})} \mathcal{N}(0, \mathbb{E}_{\theta_0}[\varphi\varphi^T]).$$

Analyzing the $n^{\frac{1}{2}}(\beta(\theta_n) - \beta(\theta_0))$ term, via a Taylor expansion around $\beta(\theta_0)$, we can write

$$\beta(\theta_n) \approx \beta(\theta_0) + \Gamma(\theta_0)(\theta_n - \theta_0)$$

for $\Gamma(\theta_0) = \frac{\partial \beta(\theta_0)}{\partial \theta^T}$. Hence, we have

$$n^{\frac{1}{2}}(\beta(\theta_n) - \beta(\theta_0)) \to \Gamma(\theta_0)\tau.$$

Finally, to analyze the $n^{\frac{1}{2}} \mathbb{E}_{\theta_n}[\varphi(Z)]$ term, we can write

$$n^{\frac{1}{2}} \mathbb{E}_{\theta_n}[\varphi(Z)] = n^{\frac{1}{2}} \int \varphi(z) p(z, \theta_n) d\nu(z).$$

Performing a first-order Taylor expansion of $p(z, \theta_n)$ around θ_0, we get

$$n^{\frac{1}{2}} \int \varphi(z) p(z, \theta_n) d\nu(z) = n^{\frac{1}{2}} \int \varphi(z) p(z, \theta_0) d\nu(z) + n^{\frac{1}{2}} \int \varphi(z) \left(\frac{\partial p(z, \theta_n^*)}{\partial \theta} \right)^T (\theta_n - \theta_0) d\nu(z)$$

for an intermediate value θ_n^*. As $n \to \infty$, the first term goes to 0 (as $\mathbb{E}_{\theta_0} \varphi(Z) = 0$) and the second term satisfies

$$n^{\frac{1}{2}} \int \varphi(z) \left(\frac{\partial p(z, \theta_n^*)}{\partial \theta} \right)^T (\theta_n - \theta_0) d\nu(z) \to \int \varphi(z) \left(\frac{\frac{\partial p(z, \theta_0)}{\partial \theta}}{p(z, \theta_0)} \right)^T p(z, \theta_0) d\nu(z) \tau,$$

since $\sqrt{n}(\theta_n - \theta_0) \to \tau$. This is equal to $\mathbb{E}_{\theta_0}[\varphi(Z) S_\theta^T(Z, \theta_0)]\tau$.

For regularity to hold, we require

$$\lim_{n \to \infty} n^{\frac{1}{2}} \mathbb{E}_{\theta_n}[\varphi(Z)] - n^{\frac{1}{2}}(\beta(\theta_n) - \beta(\theta_0)) = 0,$$

which implies that we need

$$(\mathbb{E}_{\theta_0}[\varphi(Z)S_\theta^T(Z,\theta_0)] - \Gamma(\theta_0))\tau = 0,$$

but since τ was an arbitrary constant, this holds for any τ, so we must have

$$\mathbb{E}_{\theta_0}[\varphi(Z)S_\theta^T(Z,\theta_0)] = \Gamma(\theta_0),$$

as desired. □

3.2 Influence Functions in Parametric Models

Let $\mathcal{H}$ be the Hilbert space of q-dimensional measurable functions of Z with mean zero and finite variance, with the standard inner product $\langle h_1, h_2\rangle$. Recall the standard result that the score vector $S_\theta(Z,\theta_0)$ has mean zero.

Definition 3.2.1 *The tangent space is defined as the set*

$$\mathcal{T} = \{BS_\theta(Z,\theta_0) : B \in \mathbb{R}^{q\times p}\}.$$

Definition 3.2.2 *If $\theta = (\beta^T, \eta^T)^T$, the nuisance tangent space is given by $\mathcal{T}_{nuis} = \{BS_\eta(Z,\theta_0) : B \in \mathbb{R}^{q\times r}\}$.*

Note that the final condition of Theorem 3.1.8 is equivalent to φ being orthogonal to the nuisance tangent space, or, equivalently, that $\varphi \in \mathcal{T}_{\text{nuis}}^{\perp}$.

Theorem 3.2.3 *A mean-zero, finite-variance function φ such that*

$$\mathbb{E}[\varphi(Z)S_\beta^T(Z,\theta_0)] = I, \mathbb{E}[\varphi(Z)S_\eta^T(Z,\theta_0)] = 0$$

is the influence function of some RAL estimator, given that there exists an estimator $\hat{\eta_n}$ for η_0 such that $\sqrt{n}(\hat{\eta_n} - \eta_0)$ is bounded in probability (e.g. the MLE for η, fixing β).

Proof: Let $m(Z,\beta,\eta) = \varphi(Z) - \mathbb{E}_{\beta,\eta}[\varphi(Z)]$, and define $\hat{\beta}_n$ be such that

$$\sum_{i=1}^{n} m(Z_i, \hat{\beta}_n, \hat{\eta_n}(\hat{\beta}_n)) = 0.$$

We will show that $\hat{\beta}_n$ is an asymptotically linear estimator of β_0 with influence function $\varphi(Z)$, which will finish.

Note that

$$\mathbb{E}_{\beta_0,\eta}[m(Z,\beta_0,\eta)] = \int m(z,\beta_0,\eta)p(z,\beta_0,\eta)d\nu(z) = 0,$$

by construction, since $\mathbb{E}_{\beta_0,\eta}[m(Z,\beta_0,\eta)] = \mathbb{E}_{\beta_0,\eta}[\varphi(Z)] - \mathbb{E}_{\beta_0,\eta}[\varphi(Z)] = 0$. Differentiating the integral form with respect to η^T, we get

$$\frac{\partial}{\partial \eta^T}\int m(z,\beta_0,\eta)p(z,\beta_0,\eta)d\nu(z)\Big|_{\eta=\eta_0} = 0.$$

Assuming appropriate regularity conditions so we can switch the integral and derivative, we can rewrite this as

$$\int \frac{\partial m(z,\beta_0,\eta_0)}{\partial \eta^T} p(z,\beta_0,\eta_0) d\nu(z) + \int m(z,\beta_0,\eta_0) S_\eta^T(z,\beta_0,\eta_0) p(z,\beta_0,\eta_0) d\nu(z) = 0,$$

where we recall that $\frac{\partial p(z,\beta_0,\eta_0)}{\partial \eta^T}|_{\eta=\eta_0} = S_\eta^T(z,\beta_0,\eta_0)p(z,\beta_0,\eta_0)$.

Since $\varphi = m(Z,\beta_0,\eta_0)$ and $\mathbb{E}[\varphi(Z)S_\eta^T(Z,\theta_0)] = 0$, we have

$$\int m(z,\beta_0,\eta_0) S_\eta^T(z,\beta_0,\eta_0) p(z,\beta_0,\eta_0) d\nu(z) = 0,$$

so we get

$$\int \frac{\partial m(z,\beta_0,\eta_0)}{\partial \eta^T} p(z,\beta_0,\eta_0) d\nu(z) = \mathbb{E}\left[\frac{\partial}{\partial \eta^T} m(Z,\beta_0,\eta_0)\right] = 0.$$

Similarly, we have

$$\left.\frac{\partial}{\partial \beta^T}\right|_{\beta=\beta_0} \int m(z,\beta,\eta_0) p(z,\beta,\eta_0) d\nu(z) = 0,$$

so

$$\int \frac{\partial m(z,\beta_0,\eta_0)}{\partial \beta^T} p(z,\beta_0,\eta_0) d\nu(z) + \int m(z,\beta_0,\eta_0) S_\beta^T(z,\beta_0,\eta_0) p(z,\beta_0,\eta_0) d\nu(z) = 0,$$

and by the condition $\mathbb{E}[\varphi(Z)S_\beta^T(Z,\theta_0)] = I$, we can write

$$\mathbb{E}\left[\frac{\partial}{\partial \beta^T} m(Z,\beta_0,\eta_0)\right] = -I.$$

Returning to the defining equation for $\hat{\beta}_n$, performing a first-order Taylor expansion around β_0 gives that, for some value β_n^* between $\beta_0, \hat{\beta}_n$,

$$0 = \sum_{i=1}^n m(Z_i,\hat{\beta}_n,\hat{\eta_n}(\hat{\beta}_n)) = \sum_{i=1}^n m(Z_i,\beta_0,\hat{\eta_n}(\hat{\beta}_n)) + \left[\sum_{i=1}^n \frac{\partial m}{\partial \beta^T}(Z_i,\beta_n^*,\hat{\eta_n}(\hat{\beta}_n))\right](\hat{\beta}_n - \beta_0).$$

Thus, isolating $\hat{\beta}_n - \beta_0$, we get

$$n^{\frac{1}{2}}(\hat{\beta}_n - \beta_0) = -\left[n^{-1}\sum_{i=1}^n \frac{\partial}{\partial \beta^T} m(Z_i,\beta_n^*,\hat{\eta_n}(\hat{\beta}_n))\right]^{-1}\left[n^{-\frac{1}{2}}\sum_{i=1}^n m(Z_i,\beta_0,\hat{\eta_n}(\hat{\beta}_n))\right].$$

We have

$$n^{-1}\sum_{i=1}^n \frac{\partial}{\partial \beta^T} m(Z_i,\beta_n^*,\hat{\eta_n}(\hat{\beta}_n)) \xrightarrow{p} \mathbb{E}\left[\frac{\partial}{\partial \beta^T} m(Z,\beta_0,\eta_0)\right]$$

by the LLN - furthermore, this implies that

$$\left[n^{-1}\sum_{i=1}^n \frac{\partial}{\partial \beta^T} m(Z_i,\beta_n^*,\hat{\eta_n}(\hat{\beta}_n))\right]^{-1} \xrightarrow{p} \left[\mathbb{E}\left[\frac{\partial}{\partial \beta^T} m(Z,\beta_0,\eta_0)\right]\right]^{-1} = -I.$$

For the second term, expanding around η_0, for some intermediate value η_n^*, we get

$$n^{-\frac{1}{2}}\sum_{i=1}^{n} m(Z_i,\beta_0,\hat{\eta_n}(\hat{\beta}_n)) = n^{-\frac{1}{2}}\sum_{i=1}^{n} m(Z_i,\beta_0,\eta_0) + \left[n^{-1}\sum_{i=1}^{n}\frac{\partial m(Z_i,\beta_0,\eta_n^*)}{\partial \eta^T}\right]\left[n^{-\frac{1}{2}}(\hat{\eta_n}(\hat{\beta}_n)-\eta_0)\right].$$

Again, we have

$$n^{-1}\sum_{i=1}^{n}\frac{\partial m(Z_i,\beta_0,\eta_n^*)}{\partial \eta^T} \xrightarrow{p} \mathbb{E}\left[\frac{\partial}{\partial \eta^T} m(Z,\beta_0,\eta_0)\right] = 0,$$

and by assumption, $n^{\frac{1}{2}}(\hat{\eta_n}(\hat{\beta}_n)-\eta_0)$ is bounded in probability.

Putting everything together, we get

$$n^{\frac{1}{2}}(\hat{\beta}_n-\beta_0) = n^{-\frac{1}{2}}\sum_{i=1}^{n} m(Z_i,\beta_0,\eta_0) + o_P(1) = n^{-\frac{1}{2}}\sum_{i=1}^{n}\varphi(Z_i) + o_P(1),$$

which finishes. □

Characterizing the geometry of influence functions in this manner is quite nice because, as seen earlier, the variance of the influence function determines the asymptotic variance of a RAL estimator for β. However, the variance of an influence function is the distance to the origin in a subspace of a Hilbert space, so searching for the estimator with the lowest asymptotic variance is equivalent to searching for the influence function with the shortest distance to the origin.

To actually identify the EIF, we can note that, to identify elements of $\mathcal{T}$ orthogonal to $\mathcal{T}_{\text{nuis}}$, we can consider the set $h - \Pi(h|\mathcal{T}_{\text{nuis}})$ for $h \in \mathcal{H}$, where Π is the projection operator.

Theorem 3.2.4 *We have, for $h \in \mathcal{H}$,*

$$\Pi(h|\mathcal{T}_{nuis}) = \mathbb{E}[hS_\eta^T](\mathbb{E}[S_\eta S_\eta^T])^{-1}S_\eta(Z,\theta_0).$$

We also note that, letting $\mathcal{T}_\beta = \{BS_\beta(Z,\theta_0) : B \in \mathbb{R}^{q\times q}\}$, we can write $\mathcal{T} = \mathcal{T}_\beta \oplus \mathcal{T}_{\text{nuis}}$.

As a side note, we will use the Pythagorean theorem in what follows - however, for dimensions greater than 1, the Pythagorean theorem for variances does not necessarily hold for orthogonal elements of $\mathcal{H}$. However, it does hold in the following special case, which is enough for our purposes:

Definition 3.2.5 *A linear subspace $\mathcal{U} \subset \mathcal{H}$ is a q-replicating linear space if $\mathcal{U} = \mathcal{U}^{(1)} \times \cdots \times \mathcal{U}^{(1)}$ (q copies), where $\mathcal{U}^{(1)}$ is a linear subspace in the 1-dimensional Hilbert space $\mathcal{H}^{(1)}$.*

Note that the linear subspace $\{Bv(Z) : B \in \mathbb{R}^{q\times r}\}$ is q-replicating if v is mean-zero finite variance, which we can see by defining $\mathcal{U}^{(1)} = \{b^T v(Z) : b \in \mathbb{R}^{r\times 1}\}$ - this is the structure of tangent and nuisance tangent spaces, so we can feel free to use the Pythagorean theorem here:

Theorem 3.2.6 *If $h \in \mathcal{U}$ and $\mathcal{U} \subset \mathcal{H}$ is q-replicating, and $l \in \mathcal{H}$ with $l \perp \mathcal{U}$, then $Var(l+h) = Var(l) + Var(h)$.*

Proof: Note that l is orthogonal to $\mathcal{U}$ if and only if each component is orthogonal to $\mathcal{U}^{(1)}$. Hence, if $l = (l_1,\ldots,l_q)^T$, then each l_j is orthogonal to $\mathcal{U}^{(1)}$. Letting $h = (h_1,\ldots,h_q)^T$, since each $h_k \in \mathcal{U}^{(1)}$, we have

$\mathbb{E}[l_j h_k] = 0$ for all j, k. Hence, we have that $\mathbb{E}[lh^T] = \mathbb{E}[hl^T] = 0$, so

$$\text{Var}(l+h) = \mathbb{E}[(l+h)(l+h)^T] = \mathbb{E}[ll^T] + \mathbb{E}[lh^T] + \mathbb{E}[hl^T] + \mathbb{E}[hh^T] = \mathbb{E}[ll^T] + \mathbb{E}[hh^T] = \text{Var}(l) + \text{Var}(h).$$

□

Theorem 3.2.7 *If $\varphi^*(Z)$ is an arbitrary influence function, then the set of all influence functions is given by the linear variety $\varphi^*(Z) + \mathcal{T}^\perp$.*

Proof: Let $\varphi(Z) = \varphi^*(Z) + l(Z)$, for $l \in \mathcal{T}^\perp$ and φ^* an arbitrary function. Then, we can write

$$\mathbb{E}[\varphi(Z) S_\theta^T(Z, \theta_0)] = \mathbb{E}[\varphi^*(Z) S_\theta^T(Z, \theta_0)] + \mathbb{E}[l(Z) S_\theta^T(Z, \theta_0)] = \Gamma(\theta_0).$$

On the other hand, if $\mathbb{E}[\varphi(Z) S_\theta^T(Z, \theta_0)] = \Gamma(\theta_0)$, then we can write $\varphi(Z) = \varphi^*(Z) + (\varphi(Z) - \varphi^*(Z))$, where $\varphi - \varphi^* \in \mathcal{T}^\perp$. □

Theorem 3.2.8 *The EIF is given by*

$$\varphi_{\mathit{eff}}(Z) = \varphi^*(Z) - \Pi(\varphi^*(Z) | \mathcal{T}^\perp) = \Pi(\varphi^*(Z) | \mathcal{T})$$

for φ^* an arbitrary influence function. More explicitly, we can write

$$\varphi_{\mathit{eff}}(Z) = \Gamma(\theta_0) I(\theta_0)^{-1} S_\theta(Z, \theta_0),$$

for Fisher information $I(\theta_0) = \mathbb{E}[S_\theta(Z, \theta_0) S_\theta^T(Z, \theta_0)]$.

Proof: By Theorem 3.2.7, since $\Pi(\varphi^* | \mathcal{T}^\perp) \in \mathcal{T}^\perp$, $\varphi_{\text{eff}} = \varphi^* - \Pi(\varphi^* | \mathcal{T}^\perp)$ is an influence function. Furthermore, by construction, φ_{eff} is orthogonal to $\mathcal{T}^\perp$, so we can write any other influence function φ as $\varphi = \varphi_{\text{eff}} + l$, for $l \in \mathcal{T}^\perp$. By the Pythagorean Theorem, we can write $\text{Var}(\varphi) = \text{Var}(\varphi_{\text{eff}}) + \text{Var}(l)$, which shows that $\text{Var}(\varphi_{\text{eff}}) \leq \text{Var}(\varphi)$ for any other influence function φ.

Since φ_{eff} is a projection onto $\mathcal{T}$, it can be written as $\varphi_{\text{eff}}(Z) = B_{\text{eff}} S_\theta(Z, \theta_0)$, which implies that

$$\mathbb{E}[\varphi_{\text{eff}}(Z) S_\theta^T(Z, \theta_0)] = \Gamma(\theta_0) \implies B_{\text{eff}} \mathbb{E}[S_\theta(Z, \theta_0) S_\theta^T(Z, \theta_0)] = \Gamma(\theta_0),$$

which implies the exact form of the EIF in the second part of the theorem. □

In the case where $\theta = (\beta^T, \eta^T)^T$, we can write the EIF alternatively in terms of the *efficient score*:

Definition 3.2.9 *The efficient score is written as*

$$S_{\mathit{eff}}(Z, \theta_0) = S_\beta(Z, \theta_0) - \Pi(S_\beta(Z, \theta_0) | \mathcal{T}_{\mathit{nuis}}).$$

Corollary 3.2.9.1 *If $\theta = (\beta^T, \eta^T)^T$, then we can write*

$$\varphi_{\mathit{eff}}(Z, \theta_0) = (\mathbb{E}[S_{\mathit{eff}} S_{\mathit{eff}}^T])^{-1} S_{\mathit{eff}}(Z, \theta_0).$$

Proof: Note that the efficient score vector is orthogonal to $\mathcal{T}_{\text{nuis}}$ by construction. This then implies

$$\mathbb{E}[S_{\text{eff}}(Z,\theta_0)S_\beta^T(Z,\theta_0)] = \mathbb{E}[S_{\text{eff}}(Z,\theta_0)S_{\text{eff}}^T(Z,\theta_0)] + \mathbb{E}[S_{\text{eff}}(Z,\theta_0)\Pi(S_\beta|\mathcal{T}_{\text{nuis}})^T],$$

and the second term is 0 by orthogonality. Hence,

$$\mathbb{E}[S_{\text{eff}}(Z,\theta_0)S_\beta^T(Z,\theta_0)] = \mathbb{E}[S_{\text{eff}}(Z,\theta_0)S_{\text{eff}}^T(Z,\theta_0)],$$

which implies that $\varphi_{\text{eff}}(Z,\theta_0) = (\mathbb{E}[S_{\text{eff}}S_{\text{eff}}^T])^{-1}S_{\text{eff}}(Z,\theta_0)$ satisfies the two conditions required to be an influence function.

To show efficiency, recall that the EIF is the unique influence function in $\mathcal{T}$, but since $S_\beta, \Pi(S_\beta|\mathcal{T}_{\text{nuis}})$ are both in $\mathcal{T}$, so is the efficient score, so we are done. □

Remark 3.2.10 *Note that the variance of φ_{eff}, using the efficient score representation, is $(\mathbb{E}[S_{eff}S_{eff}^T])^{-1}$. We will use this fact in what follows.*

3.3 Semiparametric Influence Functions

Having developed this theory for parametric models, we now move to semiparametric models, where η can be infinite-dimensional. We call our semiparametric model $\mathcal{P}$. The notion of a parametric submodel from the previous section remains useful. In particular, an estimator for β is RAL for a semiparametric model if it is RAL for every parametric submodel, which implies any influence function in the semiparametric model must be an influence function within a parametric submodel. In what follows, let $\mathcal{P}_{\beta,\gamma}$ be a parametric submodel of $\mathcal{P}$ indexed by γ. This gives us a few corollaries:

1. An influence function of an RAL semiparametric estimator for β is orthogonal to all parametric submodel nuisance tangent spaces.

2. For all parametric submodels $\mathcal{P}_{\beta,\gamma}$, the variance of the RAL semiparametric influence function must be greater than or equal to the parametric submodel lower bound $(\mathbb{E}[S_{\beta,\gamma}^{\text{eff}}(Z)S_{\beta,\gamma}^{\text{eff}^T}(Z)])^{-1}$.

This motivates us to define the following:

Theorem 3.3.1 *The semiparametric efficiency bound is the supremum, over all parametric submodels, of $(\mathbb{E}[S_{\beta,\gamma}^{eff}(Z)S_{\beta,\gamma}^{eff^T}(Z)])^{-1}$.*

Definition 3.3.2 *The nuisance tangent space for a semiparametric model, which we also denote $\mathcal{T}_{nuis}$, is the mean-square closure of all parametric submodel nuisance tangent spaces $\{BS_\gamma(Z,\beta_0,\eta_0) : B \in \mathbb{R}^{q\times r}\}$, where the mean-square closure is defined as the space of functions $h \in \mathcal{H}$ with $\mathbb{E}[h(Z)^T h(Z)] < \infty$ and for which there exists a sequence $B_j S_{\gamma j}(Z)$ such that $||h(Z) - B_j S_{\gamma j}(Z)||^2 \to 0$ as $j \to \infty$.*

Remark 3.3.3 *It is not necessarily generally true that $\mathcal{T}_{nuis}$, as defined above, is linear. However, for most applications, it turns out to be, and it is safe to assume $\mathcal{T}_{nuis}$ is a closed linear subspace of $\mathcal{H}$ for future results.*

With this new definition of nuisance tangent space in hand, we can define the semiparametric efficient score for β in a similar way, as

$$S_{\text{eff}}(Z, \beta_0, \eta_0) = S_\beta(Z, \beta_0, \eta_0) - \Pi(S_\beta(Z, \beta_0, \eta_0)|\mathcal{T}_{\text{nuis}}).$$

Theorem 3.3.4 *The semiparametric bound can be written as* $(\mathbb{E}[S_{\mathit{eff}}(Z)S_{\mathit{eff}}(Z)^T])^{-1}$.

Proof: For simplicity, let β be a scalar, although the argument can be easily extended to higher-dimensional β. In this case, the semiparametric efficiency bound $\mathcal{V}$ can be written as

$$\mathcal{V} = \sup_{\mathcal{P}_{\beta,\gamma}} ||S^{\text{eff}}_{\beta,\gamma}(Z)||^{-2},$$

where

$$S^{\text{eff}}_{\beta,\gamma}(Z) = S_\beta(Z) - \Pi(S_\beta(Z)|\mathcal{T}^{\gamma}_{\text{nuis}}).$$

Since $\mathcal{T}^{\text{nuis}}_{\gamma} \subset \mathcal{T}_{\text{nuis}}$, we have, for all parametric submodels $\mathcal{P}_{\beta,\gamma}$, $||S_{\text{eff}}(Z)|| \leq ||S^{\text{eff}}_{\beta,\gamma}(Z)||$. This implies

$$\mathcal{V} \leq ||S_{\text{eff}}(Z)||^{-2}.$$

To show the inequality in the other direction, by definition, since $\Pi(S_\beta(Z)|\mathcal{T}_{\text{nuis}}) \in \mathcal{T}_{\text{nuis}}$, there exists a sequence of parametric submodels $\mathcal{P}_{\beta,\gamma_j}$ and nuisance score vectors $S_{\gamma_j}(Z)$ such that

$$||\Pi(S_\beta(Z)|\mathcal{T}_{\text{nuis}}) - B_j S_{\gamma j}(Z)||^2 \xrightarrow{j\to\infty} 0$$

for matrices B_j. We can write:

$$\begin{aligned}
\mathcal{V}^{-1} &\leq ||S^{\text{eff}}_{\beta,\gamma_j}(Z)||^2 \\
&= ||S_\beta(Z) - \Pi(S_\beta(Z)|\mathcal{T}^{\text{nuis}}_{\gamma_j})||^2 \\
&\leq ||S_\beta(Z) - B_j S_{\gamma_j}(Z)||^2 \\
&= ||S_\beta(Z) - \Pi(S_\beta(Z)|\mathcal{T}_{\text{nuis}})||^2 + ||\Pi(S_\beta(Z)|\mathcal{T}_{\text{nuis}}) - B_j S_{\gamma_j}(Z)||^2,
\end{aligned}$$

where the last inequality is due to the Pythagorean Theorem. Taking $j \to \infty$, we get $\mathcal{V}^{-1} \leq ||S_{\text{eff}}(Z)||^2$, which finishes. □

Definition 3.3.5 *The efficient influence function is the influence function of a semiparametric RAL estimator, if it exists.*

Theorem 3.3.6 *A semiparametric RAL estimator for β has an influence function $\varphi(Z)$ which satisfies*

$$\mathbb{E}[\varphi(Z)S^T_\beta(Z, \beta_0, \eta_0)] = \mathbb{E}[\varphi(Z)S^T_{\mathit{eff}}(Z, \beta_0, \eta_0)] = I, \Pi(\varphi(Z)|\mathcal{T}_{\mathit{nuis}}) = 0.$$

The EIF is the unique element of the tangent space satisfying these two conditions with variance matrix

equaling the efficiency bound. We can write

$$\varphi_{eff}(Z, \beta_0, \eta_0) = (\mathbb{E}[S_{eff} S_{eff}^T])^{-1} S_{eff}(Z, \beta_0, \eta_0).$$

Proof: We first show $\varphi(Z)$ is orthogonal to $\mathcal{T}_{\text{nuis}}$. For any $h \in \mathcal{T}_{\text{nuis}}$, by definition, there exists a sequence $B_j S_{\gamma j}(Z)$ such that

$$||h(Z) - B_j S_{\gamma j}(Z)|| \xrightarrow{j \to \infty} 0.$$

We can write

$$\langle \varphi, h \rangle = \langle \varphi, B_j S_{\gamma j} \rangle + \langle \varphi, h - B_j S_{\gamma j} \rangle.$$

The first term in the sum is 0 since φ is also an influence function for an RAL estimator in the parametric submodel, so φ is orthogonal to $\mathcal{T}_{\gamma_j}^{\text{nuis}}$. Applying Cauchy-Schwarz to the second term, we get

$$|\langle \varphi, h \rangle| \leq ||\varphi|| ||h - B_j S_{\gamma j}|| \xrightarrow{j \to \infty} 0.$$

For the first condition, by the analogous theorem for the parametric model, we have

$$\mathbb{E}[\varphi(Z) S_\beta^T(Z, \beta_0, \eta_0)] = I.$$

We then have

$$\mathbb{E}[\varphi(Z) S_{\text{eff}}^T(Z, \beta_0, \eta_0)] = \mathbb{E}[\varphi(Z) S_\beta^T(Z, \beta_0, \eta_0)] - \mathbb{E}[\varphi(Z) \Pi(S_\beta(Z, \beta_0, \eta_0) | \mathcal{T}_{\text{nuis}})^T].$$

The second term in the difference is 0 as we just showed.

The rest of the proof follows similarly to the parametric case once these two conditions are proved. □

We also have the following theorem, with proof exactly the same as before:

Theorem 3.3.7 *If a semiparametric RAL estimator for β exists, then the influence function belongs to the linear variety $\{\varphi^*(Z) + \mathcal{T}^\perp\}$ for arbitrary influence function φ^* and tangent space $\mathcal{T}$ - this is also the space of influence functions. If an estimator exists that achieves the semiparametric efficiency bound, then the influence function is unique and an element of the tangent space:*

$$\varphi_{eff}(Z) = \varphi(Z) - \Pi(\varphi(Z) | \mathcal{T}^\perp) = \Pi(\varphi(Z) | \mathcal{T}).$$

Connecting back to the nonparametric case, we exhibit the following theorem:

Theorem 3.3.8 *The tangent space of a nonparametric model $\mathcal{P}$ (densities p such that $\int p(z) d\nu(z) = 1$ with respect to some dominating measure ν) is the entire Hilbert space $\mathcal{H}$.*

Proof: Let $\mathcal{P}_\theta$ be an arbitrary parametric submodel of $\mathcal{P}$. The parametric submodel tangent space is then

$$\mathcal{T}_\theta = \{B S_\theta(Z) : B \in \mathbb{R}^{q \times s}\},$$

where θ is s-dimensional and $S_\theta(Z)$ is the score vector. We know the score vector is mean-zero and is hence an element of $\mathcal{H}$ - thus, $\mathcal{T}_\theta \subset \mathcal{H}$.

It turns out that any element of $\mathcal{H}$ is an element of $\mathcal{T}_\theta$ for some θ or is a limit of elements of $\mathcal{T}_\theta$ for varying θ. To see this, choose an arbitrary $h \in \mathcal{H}$ that is bounded, mean-zero, and finite-variance, and consider the parametric submodel $p(z,\theta) = p(z,\theta_0)(1+\theta^T h(z))$ where θ is such that $1+\theta^T h(z) \geq 0$ for all z, so that p is a valid density. We can then write, for this parametric submodel:

$$\int p(z,\theta)d\nu(z) = \int p(z,\theta_0)(1+\theta^T h(z))d\nu(z) = \int p(z,\theta_0)d\nu(z) + \int \theta^T h(z)p(z,\theta_0)d\nu(z) = 1,$$

ensuring that in a neighborhood of θ_0, $p(z,\theta)$ is a proper density. As calculated earlier, the score vector is $h(z)$ as well, so choosing $B = I$, then $h(Z)$ is an element of the parametric submodel tangent space $\mathcal{T}_\theta$.

Hence, the tangent space contains all bounded mean-zero random vectors. However, any element of $\mathcal{H}$ is a limit of bounded h, which finishes. □

Note that this theorem tells us that, in the nonparametric case, the canonical gradient is, in fact, unique, a fact which we took for granted in the previous section. In particular, the orthogonal complement of the tangent space is the zero space, so there is only one possible influence function in the nonparametric model, which must be the EIF.

Remark 3.3.9 *Relating this to pathwise differentiability from the previous section, we recall that we stated the pathwise derivative φ in the expression $\frac{\partial}{\partial\varepsilon}\psi(P_\varepsilon)|_{\varepsilon=0} = \int \varphi(z;\mathbb{P})s_\varepsilon(z)d\mathbb{P}(z)$ was unique in the case of the nonparametric model. By the preceding theorem, the tangent space $\mathcal{T}$ of the nonparametric model is the entire Hilbert space $\mathcal{H}$, so $\mathcal{T}^\perp = \{0\}$ and there is a unique influence function, which must be the EIF.*

3.4 Example: Restricted Moments Model

As an example of all of this theory, we will analyze the *restricted moment model*, given by

$$Y = \mu(X,\beta) + \varepsilon, \mathbb{E}[\varepsilon|X] = 0,$$

where $\mu(x,\beta)$ is a d-dimensional function of X and q-dimensional β. We observe $(Z_1,\ldots,Z_n)$ i.i.d., where $Z_i = (Y_i, X_i)$ and aim to identify semiparametric estimators of β, where the density of an observation $p(z)$ belongs to the semiparametric model $\mathcal{P} = \{p(z,\beta,\eta(\cdot)), z = (y,x)\}$. Denote the truth as $p_0(z) = p(z,\beta_0,\eta_0(\cdot))$.

Since there is a one-to-one transformation between (Y,X) and (ε,X), it will be easier for us to work with the density $p_{Y,X}(y,x) = p_{\varepsilon,X}(y-\mu(x,\beta),x)$. Now, we factorize the density of $p_{\varepsilon,X}(\varepsilon,x)$ as

$$p_{\varepsilon,X}(\varepsilon,x) = \eta_1(\varepsilon,x)\eta_2(x)$$

for $\eta_1(\varepsilon,x) = p_{\varepsilon|X}(\varepsilon|x)$. Then, the condition $\mathbb{E}[\varepsilon|X] = 0$ translates to

$$\int \varepsilon\eta_1(\varepsilon,x)d\varepsilon = 0.$$

We also have conditions to ensure η_1,η_2 are valid densities, which are given by

$$\int \eta_1(\varepsilon,x)d\varepsilon = 1, \int \eta_2(x)d\nu(x) = 1$$

where ν is a dominating measure for X. Note that our semiparametric model can then be given by

$$p(z, \beta, \eta_1(\cdot), \eta_2(\cdot)) = \eta_1(y - \mu(x, \beta), x)\eta_2(x),$$

with true density $p_0(z) = \eta_{10}(y - \mu(x, \beta_0), x)\eta_{20}(x)$.

To get a better grasp on the tangent spaces, we consider parametric submodels $p_{\varepsilon|X}(\varepsilon|x, \gamma_1), p_X(x, \gamma_2)$ for γ_1 an r_1-dimensional vector and γ_2 an r_2-dimensional vector. Then, our parametric submodel is given by

$$\mathcal{P}_{\beta,\gamma} = p(z, \beta, \gamma_1, \gamma_2) = p_{\varepsilon|X}(y - \mu(x, \beta)|x, \gamma_1)p_X(x, \gamma_2),$$

where it also contains the truth $p_0(z) = p_{\varepsilon|X}(y - \mu(x, \beta_0)|x, \gamma_{10})p_X(x, \gamma_{20})$.

We now focus on defining the semiparametric nuisance tangent space. The nuisance score vector of the parametric submodel is given by

$$S_\gamma(z, \beta_0, \gamma_0) = \left\{ \left(\frac{\partial \log p(z, \beta, \gamma)}{\partial \gamma_1} \right)^T, \left(\frac{\partial \log p(z, \beta, \gamma)}{\partial \gamma_2} \right)^T \right\}^T \Big|_{\beta=\beta_0, \gamma=\gamma_0} = \{S^T_{\gamma_1}(z, \beta_0, \gamma_0), S^T_{\gamma_2}(z, \beta_0, \gamma_0)\}.$$

Since

$$\log p(z, \beta, \gamma_1, \gamma_2) = \log p_{\varepsilon|X}(y - \mu(x, \beta)|x, \gamma_1) + \log p_X(x, \gamma_2),$$

we can write

$$S_{\gamma_1}(z, \beta_0, \gamma_0) = \left. \frac{\partial \log p_{\varepsilon|X}(y - \mu(x, \beta_0)|x, \gamma_1)}{\partial \gamma_1} \right|_{\gamma_1=\gamma_{10}}$$

and

$$S_{\gamma_2}(z, \beta_0, \gamma_0(z, \beta_0, \gamma_0)) = \left. \frac{\partial \log p_X(x, \gamma_2)}{\partial \gamma_2} \right|_{\gamma_2=\gamma_{20}}.$$

The usefulness of the conditional factorization now becomes clear, as the score vectors split. Letting $\varepsilon = y - \mu(x, \beta_0)$, we can rewrite the score vector with respect to γ_1 as $S_{\gamma_1}(\varepsilon, x)$.

The parametric submodel's nuisance tangent space is given by

$$\Lambda_\gamma = \{BS_\gamma : B \in \mathbb{R}^{q \times r}\}$$

and can be written as

$$\Lambda_\gamma = \Lambda_{\gamma_1} \oplus \Lambda_{\gamma_2}, \Lambda_{\gamma_1} = \{BS_{\gamma_1}(\varepsilon, X) : B \in \mathbb{R}^{q \times r_1}\}, \Lambda_{\gamma_2} = \{BS_{\gamma_2}(X) : B \in \mathbb{R}^{q \times r_2}\}.$$

Lemma 3.4.1 *Λ_{γ_1} and Λ_{γ_2} are orthogonal.*

Proof: Since $\int p_{\varepsilon|X}(\varepsilon|x, \gamma_1)d\varepsilon = 1$ for all x, γ_1, we have

$$\frac{\partial}{\partial \gamma_1} \int p_{\varepsilon|X}(\varepsilon|x, \gamma_1)d\varepsilon = 0 \implies \int \frac{\partial}{\partial \gamma_1} p_{\varepsilon|X}(\varepsilon|x, \gamma_1)d\varepsilon = 0.$$

If we evaluate at $\gamma_1 = \gamma_{10}$, we can write

$$\int \frac{\frac{\partial p_{\varepsilon|X}(\varepsilon|x,\gamma_{10})}{\partial \gamma_1}}{p_{\varepsilon|X}(\varepsilon|x,\gamma_{10})} p_{\varepsilon|X}(\varepsilon|x,\gamma_{10}) d\varepsilon = 0$$

for all x, so

$$\mathbb{E}[S_{\gamma_1}(\varepsilon, X)|X] = 0.$$

In a similar fashion, we can derive $\mathbb{E}[S_{\gamma_2}(X)] = 0$.

Thus, we can write

$$\mathbb{E}[S_{\gamma_1}(\varepsilon, X) S_{\gamma_2}^T(X)] = \mathbb{E}[\mathbb{E}[S_{\gamma_1}(\varepsilon, X) S_{\gamma_2}^T(X)|X]] = \mathbb{E}[\mathbb{E}[S_{\gamma_1}(\varepsilon, X)|X] S_{\gamma_2}^T(X)] = 0,$$

which finishes. □

The semiparametric nuisance tangent space is the mean-square closure of $\Lambda_{\gamma_1} \oplus \Lambda_{\gamma_2}$, which we denote Λ. However, we also have $\Lambda = \Lambda_{1s} \oplus \Lambda_{2s}$, where Λ_{1s} is the mean-square closure of Λ_{γ_1} and Λ_{2s} is defined similarly.

Lemma 3.4.2 *Λ_{2s} is the space of all q-dimensional mean-zero functions of X with finite variance.*

Proof: This is very similar to our earlier derivation. Since the score vector $S_{\gamma_2}(X)$ has mean zero for any γ_2, any element of Λ_{γ_2} is a q-dimensional function of X with mean zero. Hence, it suffices to show that any q-dimensional mean zero function of X is either contained in some parametric submodel Λ_{γ_2} or is the limit of elements of parametric submodels. Choose an arbitrary $\alpha(X)$ with mean zero, and consider the parametric submodel $p_X(x, \gamma_2) = p_0(x)(1 + \gamma_2^T \alpha(x))$, where γ_2 is sufficiently small such that $(1 + \gamma_2^T \alpha(x)) \geq 0$ for all x. As calculated earlier, the score vector of this parametric submodel is simply $\alpha(X)$, and since any α is the limit of bounded mean-zero functions of X, we are done. □

Lemma 3.4.3 *Λ_{1s} is the space of all q-dimensional random functions $a(\varepsilon, x)$ such that*

$$\mathbb{E}[a(\varepsilon, X)|X] = 0, \mathbb{E}[a(\varepsilon, X)\varepsilon^T|X] = 0.$$

Proof: As before, we have $\mathbb{E}[S_{\gamma_1}(\varepsilon, X)|X] = 0$. Furthermore, since $\mathbb{E}[\varepsilon|X] = 0$, we have

$$\int p_{\varepsilon|X}(\varepsilon|x, \gamma_1) \varepsilon^T d\varepsilon = 0,$$

so by essentially the same argument, we have $\mathbb{E}[S_{\gamma_1}(\varepsilon, X)\varepsilon^T|X] = 0$.

Hence, for any γ_1, any element $a(\varepsilon, X) \in \Lambda_{\gamma_1}$ must satisfy $\mathbb{E}[a(\varepsilon, X)|X] = 0, \mathbb{E}[a(\varepsilon, X)\varepsilon^T|X] = 0$. On the other hand, to show that every element of Λ_{1s} is either part of some Λ_{γ_1} or the limit of such elements, we can consider the parametric submodel

$$p_{\varepsilon|X}(\varepsilon|x, \gamma_1) = p_{0\varepsilon|X}(\varepsilon|x)(1 + \gamma_1^T a(\varepsilon, x)).$$

The score vector here is $S_{\gamma_1}(\varepsilon, x) = a(\varepsilon, x)$, which finishes. □

Lemma 3.4.4 *Λ_{1s} is orthogonal to Λ_{2s}.*

Proof: By a similar argument to before, we have, for any $\alpha(X) \in \Lambda_{2s}$ and $a(\varepsilon, X) \in \Lambda_{1s}$,

$$\mathbb{E}[\alpha^T(X)a(\varepsilon, X)] = \mathbb{E}[\alpha^T(X)\mathbb{E}[a(\varepsilon, X)|X]] = 0.$$

□

Going forward, we will write $\Lambda_{1s} = \Lambda_{1sa} \cap \Lambda_{1sb}$, where

$$\Lambda_{1sa} = \{a_a(\varepsilon, X) : \mathbb{E}[a_a(\varepsilon, X)|X] = 0\}, \Lambda_{1sb} = \{a_b(\varepsilon, X) : \mathbb{E}[a_b(\varepsilon, X)\varepsilon^T|X] = 0\}.$$

Lemma 3.4.5 *We have* $\Lambda_{1sa} = \Lambda_{2s}^{\perp}$.

Proof: By the usual proof, we have that Λ_{1sa} is orthogonal to Λ_{2s}. Now, take any $h \in \mathcal{H}$. We have $\mathbb{E}[h|X] \in \Lambda_{2s}$ and $h - \mathbb{E}[h|X] \in \Lambda_{1sa}$, which implies $h = \mathbb{E}[h|X] + (h - \mathbb{E}[h|X])$ can be written as the sum of an element of Λ_{2s} and Λ_{1sa}, and since $h \in \mathcal{H}$ was arbitrary, this finishes. □

Remark 3.4.6 *Note that, from the above proof, we also have* $\Pi(h|\Lambda_{2s}) = \mathbb{E}[h|X]$ *and* $\Pi(h|\Lambda_{1sa}) = h - \mathbb{E}[h|X]$.

Lemma 3.4.7 *We have* $\Lambda_{2s} \subseteq \Lambda_{1sb}$.

Proof: We have, for any $\alpha(X) \in \Lambda_{2s}$,

$$\mathbb{E}[\alpha(X)\varepsilon^T|X] = \alpha(X)\mathbb{E}[\varepsilon^T|X] = 0,$$

so $\alpha(X) \in \Lambda_{1sb}$. □

Lemma 3.4.8 *We have*

$$\Lambda = \Lambda_{2s} \oplus (\Lambda_{1sa} \cap \Lambda_{1sb}) = \Lambda_{1sb}.$$

Proof: For any $h_1 \in \Lambda_{2s}$, by the previous lemma, we have $h_1 \in \Lambda_{1sb}$. Also, for any $h_2 \in (\Lambda_{1sa} \cap \Lambda_{1sb})$, we clearly have $h_2 \in \Lambda_{1sb}$ as well. Hence, $h_1 + h_2 \in \Lambda_{1sb}$.

On the other hand, for any $h \in \Lambda_{1sb}$, we have $\mathbb{E}[h|X] \in \Lambda_{2s} \subseteq \Lambda_{1sb}$, which implies $h - \mathbb{E}[h|X] \in \Lambda_{1sb}$. However, we also have $h - \mathbb{E}[h|X] \in (\Lambda_{1sa} \cap \Lambda_{1sb})$. □

Putting everything together, we have

$$\Lambda = \Lambda_{1sb} = \{h(\varepsilon, X) : \mathbb{E}[h(\varepsilon, X)\varepsilon^T|X] = 0\}.$$

Theorem 3.4.9 *We have*

$$\Lambda^{\perp} = \Lambda_{1sb}^{\perp} = \{A(X)\varepsilon\},$$

where $A(X)$ *is a matrix of arbitrary* $q \times d$*-dimensional functions of* X*. We also have*

$$h(\varepsilon, X) - \Pi(h|\Lambda_{1sb}) = g(X)\varepsilon, g(X) = \mathbb{E}[h(\varepsilon, X)\varepsilon^T|X](\mathbb{E}[\varepsilon\varepsilon^T|X])^{-1},$$

so

$$\Pi(h|\Lambda_{1sb}) = h - \mathbb{E}[h\varepsilon^T|X](\mathbb{E}[\varepsilon\varepsilon^T|X])^{-1}\varepsilon.$$

Proof: First, take some arbitrary $a_b \in \Lambda_{1sb}$. Then, we have

$$\mathbb{E}[a_b^T(\varepsilon, X)A(X)\varepsilon] = \mathbb{E}[\mathbb{E}[a_b^T(\varepsilon, X)A(X)\varepsilon|X]] = 0.$$

By the definition of Λ_{1sb}, $\mathbb{E}[a_b(\varepsilon, X)\varepsilon^T|X] = 0$, which implies that, for all $j = 1, \ldots, q, j' = 1, \ldots, d$, $\mathbb{E}[a_{bj}(\varepsilon, X)\varepsilon_{j'}|X] = 0$. Thus, we have

$$\mathbb{E}[a_b^T(\varepsilon, X)A(X)\varepsilon|X] = \sum_{j,j'} A_{jj'}(X)\mathbb{E}[a_{bj}(\varepsilon, X)\varepsilon_{j'}|X] = 0,$$

which implies $\{A(X)\varepsilon\}$ and Λ_{1sb} are orthogonal.

To show they are orthogonal complements, let $h \in \mathcal{H}$ be arbitrary. We wish to show that there exists some $g(X)$ such that $h(\varepsilon, X) - g(X)\varepsilon \in \Lambda_{1sb}$. By solving the restriction equation $\mathbb{E}[(h - g(X)\varepsilon)\varepsilon^T|X] = 0$, we get

$$g(X) = \mathbb{E}[h\varepsilon^T|X](\mathbb{E}[\varepsilon\varepsilon^T|X])^{-1},$$

which finishes. □

Recall that influence functions $\varphi(\varepsilon, X)$ of RAL estimators for β are elements of the orthogonal complement of the nuisance tangent space such that $\mathbb{E}[\varphi(\varepsilon, X)S_\beta^T(\varepsilon, X)] = I$. Hence, beginning with any $A(X)$, if we let $\varphi(\varepsilon, X) = CA(X)\varepsilon$ for some normalization C, if we set $C = [\mathbb{E}[A(X)\varepsilon S_\beta^T(\varepsilon, X)]]^{-1}$, we will have the desired property. Hence, we consider an m-estimator for β of the form $\sum_{i=1}^n CA(X_i)(Y_i - \mu(X_i, \beta)) = 0$, or, equivalently,

$$\sum_{i=1}^n A(X_i)(Y_i - \mu(X_i, \beta)) = 0.$$

To derive the EIF, we examine the efficient score, which is given by

$$S_{\text{eff}}(\varepsilon, X) = S_\beta(\varepsilon, X) - \Pi(S_\beta(\varepsilon, X)|\Lambda) = \mathbb{E}[S_\beta(\varepsilon, X)\varepsilon^T|X]V^{-1}(X)\varepsilon,$$

for $V(X) = \mathbb{E}[\varepsilon\varepsilon^T|X]$.

To calculate S_β, recall that our density was given as $\eta_1(y - \mu(x, \beta), x)\eta_2(x)$, for $\eta_1(\varepsilon, x) = p_{\varepsilon|X}(\varepsilon|x)$ and $\eta_2(x) = p_X(x)$. Fixing the nuisance parameter at the truth, we get

$$S_\beta(y, x, \beta_0, \eta_0(\cdot)) = \left.\frac{\frac{\partial \eta_{10}(\varepsilon, x)}{\partial \beta}}{\eta_{10}(\varepsilon, x)}\right|_{\beta=\beta_0}.$$

From the model restriction, we have

$$\int (y - \mu(x, \beta))\eta_{10}(y - \mu(x, \beta), x)dy = 0,$$

so

$$\frac{\partial}{\partial \beta^T}\int (y - \mu(x, \beta))\eta_{10}(y - \mu(x, \beta), x)dy\Big|_{\beta=\beta_0} = 0.$$

Moving the derivative into the integral gives

$$\int -D(x)\eta_{10}(\varepsilon, x)dy + \int (y - \mu(x, \beta_0))\frac{\partial \eta_{10}(\varepsilon, x)}{\partial \beta^T}dy = 0,$$

where $D(x) = \frac{\partial \mu(x,\beta_0)}{\partial \beta^T}$. Since η_{10} is a density, it integrates to 1, so the first integral integrates to $-D(x)$. For the second integral, we have

$$\frac{\partial \eta_{10}(\varepsilon, x)}{\partial \beta^T} = \eta_{10}(\varepsilon, x) S_\beta^T(\varepsilon, x),$$

so we have

$$-D(x) + \mathbb{E}[\varepsilon S_\beta^T(\varepsilon, X) | X = s] = 0,$$

or

$$D^T(x) = \mathbb{E}[S_\beta(\varepsilon, X)\varepsilon^T | X = x].$$

Solving, we get

$$S_{\text{eff}}(\varepsilon, X) = D^T(X) V^{-1}(X) \varepsilon,$$

where $D(X) = \frac{\partial \mu(X,\beta_0)}{\partial \beta^T}$. Hence, we get the optimal estimator by solving the equation

$$\sum_{i=1}^{n} D^T(X_i) V^{-1}(X_i)(Y_i - \mu(X_i, \beta)) = 0.$$

Chapter 4

Deriving Influence Functions

We have now characterized efficient influence functions, but to actually use them in practice, we need to be able to derive them for a given causal parameter $\Psi(P)$. Often, via prior work, the EIF (efficient influence function) has been derived for many standard causal parameters, but this is not always the case. In this section, we detail some methods for calculating influence functions. There are two methods one could use. The first is to pretend that data are discrete, calculate an influence function, and then convert this to a continuous influence function and verify that pathwise differentiability still holds.

While this approach is useful, it is often easier to take advantage of the fact that the EIFs of many causal effects have been calculated before. Hence, given a new causal parameter, if one can write it in terms of causal parameters with known influence functions, it is possible to use the rules of calculus to calculate influence functions of new parameters, as we illustrate.

These two methods are taken from (Kennedy, 2024) and provide a robust toolkit to derive influence functions in general. (Y.-C. Chen, 2022) showcases an additional example of using influence function building blocks to calculate the LATE functional in an instrumental variables regression without covariates (a less general case than Example 4.1.6).

4.1 Methods

In the case of discrete data, we compute the pathwise, or Gâteaux, derivative of the parameter in the direction of a point mass. In particular, if $\delta_z = I(Z = z)$, we consider the submodel $(1-\varepsilon)d\mathbb{P}(z) + \varepsilon\delta_{z'}$, the score is

$$\frac{\partial}{\partial\varepsilon}\log\left((1-\varepsilon)d\mathbb{P}(z) + \varepsilon\delta_{z'}\right)|_{\varepsilon=0} = \frac{\delta_{z'} - d\mathbb{P}(z)}{(1-\varepsilon)d\mathbb{P}(z) + \varepsilon\delta_{z'}}|_{\varepsilon=0} = \frac{\delta_{z'}}{d\mathbb{P}(z)} - 1,$$

hence

$$\int \varphi(z;\mathbb{P})s_\varepsilon(z)d\mathbb{P}(z) = \varphi(z';\mathbb{P}).$$

In particular, this implies the pathwise derivative

$$\frac{\partial}{\partial\varepsilon}\psi((1-\varepsilon)d\mathbb{P}(z) + \varepsilon\delta_{z'})\bigg|_{\varepsilon=0} = \varphi(z';\mathbb{P}),$$

which lets us calculate the influence function.

Example 4.1.1 *We examine the very simple functional $\psi = \mathbb{E}[Y|X = x]$. In the discrete case, again let $\delta_z = I(Z = z)$ and consider the submodel $\mathbb{P}_\varepsilon(Z = z) = (1-\varepsilon)\mathbb{P}(Z = z) + \varepsilon\delta_{z'}$ for some fixed z'. Then, we can calculate*

$$\mathbb{P}_\varepsilon(Y = y|X = x) = \frac{\mathbb{P}_\varepsilon(Z = z)}{\mathbb{P}_\varepsilon(X = x)} = \frac{(1-\varepsilon)\mathbb{P}(Z = z) + \varepsilon I(z = z')}{(1-\varepsilon)\mathbb{P}(X = x) + \varepsilon I(x = x')}.$$

Hence, we can calculate the Gâteaux derivative:

$$\begin{aligned}
\frac{d}{d\varepsilon}\psi((1-\varepsilon)\mathbb{P}(z) + \varepsilon\delta_{z'})|_{\varepsilon=0} &= \frac{d}{d\varepsilon}\sum_y y\frac{(1-\varepsilon)\mathbb{P}(Z = z) + \varepsilon I(z = z')}{(1-\varepsilon)\mathbb{P}(X = x) + \varepsilon I(x = x')}\Big|_{\varepsilon=0} \\
&= \sum_y y\frac{\mathbb{P}(X = x)(I(Z = z') - \mathbb{P}(Z = z)) - \mathbb{P}(Z = z)(I(X = x') - \mathbb{P}(X = x))}{\mathbb{P}(X = x)^2} \\
&= \sum_y y\left(\frac{I(Z = z') - \mathbb{P}(Z = z)}{\mathbb{P}(X = x)} - \frac{I(X = x') - \mathbb{P}(X = x)}{\mathbb{P}(X = x)}\mathbb{P}(Y = y|X = x)\right) \\
&= \sum_y y\left(\frac{I(z = z') - I(x = x')}{\mathbb{P}(X = x)}\mathbb{P}(Y = y|X = x)\right) \\
&= \frac{y'I(x = x')}{\mathbb{P}(X = x)} - \frac{I(x = x')\mathbb{E}[Y|X = x]}{\mathbb{P}(X = x)}.
\end{aligned}$$

Converting this to a continuous functional, we get the influence function of $\mathbb{E}[Y|X = x]$ is

$$\varphi(Z;\mathbb{P}) = \frac{I(X = x)}{\mathbb{P}(X = x)}(Y - \mathbb{E}[Y|X = x]).$$

When using this method, one should take care to check that pathwise differentiability holds with the conjectured influence function. In this case, pathwise differentiability can be checked to hold, although we omit the proof here.

We can also build influence functions from smaller influence functions. As one would expect, the rules of calculus apply to influence functions. In particular, if we consider the operator $\mathrm{IF} : \Psi \to L_2(\mathbb{P})$ that maps $\psi \mapsto \varphi(z)$, we can use:

Lemma 4.1.2 *We have*

$$IF(\psi_1\psi_2) = IF(\psi_1)\psi_2 + \psi_1 IF(\psi_2).$$

Proof: Let $\varphi_1 = \mathrm{IF}(\psi_1)$ and $\varphi_2 = \mathrm{IF}(\psi_2)$, so by definition, we have

$$\frac{\partial}{\partial\varepsilon}\psi_i(P_\varepsilon)|_{\varepsilon=0} = \int \varphi_i(z;\mathbb{P})s(z)d\mathbb{P}(z).$$

Then, we have, by the product rule,

$$\begin{aligned}\frac{\partial}{\partial\varepsilon}\psi_1\psi_2(P_\varepsilon)|_{\varepsilon=0} &= \psi_1(P_0)\frac{\partial}{\partial\varepsilon}\psi_2(P_\varepsilon)|_{\varepsilon=0} + \frac{\partial}{\partial\varepsilon}\psi_1(P_\varepsilon)|_{\varepsilon=0}\psi_2(P_0)\\ &= \psi_1(P_0)\int \varphi_2(z;\mathbb{P})s(z)d\mathbb{P}(z) + \psi_2(P_0)\int \varphi_1(z;\mathbb{P})s(z)d\mathbb{P}(z)\\ &= \int(\psi_1\varphi_2 + \psi_2\varphi_1)s(z)d\mathbb{P}(z),\end{aligned}$$

as desired. □

In a similar manner, other standard derivative rules like the quotient and chain rules can be applied to the influence function operator.

Lemma 4.1.3 *We have*

$$IF\left(\frac{\psi_1}{\psi_2}\right) = \frac{IF(\psi_1)\psi_2 - \psi_1 IF(\psi_2)}{\psi_2^2}.$$

Lemma 4.1.4 *We have*

$$IF(f(\psi)) = f'(\psi)IF(\psi).$$

We can use this to calculate the influence function of more complex functionals:

Example 4.1.5 *For covariates X, binary treatment D, and outcome Y, consider the functional given by $\psi = \mathbb{E}[\mathbb{E}[Y|X, D = 1]]$. Let $\mu(x) = \mathbb{E}[Y|X = x, D = 1]$, $\pi(x) = \mathbb{P}(D = 1|X = x)$ and $p(x) = \mathbb{P}(X = x)$. Then, we have:*

$$\begin{aligned}IF(\psi) &= IF\left(\sum_x \mu(x)p(x)\right)\\ &= \sum_x IF(\mu(x))p(x) + \mu(x)IF(p(x))\\ &= \sum_x \frac{I(X = x, D = 1)}{p(1, x)}(Y - \mu(x))p(x) + \mu(x)(I(X = x) - p(x))\\ &= \frac{D}{\pi(X)}(Y - \mu(X)) + \mu(X) - \psi.\end{aligned}$$

In a similar manner, one can derive that the influence function of the functional $\psi' = \mathbb{E}[\mathbb{E}[Y|X, D = 0]]$ is given by

$$\frac{1 - D}{\pi'(X)}(Y - \mu'(X)) + \mu'(X) - \psi',$$

where $\pi'(x) = \mathbb{P}(D = 0|X = x)$ and $\mu'(x) = \mathbb{E}[Y|X = x, D = 0]$.

From the above example as well as noting that influence functions combine linearly, we can derive that the influence function of the average treatment effect functional

$$\psi = \mathbb{E}[\mathbb{E}[Y|X, D = 1] - \mathbb{E}[Y|X, D = 0]]$$

is given by

$$\varphi(X, D, Y) = \frac{D}{\pi(X)}(Y - \mu(X, 1)) + \mu(X, 1) - \frac{1 - D}{1 - \pi(X)}(Y - \mu(X, 0)) - \mu(X, 0) - \psi,$$

which can more succinctly be written as

$$\varphi(X, D, Y) = \frac{2D-1}{\pi(D, X)}(Y - \mu(D, X)) + \mu(1, X) - \mu(0, X) - \psi,$$

where $\pi(d, X) = \mathbb{P}(D = d|X)$ and $\mu(D, X) = \mathbb{E}[Y|X, D]$.

To illustrate how this method quickly builds up to influence functions of complex functionals, we can then use this influence function to calculate the influence function for the local average treatment effect among compliers in an instrumental variables regression:

Example 4.1.6 *Consider the setup of an instrumental variables regression, so in particular, we have the data $W = (X, Z, D, Y)$ where X are covariates, D the treatment, Z the instrumental variable, and Y the outcome. Under necessary assumptions (given, for example, in Chapter 13 of (Chernozhukov, Hansen, Kallus, Spindler, & Syrgkanis, 2024)), the average treatment effect among the compliers (those who followed the recommendation of the instrument) is given by*

$$\psi = \frac{\mathbb{E}\{\mathbb{E}[Y|X, Z = 1] - \mathbb{E}[Y|X, Z = 0]\}}{\mathbb{E}\{\mathbb{E}[D|X, Z = 1] - \mathbb{E}[D|X, Z = 0]\}}.$$

The numerator and denominator are both average treatment effects. Write

$$\psi = \frac{\psi_1}{\psi_2}, \psi_1 = \mathbb{E}[\mathbb{E}[Y|X, Z = 1] - \mathbb{E}[Y|X, Z = 0]], \psi_2 = \mathbb{E}[\mathbb{E}[D|X, Z = 1] - \mathbb{E}[D|X, Z = 0]].$$

The influence functions φ_1 and φ_2 for ψ_1 and ψ_2 are given by

$$\varphi_1 = \frac{2Z-1}{\pi(Z, X)}(Y - \mu(Z, X)) + \mu(1, X) - \mu(0, X) - \psi_1$$

and

$$\varphi_2 = \frac{2Z-1}{\pi(Z, X)}(D - \eta(Z, X)) + \eta(1, X) - \eta(0, X) - \psi_2,$$

where $\mu(Z, X) = \mathbb{E}[Y|Z, X]$ and $\eta(Z, X) = \mathbb{E}[D|Z, X]$. Then, using the quotient rule, we have:

$$\begin{aligned}
\varphi &= IF\left(\frac{\psi_1}{\psi_2}\right) \\
&= \frac{IF(\psi_1)}{\psi_2} - \frac{\psi_1}{\psi_2}\frac{IF(\psi_2)}{\psi_2} \\
&= \frac{1}{\psi_2}\left(\frac{2Z-1}{\pi(Z, X)}(Y - \mu(Z, X)) + \mu(1, X) - \mu(0, X) - \psi_1\right) \\
&- \frac{\psi_1}{\psi_2}\left(\frac{2Z-1}{\pi(Z, X)}(D - \eta(Z, X)) + \eta(1, X) - \eta(0, X) - \psi_2\right) \\
&= \frac{1}{\psi_2}\left(\frac{2Z-1}{\pi(Z, X)}(Y - \mu(Z, X)) + \mu(1, X) - \mu(0, X) - \psi\left(\frac{2Z-1}{\pi(Z, X)}(D - \eta(Z, X)) + \eta(1, X) - \eta(0, X)\right)\right)
\end{aligned}$$

Derivations for even more common influence functions using these methods are given in Section 3 of (Kennedy, 2024).

Chapter 5

Targeted Maximum Likelihood Estimators

Having developed the theory of influence functions to help characterize efficient estimators, we now turn to actually constructing estimators that will achieve our desired statistical properties. In particular, given an initial estimate of the data-generating distribution, we want to modify it so that the plug-in estimator on the modified distribution has an asymptotically linear expansion in the EIF. In this chapter, we introduce the *Targeted Maximum Likelihood Estimator* (or *TMLE*), which works by fluctuating an initial estimate of the data-generating distribution along a direction chosen to yield an asymptotically linear expansion with the efficient influence function.

To motivate this further, recall that, given some estimate $\hat{\mathbb{P}}$ of the true probability distribution $\mathbb{P}$, the von Mises expansion at $(\hat{\mathbb{P}}, \mathbb{P})$ gives

$$\psi(\hat{\mathbb{P}}) - \psi(\mathbb{P}) = -\int \varphi(z; \hat{\mathbb{P}}) d\mathbb{P}(z) + R_2(\hat{\mathbb{P}}, \mathbb{P}),$$

since $\int \varphi(z; \hat{\mathbb{P}} d\hat{\mathbb{P}}(z) = 0$ as φ is mean-zero. In particular, the plug-in estimator $\hat{\psi}_{\text{plug}} = \psi(\hat{\mathbb{P}})$ has a first-order bias $-\int \varphi(z; \hat{\mathbb{P}}) d\mathbb{P}(z)$, which leads to the natural *one-step estimator*

$$\hat{\psi}_{\text{one-step}} = \hat{\psi}_{\text{plug}} + \mathbb{P}_n(\varphi(Z; \hat{\mathbb{P}})),$$

where $\mathbb{P}_n$ is the empirical probability distribution from data. However, one can imagine instead fluctuating the estimated distribution $\hat{\mathbb{P}}$ so that the first-order bias term is naturally removed - in particular, constructing a fluctuated estimate $\hat{\mathbb{P}}^*$ such that $\mathbb{P}_n(\varphi(Z; \hat{\mathbb{P}}^*)) \approx 0$, so that

$$\psi(\hat{\mathbb{P}}^*) \approx \psi(\hat{\mathbb{P}}^*) + \mathbb{P}_n(\varphi(Z; \hat{\mathbb{P}}^*))$$

and the targeted estimator behaves the same as the one-step estimator asymptotically, removing first-order bias.

The seminal paper for the TMLE is (Van Der Laan & Rubin, 2006), where many of the asymptotic results in this section are taken. We refer those more interested in the applications of the TMLE to (Van der Laan,

Rose, et al., 2011), with a more modern treatment in (Van der Laan & Rose, 2018).

5.1 Defining the TMLE

Remark 5.1.1 *In this chapter and going forward, we will use D^* to refer to the EIF, or canonical gradient, matching the typical notation in the TMLE literature.*

The TMLE procedure, as defined in (Van Der Laan & Rubin, 2006), proceeds as follows. Say we want to estimate some parameter of interest Ψ at p_n^0.

1. Begin with an initial density estimator p_n^0 based on the empirical probability distribution P_n and choose a parametric submodel $p_n^0(\varepsilon)$ with score function $D^*(p_n^0)$.

2. Define
$$\varepsilon(P_n|p_n^0) = \text{argmax}_\varepsilon \sum_{i=1}^{n} \log p_n^0(\varepsilon)(O_i), p_n^1 = p_n^0(\varepsilon(P_n|p_n^0)).$$

3. Iterate to get
$$p_n^{k+1} = p_n^k(\varepsilon(P_n|p_n^k))$$
and define
$$\psi_n = \lim_{k\to\infty} \Psi(p_n^k).$$

The fact that this procedure actually does the fluctuating we want it to do is elucidated by the following claim:

Theorem 5.1.2 *Let*
$$\lim_{\varepsilon\to 0} \limsup_{k\to\infty} \left| P_n \frac{\frac{\partial}{\partial\varepsilon} p_n^k(\varepsilon)}{p_n^k(\varepsilon)} - P_n \frac{p_n^{k'}(0)}{p_n^k(0)} \right| \to 0$$
and let there exist a constant matrix A_k for each k such that $A_k \frac{p_n^{k'}}{p_n^k} = D(p_n^k)$ with $\limsup_{k\to\infty} ||A_k|| < \infty$.
If $\varepsilon(P_n|p_n^k)$ is such that $P_n \frac{\frac{\partial}{\partial\varepsilon} p_n^k(\varepsilon)}{p_n^k(\varepsilon)} = 0$ for all k and $\varepsilon(P_n|p_n^k) \to 0$ for $k \to \infty$, then
$$P_n D(p_n^k) \to 0$$
for $k \to \infty$.

Proof: If $\varepsilon(P_n|p_n^k) \xrightarrow{k\to\infty} 0$, then by the given limsup condition, we have
$$P_n \left[\frac{\frac{\partial}{\partial\varepsilon(P_n|p_n^k)} p_n^k(\varepsilon(P_n|p_n^k))}{p_n^k(\varepsilon(P_n|p_n^k))} \right] - P_n \left[\frac{p_n^{k'}(0)}{p_n^k(0)} \right] \xrightarrow{k\to\infty} 0.$$

From the definition of $\varepsilon(P_n|p_n^k)$, the first term is 0, so we have $P_n \frac{p_n^{k'}(0)}{p_n^k(0)} \to 0$ as $k \to \infty$. Defining A_k as in the theorem statement, since its norm is uniformly bounded in k, we can apply A_k to this limit and maintain it, getting $P_n D(p_n^k) \to 0$, as desired.

□

Roughly, Theorem 5.1.2 tells us that as long as our ε estimates go to 0, we will also eventually solve the equation $P_n D(p_n^k) = 0$. This is, of course, the eventual goal of the TMLE estimator.

Definition 5.1.3 *Let $\mathcal{F}$ be a collection of square integrable functions on some probability space with distribution P, and define the empirical process $\mathbb{G}_n$ via*

$$\mathbb{G}_n(f) = \sqrt{n}(\mathbb{P}_n - P)(f),$$

*where $\mathbb{P}_n$ is the empirical measure of an i.i.d. sample from P. Then, $\mathcal{F}$ is a P-**Donsker class** if $(\mathbb{G}_n)_{n=1}^{\infty}$ converges in distribution to a tight Borel measurable element $\mathbb{G}$ in $\ell^{\infty}(\mathcal{F})$.*

The following result, which we state without proof, will be used repeatedly. We refer the reader to Section 2.1.2 of (Van Der Vaart & Wellner, 1996) for further exposition.

Definition 5.1.4 *Define the seminorm*

$$\rho_P(f) = (P(f - Pf)^2)^{\frac{1}{2}}.$$

*An empirical process $\mathbb{G}_n$ is **asymptotically equicontinuous** if, for every $\varepsilon > 0$,*

$$\lim_{\delta \downarrow 0} \limsup_{n \to \infty} P^* \left(\sup_{\rho_P(f-g)<\delta} |\mathbb{G}_n(f-g)| > \varepsilon \right) = 0$$

(where P^* denotes outer probability).

Theorem 5.1.5 *A class $\mathcal{F}$ is Donsker if and only if $\mathcal{F}$ is totally bounded in $L_2(P)$ and that the associated empirical process $\mathbb{G}_n$ is asymptotically equicontinuous.*

5.2 Efficiency & Asymptotic Linearity

In terms of efficiency theorems, we have the following particularly strong result in the case of a convex model and linear causal parameter:

Theorem 5.2.1 *If $P_n D(p_n^k) \to 0$ for $k \to \infty$, then we can pick $K = K(n)$ large enough such that the TMLE $p_n = p_n^K$ satisfies*

$$P_n D(p_n) = R(n, K(n)) = o_P(\frac{1}{\sqrt{n}}).$$

Let $\frac{p_0}{p_n} < \infty$ uniformly on the support of p_0, $\mathcal{M}$ be convex, Ψ be linear. Then, we can write

$$\Psi(p_n) - \Psi(p_0) = (P_n - P_0)D(p_n) + R(n, K(n)).$$

If $D(p_n)$ is in a P_0-Donsker class w.p. tending to 1, then it follows that we can write

$$\Psi(p_n) - \psi_0 = O_P\left(\frac{1}{\sqrt{n}}\right).$$

To prove this, we first take a detour to the convex/linear case.

5.2.1 The Convex/Linear Case

Say that T is a random variable with distribution function F, and we only observe $X = \Phi(T, C)$, for some mapping Φ and censoring variable C. The probability distribution of X can be indexed by the distribution F of X and the conditional distribution G of $C|X$, denoted as $P_{F,G}$.

Definition 5.2.2 *The conditional distribution $G(\cdot|T)$ of C satisfies coarsening at random, which means that, for every t, t',*

$$P_{X|T=t}(dx) = P_{X|T=t'}(dx) \text{ on } \{x : t, t' \in C(x)\}.$$

In other words, $p(x|t) = h(x)$ for some function h.

In this setting, we say $\mu = F\kappa = \int \kappa(t)dF(t)$ is a parameter of interest, with $\mu_n = F_n\kappa = \int \kappa(t)dF_n(t)$ as an estimator. The key ingredient here is the following lemma from (M. J. van der Laan, 1998).

Lemma 5.2.3 *Let F_n be an estimator of F, and for $\alpha \in [0,1]$, let $F_n(\alpha) = (1-\alpha)F_n + \alpha F$. For each F, denote by $\tilde{\ell}(\cdot|F, G, \kappa)$ the EIF of $F\kappa$ at $P_{F,G}$. Suppose, in addition:*

1. *$F\kappa$ is pathwise differentiable at every $F \in \{F_n(\alpha) : \alpha \in [0,1]\}$.*
2. *$F \ll_b F_n$ or $\tilde{\ell}(\cdot|F_n(\alpha), G, \kappa) \to \tilde{\ell}(\cdot|F_n, G, \kappa)$ for $\alpha \to 0$ with respect to the $L^1(P_{F,G})$ norm.*

Then,

$$F_n\kappa - F\kappa = -\int \tilde{\ell}(x|F_n, G, \kappa)dP_{F,G}(x).$$

Proof: For each $F_1 \ll_b F$, define a line from F_1 to F with densities

$$f_h(\varepsilon) = (1 + \varepsilon h)f, h = \frac{f_1 - f}{f} \in L_0^2(F).$$

Since the model is convex, restricting $\varepsilon \in [0,1]$ means these lines form submodels. The scores are just given by the set of $h = \frac{dF_1 - dF}{dF}$. We have

$$\frac{d}{d\varepsilon} \log(p_{f_h(\varepsilon),G})\Big|_{\varepsilon=0}(x) = A_F(h)(x)$$

where $A_F : L_0^2(F) \to L_0^2(P_{F,G}), h \mapsto \mathbb{E}_F[h(T)|X]$ is the *score operator* of F (as seen in the first section).

We have

$$\langle \kappa, h \rangle = \frac{1}{\varepsilon}\left(F_h(\varepsilon)\kappa - F\kappa\right),$$

since the pathwise derivative of $F\kappa$ along $F_h(\varepsilon)$ is $\langle \kappa, h \rangle$, which is the right side by definition. We have $dF_h(\varepsilon) = (1 + \varepsilon h)dF$, so we have

$$F_h(\varepsilon)\kappa = \int \kappa dF_h(\varepsilon) = \int \kappa \left(1 + \varepsilon\left(\frac{dF_1 - dF}{dF}\right)\right)dF = (1-\varepsilon)\int \kappa dF + \varepsilon \int \kappa dF_1 = (1-\varepsilon)F\kappa + \varepsilon F_1\kappa.$$

Hence,

$$F_h(\varepsilon)\kappa - F\kappa = \varepsilon(F_1\kappa - F\kappa),$$

so

$$\langle \kappa, h \rangle = F_1\kappa - F\kappa.$$

Since $F \to P_{F,G}$ is linear, we have

$$P_{F_h(\varepsilon),g} = (1-\varepsilon)P_{F,G} + \varepsilon P_{F_1,G},$$

by similar logic as the previous. Hence,

$$\frac{d}{d\varepsilon} dP_{F_h(\varepsilon),G}\Big|_{\varepsilon=0} = dP_{F_1,G} - dP_{F,G}.$$

Thus,

$$A_F(h) = \frac{d}{d\varepsilon} \log dP_{F_h(\varepsilon),G}\Big|_{\varepsilon=0} = \frac{\frac{d}{d\varepsilon} dP_{F_h(\varepsilon),G}|_{\varepsilon=0}}{dP_{F,G}} = \frac{dP_{F_1,G} - dP_{F,G}}{dP_{F,G}}.$$

Hence,

$$\int \tilde{\ell}(x|F,G,\kappa) A_F(h)(x) dP_{F,G}(x) = \int \tilde{\ell}(x|F,G,\kappa) dP_{F_1}(x) - \int \tilde{\ell}(x|F,G,\kappa) dP_{F,G}(x).$$

Since $\tilde{\ell}$ has mean zero with respect to $P_{F,G}$, we have

$$\int \tilde{\ell}(x|F,G,\kappa) A_F(h)(x) dP_{F,G}(x) = \int \tilde{\ell}(x|F,G,\kappa) dP_{F_1}(x).$$

We have

$$\langle \kappa, h \rangle = \langle \tilde{\ell}(\cdot|F,G,\kappa), A_F(h) \rangle_{P_{F,G}},$$

which implies that

$$F_1\kappa - F\kappa = \int \tilde{\ell}(x|F,G,\kappa) dP_{F_1,G}(x).$$

This holds for any $F_1 \ll_b F$. Swapping F and F_1, we get that, for any $F \ll_b F_1$,

$$F\kappa - F_1\kappa = \int \tilde{\ell}(x|F_1,G,\kappa) dP_{F,G}(x)$$

or, equivalently,

$$F_1\kappa - F\kappa = -\int \tilde{\ell}(x|F_1,G,\kappa) dP_{F,G}(x).$$

If $F \ll_b F_n$, we can set $F_1 = F_n$ and finish. However, if not, we have $F \ll_b F_n(\alpha)$ for $\alpha \neq 0$ (since $F_n(\alpha) = (1-\alpha)F_n + \alpha F = 0 \implies F = 0$ for $\alpha \neq 0$). Thus, if $\tilde{\ell}(\cdot|F_n(\alpha),G,\kappa) \to \tilde{\ell}(F_n,G,\kappa)$ as $\alpha \to 0$, the identity will also hold at $F_1 = F_n$ by continuity, which finishes. □

We now fill in the proof of Theorem 5.2.1 from (Van Der Laan & Rubin, 2006), which only provides a sketch of this proof.

Proof: In the case of a convex $\mathcal{M}$ and linear Ψ, we can write $\Psi(p) - \Psi(p_0) = -P_0 D(p)$ for any $p, p_0 \in \mathcal{M}$ with $p \ll_b p_0$ (equivalently, $p_0/p < \infty$). By the previous result, if $p_n^\infty \in \mathcal{M}$, we have $P_n D(p_n^\infty) = 0$, using p_n^∞ in place of p gives

$$\Psi(p_n^\infty) - \Psi(p_0) = (P_n - P_0) D(p_n^\infty).$$

Now, define the process $\mathbb{G}_n(f) = \sqrt{n}(P_n - P_0)f$, so the previous becomes

$$\sqrt{n}(\Psi(p_n^\infty) - \Psi(p_0)) = \mathbb{G}_n(D(p_n^\infty)).$$

Let $\mathcal{F}$ be a P_0-Donsker class such that, with probability tending to 1, $D(p_n^\infty) \in \mathcal{F}$. Since $\mathcal{F}$ is P_0-Donsker, $\mathbb{G}_n$ converges in distribution in $\ell^\infty(\mathcal{F})$ to a tight limit $\mathbb{G}$, which implies that $\sup_{f\in\mathcal{F}} |\mathbb{G}_n(f)| = O_P(1)$. Hence, if $D(p_n^\infty) \in \mathcal{F}$, we have $|\mathbb{G}_n(D(p_n^\infty))| = O_P(1)$, which implies

$$(P_n - P_0)D(p_n^\infty) = O_P(n^{-\frac{1}{2}}),$$

with probability tending towards one. Hence,

$$\Psi(p_n^\infty) - \Psi(p_0) = O_P(n^{-\frac{1}{2}}),$$

which proves the second statement when noting $\Psi(p_0) = \psi_0$. □

Corollary 5.2.3.1 *If $P_0(D(p_n^\infty) - D(p_1))^2 \xrightarrow{p} 0$ for some $p_1 \in \mathcal{M}$, then*

$$\Psi(p_n^\infty) - \Psi(p_0) = (P_n - P_0)D(p_1) + o_P(n^{-\frac{1}{2}}).$$

Hence, $\Psi(p_n^\infty)$ is asymptotically linear with influence function $D_0(p_1) := D(p_1) - P_0D(p_1)$, and if $D(p_1) = D(p_0)$, then the TMLE is asymptotically efficient.

Proof: We begin with the identity

$$\Psi(p_n^\infty) - \Psi(p_0) = (P_n - P_0)D(p_n^\infty).$$

From this, we can write

$$\Psi(p_n^\infty) - \Psi(p_0) = (P_n - P_0)D(p_1) + (P_n - P_0)(D(p_n^\infty) - D(p_1)),$$

and noting that $P_0D(p_1)$ is a constant and therefore $(P_n - P_0)P_0D(p_1) = 0$, we can write

$$\Psi(p_n^\infty) - \Psi(p_0) = (P_n - P_0)D_0(p_1) + (P_n - P_0)(D(p_n^\infty) - D(p_1)).$$

Hence, it remains to show that $(P_n - P_0)(D(p_n^\infty) - D(p_1)) = o_P(n^{-\frac{1}{2}})$, or, equivalently, that $\mathbb{G}_n(D(p_n^\infty) - D(p_1)) = o_P(1)$.

Since $D(p_n^\infty)$ is contained in a P_0-Donsker class $\mathcal{F}$ with probability tending towards 1, then, with probability tending to 1, by Theorem 5.1.5, $\mathbb{G}_n$ is asymptotically equicontinuous with respect to $\rho(f,g) := (P_0(f-g)^2)^{\frac{1}{2}}$ on $\mathcal{F}$. In particular,

$$\sup_{g,h\in\mathcal{F},\rho(g,h)<\delta} |\mathbb{G}_n(g-h)| \xrightarrow{p} 0,$$

and since $\rho(D(p_n^\infty) - D(p_1)) \xrightarrow{p} 0$ (by assumption), we have

$$\mathbb{G}_n(D(p_n^\infty) - D(p_1)) = o_P(1),$$

which implies

$$(P_n - P_0)(D(p_n^\infty) - D(p_1)) = o_P\left(\frac{1}{\sqrt{n}}\right).$$

Plugging this back into the initial identity yields

$$\begin{aligned}\Psi(p_n^\infty) - \Psi(p_0) = (P_n - P_0)D(p_n^\infty) &= (P_n - P_0)D(p_1) + (P_n - P_0)(D(p_n^\infty) - D(p_1)) \\ &= (P_n - P_0)D(p_1) + o_P\left(\frac{1}{\sqrt{n}}\right),\end{aligned}$$

as desired. □

5.2.2 The General Case

This is quite a strong consistency result in that it is robust to the starting density p_n^0 - this is mainly due to the strong assumptions of the convexity of $\mathcal{M}$ and linearity of Ψ. When this is not the case, we can still get the following result:

Theorem 5.2.4 *Let $p_n^\infty \in \mathcal{M}$ denote the limit of the TMLE algorithm if it exists, and otherwise a p_n^k for large enough k such that $P_n D(p_n^\infty) = 0$. Assume Ψ is pathwise differentiable and write*

$$\Psi(p) - \Psi(p_0) = -P_0 D(p) + R(p, p_0)$$

for any $p \in \mathcal{M}$. If $D(p_n^\infty)$ is in a P_0-Donsker class with probability tending to 1 and $R(p_n^\infty, p_0) = o_P(n^{-\frac{1}{2}})$, then

$$\Psi(p_n^\infty) - \Psi(p_0) = O_P\left(\frac{1}{\sqrt{n}}\right),$$

so $\Psi(p_n^\infty)$ is $\sqrt{n}$-consistent.

Proof: The proof proceeds roughly the same as in the convex/linear case. Since $P_n D(p_n^\infty) = 0$, we have

$$\Psi(p_n^\infty) - \Psi(p_0) = (P_n - P_0)D(p_n^\infty) + R(p_n^\infty, p_0).$$

As before, define the empirical process $\mathbb{G}_n(f) = \sqrt{n}(P_n - P_0)f$, so this identity becomes

$$\sqrt{n}(\Psi(p_n^\infty) - \Psi(p_0)) = \mathbb{G}_n(D(p_n^\infty)) + \sqrt{n}R(p_n^\infty, p_0) = \mathbb{G}_n(D(p_n^\infty)) + o_P(1),$$

using the assumption that $R(p_n^\infty, p_0) = o_P(n^{-\frac{1}{2}})$. Let $\mathcal{F}$ be a P_0-Donsker class such that $D(p_n^\infty) \in \mathcal{F}$ with probability tending to 1. This implies that $\sup_{f \in \mathcal{F}} |\mathbb{G}_n(f)| = O_P(1)$, which implies that, with probability tending to 1,

$$(P_n - P_0)D(p_n^\infty) = O_P(n^{-\frac{1}{2}}),$$

which implies that

$$\Psi(p_n^\infty) - \Psi(p_0) = O_P(n^{-\frac{1}{2}})$$

when plugging back into the previous. □

Remark 5.2.5 *The added assumption that $R(p_n^\infty, p_0) = o_P(\frac{1}{\sqrt{n}})$ is essentially the cost of not assuming the strong structure of a convex model and linear parameter. Assuming this structure sets this remainder term to 0, as shown via the identity of (M. J. van der Laan, 1998), so if this is not the case, we need some assumption to control this remainder term.*

Using similar reasoning, we have a similar corollary to the convex model and linear target parameter case as well:

Corollary 5.2.5.1 *If, in addition to the previous assumptions, $P_0(D(p_n^\infty) - D(p_1))^2 \xrightarrow{p} 0$ for some $p_1 \in \mathcal{M}$, then $\Psi(p_n^\infty)$ is asymptotically linear with expansion*

$$\Psi(p_n^\infty) - \Psi(p_0) = (P_n - P_0)D(p_1) + o_P\left(\frac{1}{\sqrt{n}}\right).$$

Hence, if $D(p_1) = D(p_0)$, we have asymptotic efficiency of the TMLE.

This proof exactly mirrors the proof of Corollary 5.2.3.1, so we omit it.

5.3 Targeted Minimum Loss-Based Estimation

A simple modification to TMLE (also aptly named TMLE) is *targeted minimum loss-based estimation*, where the parameter of interest $\Psi(P)$ is represented as $\Psi_1(Q(P))$, where $Q(P) = \text{argmin}_Q PL(Q)$ for some loss function $L(Q)$ that is dependent on the data O. The canonical gradient $D^*(P)$ can then be written as $D^*(P) = D_1^*(Q(P), G(P))$ for a nuisance parameter G.

The full targeted minimum loss-based estimation procedure proceeds as follows:

1. Begin with an initial estimator (Q_n^0, G_n^0).
2. Define a local least favorable submodel $Q_{n,\varepsilon}^0$ such that $\frac{d}{d\varepsilon}L(Q_{n,\varepsilon}^0)|_{\varepsilon=0}$ spans $D^*(Q_n^0, G_n^0)$ (in other words, $\langle D^*(Q_n^0, G_n^0)\rangle \subseteq \langle \frac{d}{d\varepsilon}L(Q_{n,\varepsilon}^0)|_{\varepsilon=0}\rangle$.
3. Update $Q_n^{k+1} = Q_{n,\varepsilon^k}^k$, where $\varepsilon^k = \text{argmin}_\varepsilon P_n L(Q_{n,\varepsilon}^k)$.

In the end, the resulting TMLE will satisfy $P_n(D^*(Q_n^*, G_n^0)) = 0$, and we will obtain a corresponding TMLE $\Psi_1(Q_n^*)$ for ψ_0. The main difference here is that we use a generic loss function L rather than log-likelihood for the update step. This is a simple change that allows for more generality in the framework - an example will be given in the next section.

5.4 Example: Mean Missing at Random

A good example of implementing the TMLE in practice is given by (Díaz & Rosenblum, 2015), which implements several different TMLEs, using different parametric submodels, for the mean missing at random problem. This shows the amount of variety and choice that even the framework of the basic TMLE gives.

We begin with the setup. Say we have observations $O_1, O_2, \ldots, O_n$ observed as $O = (X, M, MY) \sim P_0$. Here, the covariates are represented as X and the binary missing flag is M - if $M = 0$, we always observe $MY = 0$, so we only observe Y if $M = 1$. We assume that the binary outcome Y is "missing at random," so M is independent of Y given X. We also assume overlap/positivity for identification of causal effects.

Throughout, denote the marginal density of the covariates by $p_X(x)$, the propensity as $p_M(X) = P(M = 1|X)$, and the outcome regression as $\mu(X) = \mathbb{E}_P[Y|M = 1, X] = P(Y = 1|M = 1, X)$ (here, P refers to the overall density). Our functional of interest is

$$\Psi(\mu, p_X) = \mathbb{E}_{p(X)}[Y] = \mathbb{E}_{p_X}[\mu(X)].$$

The efficient influence function for this scenario was calculated in Example 4.1.5, and is given by

$$D(p, O) = \frac{M}{p_M(X)}(Y - \mu(X)) + \mu(X) - \Psi(\mu, p_X).$$

Note that via conditional factorization, we can write the full density $p(x, m, my)$ as

$$p(x, m, my) = p_1(x)p_2(m|x)p_3(my|x, m).$$

The density p_1 is exactly p_X. Since M is binary, the density $p_2(m|x)$ is exactly characterized by $p_M(x) = P(M = 1|X)$. Finally, since Y is binary, $p_3(my|m, x)$ is fully characterized by $P(Y = 1|M, X) = \mathbb{E}_P[Y|M = 1, X]$, which is exactly $\mu(X)$. Hence, we can equivalently write $D(p, O)$ as a function of p_X, p_M, μ, O, which we may find helpful to do for implementations of the TMLE where we choose to fluctuate p_X, p_M, μ as opposed to the whole density p at once (this may be more tractable).

5.4.1 One-Step Implementation

We begin with an implementation that converges in one step - more on this kind of TMLE will be discussed in Chapter 6.

We begin with an initial estimate of p_X to be p_X^0, the empirical distribution of the covariates X from the given dataset. Initial estimators μ^0 and p_M^0 for μ and p_M are also defined.

Owing to our decomposition of the density p into p_X, p_M, μ, we describe our fluctuation p_ε in terms of fluctuations to p_X, p_M, μ. Given μ^k, we define the parametric submodel

$$\text{logit } (\mu_\varepsilon^k(X)) = \text{logit } (\mu^k(X)) + \varepsilon H_Y(X), H_Y(X) = \frac{1}{p_M^k(X)}.$$

We keep the propensity p_M^k unchanged from its initial estimator (we do not define a fluctuation for it), and for p_X, we define the parametric submodel

$$p_{X,\theta}^k(X) \propto \exp(\log p_X^k(X) + \theta H_X^k(X)), H_X^k(X) = \mu^k(X) - \mathbb{E}_{p_X^k}[\mu^k(X)].$$

To show that this submodel fits into the broader TMLE framework, we must ensure that, under these

submodels defined for p_X, p_M, μ,

$$D(p^k, o) = D(p_X^k, p_M^k, \mu^k, o) = \frac{\partial}{\partial \varepsilon} \log p_\varepsilon^k(o)\bigg|_{\varepsilon=0} = \frac{\partial}{\partial \theta} \log p_{X,\theta}^k(o)\bigg|_{\theta=0} + \frac{\partial}{\partial \varepsilon} \log p_{3,\varepsilon}^k(o)\bigg|_{\varepsilon=0},$$

where we can ignore the term from p_M on the RHS as the estimator of p_M is not modified from its initial state.

Beginning with μ^k, we have

$$\mu_\varepsilon^k(X) = \text{expit}(\text{logit } \mu_k(X) + \varepsilon H_Y(X)).$$

Then, we have

$$\frac{\partial}{\partial \varepsilon} \mu_\varepsilon^k(X) = H_Y(X)\mu_\varepsilon^k(X)(1 - \mu_\varepsilon^k(X)),$$

since $\sigma'(z) = \sigma(z)(1 - \sigma(z))$, where σ is the sigmoid (or expit) function. Setting $\varepsilon = 0$, we get

$$\frac{\partial}{\partial \varepsilon} \mu_\varepsilon^k(X)\bigg|_{\varepsilon=0} = H_Y(X)\mu^k(X)(1 - \mu^k(X)).$$

Now, $p_{3,\varepsilon}$ can be viewed as a binary distribution with success probability μ_ε^k if $m = 1$, and always equals 0 otherwise. In particular (conditioned on m, x),

$$p_{3,\varepsilon}(my) = (\mu_\varepsilon^k(x)^y (1 - \mu_\varepsilon^k(x))^{1-y})^m,$$

where the probability should be 1 if m is 0 since we are conditioning on m. Hence, the log-likelihood is given by

$$m(y \log \mu_\varepsilon^k(x) + (1 - y) \log(1 - \mu_\varepsilon^k(x)).$$

Differentiating with respect to ε, we get

$$\frac{my}{\mu_\varepsilon^k(x)} \frac{\partial}{\partial \varepsilon} \mu_\varepsilon^k(x) - \frac{m(1-y)}{1 - \mu_\varepsilon^k(x)} \frac{\partial}{\partial \varepsilon} \mu_\varepsilon^k(x) = M H_Y(X)\mu_\varepsilon^k(X)(1 - \mu_\varepsilon^k(X)) \left(\frac{Y}{\mu_\varepsilon^k(X)} - \frac{1 - Y}{1 - \mu_\varepsilon^k(X)} \right).$$

Setting $\varepsilon = 0$ and simplifying, we get

$$\frac{\partial}{\partial \varepsilon} \log p_{3,\varepsilon}^k(o)\bigg|_{\varepsilon=0} = M H_Y(X)(Y - \mu^k(X)) = \frac{M}{p_M^k(X)}(Y - \mu^k(X)).$$

For the $p_{X,\theta}^k$ term, we can write

$$p_{X,\theta}^k(X) = Z(\theta) \exp(\log p_X^k(X) + \theta H_X^k(X)),$$

where

$$Z(\theta) = \left(\int \exp(\log p_X^k(x) + \theta H_X^k(x)) dx \right)^{-1}$$

is a normalizing constant. This means

$$\log p^k_{X,\theta}(X) = \log Z(\theta) + \log p^k_X(X) + \theta H^k_X(X).$$

Differentiating with respect to θ, we get

$$\frac{\partial}{\partial\theta}\log p^k_{X,\theta}(X) = \frac{\partial}{\partial\theta}\log Z(\theta) + H^k_X(X).$$

We can write

$$\frac{\partial}{\partial\theta}\log Z(\theta) = \frac{1}{Z(\theta)}\frac{\partial}{\partial\theta}Z(\theta) = \frac{\int \frac{\partial}{\partial\theta}\exp(\log p^k_X(x) + \theta H^k_X(x))dx}{\int \exp(\log p^k_X(x) + \theta H^k_X(x))dx} = \frac{\int H^k_X(x)\exp(\log p^k_X(x) + \theta H^k_X(x))dx}{\int \exp(\log p^k_X(x) + \theta H^k_X(x))dx}.$$

Evaluated at $\theta = 0$, we get

$$\frac{\partial}{\partial\theta}\log Z(\theta)\bigg|_{\theta=0} = \frac{\int H^k_X(x)p^k_X(x)dx}{\int p^k_X(x)dx} = \mathbb{E}_{p^k_X}[H^k_X(X)] = \mathbb{E}_{p^k_X}[\mu^k(X) - \mathbb{E}_{p^k_X}[\mu^k(X)]] = 0.$$

Thus, we have

$$\frac{\partial}{\partial\theta}\log p^k_{X,\theta}(X) = H^k_X(X) = \mu^k(X) - \mathbb{E}_{p^k_X}[\mu^k(X)] = \mu^k(X) - \Psi(p^k).$$

Putting everything together, we see that

$$D(p^k, O) = \frac{\partial}{\partial\theta}p^k_{X,\theta}(O)\bigg|_{\theta=0} + \frac{\partial}{\partial\varepsilon}p^k_{3,\varepsilon}(O)\bigg|_{\varepsilon=0},$$

which means these parametric submodels comprise a valid use of TMLE.

We can make this TMLE converge in one step by implementing it as follows. Note that the MLE of p_X, in the nonparametric model, is exactly p^0_X, so the fluctuation parameter $\hat{\theta}$ for p^0_X will be fit to 0 in the first iteration of the TMLE.

To fit the fluctuation parameter $\hat{\varepsilon}$ for the logit model for μ^0, we can fit a logistic regression of the outcome Y on the covariate $H_Y(X) = \frac{1}{p^0_M(X)}$ among data with $M = 1$, with offset logit $\mu^0(X)$. The associated score equation for this logistic regression is derived by deriving the loss function

$$L(\varepsilon) = \sum_{i=1}^{n} M_i[Y_i \log \mu^0_\varepsilon(X_i) + (1 - Y_i)\log(1 - \mu^0_\varepsilon(X_i))]$$

with respect to ε. We have:

$$\begin{aligned}
\frac{d}{d\varepsilon}L(\varepsilon) &= \sum_{i=1}^{n} M_i \frac{d}{d\varepsilon}[Y_i \log \mu_\varepsilon^0(X_i) + (1-Y_i)\log(1-\mu_\varepsilon^0(X_i))] \\
&= \sum_{i=1}^{n} M_i \left[\frac{Y_i}{\mu_\varepsilon^0(X_i)}\frac{d\mu_\varepsilon^0(X_i)}{d\varepsilon} - \frac{1-Y_i}{1-\mu_\varepsilon^0(X_i)}\frac{d\mu_\varepsilon^0(X_i)}{d\varepsilon}\right] \\
&= \sum_{i=1}^{n} \frac{M_i(Y_i - \mu_\varepsilon^0(X_i))}{\mu_\varepsilon^0(X_i)(1-\mu_\varepsilon^0(X_i))}\frac{d\mu_\varepsilon^0(X_i)}{d\varepsilon} \\
&= \sum_{i=1}^{n} \frac{M_i(Y_i - \mu_\varepsilon^0(X_i))}{\mu_\varepsilon^0(X_i)(1-\mu_\varepsilon^0(X_i))}\mu_\varepsilon^0(X_i)(1-\mu_\varepsilon^0(X_i))H_Y(X_i) \\
&= \sum_{i=1}^{n} \frac{M_i}{p_M^0(X_i)}(Y_i - \mu_\varepsilon^0(X_i)),
\end{aligned}$$

where we use the fact that

$$\frac{d\mu_\varepsilon^0(X_i)}{d\varepsilon} = \mu_\varepsilon^0(X_i)(1-\mu_\varepsilon^0(X_i))H_Y(X_i).$$

At $\hat{\varepsilon}$, we have

$$\sum_{i=1}^{n} \frac{M_i}{p_M^0(X_i)}(Y_i - \mu_{\hat{\varepsilon}}^0(X_i)) = 0.$$

Setting

$$\mu^1(X_i) = \mu_{\hat{\varepsilon}}^0(X_i) = \text{expit}\left(\text{logit } \mu^0(X_i) + \frac{\hat{\varepsilon}}{p_M^0(X_i)}\right),$$

we get

$$\sum_{i=1}^{n} \frac{M_i}{p_M^0(X_i)}(Y_i - \mu^1(X_i)) = 0.$$

On the second iteration, the score equation being solved by the same logistic regression is

$$\sum_{i=1}^{n} \frac{M_i}{p_M^0(X_i)}\left(Y_i - \text{expit}\left(\text{logit } \mu^1(X_i) + \frac{\varepsilon}{p_M^0(X_i)}\right)\right) = 0.$$

This is clearly solved at $\varepsilon = 0$, so by convexity of the objective, the algorithm terminates in one step, with the point estimate TMLE being $\hat{\psi} = \overline{\mu^1(X_i)}$.

One can imagine implementing the fluctuation for μ as

$$\text{logit } \mu_\varepsilon^k(X) = \text{logit } \mu^k(X) + \varepsilon,$$

and instead running a weighted logistic regression with weights $\frac{1}{p_M^0(X)}$ and offset logit $\mu^k(X)$. This implementation also converges in one step by essentially the same argument as before, but this is an example of *targeted minimum loss-based estimation*, using the weighted log-likelihood loss as opposed to the standard log-likelihood to fit $\hat{\varepsilon}$. (Díaz & Rosenblum, 2015) finds that, in large samples under misspecified data, this method may perform worse than standard TMLE, although it performs similarly in other settings.

5.4.2 General TMLE Implementations

One can also adhere more closely to the TMLE framework by fluctuating the entire density. So far, the submodels we have seen have been linear fluctuations, but it is also possible to use exponential tilts as in (Díaz & Rosenblum, 2015), which we showcase here.

One possible exponential tilt to use is, at iteration k,

$$p_\varepsilon^k(O) = c(\varepsilon, p^k)\exp(\varepsilon D(p^k, O))p^k(O),$$

where

$$c(\varepsilon, p^k) = \left[\int \exp\{\varepsilon D(p^k, o)\}p^k(o)d\nu(o)\right]^{-1}$$

is a normalizing constant. At $\varepsilon = 0$, we have $c(0, p^k) = 1$, so the parametric submodel at $\varepsilon = 0$ passes through the truth.

To verify the score requirement of a parametric submodel, we calculate:

$$\begin{aligned}
\frac{d}{d\varepsilon}\log p_\varepsilon^k(o)|_{\varepsilon=0} &= \frac{d}{d\varepsilon}\log[\exp[\varepsilon D(p^k, o)]]\Big|_{\varepsilon=0} + \frac{d}{d\varepsilon}\log c(\varepsilon, p^k)\Big|_{\varepsilon=0} \\
&= D(p^k, o) + \frac{d}{d\varepsilon}\left[\log\left[\int \exp\{\varepsilon D(p^k, o)\}p^k(o)d\nu(o)\right]^{-1}\right]\Bigg|_{\varepsilon=0} \\
&= D(p^k, o) - \frac{d}{d\varepsilon}\left[\log\left[\int \exp\{\varepsilon D(p^k, o)\}p^k(o)d\nu(o)\right]\right]|_{\varepsilon=0} \\
&= D(p^k, o) - \left[\left[\int \exp\{\varepsilon D(p^k, o)\}p^k(o)d\nu(o)\right]^{-1}\frac{d}{d\varepsilon}\int \exp\{\varepsilon D(p^k, o)\}p^k(o)d\nu(o)\right]|_{\varepsilon=0} \\
&= D(p^k, o) - \left[c(\varepsilon, p^k)\int \frac{d}{d\varepsilon}\exp\{\varepsilon D(p^k, o)\}p^k(o)d\nu(o)\right]|_{\varepsilon=0} \\
&= D(p^k, o) - c(0, p^k)\int D(p^k, o)p^k(o)d\nu(o) = D(p^k, o),
\end{aligned}$$

Several exponential tilts are possible. Another tilt considered by (Díaz & Rosenblum, 2015) is

$$p_\varepsilon^k(O) = c(\varepsilon, p^k)(1 + \exp(-2\varepsilon D(p^k, O)))^{-1}p^k(O),$$

where

$$c(\varepsilon, p^k) = \left[\int (1 + \exp(-2\varepsilon D(p^k, O))^{-1}p^k(O)d\nu(O)\right]^{-1}$$

is a normalizing constant. Again, at $\varepsilon = 0$, this normalizing constant evaluates to 1, so it remains to check the score requirement to verify that this is a valid parametric submodel:

$$
\begin{aligned}
\frac{d}{d\varepsilon}\log p_\varepsilon^k(o)|_{\varepsilon=0} &= \frac{d}{d\varepsilon}\log c(\varepsilon,p^k)\Big|_{\varepsilon=0} - \frac{d}{d\varepsilon}\log(1+\exp(-2\varepsilon D(p^k,O)))\Big|_{\varepsilon=0} \\
&= \frac{d}{d\varepsilon}\left[\log\left[\int(1+\exp(-2\varepsilon D(p^k,O))^{-1}p^k(O)d\nu(O)\right]^{-1}\right]\Big|_{\varepsilon=0} + D(p^k,O) \\
&= D(p^k,O) - \left[\int(1+\exp(-2\varepsilon D(p^k,O))^{-1}p^k(O)d\nu(O)\right]^{-1} \\
&\cdot \frac{d}{d\varepsilon}\int(1+\exp(-2\varepsilon D(p^k,O)))^{-1}p^k(O)d\nu(O)\Big|_{\varepsilon=0} \\
&= D(p^k,O) - \left[c(\varepsilon,p^k)\int\frac{d}{d\varepsilon}(1+\exp(-2\varepsilon D(p^k,O)))^{-1}p^k(O)d\nu(O)\right]\Big|_{\varepsilon=0} \\
&= D(p^k,O) - c(0,p^k)\int D(p^k,O)p^k,Od\nu(O) = D(p^k,O),
\end{aligned}
$$

showing that this is an alternative valid parametric submodel as well. This example shows that many possible parametric submodels or fluctuations are possible when implementing TMLE (even in its base form).

The remaining chapters study variants of TMLE that modify this basic construction in different ways. One-Step TMLE focuses on the algorithmic question of whether targeting can be achieved in a single update rather than through iteration. Cross-Validated TMLE addresses empirical process restrictions by using sample splitting to weaken Donsker-type conditions. Collaborative TMLE changes how nuisance estimators are selected, allowing the nuisance estimation procedure itself to be targeted toward the parameter of interest. Higher-Order TMLE extends the bias correction idea beyond the first-order expansion by incorporating higher-order influence functions. Thus, each extension preserves the central TMLE principle while addressing a different limitation of the basic estimator.

Chapter 6

One-Step TMLE

The standard TMLE is presented as an iterated method, but it turns out that there exist versions of the TMLE that allow us to achieve the same first-order correction in a single update, which can be more efficient. This is done by picking a parametric submodel such that the score function equals the efficient influence function *everywhere*, not just at $\varepsilon = 0$ (which is the usual requirement).

Much of the exposition for this section is taken from (M. van der Laan & Gruber, 2016).

6.1 Locally Least Favorable Models

Definition 6.1.1 *A **least favorable model** at P is some parametric submodel $\{P_{\varepsilon,h^*} : \varepsilon\}$, dominated by P, such that the Cramér-Rao lower bound*

$$CR(h|P) = \frac{\left(\frac{d}{d\varepsilon}\Psi(P_{\varepsilon,h}|_{\varepsilon=0}\right)^2}{-P\frac{d^2}{d\varepsilon^2}\log\frac{dP_{\varepsilon,h}}{dP}|_{\varepsilon=0}}$$

is maximized over all parametric submodels $\{P_{\varepsilon,h} : \varepsilon\}$, where the function h varies over some set whose mean square closure generates the full tangent space.

As in Chapter 2, letting S_h be the score function of $P_{\varepsilon,h}$ at $\varepsilon = 0$, we have

$$\mathrm{CR}(h|P) = \frac{(PD^*(P)S_h)^2}{PS_h^2}.$$

Using Cauchy-Schwarz and taking the equality case, we see that the choice of S in the tangent space which maximizes this is given by $S = D^*(P)$, meaning a least favorable model is simply one that has score at P equal to $D^*(P)$.

Definition 6.1.2 *A **locally least favorable submodel** is a submodel such that the optimality of the Cramér-Rao lower bound holds at $\varepsilon = 0$, or, equivalently, that the score at $P_{0,h}$ is equal to $D^*(P)$:*

$$\frac{\partial}{\partial\varepsilon}\log\frac{dP_{\varepsilon,h}}{dP}\bigg|_{\varepsilon=0} = D^*(P)$$

The following lemma provides some motivation on why a locally least favorable submodel is sensible to consider in the case of TMLE.

Lemma 6.1.3 *Under standard smoothness assumptions, we have*

$$CR(h|P) = \lim_{\varepsilon \to 0} \frac{(\Psi(P_{\varepsilon,h}) - \Psi(P))^2}{-2P \log \frac{dP_{\varepsilon,h}}{dP}}.$$

Proof: This is essentially just a second-order Taylor expansion. Let $\ell(\varepsilon) = P \log \frac{dP_{\varepsilon,h}}{dP}$, and consider performing a second-order Taylor expansion of ℓ around $\varepsilon = 0$. Note that $\ell(0) = 0$. The first derivative of ℓ is, by definition, the score S_h. Hence, $\ell'(0) = PS_h = 0$ since the score has mean zero. The second derivative is

$$\ell''(0) = -PS_h^2,$$

for I the Fisher information of the submodel. Hence, we have

$$\ell(\varepsilon) = -\frac{\varepsilon^2}{2} PS_h^2 + o(\varepsilon^2),$$

so

$$-2P \log \frac{dP_{\varepsilon,h}}{dP} = \varepsilon^2 PS_h^2 + o(\varepsilon^2),$$

which means

$$PS_h^2 = \frac{-2P \log \left(\frac{dP_{\varepsilon,h}}{dP} \right)}{\varepsilon^2} + o(1).$$

To get the desired expression, we can write

$$\frac{d}{d\varepsilon} \Psi(P_{\varepsilon,h}) \bigg|_{\varepsilon=0} = \lim_{\varepsilon \to 0} \frac{\Psi(P_{\varepsilon,h}) - \Psi(P)}{\varepsilon},$$

so

$$\mathrm{CR}(h|P) = \frac{\left(\frac{d}{d\varepsilon} \Psi(P_{\varepsilon,h}) \Big|_{\varepsilon=0} \right)}{PS_h^2} = \lim_{\varepsilon \to 0} \frac{\frac{(\Psi(P_{\varepsilon,h}) - \Psi(P))^2}{\varepsilon^2}}{\frac{-2P \log \left(\frac{dP_{\varepsilon,h}}{dP} \right)}{\varepsilon^2}} = \lim_{\varepsilon \to 0} \frac{(\Psi(P_{\varepsilon,h}) - \Psi(P))^2}{-2P \log \frac{dP_{\varepsilon,h}}{dP}},$$

as desired. □

In particular, the Cramér-Rao bound equals the squared change in the parameter of interest divided by an infinitesimal change in the log-likelihood at P - hence, maximizing this bound directly affects the targeting ability. Of course, it would be particularly nice if this property held not just at $\varepsilon = 0$, but in a more global setting. The extension to a universal setting allows us to actually define the one-step TMLE.

6.2 Universally Least Favorable Models

Definition 6.2.1 *Let $\mathcal{M}$ be some model, and fix some a. Consider the parametric submodel $U(P) = \{P_\varepsilon : \varepsilon \in (-a, a)\}$ such that $P_0 = P$. We say that $U(P)$ is a* ***universal least favorable submodel*** *if, for all $\varepsilon \in (-a, a)$,*

$$\frac{\partial}{\partial \varepsilon} \log \frac{dP_\varepsilon}{dP} = D^*(P_\varepsilon).$$

Theorem 6.2.2 *Let P_n^0 be some initial estimator of P_0, and let $U(P)$ be a universal least favorable submodel for a given P. Then, TMLE on this submodel converges in one step.*

Proof: In the first step of TMLE, we set our fluctuation

$$\varepsilon_n^0 = \text{argmax}_\varepsilon P_n \log \frac{dP_{n,\varepsilon}^0}{dP_n^0}$$

and set $P_n^1 = P_{n,\varepsilon_n^0}^0$. Since ε_n^0 is a local maximum, by the universality of the submodel, we have $P_n D^*(P_n^1) = 0$. In particular, further iteration will not give any more updates, since

$$\varepsilon_n^1 = \text{argmax}_\varepsilon P_n \log \frac{dP_{n,\varepsilon}^1}{dP_n^1} = 0,$$

which implies the TMLE converges in one step. □

Of course, this result is only of use if such a universal least favorable submodel actually exists. Fortunately, we can actually produce an explicit one:

Theorem 6.2.3 *Given a starting density p (we can take $p = 1$), for all $\varepsilon \geq 0$, define*

$$p_\varepsilon = p \exp\left(\int_0^\varepsilon D^*(P_x)dx\right),$$

and similarly, for $\varepsilon < 0$, we recursively define

$$p_\varepsilon = p \exp\left(-\int_\varepsilon^0 D^*(P_x)dx\right).$$

Then, we have $\{P_\varepsilon : \varepsilon \in (-a, a)\}$ is a set of probability distributions dominated by P with $P_0 = P$ and, for each $\varepsilon \in (-a, a)$,

$$\frac{\partial}{\partial \varepsilon} \log \frac{dP_\varepsilon}{dP} = D^*(P_\varepsilon),$$

so $\{P_\varepsilon : \varepsilon \in (-a, a)\}$ is a universal least favorable model.

Proof: It is essentially by definition that $\frac{d}{d\varepsilon} \log p_\varepsilon = D^*(P_\varepsilon)$, so we just have to check that $p_\varepsilon \geq 0$ and that it integrates to 1 for all ε. Nonnegativity clearly holds, so we check the second condition.

Let $C(\varepsilon, P) = \int p_\varepsilon dP$ for each ε, and define $p'_\varepsilon = C(\varepsilon, P)^{-1} p_\varepsilon$. Then, the score is given by

$$S(\varepsilon, P) = \frac{1}{C(\varepsilon, P)} \frac{\partial}{\partial \varepsilon} C(\varepsilon, P) + D^*(P_\varepsilon).$$

Using the well-known identity $P_\varepsilon S(\varepsilon, P) = 0$, in addition to the fact that $P_\varepsilon D^*(P_\varepsilon) = 0$, we must have $\frac{\partial}{\partial \varepsilon} C(\varepsilon, P) = 0$. Hence, $C(\varepsilon, P)$ is constant and $C(0, P) = \int p dP = 1$, which finishes. □

Hence, we can always compute a universal least favorable model from any starting density! This submodel can be practically constructed via discretization. In particular, we have, for $\varepsilon > 0$,

$$\frac{p_{\varepsilon + d\varepsilon}}{p_\varepsilon} = \exp\left(\int_\varepsilon^{\varepsilon + d\varepsilon} D^*(P_x)dx\right) \approx \exp(D^*(p_\varepsilon)d\varepsilon) \approx 1 + d\varepsilon D^*(P_\varepsilon),$$

where the last approximation is via a first-order approximation.

Hence, a practical construction for this universal least favorable model is to create a grid $0 = x_0 < x_1 < \cdots < x_N = a$, and setting

$$p_{x_j} = p_{x_{j-1}}(1 + (x_j - x_{j-1})D^*(P_{x_{j-1}}))$$

for $j = 1, \ldots, N$. Of course, if $\mathcal{M}$ is fully nonparametric, this discretized version will still be a submodel of $\mathcal{M}$, but otherwise this is seemingly not guaranteed since the process could select probability distributions that are not elements of $\mathcal{M}$.

To remedy this, we can instead phrase the construction in terms of a limit of the discretized versions. Assume we have access to a mapping $P \mapsto \{P_\varepsilon^{\text{lfm}} : \varepsilon \in (-a, a)\} \subseteq \mathcal{M}$ $(a > 0)$ which maps $P \in \mathcal{M}$ to a locally least favorable submodel of $\mathcal{M}$ through P at $\varepsilon = 0$ with score $D^*(P)$. Consider an equally spaced grid $0 = x_0 < x_1 < \cdots < x_N = a$, and starting from $P_0 = P$, we define $p_{x_{j+1}} = p_{x_j,h}^{\text{lfm}}$ for $j = 1, \ldots, N-1$. Similarly, we set $p_{-x_{j+1}} = p_{-x_j,-h}^{\text{lfm}}$ for $-a = -x_N < -x_{N-1} < \cdots < -x_1 < -x_0 = 0$. By construction, all the P_i are in $\mathcal{M}$. Letting $N \to \infty$ gives an alternative construction of the universal least favorable model, with the guarantee that we never leave our overarching model $\mathcal{M}$.

Theorem 6.2.4 *Assuming the Taylor expansion*

$$p_{\varepsilon,d\varepsilon}^{lfm} = p_\varepsilon + \frac{d}{dh} p_{\varepsilon,h}^{lfm}\bigg|_{\varepsilon=0} d\varepsilon + R_2(p_\varepsilon, d\varepsilon),$$

where we have the universal bounds

$$\sup_\varepsilon \sup_o |R_2(p_\varepsilon, d\varepsilon)(o)| = O((d\varepsilon)^2), \sup_\varepsilon \sup_o |D^*(P_\varepsilon)p_\varepsilon|(o) < \infty,$$

the construction of the universal least favorable submodel given in the preceding paragraph aligns with the definition in Definition 6.2.1 (so, in particular, the analytic version is an element of $\mathcal{M}$).

Proof: We can write the Taylor expansion as

$$p_{\varepsilon,d\varepsilon}^{\text{lfm}} = p_\varepsilon + \frac{d}{dh} p_{\varepsilon,h}^{\text{lfm}}\bigg|_{\varepsilon=0} d\varepsilon + R_2(p_\varepsilon, d\varepsilon) = p_\varepsilon(1 + d\varepsilon D^*(P_\varepsilon)) + O((d\varepsilon)^2),$$

so

$$p_\varepsilon = p \exp\left(\int_0^\varepsilon D^*(P_x)dx\right)$$

for $\varepsilon > 0$, as desired. The proof for $\varepsilon < 0$ is identical. □

6.3 Universal Least Favorable Models for Loss-Based Estimation

The preceding construction can be fairly easily extended to more general loss functions. To reiterate the setup, let our functional be of the form $\Psi(P) = \Psi_1(Q(P))$ for some parameter Q, where $Q : \mathcal{M} \to Q(\mathcal{M}) := \{Q(P) : P \in \mathcal{M}\}$, and let $L(Q)$ be a loss function for Q, so $Q(P) = \text{argmin}_{Q \in Q(\mathcal{M})} PL(Q)$. The efficient influence curve $D^*(P)$ is a function of $Q(P)$ and a nuisance parameter $G(P)$, and we can assume, WLOG,

that the efficient influence curve lies in the tangent space of Q (otherwise, we can include G in the definition of Q so that this holds).

Given some (Q, G), as before, we begin with a local least favorable model $\{Q_\varepsilon^{\text{lfm}} : \varepsilon \in (-q, q)\}$ such that

$$\frac{d}{d\varepsilon} L(Q_\varepsilon^{\text{lfm}})\bigg|_{\varepsilon=0} = D^*(Q, G).$$

Our construction of a universal least favorable model would then proceed by defining a grid $0 = x_0 < x_1 < \cdots < x_N = a$ of equally spaced points. We define $Q_{x_{j+1}} = Q^{\text{lfm}}_{x_j, x_{j+1}-x_j}$ for $j = 1, 2, \ldots, N-1$. Similarly, we define a grid $-a = -x_N < -x_{N-1} < \cdots < -x_0 = 0$ and set $Q_{x_{j+1}} = Q^{\text{lfm}}_{-x_j, -(x_{j+1}-x_j)}$ for $j = 1, 2, \ldots, N-1$. By similar logic as before, all the elements Q_{x_j}, Q_{-x_j} are part of the parameter space $Q(\mathcal{M})$.

As in the previous section, we have the following theorem which shows that this construction, as we take the grid to be finer and finer, implements an analytic construction that is an element of $Q(\mathcal{M})$. The analytic construction in question is as follows.

Theorem 6.3.1 *Let $\{Q_\varepsilon^{lfm} : \varepsilon \in (-a, a)\}$ be a local least favorable model with respect to some loss function $L(Q)$ such that*

$$\frac{d}{d\varepsilon} L(Q_\varepsilon^{lfm})\bigg|_{\varepsilon=0} = D^*(Q, G).$$

Furthermore, assume there exists a function $\dot{L}$ such that

$$\frac{d}{d\varepsilon} L(Q_\varepsilon^{lfm})\bigg|_{\varepsilon=0} = \dot{L}(Q) \frac{d}{d\varepsilon} Q_\varepsilon^{lfm}\bigg|_{\varepsilon=0}.$$

Define the analytic construction

$$Q_\varepsilon = Q + \int_0^\varepsilon \frac{D^*(Q_x, G)}{\dot{L}(Q_x)} dx, Q_{-\varepsilon} = Q - \int_{-\varepsilon}^0 \frac{D^*(Q_x, G)}{\dot{L}(Q_x)} dx.$$

Assume the second order Taylor expansion

$$Q_{\varepsilon, d\varepsilon}^{lfm} = Q_\varepsilon + \frac{d}{dh} Q_{\varepsilon, h}^{lfm}\bigg|_{\varepsilon=0} d\varepsilon + R_2(Q_\varepsilon, G, d\varepsilon).$$

Also assume the bounds

$$\sup_\varepsilon \sup_o |R_2(Q_\varepsilon, G, d\varepsilon)(o)| = O((d\varepsilon)^2), \sup_\varepsilon \sup_o \left| \frac{D^*(Q_\varepsilon, G)}{\dot{L}(Q_\varepsilon)}(o) \right| < \infty.$$

Then, the analytic construction $\{Q_\varepsilon : \varepsilon\}$ is a subset of $Q(\mathcal{M})$.

Proof: Let $\varepsilon > 0$ be arbitrary - the proof for $\varepsilon < 0$ is similar. The Taylor expansion is

$$Q_{\varepsilon, d\varepsilon}^{\text{lfm}} = Q_\varepsilon + \frac{d}{dh} Q_{\varepsilon, h}^{\text{lfm}}\bigg|_{\varepsilon=0} d\varepsilon + R_2(Q_\varepsilon, G, d\varepsilon) = Q_\varepsilon + \frac{D^*(Q_\varepsilon, G)}{\dot{L}(Q_\varepsilon)} d\varepsilon + R_2(Q_\varepsilon, G, d\varepsilon).$$

This implies the analytic characterization

$$Q_\varepsilon = Q + \int_0^\varepsilon \frac{D^*(Q_x, G)}{\dot{L}(Q_x)} dx.$$

To check universality, we note that

$$\frac{d}{d\varepsilon} L(Q_\varepsilon) = \dot{L}(Q_\varepsilon) \frac{d}{d\varepsilon} Q_\varepsilon = \dot{L}(Q_\varepsilon) \frac{d}{d\varepsilon} L \left(Q + \int_0^\varepsilon \frac{D^*(Q_x, G)}{\dot{L}(Q_x)} dx \right) = \dot{L}(Q_\varepsilon) \frac{D^*(Q_\varepsilon, G)}{\dot{L}(Q_\varepsilon)} = D^*(Q_\varepsilon, G),$$

and since this is true for arbitrary ε, we have the desired universality property, so we are done. □

Chapter 7

Cross-Validated TMLE

Our results on asymptotic linearity, thus far, have imposed Donsker conditions. These conditions, while useful for proofs to control terms of the form $(P_n - P_0)(D^*(Q_n^*, g_n) - D^*(Q_0, g_0))$, can be difficult to verify in practice. Such conditions are necessary because Q_n^* and g_n are estimated from the same data over which the difference $P_n - P_0$ is evaluated.

A way to get around the use of such conditions is to impose a *sample-splitting* assumption, typically cross-validation in practice. Then, the estimated influence function is fit on different data than the empirical distribution is evaluated on, and the aforementioned empirical process term can then be controlled without Donsker conditions.

Much of the exposition of this section is from (Zheng & Van Der Laan, 2010).

7.1 Defining the CV-TMLE

We again work in the setting where we observe n observations $O_1, \ldots, O_n$ i.i.d. from $P_0 \in \mathcal{M}$ and want to estimate some target parameter $\Psi(P_0)$. Say that Ψ is pathwise differentiable, for each $P \in \mathcal{M}$ along the set of 1-dimensional submodels $\{P_h(\varepsilon) : \varepsilon\}$ such that, for all $h \in \mathcal{H}$,

$$\frac{d}{d\varepsilon}\Psi(P_h(\varepsilon))\Big|_{\varepsilon=0} = PD(P)S(h),$$

where $D(P)$ is the canonical gradient and $S(h)$ is the score of $P_h(0)$.

Work in the setting of targeted minimum loss-based estimation, so our parameter is written as $\Psi(Q(P_0))$ and $D^*(P) = D^*(Q(P), G(P))$ for some relevant parameter $Q \in \mathcal{Q}$ and nuisance parameter G and there exists some uniformly bounded loss function L such that $Q(P_0) = \text{argmin}_{Q \in \mathcal{Q}} P_0 L(Q)$.

To define a TMLE, as usual, we begin with an initial estimator $\hat{Q}(P_n)$ of $Q(P_0)$, as well as an initial estimator $\hat{g}(P_n)$ of $g(P_0)$. Then, given $\hat{Q}, \hat{g}$, we consider fluctuations $\hat{Q}(P_n)(\varepsilon)$ such that

$$\langle D^*(\hat{Q}(P_n), \hat{g}(P_n))\rangle \subseteq \left\langle \frac{d}{d\varepsilon} L(\hat{Q}(P_n)(\varepsilon))\Big|_{\varepsilon=0} \right\rangle.$$

In particular, we specify a submodel through $\hat{Q}(P_n)$, indexed by ε, such that the score at $\varepsilon = 0$ is a function which spans the EIF at $(\hat{Q}(P_n), \hat{g}(P_n))$.

Instead of fitting ε on the entire data, as in standard TMLE, we randomly split the samples into training and validation sets $\mathcal{T}$ and $\mathcal{V}$, respectively, split by some random vector $B_n = \{0,1\}^n$ (so the training set is samples O_i where $B_n(i) = 0$ and the validation set is samples O_i where $B_n(i) = 1$). Let P^0_{n,B_n} and P^1_{n,B_n} are the empirical distributions of the train and validation sets, respectively, so that

$$P^0_{n,B_n} f = \frac{1}{\#\{i : B_n(i) = 0\}} \sum_{i:B_n(i)=0} f(O_i), P^1_{n,B_n} f = \frac{1}{\#\{i : B_n(i) = 1\}} \sum_{i:B_n(i)=1} f(O_i).$$

Then, we can estimate

$$\varepsilon^0_n = \hat{\varepsilon}(P_n) = \text{argmin}_\varepsilon P^1_{n,B_n} L(\hat{Q}(P^0_{n,B_n})(\varepsilon)).$$

Note that this is equivalent to $\text{argmin}_\varepsilon L(\hat{Q}(P^0_{n,B_n}(\varepsilon))$, which is finding the optimal ε on the validation set after fitting $\hat{Q}$ on the training set. The quantity $P^1_{n,B_n} L(\hat{Q}(P^0_{n,B_n}(\varepsilon^0_n))$ is then the validation error of this optimal ε.

In this way, one can derive an update ε^0_{n,B_n} for each split B_n of the data. One can iterate this process until $\varepsilon^k_n \approx 0$, in which case the TMLE can be defined as

$$\hat{\Psi}(P_n) = \mathbb{E}_{B_n}[\Psi(\hat{Q}^*(P^0_{n,B_n}))].$$

7.2 Asymptotic Linearity

We develop the asymptotic linearity result for the CV-TMLE in a series of lemmas.

Lemma 7.2.1 *Let $\hat{Q}(P_n), \hat{g}(P_n)$ be initial estimators of Q_0, g_0 respectively. Assume that $\{\hat{Q}(P_n)(\varepsilon) : \varepsilon\} \in \mathcal{Q}$ with probability 1, the loss function $L(Q)$ is uniformly bounded in Q, and we also have the double-uniform bound*

$$M_1 = \sup_Q \sup_O |L(Q)(O)| < \infty$$

over some support $O \sim P_0$.

Let

$$\hat{\Psi}(P_n) = \mathbb{E}_{B_n}[\Psi(\hat{Q}(P^0_{n,B_n})(\varepsilon_n))].$$

Assuming that

$$\Psi(Q(P)) - \Psi(Q(P_0)) = -P_0 D^*(Q(P), g_0) + O_P(||\Psi(Q(P)) - \Psi(Q(P_0))||^2),$$

we have

$$\begin{aligned}\hat{\Psi}(P_n) - \psi_0 = &\mathbb{E}_{B_n}[(P^1_{n,B_n} - P_0) D^*(\hat{Q}(P^0_{n,B_n})(\varepsilon_n), \hat{g}(P^0_{n,B_n}))] \\ &+ \mathbb{E}_{B_n}[P_0(D^*(\hat{Q}(P^0_{n,B_n})(\varepsilon_n), \hat{g}(P^0_{n,B_n})) - D^*(\hat{Q}(P^0_{n,B_n})(\varepsilon_n), g_0)] \\ &- \mathbb{E}_{B_n}[P_0(D^*(Q_0, \hat{g}(P^0_{n,B_n})) - D^*(Q_0, g_0))] \\ &+ O_P(||\hat{\Psi}(P_n) - \psi_0||^2).\end{aligned}$$

Remark 7.2.2 *For the above lemma, we must also assume the "double robustness" of D^*, or that $P_0 D^*(Q_0, g) =$*

0 for all nuisances g, which typically holds for causal parameters of interest - this assumption allows us to subtract out the third term of the expansion.

To prove this lemma, we first show an auxiliary lemma, based on the following definitions:

Definition 7.2.3 *Let $\mathcal{F}$ be a class of functions $f : O \to \mathbb{R}$. The covering number $N(\varepsilon, \mathcal{F}, L_2(Q))$ is the minimum number of ε-balls, defined with respect to the $L_2(Q)$ norm, needed to cover $\mathcal{F}$.*

Definition 7.2.4 *Given a function class $\mathcal{F}$, its envelope F is a function such that $|f| \leq F$ for all $f \in \mathcal{F}$ (again, with respect to the $L_2(Q)$ norm).*

Definition 7.2.5 *Let $\mathcal{F}$ be a class of functions $f : O \to \mathbb{R}$. The entropy integral is given by*

$$Entropy(\mathcal{F}) = \int_0^\infty \sqrt{\log \sup_Q N(\varepsilon ||F||_{Q,2}, \mathcal{F}, L_2(Q))} d\varepsilon,$$

where $||F||_{Q,2}$ is the $L_2(Q)$ norm of the envelope F of $\mathcal{F}$.

The following lemma, Lemma 2.14.1 of (Van Der Vaart & Wellner, 1996), will be useful for us.

Lemma 7.2.6 *If $\mathcal{F}$ is a class of measurable functions of O and $G_n = \sqrt{n}(P_n - P_0)$ the standard empirical process, then*

$$\mathbb{E}[\sup_{f \in \mathcal{F}} |G_n f|] \leq Entropy(\mathcal{F}) \sqrt{P_0 F^2}.$$

From Lemma 7.2.6, we can prove the following result which will be of use to us:

Lemma 7.2.7 *Let $||\varepsilon_n - \varepsilon_0|| \xrightarrow{p} 0$. For each sample split B_n, condition on P^0_{n,B_n} and define the function class of measurable functions f,*

$$\mathcal{F}(P^0_{n,B_n}) = \{f_\varepsilon(P^0_{n,B_n}) := f(\varepsilon, P^0_{n,B_n}) - f(\varepsilon_0, P_0) : \varepsilon\}.$$

Assume the indexing set of ε contains ε_n with probability approaching 1.

Fix some sequence $\delta_n \to 0$ and define, for each n,

$$\mathcal{F}_{\delta_n}(P^0_{n,B_n}) = \{f_\varepsilon \in \mathcal{F}(P^0_{n,B_n}) : ||\varepsilon - \varepsilon_0|| < \delta_n\}.$$

Let $F(\delta_n, P^0_{n,B_n})$ denote the envelope of $\mathcal{F}_{\delta_n}(P^0_{n,B_n})$.

If

$$\mathbb{E}[Entropy(\mathcal{F}_{\delta_n}(P^0_{n,B_n}))\sqrt{P_0 F(\delta_n, P^0_{n,B_n})^2}] \xrightarrow{n \to \infty} 0,$$

then

$$\sqrt{n}(P^1_{n,B_n} - P_0)(f(\varepsilon_n, P^0_{n,B_n}) - f(\varepsilon_0, P_0)) = o_P(1).$$

Proof: Define

$$G_{n,B_n} = \sqrt{n}(P^1_{n,B_n} - P_0),$$

which is an empirical process based on $\mathcal{V}$ after conditioning on P^0_{n,B_n}. Hence, conditional on P^0_{n,B_n}, applying Lemma 7.2.6 to the function class $\mathcal{F}_{\delta_n}(P^0_{n,B_n})$ yields

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}_{\delta_n}(P^0_{n,B_n})}|G_{n,B_n}f||P^0_{n,B_n}\right] \leq \text{Entropy}(\mathcal{F}_{\delta_n}(P^0_{n,B_n}))\sqrt{P_0F(\delta_n,P^0_{n,B_n})^2}.$$

Taking the expectation of the right side, we have, by assumption,

$$\mathbb{E}[\text{Entropy}(\mathcal{F}_{\delta_n}(P^0_{n,B_n}))\sqrt{P_0F(\delta_n,P^0_{n,B_n})^2}] \xrightarrow{n\to\infty} 0,$$

so

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}_{\delta_n}(P^0_{n,B_n})}|G_{n,B_n}f||P^0_{n,B_n}\right] \xrightarrow{n\to\infty} 0.$$

By Markov's inequality, this implies

$$\sup_{f\in\mathcal{F}_{\delta_n}(P^0_{n,B_n})}|G_{n,B_n}f| = o_P(1).$$

Since we assume $\varepsilon_n \xrightarrow{p} \varepsilon_0$, there exists a fixed sequence $\delta_n \to 0$ such that $P(||\varepsilon_n - \varepsilon_0|| < \delta_n) \to 1$. By definition, when $||\varepsilon_n - \varepsilon_0|| < \delta_n$, we have $f(\varepsilon_n, P^0_{n,B_n}) - f(\varepsilon_0, P_0) \in \mathcal{F}_{\delta_n}(P^0_{n,B_n})$. Hence, with probability approaching 1,

$$f(\varepsilon_n, P^0_{n,B_n}) - f(\varepsilon_0, P_0) \in \mathcal{F}_{\delta_n}(P^0_{n,B_n}).$$

Thus, putting everything together,

$$G_{n,B_n}(f(\varepsilon_n, P^0_{n,B_n}) - f(\varepsilon_0, P_0)) = \sqrt{n}(P^1_{n,B_n} - P_0)(f(\varepsilon_n, P^0_{n,B_n}) - f(\varepsilon_0, P_0)) = o_P(1),$$

as desired. □

We can now show Lemma 7.2.1 using Lemma 7.2.7.

Proof: For each split B_n, we can write

$$\Psi(\hat{Q}(P^0_{n,B_n})(\varepsilon_n)) - \Psi(Q(P_0)) = -P_0D^*(\hat{Q}(P^0_{n,B_n})(\varepsilon_n), g_0) + O_P(||\Psi(\hat{Q}(P^0_{n,B_n})(\varepsilon_n)) - \Psi(Q(P_0))||^2).$$

Taking expectation over B_n gives

$$\mathbb{E}_{B_n}[\Psi(\hat{Q}(P^0_{n,B_n})(\varepsilon_n))] - \Psi(Q(P_0)) = -\mathbb{E}_{B_n}[P_0D^*(\hat{Q}(P^0_{n,B_n})(\varepsilon_n), g_0)] + O_P(||\Psi(\hat{Q}(P^0_{n,B_n})(\varepsilon_n)) - \Psi(Q(P_0))||^2).$$

Due to one-step convergence, we have

$$\mathbb{E}_{B_n}[P^1_{n,B_n}D^*(\hat{Q}(P_{n,B_n})(\varepsilon_n), \hat{g}(P^0_{n,B_n})] = 0.$$

Thus, we have

$$\Psi(\hat{Q}(P^0_{n,B_n})(\varepsilon_n)) - \Psi(Q(P_0)) = -\mathbb{E}_{B_n}[P_0 D^*(\hat{Q}(P^0_{n,B_n})(\varepsilon_n), \hat{g}(P^0_{n,B_n})) - P^1_{n,B_n} D^*(\hat{Q}(P_{n,B_n})(\varepsilon_n), g_0)] + O_P(||\Psi(\hat{Q}(P^0_{n,B_n})(\varepsilon_n)) - \Psi(Q(P_0))||^2).$$

Now, add and subtract $P_0 D^*(\hat{Q}(P^0_{n,B_n})(\varepsilon_n), \hat{g}(P^0_{n,B_n}))$ inside the expectation to get

$$\begin{aligned}\Psi(\hat{Q}(P^0_{n,B_n})(\varepsilon_n)) - \Psi(Q(P_0)) = &-\mathbb{E}_{B_n}[(P^1_{n,B_n} - P_0) D^*(\hat{Q}(P^0_{n,B_n})(\varepsilon_n), \hat{g}(P^0_{n,B_n})) \\ &+ P_0(D^*(\hat{Q}(P^0_{n,B_n})(\varepsilon_n), \hat{g}(P^0_{n,B_n})) - D^*(\hat{Q}(P^0_{n,B_n})(\varepsilon_n), g_0))] \\ &+ O_P(||\Psi(\hat{Q}(P^0_{n,B_n})(\varepsilon_n)) - \Psi(Q(P_0))||^2).\end{aligned}$$

Finally, by the double robustness of D^*, $P_0 D^*(Q_0, g) = 0$ for all nuisances g. Hence, we have

$$\mathbb{E}_{B_n}[P_0((D^*(Q_0, \hat{g}(P^0_{n,B_n})) - D^*(Q_0, g_0))] = 0,$$

and adding this term in yields the result. □

Lemma 7.2.8 *Say that, in addition, we have $\varepsilon_0 = \varepsilon(P_0)$ with $||\varepsilon_n - \varepsilon_0|| \xrightarrow{p} 0$. Assume that, for each sample split B_n, conditional on P^0_{n,B_n}, the function class*

$$\mathcal{F}(P^0_{n,B_n}) = \{O \mapsto D^*(\hat{Q}(P^0_{n,B_n})(\varepsilon), \hat{g}(P^0_{n,B_n})) - D^*(\hat{Q}(P_0)(\varepsilon_0), \hat{g}(P_0)) : \varepsilon\}$$

contains ε_n with probability approaching 1. Also, take some fixed sequence $\delta_n \xrightarrow{n\to\infty} 0$ and define the subclasses

$$\mathcal{F}_{\delta_n}(P^0_{n,B_n}) = \{f(\varepsilon) \in \mathcal{F}(P^0_{n,B_n}) : ||\varepsilon - \varepsilon_0|| < \delta_n\}.$$

Also assume that

$$\mathbb{E}[\mathit{Entropy}(\mathcal{F}_{\delta_n}(P^0_{n,B_n}))\sqrt{P_0 F^2(\delta_n, P^0_{n,B_n})}] \xrightarrow{n\to\infty} 0,$$

where $F(\delta_n, P^0_{n,B_n})$ is the envelope of $\mathcal{F}_{\delta_n}(P^0_{n,B_n})$. Then, we have the expansion:

$$\begin{aligned}\hat{\Psi}(P_n) - \Psi(Q(P_0)) = &(P_n - P_0) D^*(\hat{Q}(P_0)(\varepsilon_0), \hat{g}(P_0)) + o_P(\frac{1}{\sqrt{n}}) \\ &+ \mathbb{E}_{B_n}[P_0(D^*(\hat{Q}(P^0_{n,B_n})(\varepsilon), \hat{g}(P^0_{n,B_n})) - D^*(\hat{Q}(P^0_{n,B_n})(\varepsilon), g_0))] \\ &- \mathbb{E}_{B_n}[P_0(D^*(Q_0, \hat{g}(P^0_{n,B_n})) - D^*(Q_0, g_0))] \\ &+ O_P(||\hat{\Psi}(P_n) - \Psi(Q(P_0))||^2).\end{aligned}$$

Proof: The assumptions of this lemma are exactly the assumptions in Lemma 7.2.7. Hence, applying Lemma 7.2.7 with D^* in the place of f, we get, for each sample split B_n,

$$(P^1_{n,B_n} - P_0)(D^*(\hat{Q}(P^0_{n,B_n})(\varepsilon_n), \hat{g}(P^0_{n,B_n})) - D^*(\hat{Q}(P_0)(\varepsilon_0), \hat{g}(P_0))) = o_P(\frac{1}{\sqrt{n}}).$$

Taking expectation with respect to B_n and rearranging, we get

$$\mathbb{E}_{B_n}[(P^1_{n,B_n} - P_0)(D^*(\hat{Q}(P^0_{n,B_n})(\varepsilon_n), \hat{g}(P^0_{n,B_n}))] = \mathbb{E}_{B_n}[(P^1_{n,B_n} - P_0)(D^*(\hat{Q}(P_0)(\varepsilon_0), \hat{g}(P_0))] + o_P(\frac{1}{\sqrt{n}}).$$

□

If $\hat{g}(P_n) = g_0$, the terms after the first cancel to give:

Corollary 7.2.8.1 *If, in addition to the previous assumptions, we assume the consistency of $\hat{g}$, so $\hat{g}(P_n) = g_0$, and the consistency of the TMLE itself (so $\hat{\Psi}(P_n) \xrightarrow{p} \psi_0$), we have the linear expansion:*

$$\hat{\Psi}(P_n) - \psi_0 = (P_n - P_0)(D^*(\hat{Q}(P_0)(\varepsilon_0), g_0)) + o_P(\frac{1}{\sqrt{n}}).$$

We get asymptotic efficiency if $\hat{Q}(P_0)(\varepsilon_0) = Q_0$, so

$$\hat{\Psi}(P_n) - \psi_0 = (P_n - P_0)D^*(Q_0, g_0) + o_P(\frac{1}{\sqrt{n}}).$$

Proof: Plugging in $\hat{g}(P_n) = g_0$, we get

$$\mathbb{E}_{B_n}[P_0(D^*(\hat{Q}(P^0_{n,B_n})(\varepsilon), \hat{g}(P^0_{n,B_n})) - D^*(\hat{Q}(P^0_{n,B_n})(\varepsilon), g_0))] = 0$$

and

$$\mathbb{E}_{B_n}[P_0(D^*(Q_0, \hat{g}(P^0_{n,B_n})) - D^*(Q_0, g_0))] = 0,$$

which implies that

$$\hat{\Psi}(P_n) - \Psi(Q(P_0)) = (P_n - P_0)D^*(\hat{Q}(P_0)(\varepsilon_0), \hat{g}(P_0)) + o_P\left(\frac{1}{\sqrt{n}}\right) + O_P(||\hat{\Psi}(P_n) - \Psi(Q(P_0))||^2).$$

Taking norms of both sides, we have that

$$||(P_n - P_0)D^*(\hat{Q}(P_0)(\varepsilon_0), \hat{g}(P_0))|| = O_P\left(\frac{1}{\sqrt{n}}\right)$$

by a standard application of CLT. Hence, by applying the triangle inequality to the right-hand side, we get that

$$||\hat{\Psi}(P_n) - \Psi(Q(P_0))|| \leq O_P\left(\frac{1}{\sqrt{n}}\right) + o_P\left(\frac{1}{\sqrt{n}}\right) + O_P(||\hat{\Psi}(P_n) - \Psi(Q(P_0))||^2).$$

Let $X_n = ||\hat{\Psi}(P_n) - \Psi(Q(P_0))||$, so

$$X_n \leq O_P\left(\frac{1}{\sqrt{n}}\right) + O_P(X_n^2) \implies X_n \leq A_n + B_n X_n^2, A_n = O_P\left(\frac{1}{\sqrt{n}}\right), B_n = O_P(1).$$

Fix some $\delta > 0$ - then, there exists some finite M such that, with probability $1 - \delta$, $A_n \leq \frac{M}{\sqrt{n}}, B_n \leq M$ for n sufficiently large, so

$$X_n \leq \frac{M}{\sqrt{n}} + MX_n^2.$$

For fixed $K > 2M$, if $X_n > \frac{K}{\sqrt{n}}$, then $\frac{K}{\sqrt{n}} \leq \frac{M}{\sqrt{n}} + MX_n^2$, but if $X_n \leq \frac{1}{2M}$, then $MX_n^2 \leq \frac{X_n}{2}$, so $X_n \leq \frac{2M}{\sqrt{n}}$,

which contradicts $X_n > \frac{K}{\sqrt{n}}$. Hence (assuming consistency of X_n so that $X_n > \frac{1}{2M}$ occurs with small probability), we have $X_n = O_P(\frac{1}{\sqrt{n}})$.

Since $X_n = O_P(\frac{1}{\sqrt{n}})$, $X_n^2 = o_P(\frac{1}{\sqrt{n}})$, which implies the desired asymptotic linearity as $O_P(||X_n||^2)$ combines with the $o_P(\frac{1}{\sqrt{n}})$ in our previous expansion to get

$$\hat{\Psi}(P_n) - \Psi(Q(P_0)) = (P_n - P_0)D^*(\hat{Q}(P_0)(\varepsilon_0), \hat{g}(P_0)) + o_P\left(\frac{1}{\sqrt{n}}\right).$$

The asymptotic efficiency statement for if $\hat{Q}(P_0)(\varepsilon_0) = Q_0$ follows from a substitution. □

Finally, if we do not have $\hat{g}(P_n) = g_0$, but rather the weaker statement that $\hat{g}(P_0) = g_0$, we can still get an asymptotic linearity statement with a slightly different influence function, under a few qualifying assumptions:

Theorem 7.2.9 *Say that $\hat{g}(P_0) = g_0$ and that $\hat{Q}(P_n)(\varepsilon_n)$ converges to some limit $\overline{Q}$, which is not necessarily equal to Q_0. Assume that*

$$\begin{aligned}\mathbb{E}_{B_n}[P_0(D^*(\hat{Q}(P^0_{n,B_n})(\varepsilon_n), \hat{g}(P^0_{n,B_n})) - D^*(\hat{Q}(P^0_{n,B_n})(\varepsilon_n), g_0))] \\ -\mathbb{E}_{B_n}[P_0(D^*(\overline{Q}, \hat{g}(P^0_{n,B_n})) - D^*(\overline{Q}, g_0))] = o_P\left(\frac{1}{\sqrt{n}}\right)\end{aligned}$$

and that there exists some mean zero function $IC'(P_0) \in L_0^2(P_0)$ such that

$$\begin{aligned}\mathbb{E}_{B_n}[P_0(D^*(\overline{Q}, \hat{g}(P^0_{n,B_n})) - D^*(\overline{Q}, g_0))] - \mathbb{E}_{B_n}[P_0(D^*(Q_0, \hat{g}(P^0_{n,B_n})) - D^*(Q_0, g_0))] \\ = (P_n - P_0)IC'(P_0) + o_P\left(\frac{1}{\sqrt{n}}\right).\end{aligned}$$

Then, we still have asymptotic linearity of $\hat{\Psi}(P_n)$, with the expansion

$$\hat{\Psi}(P_n) - \psi_0 = (P_n - P_0)(D^*(\hat{Q}(P_0)(\varepsilon_0), g_0) + IC'(P_0)) + o_P\left(\frac{1}{\sqrt{n}}\right).$$

Proof: We begin with the expansion of $\hat{\Psi}(P_n) - \Psi(Q(P_0))$ from Lemma 7.2.8, but we also add and subtract $\mathbb{E}_{B_n}[P_0(D^*(\overline{Q}, \hat{g}(P^0_{n,B_n})) - D^*(\overline{Q}, g_0))]$ from the right-hand side to get:

$$\begin{aligned}\hat{\Psi}(P_n) - \Psi(Q(P_0)) = (P_n - P_0)D^*(\hat{Q}(P_0)(\varepsilon_0), \hat{g}(P_0)) + o_P(\frac{1}{\sqrt{n}}) \\ + \mathbb{E}_{B_n}[P_0(D^*(\hat{Q}(P^0_{n,B_n})(\varepsilon), \hat{g}(P^0_{n,B_n})) - D^*(\hat{Q}(P^0_{n,B_n})(\varepsilon), g_0))] \\ - \mathbb{E}_{B_n}[P_0(D^*(\overline{Q}, \hat{g}(P^0_{n,B_n})) - D^*(\overline{Q}, g_0))] \\ + \mathbb{E}_{B_n}[P_0(D^*(\overline{Q}, \hat{g}(P^0_{n,B_n})) - D^*(\overline{Q}, g_0))] \\ - \mathbb{E}_{B_n}[P_0(D^*(Q_0, \hat{g}(P^0_{n,B_n})) - D^*(Q_0, g_0))] \\ + O_P(||\hat{\Psi}(P_n) - \Psi(Q(P_0))||^2).\end{aligned}$$

Our first assumption tells us that

$$\mathbb{E}_{B_n}[P_0(D^*(\hat{Q}(P^0_{n,B_n})(\varepsilon), \hat{g}(P^0_{n,B_n})) - D^*(\hat{Q}(P^0_{n,B_n})(\varepsilon), g_0))] \\ -\mathbb{E}_{B_n}[P_0(D^*(\overline{Q}, \hat{g}(P^0_{n,B_n})) - D^*(\overline{Q}, g_0))] = o_P\left(\frac{1}{\sqrt{n}}\right),$$

and our second assumption tells us that

$$\mathbb{E}_{B_n}[P_0(D^*(\overline{Q}, \hat{g}(P^0_{n,B_n})) - D^*(\overline{Q}, g_0))] - \mathbb{E}_{B_n}[P_0(D^*(Q_0, \hat{g}(P^0_{n,B_n})) - D^*(Q_0, g_0))] \\ = (P_n - P_0)\mathrm{IC}'(P_0) + o_P\left(\frac{1}{\sqrt{n}}\right).$$

Plugging these two assumptions into our initial expansion, we get

$$\hat{\Psi}(P_n) - \psi_0 = (P_n - P_0)(D^*(\hat{Q}(P_0)(\varepsilon_0), \hat{g}(P_0)) + \mathrm{IC}'(P_0)) + o_P\left(\frac{1}{\sqrt{n}}\right) + O_P\left(||\hat{\Psi}(P_n) - \psi_0||^2\right).$$

Then, by taking norms on both sides, by an identical argument to the previous lemma, we get the desired asymptotically linear expansion. □

Chapter 8

Collaborative TMLE

The first extension to the TMLE we will discuss is the *collaborative TMLE* (C-TMLE), which is a refinement of the regular TMLE that focuses on estimating the nuisance in such a way that we most effectively estimate the causal parameter of interest as well. In standard TMLE, the nuisance is estimated and the debiasing comes from fluctuations to better estimate the causal parameter. However, the choice of the nuisance can affect our choice of fluctuation that we use to choose targeting directions. It stands to reason that, by estimating the nuisance parameter well, we may be able to further decrease bias. This is the idea behind the collaborative TMLE.

The exposition of this section largely comes from (M. J. van der Laan & Gruber, 2010). An example of the C-TMLE used in practice, to genomic data, is given in (Gruber & van der Laan, 2010).

8.1 Motivating C-TMLE

We begin by describing the TMLE for a standard censored data model which satisfies coarsening at random. More on this particular setting can be found in (Laan & Robins, 2003). Let O be our observations based on full data X and censoring mechanism C, so $O = \Phi(C, X)$, and say $O \sim P_0$. Via coarsening at random, we can assume the density factors as $dP_0(O) = Q_0(X)g_0(O|X)$, where $g_0(O|X) = h(O)$ for some measurable function h.

The factorization of our density implies that any semiparametric model $\mathcal{M}$ for P_0 can be created from a model $\mathcal{Q}$ for Q_0 and a model $\mathcal{G}$ for g_0. In what follows, we will use g_0 to denote both the distribution of O given X and the distribution of C given X, since the distribution of C given X identifies the distribution of O given X.

Suppose our parameter of interest is $\Psi : \mathcal{M} \to \mathbb{R}^d$, and assume it is pathwise differentiable with canonical gradient $D^*(P) \in L_0^2(P)$, so that, for a rich family of parametric submodels $\{P_\varepsilon : \varepsilon\}$, there exists some score $S \in L_0^2(P)$ such that

$$\frac{\partial}{\partial \varepsilon}\Psi(P_\varepsilon)\bigg|_{\varepsilon=0} = \mathbb{E}_P[D^*(P)S].$$

Since $dP = Qg$ in our model, we write $D^*(P) = D^*(Q, g)$ as well. Also assume that Ψ is a parameter of Q alone, so it can be determined from the full data distribution.

A consequence of this assumption is that, when performing TMLE, we can choose a fluctuation model

$\{P_\varepsilon : \varepsilon\}$ which only fluctuates Q - hence, we can choose a submodel $\{Q_{g,\varepsilon} : \varepsilon\}$ and define $dP_\varepsilon = Q_{g,\varepsilon}g$. We index $Q_{g,\varepsilon}$ to denote that the fluctuation for Q is also indexed by g. The standard TMLE proceeds as follows:

1. Begin with an initial estimator Q^0 of Q_0.
2. Define $Q^1 = Q_{g,\varepsilon^1}$, where $\varepsilon^1 = \text{argmin}_\varepsilon P_n L(Q^0_{g,\varepsilon}(P_n))$ and $O(Q) = -\log Q$ is the log-likelihood loss.
3. Keep repeating this, so at step k, we define $\varepsilon^k = \text{argmin}_\varepsilon P_n L(Q^{k-1}_{g,\varepsilon}(P_n))$. Repeat until $\varepsilon^k \approx 0$.
4. The resulting TMLE is calculated by plugging in the empirical distribution to the final estimate $Q^* = Q^k$.

Hence, given an initial estimator Q^0_n for Q_0 and an estimator g_n for g_0, running the TMLE to convergence will produce an estimator (Q^*_n, g_n) which solves

$$P_n D^*(Q^*_n, g_n) = 0,$$

as in Theorem 5.1.2.

In particular, we have many different choices for our TMLE, indexed by different initial estimators, and we want to know which TMLE we should pick. We can perform this selection via a loss-based cross-validation - in particular, as in CV-TMLE, let $B_n \in \{0,1\}^n$ be a random variable that splits the data into a training sample (where $B_n(i) = 0$) and a validation sample (where $B_n(i) = 1$), with associated empirical distributions P^0_{n,B_n} and P^1_{n,B_n}, respectively. Assuming access to some loss function $L^*(Q)$ for Q_0 such that $Q_0 = \text{argmin}_Q P_0 L^*(Q)$ (for example, the log-likelihood loss suffices), if our different TMLEs are indexed as $\hat{Q}^*_k(P_n)$ across k, we can select a TMLE by selecting the one which minimizes the loss L^* on the held-out validation set:

$$\hat{k}(P_n) = \text{argmin}_k \mathbb{E}_{B_n} P^1_{n,B_n} L^*(\hat{Q}^*_k(P^0_{n,B_n})),$$

with the selected TMLE then being $\hat{Q}^*_n = \hat{Q}^*_{\hat{k}(P_n)}(P_n)$ (plugging in the empirical distribution).

To see that this cross-validation is actually useful, we note the following result, due to (Dudoit & van der Laan, 2005) (we do not prove this here):

Theorem 8.1.1 *Let $L^*(Q)$ be a loss function such that:*

1. *There exists some universal constant M_1 such that $\sup_{O,Q} |L^*(Q) - L^*(Q_0)|(O) < M_1 < \infty$.*
2. *There exists some universal constant M_2 such that $\sup_Q \frac{Var_{P_0}(L^*(Q) - L^*(Q_0))}{P_0(L^*(Q) - L^*(Q_0))} \leq M_2$.*

Then, we have the following oracle inequality. For any $\delta > 0$,

$$\mathbb{E}_{B_n}[P_0(L(\hat{Q}_{\hat{k}}(P^0_{n,B_n}) - L(Q_0)))] \leq (1+2\delta)\mathbb{E}_{B_n}[\min_k P_0(L\hat{Q}_k(P^0_{n,B_n})) - L(Q_0)] + 2C(M_1, M_2, \delta)\frac{1 + \log K(n)}{np},$$

where

$$C(M_1, M_2, \delta) = 2(1+\delta)^2\left(\frac{M_1}{3} + \frac{M_2}{\delta}\right)$$

is a constant and $K(n)$ is the number of cross-validation candidates and np is the expected size of the validation set.

Remark 8.1.2 *The second assumption in the preceding theorem is essentially equivalent to the fact that L^* has associated dissimilarity $d(Q, Q_0) := P_0(L^*(Q) - L^*(Q_0))$ that is quadratic in distance between Q and Q_0. This is known to hold, for instance, for log-likelihood and weighted square-loss.*

The upshot of the preceding theorem is that, assuming $K(n)$ is polynomial in n, so that $\log K(n) = O(\log n)$, then, if the oracle term (first term) dominates the second, we get that the cross-validated selector is asymptotically equivalent to the oracle selector, and if not, then the cross-validated selector achieves $\frac{\log n}{n}$ convergence rates for Q with respect to the dissimilarity $d(Q, Q_0)$ induced by the loss L (when noting that, by assumption, the dissimilarity is quadratic in $||Q - Q_0||$, by the preceding remark).

Remark 8.1.3 *The preceding result on cross-validation is particularly strong in that generalizations hold for estimated losses L_n which approximate some fixed loss L.*

Definition 8.1.4 *A **targeted loss function** is a modification to some loss function $L(Q)$ such that we apply it to the targeted estimate of Q - in particular, we replace $L(\hat{Q})$ with $L(\hat{Q}^*)$, where $\hat{Q}^*$ is the TMLE beginning with initial estimator $\hat{Q}$.*

Definition 8.1.5 *A **collaborative targeted maximum likelihood estimator** (C-TMLE) is one in which the nuisance g_n is estimated in conjunction (collaboration) with the estimate for Q_n, instead of separately.*

One can now imagine using a targeted loss function to select among different collaborative TMLEs that are indexed by different initial estimators (for both Q and g). We will describe this procedure more formally in the next section.

Definition 8.1.6 *The **entropy** of the fit $\hat{Q}(P_n)$ with respect to some loss function $L(Q)$ is defined as $P_n L(\hat{Q}(P_n))$. Entropy can be defined similarly for estimates $\hat{g}(P_n)$ as well, with respect to some loss L_1 for g.*

In essence, we would like our loss function L for Q to denote how well $\Psi(Q)$ approximates $\Psi(Q_0)$, and our loss function L_1 to denote how close we are to the optimal parametric submodel implied by g_0.

How the C-TMLE procedure works, roughly, is to couple decreases in L-entropy for Q with decreases in L_1-entropy for g, resulting in targeting and fluctuations that are closer and closer to optimal. The quality of the targeted estimator can be improved by ensuring the nuisance estimator g also helps identify which directions Q should be updated. Hence, by updating the two in "collaboration," estimation can be improved.

8.2 The C-TMLE Procedure

With the preceding motivation in mind, the C-TMLE procedure for censored data mechanisms then proceeds, formally, as follows:

1. Estimate Q_0 via some estimate Q_n based on a loss function $L(Q)$.
2. Let $L^*(Q)$ be the associated targeted loss function for $L(Q)$, and $L_1(g)$ be a loss function for g_0.
3. Choose some index i, and for each $i \in I$, let $g_{n,i}$ be a candidate estimator for g_0. Define $d(i) = P_n L_1(g_{n,i})$ as the entropy of $g_{n,i}$.

4. Select some deterministic sequence $d^0 > d^1 > \cdots > d^K$ of entropy values, and begin with an estimator g_n^0 that has entropy larger than d^0. Calculate the corresponding TMLE Q_n^{0*} from initial estimator Q_n, based on the fluctuation induced by g_n^0.

5. At step k, say we currently have some TMLE (g_n^k, Q_n^{k*}), with g^k having entropy larger than d^k and Q_n^{k*} being the TMLE from initial estimate Q_n^k with fluctuation defined by g_n^k. We search among a set of candidate estimators $g_{n,i}$ such that the entropy of $g_{n,i}$ is between d^k and d^{k+1} and choose the estimator $g_{n,i}$ which minimizes the targeted loss $P_n L^*(Q^{k*}_{n,g_{n,i}})$, using the initial estimator Q_n^k.

   (a) If $P_n L^*(Q^{k*}_{n,g_{n,i}}) \leq P_n L^*(Q_n^{k*})$ or $P_n L(Q^{k*}_{n,g_{n,i}}) \leq P_n L(Q_n^{k*})$ (where $L(Q)$ is just the log-likelihood loss), so the fit of the TMLE improves, then set $g_n^{k+1} = g_{n,i}$ and $Q_n^{(k+1)*} = Q^{k*}_{n,g_{n,i}}$, keeping our initial estimator $Q_n^{k+1} = Q_n^k$ for the next stage.

   (b) Otherwise, reject $g_{n,i}$ as a fluctuation based on the old initial estimator, and reset the initial estimator to the current TMLE, so $Q_n^k := Q_n^{k*}$. Then, rerun the same search over candidates $g_{n,i}$ in the same entropy range. Since the targeting step now fluctuates through Q_n^{k*}, and since $\varepsilon = 0$ corresponds to Q_n^{k*}, the empirical likelihood maximizer must satisfy $P_n L(Q^{k*}_{n,g_{n,i}}) \geq P_n L(Q_n^{k*})$, where $L(Q)$ is the log-likelihood loss. Hence, the monotonicity condition now holds, and we can accept the update and move to the next step.

6. After running this algorithm for K steps, we get a sequence of TMLEs (g_n^k, Q_n^{k*}) for $k = 0, 1, 2, \ldots, K$. Furthermore, we know that, for each k, Q_n^{k*} either improves fit with respect to the targeted loss function L^* or with respect to log-likelihood, and by definition of our deterministic sequence d^k, we know that the entropy of g_n^k also decreases with k.

7. We use cross-validation to select which of the K estimators we actually use. In particular, using B_n to index our train-validation split as usual, we set

$$k_n = \text{argmin}_k \mathbb{E}_{B_n}[P^1_{n,B_n}(L^*(\hat{Q}^{k*}(P^0_{n,B_n})))],$$

   and our collaborative estimate $(Q_n^{k_n*}, g_n^{k_n})$ is the C-TMLE. The point estimate for our functional would then just be given by substitution $\Psi(Q_n^{k_n*})$.

We make a few remarks on this procedure:

Remark 8.2.1 *As with other TMLE templates, we can easily use a general loss function $L(Q)$ to extend this to minimum loss-based estimation rather than maximum likelihood estimation.*

Remark 8.2.2 *Since the final C-TMLE is a TMLE applied to an appropriate fluctuation, it solves the efficient influence curve equation, so*

$$P_n D^*(Q_n^*, g_n) = 0.$$

Remark 8.2.3 *One still notes that there is dependence in this procedure on the initial estimator Q_n^0 as well as the indexing sets over which we search for $g_{n,i}$ at each stage. Given candidate C-TMLEs $(Q_{nj}*, g_{nj})$ based on these choices, a natural way to choose among them is to cross-validate based on the empirical variance of the canonical gradient, picking*

$$j_n = argmin_j \mathbb{E}_{B_n}[(P^1_{n,B_n}(D^*(\hat{Q}^*_j(P^0_{n,B_n}), \hat{g}_j(P^0_{n,B_n}))))^2],$$

where B_n denotes our train-validation split.

Example 8.2.4 *As an example of a targeted loss function, one can imagine targeting the standard log-likelihood loss. Say we want to choose over many different C-TMLEs $P_n \to \hat{Q}_k^*(P_n)$. Then, we can add a term to our loss that estimates the mean-square error of the (limit of) the substitution estimator $\hat{\Psi}(\hat{Q}(P_n))$. This has the effect of keeping the log-likelihood as the dominant term while penalizing particularly bad estimates of the target parameter. One can estimate this mean-square error, for instance, through the variance of the efficient influence curve of the target parameter. This type of targeted loss function would also be used during the actual estimation sequence in step 5 of the procedure as well.*

8.3 Consistency and Asymptotic Linearity

Theorem 8.3.1 *The C-TMLE is consistent.*

Proof: Without loss of generality, we can assume that, at every step of the sequence, the initial estimator is the targeted estimator from the previous step (otherwise, we can just take the subsequence of steps for which this is the case). For $k = 1, \ldots, K$, let $g_k = \lim_{n\to\infty} g_n^k$ (so $g_K = g_0$), and consider the corresponding limits Q^*_{k,g_k} of the TMLEs Q_n^{*k} that form our sequence. Since each element in the sequence is a targeting of the previous, the sequence $P_n \log Q^*_{n,k,g_{nk}}$ is non-decreasing in k. Taking the limit as $n \to \infty$, we get $P_0 \log Q^*_{kg_k}$ is non-decreasing in k as well.

Now, assuming a uniformly bounded loss function (so $L(Q) = \log Q$ is uniformly bounded for all candidates Q), we can apply Theorem 8.1.1 to get that the cross-validation selector of k which ends the C-TMLE procedure is asymptotically equivalent to the oracle selector $\tilde{k}_n = \text{argmax}_k P_0 \log Q^*_{k,g_{nk}}$. Taking the limit, if n is large enough, this selector is $\tilde{k} = \text{argmax}_k P_0 \log Q^*_{kg_k}$.

Since the sequence $P_0 \log Q^*_{kg_k}$ is nondecreasing in k, a maximum is attained at $k = K$, giving $P_0 \log Q^*_{Kg_0}$, for which $\Psi(Q^*_{Kg_0}) = \psi_0$, implying consistency. However, say $\tilde{k} < K$, which implies

$$P_0 \log Q^*_{\tilde{k}g_{\tilde{k}}} = P_0 \log Q^*_{(\tilde{k}+1)g_{(\tilde{k}+1)}} = \cdots = P_0 \log Q^*_{Kg_K},$$

by monotonicity. Since each estimator is a TMLE of the previous, the ε fluctuations at each step must be 0, meaning

$$Q^*_{\tilde{k}} = Q^*_{\tilde{k}+1} = \cdots = Q^*_K.$$

However, Q^*_K itself is a TMLE with nuisance parameter g_0, so $\varepsilon \mapsto P_0 \log Q^*_{K,g_0}(\varepsilon)$ is maximized at $\varepsilon = 0$. Since $Q^*_K = Q^*_{\tilde{k}}$, we must then have $\varepsilon \mapsto P_0 \log Q^*_{\tilde{k},g_0}(\varepsilon)$ is also maximized at $\varepsilon = 0$. Taking the derivative at $\varepsilon = 0$, this means $P_0 D^*(Q^*_{\tilde{k}}, g_0) = 0$, so the efficient influence curve equation is solved, which implies consistency of $\Psi(Q^*_{n,\tilde{k}})$. Hence, whatever the cross-validator selects, we will have consistency. □

One can also prove an asymptotic linearity result for the C-TMLE:

Theorem 8.3.2 *Let $\psi_n = \Psi(Q_n^*)$ denote the final point estimate of the C-TMLE. Let $Q^* = \lim_{n\to\infty} Q_n^*$, and let g_n be an estimator with $g_0 = \lim_{n\to\infty} g_n$. Under the following assumptions:*

1. *The efficient influence curve estimating equation is solved, so $P_n D^*(Q_n^*, g_n, \psi_n) = 0$. (Here, we include ψ in the parameterization as the influence curve typically includes a term depending on Ψ).*

2. *We have* $P_0 D^*(Q^*, g_0, \psi_0) = 0$ *and* $P_0 D^*(Q_n^*, g_0, \psi_0) = o_P(\frac{1}{\sqrt{n}})$.

3. *The estimators are consistent, so*

$$P_0(D^*(Q_n^*, g_n, \psi_n) - D^*(Q^*, g_0, \psi_0))^2 \xrightarrow{p} 0.$$

Consistency also holds if any of Q_n^*, g_n, ψ_n *is initially replaced by its limit, as well.*

4. *The derivative* $c_0 := -\frac{\partial}{\partial \psi_0} P_0 D^*(Q^*, g_0, \psi_0)$ *exists and is invertible.*

5. *With probability tending to 1,* $\{D^*(Q, g, \Psi(Q)) : Q, g\}$ *is a* P_0*-Donsker class, where the set that* (Q, g) *vary over include* $(Q_n^*, g_n), (Q^*, g_n), (Q_n^*, g_0)$.

6. *If* $\Phi(g) = P_0 D^*(Q^*, g, \psi_0)$*, then*

$$\Phi(g_n) - \Phi(g_0) = (P_n - P_0) IC_{g_0} + o_P\left(\frac{1}{\sqrt{n}}\right)$$

for some mean-zero function $IC_{g_0} \in L_0^2(P_0)$.

7. *We have*

$$R_{n1} := P_0(D^*(Q_n^*, g_n, \psi_n) - D^*(Q^*, g_n, \psi_n)) - P_0(D^*(Q_n^*, g_0, \psi_0) - D^*(Q^*, g_0, \psi_0)) = o_P\left(\frac{1}{\sqrt{n}}\right)$$

and

$$R_{n2} := P_0(D^*(Q^*, g_n, \psi_n) - D^*(Q^*, g_0, \psi_n)) - P_0(D^*(Q_n^*, g_n, \psi_0) - D^*(Q^*, g_0, \psi_0)) = o_P\left(\frac{1}{\sqrt{n}}\right).$$

Note that R_{n1} *and* R_{n2} *are both second-order terms.*

Then, ψ_n *is asymptotically linear at* P_0 *with the relation*

$$\psi_n - \psi_0 = (P_n - P_0) IC(P_0) + o_P\left(\frac{1}{\sqrt{n}}\right),$$

where $IC(P_0) = c_0^{-1}(D^*(Q^*, g_0, \psi_0) + IC_{g_0})$.

Proof: Consider Taylor expanding $P_0 D^*(Q^*, g_0, \psi)$ in $\psi_n - \psi_0$. Noting that $P_0 D^*(Q^*, g_0, \psi_0) = 0$, this gives

$$P_0 D^*(Q^*, g_0, \psi_n) = -c_0(\psi_n - \psi_0) + o(|\psi_n - \psi_0|).$$

Since $P_n D^*(Q_n^*, g_n, \psi_n) = 0$, we can write

$$\begin{aligned} -P_0 D^*(Q^*, g_0, \psi_n) &= P_n D^*(Q_n^*, g_n, \psi_n) - P_0 D^*(Q^*, g_0, \psi_n) \\ &= (P_n - P_0) D^*(Q^*, g_0, \psi_n) \\ &+ P_n(D^*(Q_n^*, g_n, \psi_n) - D^*(Q^*, g_n, \psi_n)) \\ &+ P_n(D^*(Q^*, g_n, \psi_n) - D^*(Q^*, g_0, \psi_n)), \end{aligned}$$

where we write $P_n D^*(Q_n^*, g_n, \psi_n) - P_0 D^*(Q^*, g_0, \psi_n)$ as a telescoping sum to get the second equality.

First, consider the term $(P_n - P_0)D^*(Q^*, g_0, \psi_n)$. We can write

$$(P_n - P_0)D^*(Q^*, g_0, \psi_n) = (P_n - P_0)D^*(Q^*, g_0, \psi_0) + (P_n - P_0)(D^*(Q^*, g_0, \psi_n) - D^*(Q^*, g_0, \psi_0)).$$

Define the empirical process $\mathbb{G}_n(f) = \sqrt{n}(P_n - P_0)f$. By assumption 3,

$$P_0(D^*(Q^*, g_0, \psi_n) - D^*(Q^*, g_0, \psi_0))^2 \xrightarrow{p} 0,$$

and by assumption 5, the class $\mathcal{F} := \{D^*(Q, g, \Psi(Q)) : Q, g\}$ is P_0-Donsker and contains both $D^*(Q^*, g_0, \psi_n)$ and $D^*(Q^*, g_0, \psi_0)$. Hence, by Theorem 5.1.5, by similar logic to the proof of Corollary 5.2.3.1, we have

$$(P_n - P_0)(D^*(Q^*, g_0, \psi_n) - D^*(Q^*, g_0, \psi_0)) = o_P(n^{-\frac{1}{2}}),$$

which implies

$$(P_n - P_0)D^*(Q^*, g_0, \psi_n) = (P_n - P_0)D^*(Q^*, g_0, \psi_0) + o_P\left(\frac{1}{\sqrt{n}}\right).$$

For the second term, first write

$$\begin{aligned} P_n(D^*(Q_n^*, g_n, \psi_n) - D^*(Q^*, g_n, \psi_n)) = &(P_n - P_0)(D^*(Q_n^*, g_n, \psi_n) - D^*(Q^*, g_n, \psi_n)) \\ &+ P_0(D^*(Q_n^*, g_n, \psi_n) - D^*(Q^*, g_n, \psi_n)). \end{aligned}$$

Again, by the consistency assumption (3) and the Donsker assumption (5), combined with Theorem 5.1.5, we can write

$$(P_n - P_0)(D^*(Q_n^*, g_n, \psi_n) - D^*(Q^*, g_n, \psi_n)) = o_P\left(\frac{1}{\sqrt{n}}\right).$$

For the second part, we write

$$P_0(D^*(Q_n^*, g_n, \psi_n) - D^*(Q^*, g_n, \psi_n)) = P_0(D^*(Q_n^*, g_0, \psi_0) - D^*(Q^*, g_0, \psi_0)) + R_{n1}.$$

By assumption 2,

$$P_0(D^*(Q_n^*, g_0, \psi_0) - D^*(Q^*, g_0, \psi_0)) = o_P(\frac{1}{\sqrt{n}}),$$

and by assumption 7, $R_{n1} = o_P(\frac{1}{\sqrt{n}})$, so, putting everything together,

$$P_n(D^*(Q_n^*, g_n, \psi_n) - D^*(Q^*, g_n, \psi_n)) = o_P\left(\frac{1}{\sqrt{n}}\right).$$

The third term proceeds similarly, where we begin by writing

$$\begin{aligned} P_n(D^*(Q^*, g_n, \psi_n) - D^*(Q^*, g_0, \psi_n)) = &(P_n - P_0)(D^*(Q^*, g_n, \psi_n) - D^*(Q^*, g_0, \psi_n)) \\ &+ P_0(D^*(Q^*, g_n, \psi_n) - D^*(Q^*, g_0, \psi_n)). \end{aligned}$$

Again, by consistency with respect to g and the Donsker condition, by Theorem 5.1.5, we have

$$(P_n - P_0)(D^*(Q^*, g_n, \psi_n) - D^*(Q^*, g_0, \psi_n)) = o_P\left(\frac{1}{\sqrt{n}}\right).$$

For the second part, we can write

$$P_0(D^*(Q^*, g_n, \psi_n) - D^*(Q^*, g_0, \psi_n)) = P_0(D^*(Q^*, g_n, \psi_0) - D^*(Q^*, g_0, \psi_0)) + R_{n2},$$

where $R_{n2} = o_P(\frac{1}{\sqrt{n}})$. By assumption 6,

$$\Phi(g_n) - \Phi(g_0) = P_0(D^*(Q^*, g_n, \psi_0) - D^*(Q^*, g_0, \psi_0)) = (P_n - P_0)\mathrm{IC}_{g_0} + o_P\left(\frac{1}{\sqrt{n}}\right),$$

so putting everything together, we have

$$P_n(D^*(Q^*, g_n, \psi_n) - D^*(Q^*, g_0, \psi_n)) = (P_n - P_0)\mathrm{IC}_{g_0} + o_P\left(\frac{1}{\sqrt{n}}\right).$$

Putting these three terms together, we have

$$c_0(\psi_n - \psi_0) + o(|\psi_n - \psi_0|) = (P_n - P_0)D^*(Q^*, g_0, \psi_0) + (P_n - P_0)\mathrm{IC}_{g_0} + o_P\left(\frac{1}{\sqrt{n}}\right),$$

so isolating $\psi_n - \psi_0$ gives the desired asymptotic linearity relation

$$\psi_n - \psi_0 = (P_n - P_0)c_0^{-1}(D^*(Q^*, g_0, \psi_0) + \mathrm{IC}_{g_0}) + o_P(|\psi_n - \psi_0|) + o_P\left(\frac{1}{\sqrt{n}}\right),$$

or equivalently,

$$(c_0 + o_P(1))(\psi_n - \psi_0) = (P_n - P_0)(D^*(Q^*, g_0, \psi_0) + \mathrm{IC}_{g_0}) + o_P\left(\frac{1}{\sqrt{n}}\right).$$

Since c_0 is bounded away from 0, and the term $(P_n - P_0)(D^*(Q^*, g_0, \psi_0) + \mathrm{IC}_{g_0})$ is $O_P(\frac{1}{\sqrt{n}})$ (by the Central Limit Theorem), we get $|\psi_n - \psi_0| = O_P(\frac{1}{\sqrt{n}})$, which gives the final desired relation

$$\psi_n - \psi_0 = (P_n - P_0)c_0^{-1}(D^*(Q^*, g_0, \psi_0) + \mathrm{IC}_{g_0}) + o_P\left(\frac{1}{\sqrt{n}}\right).$$

□

Chapter 9

Higher-Order TMLE

So far, we have studied TMLEs which focus on first-order bias correction. However, if the second-order remainder is not small enough under the available nuisance convergence rates, we cannot use standard first-order TMLE methods. This motivates the development of higher-order targeted learning, where we correct higher-order bias terms, only requiring control of higher-order remainder terms. The tradeoff is requiring higher-order differentiability of the target parameter, although this can be sidestepped with approximations of higher-order derivatives, if they exist. A broader treatment of higher-order influence functions and their relation to estimation can be found in (Robins, Li, Tchetgen, van der Vaart, et al., 2008).

In the setting of standard TMLE, consider writing the canonical gradient D^* as $D^{(1)*}$, where the (1) is to emphasize the first-order. Then, our typical TMLE efficiency theorem is of the form:

Assuming that:

1. $D^{(1)*}(P_n^*)$ is in a P_0-Donsker class with probability tending to one
2. $P_0(D^{(1)*}(P_n^*) - D^{(1)*}(P_0))^2 \xrightarrow{p} 0$
3. The remainder term $R_2(P_n^*, P_0)$ is $o_P(\frac{1}{\sqrt{n}})$

we get asymptotic linearity and efficiency, so

$$\hat{\Psi}(P_n^*) - \psi_0 = (P_n - P_0)D^{(1)*}(P_0) + o_P\left(\frac{1}{\sqrt{n}}\right).$$

The idea of the higher-order TMLE (we will present the ideas behind the second-order TMLE in this section, but the results can be extended to even higher orders) is to replace the third condition on the decay rate of R_2 by a condition on the decay rate of R_3, the third-order remainder.

Much of the exposition for this section is from (Carone, Díaz, & van der Laan, 2014). We refer the interested reader to (M. van der Laan, Wang, & van der Laan, 2021) for further exposition.

9.1 Second-Order Scores and Canonical Gradients

As usual, let $L_0^2(P)$ denote the full Hilbert space of square-integrable real-valued functions with mean zero under some P, and let $T(P)$ be the tangent space of a model $\mathcal{M}$ at $P \in \mathcal{M}$.

Definition 9.1.1 *We define $L_0^{2*}(P^2)$ to be the Hilbert space of square-integrable real-valued functions f defined on $\mathcal{O}^2$ which are symmetric and satisfy, almost surely,*

$$\int f(x_1, x)dP(x_1) = \int f(x, x_2)dP(x_2) = 0.$$

Here, the inner product is the double integral

$$\langle f_1, f_2 \rangle_{P_2} = P^2(f_1 f_2) = \int f_1(x_1, x_2) f_2(x_1, x_2) dP(x_1) dP(x_2).$$

We begin with a definition of first-order pathwise differentiability and then extend this, in the natural manner, to second-order:

Definition 9.1.2 *We say that a functional Ψ is **first-order pathwise differentiable** if, for some $P \in \mathcal{M}$ and parametric submodel $\{P_\varepsilon : \varepsilon\} \subseteq \mathcal{M}$ that has **first-order score** $s^{(1)}(P)(o) = \frac{\partial}{\partial \varepsilon} \log p_\varepsilon(o)|_{\varepsilon=0}$, we can write*

$$\Psi(P_\varepsilon) - \Psi(P) = \varepsilon \int D^{(1)}(P)(o) s^{(1)}(o) dP(o) + o(\varepsilon)$$

*for some $D^{(1)}(P) \in L_0^2(P)$. We call $D^{(1)}(P)$ the **first-order gradient** of Ψ at P.*

Recall that we can also write the first-order score as

$$s^{(1)}(P)(o) = \frac{\frac{\partial}{\partial \varepsilon} p_\varepsilon(o)|_{\varepsilon=0}}{p(o)},$$

which leads to the natural extension:

Definition 9.1.3 *The **second-order score** of ε in $\{P_\varepsilon : \varepsilon\}$ at $\varepsilon = 0$ is defined as*

$$s^{(2)}(P)(o) = \frac{\frac{\partial^2}{\partial \varepsilon^2} p_\varepsilon(o)|_{\varepsilon=0}}{p(o)}.$$

Using this, we can extend the notion of first-order pathwise differentiability to second-order as well:

Definition 9.1.4 *Let Ψ be first-order pathwise differentiable, and let $D^{(1)}(P) \in L_0^2(P)$ be a first-order gradient of Ψ. Define*

$$A_1(s^{(1)})(P) = \int D^{(1)}(P)(o) s^{(1)}(o) dP(o)$$

and

$$A_2(s^{(1)}, s^{(2)})(P) = \int D^{(1)}(P)(o) s^{(2)}(o) dP(o) + \iint D^{(2)}(P)(o_1, o_2) s^{(1)}(o_1) s^{(1)}(o_2) dP(o_1) dP(o_2)$$

for some element $D^{(2)}(P) \in L_0^{2*}(P^2)$. If we can write

$$\Psi(P_\varepsilon) - \Psi(P) = \varepsilon A_1(s^{(1)})(P) + \frac{1}{2}\varepsilon^2 A_2(s^{(1)}, s^{(2)}) + o(\varepsilon^2),$$

*then Ψ is **second-order pathwise differentiable** with **second-order gradient** $D^{(2)}(P)$ at P.*

Remark 9.1.5 *As one might expect, a second-order gradient can be computed by computing the first-order gradient of a first-order gradient mapping $P \mapsto D^{(1)}(P)(o)$.*

As the representation given in the preceding definition can be a bit unwieldy, we can trim it down a bit. In particular, recall the second-order Taylor expansion

$$\frac{dP_\varepsilon}{dP} = 1 + \varepsilon s^{(1)} + \frac{\varepsilon^2}{2} s^{(2)} + o(\varepsilon^2) \implies \frac{d(P_\varepsilon - P)}{dP} = \varepsilon s^{(1)} + \frac{\varepsilon^2}{2} s^{(2)} + o(\varepsilon^2).$$

Hence, we can write, for any square-integrable f,

$$(P_\varepsilon - P)f = \int f d(P_\varepsilon - P) = \varepsilon \int f s^{(1)} dP + \frac{\varepsilon^2}{2} \int f s^{(2)} dP + o(\varepsilon^2).$$

Setting $f = D^{(1)}(P)$, we get

$$(P_\varepsilon - P)D^{(1)}(P) = \varepsilon \int D^{(1)}(P) s^{(1)} dP + \frac{\varepsilon^2}{2} \int D^{(1)}(P) s^{(2)} dP + o(\varepsilon^2).$$

Furthermore, we have

$$(P_\varepsilon - P)^2 D^{(2)}(P) = \varepsilon^2 \iint D^{(2)}(P)(o_1, o_2) s^{(1)}(o_1) s^{(1)}(o_2) dP(o_1) dP(o_2) + o(\varepsilon^2),$$

(by taking the leading term), so we can write

$$\begin{aligned}(P_\varepsilon - P)D^{(1)}(P) + \frac{1}{2}(P_\varepsilon - P)^2 D^{(2)}(P) &= \varepsilon \int D^{(1)}(P) s^{(1)} dP \\ &\quad + \frac{\varepsilon^2}{2} \left[\int D^{(1)}(P) s^{(2)} dP + \iint D^{(2)}(P) s^{(1)} s^{(1)} dP dP \right] + o(\varepsilon^2) \\ &= \varepsilon A_1(s^{(1)})(P) + \frac{1}{2}\varepsilon^2 A_2(s^{(1)}, s^{(2)}) + o(\varepsilon^2).\end{aligned}$$

Hence, we can write, for a second-order pathwise differentiable Ψ,

$$\Psi(P_\varepsilon) - \Psi(P) = (P_\varepsilon - P)D^{(1)}(P) + \frac{1}{2}(P_\varepsilon - P)^2 D^{(2)}(P) + o(\varepsilon^2).$$

Since $PD^{(1)}(P) = 0$ and $D^{(2)}(P) \in L_0^{2*}(P^2)$, we can write, even more succinctly,

$$\Psi(P_\varepsilon) - \Psi(P) = P_\varepsilon D^{(1)}(P) + \frac{1}{2} P_\varepsilon^2 D^{(2)}(P) + o(\varepsilon^2).$$

Definition 9.1.6 *We say that Ψ is **strongly second-order differentiable** if, at each $P \in \mathcal{M}$, the second-order expansion*

$$\Psi(P_\varepsilon) - \Psi(P) = P_\varepsilon D^{(1)}(P) + \frac{1}{2} P_\varepsilon^2 D^{(2)}(P) + o(\varepsilon^2)$$

holds uniformly along all parametric submodels $\{P_\varepsilon : \varepsilon\}$ through P.

If Ψ is strongly second-order differentiable, then we can write

$$\Psi(P) - \Psi(P_0) = -P_0 D^{(1)}(P) - \frac{1}{2} P_0^2 D^{(2)}(P) + R_3(P, P_0),$$

where R_3 is a third-order difference term between P and P_0.

All of the preceding theory holds for any valid pair of first- and second-order gradients. Since the eventual 2-TMLE we construct will be asymptotically linear with influence function $D^{(1)}(P_0)$, it makes sense to choose our first-order gradient to be the typical canonical gradient $D^{(1)*}(P)$ (recall that we can get such a first-order canonical gradient by projecting any first-order gradient onto the first-order tangent space $T(P)$). Hence, to define a second-order canonical gradient, it makes sense to define a second-order tangent space:

Definition 9.1.7 *The **second-order tangent space** at P is the mean-square closure of the linear span of products of the form $(o_1, o_2) \mapsto s^{(1)}(o_1)s^{(1)}(o_2)$, where $s^{(1)}$ is the first-order score. Then, the **second-order canonical gradient** is the projection of any second-order gradient onto the second-order tangent space.*

We can calculate such a projection via the following lemma.

Lemma 9.1.8 *Begin with a second-order gradient*

$$D^{(2)}(P)(o_1, o_2) = \int S_{1,x}(P)(o_1)S_{2,x}(P)(o_2)h(P)(x)d\nu(x)$$

that is symmetric, where $S_{1,x}(P), S_{2,x}(P) \in L_0^2(P)$ for each x, a function $x \mapsto h(P)(x)$, and measure ν. Let $S_{1,x}^*(P), S_{2,x}^*(P)$ denote the projections of $S_{1,x}(P), S_{2,x}(P)$ onto the first-order tangent space $T(P)$ respectively. Then,

$$D^{(2)*}(P)(o_1, o_2) = \int S_{1,x}^*(P)(o_1)S_{2,x}^*(P)(o_2)h(P)(x)d\nu(x)$$

is the second-order canonical gradient of Ψ at P.

The proof follows from the following characterization of the second-order tangent space in terms of the first-order tangent space.

Lemma 9.1.9 *Let $\{e_1, e_2, \ldots\}$ be an orthonormal basis of the first-order tangent space $T(P) \subseteq L_0^2(P)$. Then, the second-order tangent space $T^{(2)}(P)$ is*

$$T^{(2)}(P) = \left\{ (o_1, o_2) \mapsto \sum_{i_1, i_2} C(i_1, i_2)e_{i_1}(o_1)e_{i_2}(o_2) : C \in \mathcal{S} \right\},$$

where $\mathcal{S}$ is the set of all symmetric mappings from $\mathbb{N} \times \mathbb{N}$ to $\mathbb{R}$.

Proof: Let H be the aforementioned set, so we want to prove that $T^{(2)}(P) = H$. First, we show that $T^{(2)}(P) \subseteq H$. Recall that $T^{(2)}(P)$ is the mean-square closure of the linear span of products of first-order scores. Take some such product, denoted $(o_1, o_2) \mapsto \int_x S_x(o_1)S_x(o_2)d\mu(x)$. By virtue of the e_i forming a basis of $T(P)$, $S_x \in T(P)$ can be written as $S_x = \sum_i C_x(i)e_i$ for some mapping $C_x : \mathbb{N} \to \mathbb{R}$. Hence, we have

$$\int_x S_x(o_1)S_x(o_2)d\mu(x) = \sum_{i_1, i_2} \int C_x(i_1)C_x(i_2)e_{i_1}(o_1)e_{i_2}(o_2)d\mu(x),$$

meaning that, if we define $C(i_1, i_2) = \int C_x(i_1)C_x(i_2)d\mu(x)$ and noting that our choice of cross-product was arbitrary, we have that the set of products of first-order scores is a subset of H. Since H is equal to the mean-square closure of its linear span, we have $T^{(2)}(P) \subseteq H$.

On the other hand, take some arbitrary $h \in H$ corresponding to some symmetric mapping $C_0 : \mathbb{N}\times\mathbb{N} \to \mathbb{R}$. We want to construct a cross-product of scores $\int_x S_x^\delta(o_1)S_x^\delta(o_2)d\mu^\delta(x)$ that approximates h to some fixed $\delta > 0$. By closure, there exists some $k(\delta) \in \mathbb{N}$ such that

$$\left|\left| \sum_{i_1,i_2>k(\delta)} C_0(i_1,i_2)e_{i_1}(o_1)e_{i_2}(o_2) \right|\right|_{L_2(P)} < \delta.$$

For each $(i_1, i_2) \in \{1,\ldots,k(\delta)\}^2$, let μ^δ put mass $\frac{1}{2}\text{sign}(C_0(i_1,i_2))$ on (i_1,i_2) and 0 elsewhere. For $x = (i_1,i_2)$ in this set, set $C_x^\delta(j) = |C_0(i_1,i_2)|^{\frac{1}{2}}$ for $j = i_1$ or $j = i_2$ and $C_x^\delta(j) = 0$ otherwise. Then, if we let $S_x = \sum_j C_x^\delta(j)e_j$ (clearly a valid element of $T(P)$), we have, by construction:

$$\begin{aligned}
\int_x S_x^\delta(o_1)S_x^\delta(o_2)d\mu^\delta(x) &= \sum_{j_1,j_2} \int C_x^\delta(j_1)C_x^\delta(j_2)d\mu^\delta(x)e_{j_1}(o_1)e_{j_2}(o_2) \\
&= \sum_{j_1,j_2} \sum_{1\le i_1,i_2\le k(\delta)} C_{i_1,i_2}^\delta(j_1)C_{i_1,i_2}^\delta(j_2)d\mu^\delta((i_1,i_2))e_{j_1}(o_1)e_{j_2}(o_2) \\
&= \sum_{1\le i_1,i_2\le k(\delta)} C_0(i_1,i_2)e_{i_1}(o_1)e_{i_2}(o_2),
\end{aligned}$$

which implies that

$$\int_x S_x^\delta(o_1)S_x^\delta(o_2)d\mu^\delta(x) - \sum_{i_1,i_2} C_0(i_1,i_2)e_{i_1}(o_1)e_{i_2}(o_2) = -\sum_{i_1,i_2>k(\delta)} C_0(i_1,i_2)e_{i_1}(o_1)e_{i_2}(o_2),$$

so by the definition of $k(\delta)$, we have approximated h within δ. By the closure of $T^{(2)}(P)$, we have $H \subseteq T^{(2)}(P)$, so we are done. □

We can now prove Lemma 9.1.8:

Proof: Fix an orthonormal basis $\{e_1, e_2, \ldots\}$ of $T(P)$. Fix some x and consider the projection of the map $(o_1,o_2) \mapsto S_{1,x}(o_1)S_{2,x}(o_2)$ onto $T^{(2)}(P)$. This map, evaluated at (o_1,o_2) (where, in the first equality, we use the fact that $\{e_{i_1} \otimes e_{i_2}\}$ spans the second order tangent-space, so we use the projection theorem) is:

$$\begin{aligned}
\Pi((o_1,o_2) \mapsto S_{1,x}(o_1)S_{2,x}(o_2)|T^{(2)}(P))(o_1,o_2) &= \sum_{i_1,i_2} \langle S_{1,x}S_{2,x}, e_{i_1}\otimes e_{i_2}\rangle (e_{i_1}\otimes e_{i_2}) \\
&= \sum_{i_1,i_2} \mathbb{E}_{P^2}[S_{1,x}(O_1)S_{2,x}(O_2)e_{i_1}(O_1)e_{i_2}(O_2)]e_{i_1}(o_1)e_{i_2}(o_2) \\
&= \sum_{i_1,i_2} \mathbb{E}_P[S_{1,x}(O_1)e_{i_1}(O_1)]\mathbb{E}_P[S_{2,x}(O_2)e_{i_2}(O_2)]e_{i_1}(o_1)e_{i_2}(o_2) \\
&= \sum_{i_1} \mathbb{E}_P[S_{1,x}(O_1)e_{i_1}(O_1)]e_{i_1}(o_1) \\
&\cdot \sum_{i_2} \mathbb{E}_P[S_{2,x}(O_1)e_{i_2}(O_2)]e_{i_2}(o_2) \\
&= \Pi(o \mapsto S_{1,x}(o)|T(P))(o_1) \cdot \Pi(o \mapsto S_{2,x}(o)|T(P))(o_2).
\end{aligned}$$

Since projections combine linearly, we get the desired characterization. □

9.2 The 2-TMLE

For a 2-TMLE to be accurately constructed, we do not actually require a full second-order canonical gradient, but rather, we can get away with imposing a set of weaker conditions:

Definition 9.2.1 *An element $D^{(2)}(P) \in L_0^2(P^2)$ is a **second-order partial canonical gradient** of Ψ, with first-order canonical gradient $D^{(1)}$, if:*

1. $\Psi(P) - \Psi(P_0) = -P_0 D^{(1)}(P) - \frac{1}{2} P_0^2 D^{(2)}(P) + R_3(P, P_0)$
2. *Either of the restriction mappings $o_1 \mapsto D^{(2)}(P)(o_1, o)$ or $o_2 \mapsto D^{(2)}(P)(o, o_2)$ lie in the first-order tangent space $T(P)$ for each o.*

Suppose that we have access to a second-order partial canonical gradient $D^{(2)}(P)$ of Ψ, along with a first-order canonical gradient $D^{(1)}(P)$. Assume, WLOG, that $o_1 \mapsto D^{(2)}(P)(o_1, o) \in T(P)$ for each o. Then, taking a linear combination, we have

$$o \mapsto \overline{D}_n^{(2)}(P)(o) := \frac{1}{n} \sum_{j=1}^{n} D^{(2)}(P)(o, O_j) \in T(P),$$

which means that the statistic

$$\frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} D^{(2)}(P)(O_i, O_j) = \frac{1}{n} \sum_{i=1}^{n} \overline{D}_n^{(2)}(P)(O_i)$$

is the empirical mean of a score at P. Hence, if we define a TMLE with a parametric submodel through P whose score vector components span $D^{(1)*}(P)$ and $\overline{D}_n^{(2)}(P)$ (possible since both are in $T(P)$), we can create a targeted estimate P_n^* which satisfies

$$P_n D^{(1)*}(P_n^*) = P_n^2 D^{(2)}(P_n^*) = 0.$$

The final algorithm to construct a 2-TMLE is essentially the same as that for the standard TMLE, but with a larger parametric submodel:

1. Begin with some initial estimator P_n^0 of P_0.
2. At each step k, construct a 2-dimensional parametric submodel $\{P_n^k(\varepsilon_1, \varepsilon_2) : (\varepsilon_1, \varepsilon_2)\}$ such that, at $(0, 0)$, the score vectors span $D^{(1)}(P_n^k)$ and $\overline{D}_n^{(2)}(P_n^k)$.
3. Choose
$$(\varepsilon_1^k, \varepsilon_2^k) = \text{argmax}_{\varepsilon_1, \varepsilon_2} P_n \log p_n^k(\varepsilon_1, \varepsilon_2).$$
4. Set $P_n^{k+1} = P_n^k(\varepsilon_1^k, \varepsilon_2^k)$.
5. Repeat until $\varepsilon_1^k, \varepsilon_2^k \approx 0$ and, at that point, output the final substitution $\psi_n^* = \Psi(P_n^*)$.

In many cases, the target parameter may not be twice-differentiable, in which case we can use an approximation $D_h^{(2)}$, indexed by h, for the second-order canonical gradient. In particular, we define the approximation $D_h^{(2)}$ such that

$$R_2(P_n^*, P_0) = -\frac{1}{2} \lim_{h \to 0} P_0^2 D_h^{(2)}(P_n^*) + R_3(P_n^*, P_0)$$

for a third-order remainder term $R_3(P_n^*, P_0)$. Hence, defining the bias term $B_n(h)$, we have

$$R_2(P_n^*, P_0) = -\frac{1}{2} P_0^2 D_h^{(2)}(P_n^*) + B_n(h) + R_3(P_n^*, P_0), B_n(h) = \frac{1}{2}(P_0^2 D_h^{(2)}(P_n^*) - \lim_{h \to 0} P_0^2 D_h^{(2)}(P_n^*)).$$

Then, the 2-TMLE $\Psi(P_n^*)$ with approximate gradients satisfies

$$\Psi(P_n^*) - \Psi(P_0) = (P_n - P_0) D^{(1)*}(P_n^*) - \frac{1}{2} P_0^2 D_h^{(2)}(P_n^*) + B_n(h) + R_3(P_n^*, P_0).$$

9.3 Asymptotic Linearity

The asymptotic linearity theorem for the 2-TMLE proceeds under the following conditions:

Theorem 9.3.1 *Let Ψ, with first-order canonical gradient $D^{(1)*}(P)$ and second-order partial canonical gradient $D^{(2)}(P)$ satisfy the second-order expansion*

$$\Psi(P) - \Psi(P_0) = -P_0 D^{(1)*}(P) - \frac{1}{2} P_0^2 D^{(2)}(P) + R_3(P, P_0),$$

and suppose the TMLE P_n^* satisfies

$$P_n D^{(1)*}(P_n^*) = 0, P_n^2 D^{(2)}(P_n^*) = P_n \overline{D}_n^{(2)}(P_n^*) = 0.$$

Under the following assumptions:

1. *$D^{(1)*}(P_n^*)$ is in a P_0-Donsker class $\mathcal{F}$ with probability tending to one,*
2. *$P_0[D^{(1)*}(P_n^*) - D^{(1)*}(P_0)]^2 = o_P(1)$,*
3. *$(P_n^2 - P_0^2) D^{(2)}(P_n^*) = o_P\left(\frac{1}{\sqrt{n}}\right)$,*
4. *$R_3(P_n^*, P_0) = o_P\left(\frac{1}{\sqrt{n}}\right)$,*

the 2-TMLE ψ_n^* is asymptotically linear with influence function $D^{(1)*}(P_0)$ (and is hence asymptotically efficient as well).

Proof: This theorem is actually a special case of the next theorem with $B_n(h_n) = 0$, so we defer the proof. □

To actually establish the condition $(P_n^2 - P_0^2) D^{(2)}(P_n^*) = o_P(\frac{1}{\sqrt{n}})$, the following lemma is typically used instead, whose conditions may be easier to verify:

Lemma 9.3.2 *Under the following assumptions:*

1. *The mappings*
$$o \mapsto \int D^{(2)}(P_n^*)(o_1, o)dP_0(o_1), o \mapsto \int D^{(2)}(P_n^*)(o, o_2)dP_0(o_2)$$
are in a P_0-Donsker class $\mathcal{G}$ with probability tending to one,

2. *We have*
$$J_{n,1}^* := \int \left[\int D^{(2)}(P_n^*)(o_1, o_2)dP_0(o_1)\right]^2 dP_0(o_2) \xrightarrow{p} 0$$
and
$$J_{n,2}^* = \int \left[\int D^{(2)}(P_n^*)(o_1, o_2)dP_0(o_2)\right]^2 dP_0(o_1) \xrightarrow{p} 0,$$

3. $(P_n - P_0)^2 D^{(2)}(P_n^*) = o_P(\frac{1}{\sqrt{n}})$,

we have $(P_n^2 - P_0^2)D^{(2)}(P_n^*) = o_P(\frac{1}{\sqrt{n}})$.

To prove this lemma, we will use Lemma 19.24 from (Van der Vaart, 2000), which we state here:

Lemma 9.3.3 *If $\mathcal{F}$ is a P-Donsker class of measurable functions and f_n is a sequence of random functions in $\mathcal{F}$ such that, for some $f_0 \in L_2(P)$, $\int (f_n(x) - f_0(x))^2 dP(x) \xrightarrow{p} 0$, then $\mathbb{G}_n(f_n - f_0) \xrightarrow{p} 0$.*

Proof: We write

$$(P_n^2 - P_0^2)D^{(2)}(P_n^*) = (P_n - P_0 + P_0)^2 D^{(2)}(P_n^*) - P_0^2 D^{(2)}(P_n^*).$$

Expanding this out yields

$$(P_n^2 - P_0^2)D^{(2)}(P_n^*) = (P_n - P_0)^2 D^{(2)}(P_n^*) + (P_n - P_0) \otimes P_0 D^{(2)}(P_n^*) + P_0 \otimes (P_n - P_0)D^{(2)}(P_n^*).$$

For the first cross-term, we have

$$\begin{aligned}((P_n - P_0) \otimes P_0)D^{(2)}(P_n^*) &= \iint D^{(2)}(P_n^*)(o_1, o_2)d(P_n - P_0)(o_1)dP_0(o_2) \\ &= \int \left[\int D^{(2)}(P_n^*)(o_1, o_2)dP_0(o_2)\right] d(P_n - P_0)(o_1) \\ &= (P_n - P_0)D_1^{(2)}(P_n^*),\end{aligned}$$

where we define
$$D_1^{(2)}(P)(o) := \int D^{(2)}(P)(o, o_2)dP_0(o_2).$$
We use Fubini's theorem to switch the order of integration. Similarly, if we define
$$D_2^{(2)}(P)(o) := \int D^{(2)}(P)(o_1, o)dP_0(o_1),$$
we have
$$(P_0 \otimes (P_n - P_0))D^{(2)}(P_n^*) = (P_n - P_0)D_2^{(2)}(P_n^*).$$

Hence, we can write

$$(P_n^2 - P_0^2)D^{(2)}(P_n^*) = (P_n - P_0)\left[D_1^{(2)}(P_n^*) + D_2^{(2)}(P_n^*)\right] + (P_n - P_0)^2 D^{(2)}(P_n^*).$$

Consider the term $(P_n - P_0)\left[D_1^{(2)}(P_n^*) + D_2^{(2)}(P_n^*)\right]$. The first two assumptions mirror those of Lemma 9.3.3, and applying it yields

$$(P_n - P_0)D_1^{(2)}(P_n^*) = o_P\left(\frac{1}{\sqrt{n}}\right), (P_n - P_0)D_2^{(2)}(P_n^*) = o_P\left(\frac{1}{\sqrt{n}}\right).$$

Then, by the third assumption, we have

$$(P_n^2 - P_0^2)D^{(2)}(P_n^*) = o_P\left(\frac{1}{\sqrt{n}}\right) + o_P\left(\frac{1}{\sqrt{n}}\right),$$

as desired. □

If the parameter of interest does not actually admit a second-order gradient due to not being differentiable enough, we can still get asymptotic linearity of the TMLE even with just an approximate second-order partial canonical gradient, as the following theorem shows.

Definition 9.3.4 *We say $D_h^{(2)}(P)$ is an **approximate second-order partial canonical gradient** if either of the maps $o_1 \mapsto D_h^{(2)}(P)(o_1, o)$ or $o_2 \mapsto D_h^{(2)}(P)(o, o_2)$ lies in $T(P)$ and the expansion*

$$\Psi(P) - \Psi(P_0) = -P_0 D^{(1)*}(P) - \frac{1}{2}\lim_{h\to 0} P_0^2 D_h^{(2)}(P) + R_3(P, P_0)$$

holds.

Theorem 9.3.5 *Suppose we have access to an approximate second-order partial canonical gradient $D_h^{(2)}(P)$. Furthermore, say our TMLE satisfies*

$$P_n D^{(1)*}(P_n^*) = P_n^2 D_{h_n}^{(2)}(P_n^*) = 0,$$

where h_n is our final approximation of the second-order gradient. Under the following assumptions:

1. *$D^{(1)*}(P_n^*)$ is in a P_0-Donsker class $\mathcal{F}$ with probability tending to one*
2. *$P_0[D^{(1)*}(P_n^*) - D^{(1)*}(P_0)]^2 = o_P(1)$*
3. *$(P_n^2 - P_0^2)D_{h_n}^{(2)}(P_n^*) = o_P\left(\frac{1}{\sqrt{n}}\right)$*
4. *$B_n(h_n) := \frac{1}{2}\left[P_0^2 D_{h_n}^{(2)}(P_n^*) - \lim_{h\to 0} P_0^2 D_h^{(2)}(P_n^*)\right] = o_P\left(\frac{1}{\sqrt{n}}\right)$*
5. *$R_3(P_n^*, P_0) = o_P\left(\frac{1}{\sqrt{n}}\right)$,*

the TMLE ψ_n^* is asymptotically linear with influence function $D^{(1)*}(P_0)$ (and is hence also asymptotically efficient).

Proof: We can write:

$$\begin{aligned}\Psi(P_n^*) - \Psi(P_0) &= -P_0 D^{(1)*}(P_n^*) - \frac{1}{2}\lim_{h\to 0} P_0^2 D_h^{(2)}(P_n^*) + R_3(P_n^*, P_0) \\ &= -P_0 D^{(1)*}(P_n^*) - \frac{1}{2}P_0^2 D_{h_n}^{(2)}(P_n^*) + B_n(h_n) + R_3(P_n^*, P_0) \\ &= (P_n - P_0)D^{(1)*}(P_n^*) + \frac{1}{2}(P_n^2 - P_0^2)D_{h_n}^{(2)}(P_n^*) + B_n(h_n) + R_3(P_n^*, P_0),\end{aligned}$$

where for the last step we use the fact that the TMLE satisfies

$$P_n D^{(1)*}(P_n^*) = P_n^2 D_{h_n}^{(2)}(P_n^*) = 0.$$

The first two assumptions allow us to apply Lemma 9.3.3, which yields

$$(P_n - P_0)(D^{(1)*}(P_n^*) - D^{(1)*}(P_0)) = o_P\left(\frac{1}{\sqrt{n}}\right) \implies (P_n - P_0)D^{(1)*}(P_n^*) = (P_n - P_0)D^{(1)*}(P_0) + o_P\left(\frac{1}{\sqrt{n}}\right).$$

Assumptions 3, 4, and 5 then imply that the last three terms in our expansion of $\Psi(P_n^*) - \Psi(P_0)$ are also all $o_P(\frac{1}{\sqrt{n}})$, which then immediately implies

$$\Psi(P_n^*) - \Psi(P_0) = (P_n - P_0)D^{(1)*}(P_0) + o_P\left(\frac{1}{\sqrt{n}}\right),$$

as desired. □

9.3.1 Confidence Intervals

To end, it is worth mentioning the practical use of confidence intervals with the second-order TMLE. As Theorem 9.3.1 states, the influence function of the 2-TMLE ψ_n^* is $D^{(1)*}(P_0)$ (and this holds for higher-order k-TMLEs as well). Hence, as with the regular TMLE, or as with any asymptotically linear estimator, we can use the confidence interval

$$\left(\psi_n^* - \frac{1}{\sqrt{n}} z_{1-\frac{\alpha}{2}}\sqrt{P_n[D^{(1)*}(P_n^*)]^2}, \psi_n^* + \frac{1}{\sqrt{n}} z_{1-\frac{\alpha}{2}}\sqrt{P_n[D^{(1)*}(P_n^*)]^2}\right).$$

This is asymptotically correct, but does not take into account the extra expansion the 2-TMLE offers. It is plausible that finite-sample performance could be improved by including the second-order term in the confidence interval, even if this does not change asymptotic behavior. Under the assumptions of the preceding theorems (Theorem 9.3.5 and Theorem 9.3.1, depending on whether second-order gradients are approximated), we have

$$\sqrt{n}(\psi_n^* - \psi_0) = \sqrt{n}(P_n - P_0)D^{(1)*}(P_n^*) + \frac{\sqrt{n}}{2}(P_n^2 - P_0^2)D^{(2)}(P_n^*) + \sqrt{n}R_3(P_n^*, P_0).$$

If we sample $\tilde{O}_1, \tilde{O}_2, \ldots, \tilde{O}_n$ from a consistent estimator of P_0, we can estimate the first term, by bootstrap, of the expansion by

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left[D^{(1)*}(P_n^*)(\tilde{O}_i) - P_n D^{(1)*}(P_n^*) \right]$$

and we can bootstrap the second term of the expansion by

$$\frac{1}{2n^{\frac{3}{2}}} \sum_{i=1}^{n} \sum_{j=1}^{n} \left[D^{(2)}(P_n^*)(\tilde{O}_i, \tilde{O}_j) - P_n^2 D^{(2)}(P_n^*) \right],$$

so the entire bootstrapped statistic is

$$\tilde{Z}_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left[D^{(1)*}(P_n^*)(\tilde{O}_i) - P_n D^{(1)*}(P_n^*) \right] + \frac{1}{2n^{\frac{3}{2}}} \sum_{i=1}^{n} \sum_{j=1}^{n} \left[D^{(2)}(P_n^*)(\tilde{O}_i, \tilde{O}_j) - P_n^2 D^{(2)}(P_n^*) \right].$$

Letting q represent the quantile of the conditional distribution of $\tilde{Z}_n$ given the empirical distribution P_n, we get the bootstrapped confidence interval

$$\left(\psi_n^* - \frac{1}{\sqrt{n}} q_{1-\frac{\alpha}{2}}, \psi_n^* + \frac{1}{\sqrt{n}} q_{1-\frac{\alpha}{2}} \right)$$

that takes into account the second-order expansion.

Bibliography

Bickel, P. J., Klaassen, C. A., Bickel, P. J., Ritov, Y., Klaassen, J., Wellner, J. A., & Ritov, Y. (1993). *Efficient and adaptive estimation for semiparametric models* (Vol. 4). Springer.

Carone, M., Díaz, I., & van der Laan, M. J. (2014). Higher-order targeted minimum loss-based estimation.

Chen, Y., Kennedy, E. H., & Balakrishnan, S. (2026). On the equivalence between neyman orthogonality and pathwise differentiability. *arXiv preprint arXiv:2603.15817*.

Chen, Y.-C. (2022). A note on the instrumental variable and local average treatment effect.

Chen, Y.-C. (2024). A note on semi-parametric estimators.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., & Robins, J. (2018). *Double/debiased machine learning for treatment and structural parameters.* Oxford University Press Oxford, UK.

Chernozhukov, V., Hansen, C., Kallus, N., Spindler, M., & Syrgkanis, V. (2024). Applied causal inference powered by ml and ai. *arXiv preprint arXiv:2403.02467*.

Díaz, I., & Rosenblum, M. (2015). Targeted maximum likelihood estimation using exponential families. *The international journal of biostatistics*, *11*(2), 233–251.

Dudoit, S., & van der Laan, M. J. (2005). Asymptotics of cross-validated risk estimation in estimator selection and performance assessment. *Statistical methodology*, *2*(2), 131–154.

Foster, D. J., & Syrgkanis, V. (2023). Orthogonal statistical learning. *The Annals of Statistics*, *51*(3), 879–908.

Gruber, S., & van der Laan, M. J. (2010). An application of collaborative targeted maximum likelihood estimation in causal inference and genomics. *The International Journal of Biostatistics*, *6*(1), 18.

Kennedy, E. H. (2024). Semiparametric doubly robust targeted double machine learning: a review. *Handbook of statistical methods for precision medicine*, 207–236.

Laan, M. J., & Robins, J. M. (2003). *Unified methods for censored longitudinal data and causality*. Springer.

Robins, J., Li, L., Tchetgen, E., van der Vaart, A., et al. (2008). Higher order influence functions and minimax estimation of nonlinear functionals. In *Probability and statistics: essays in honor of david a. freedman* (Vol. 2, pp. 335–422). Institute of Mathematical Statistics.

Sen, B. (2018). Semiparametric statistics.

Tsiatis, A. A. (2006). *Semiparametric theory and missing data*. Springer.

van der Laan, M., & Gruber, S. (2016). One-step targeted minimum loss-based estimation based on universal least favorable one-dimensional submodels. *The international journal of biostatistics*, *12*(1), 351–378.

van der Laan, M., Wang, Z., & van der Laan, L. (2021). Higher order targeted maximum likelihood estimation. *arXiv preprint arXiv:2101.06290*.

van der Laan, M. J. (1998). Identity for the npmle in censored data models. *Lifetime Data Analysis*, *4*(1), 83–102.

van der Laan, M. J., & Gruber, S. (2010). Collaborative double robust targeted maximum likelihood estimation. *The international journal of biostatistics*, *6*(1), 17.

Van der Laan, M. J., & Rose, S. (2018). *Targeted learning in data science*. Springer.

Van der Laan, M. J., Rose, S., et al. (2011). *Targeted learning: causal inference for observational and experimental data* (Vol. 4). Springer.

Van Der Laan, M. J., & Rubin, D. (2006). Targeted maximum likelihood learning.

Van der Vaart, A. W. (2000). *Asymptotic statistics* (Vol. 3). Cambridge university press.

Van Der Vaart, A. W., & Wellner, J. A. (1996). Weak convergence. In *Weak convergence and empirical processes: with applications to statistics* (pp. 16–28). Springer.

Zheng, W., & Van Der Laan, M. J. (2010). Asymptotic theory for cross-validated targeted maximum likelihood estimation.