May 10, 2021

Influence functions and functional derivatives

Helene Rytgaard and Anders Munch

In this note we provide some motivation for studying influence functions and semiparametric efficiency theory, and we briefly introduce the main theoretical concepts needed to study these topics. For concreteness we relate the discussion to the specific problem of estimating the average treatment effect (ATE), which we introduce in section 1. In section 2 we demonstrate how a naive (non-targeted) estimation procedure can fail, and then give some heuristics that indicate a strategy for remedying this; in particular, we will see that we would like to be able to talk about the *derivative of a functional* defined on a collection of probability measures. Hence in subsection 3.1 we first make some general considerations about how to define a functional derivative, and then zoom in on the statistical setting in subsections 3.2 and 3.3. These subsections of section 3 are slightly technical; this is because the goal is to collect mathematically precise definitions and statements from the semiparametric literature, which can be hard to find in compact form elsewhere. The main point to take away from this section is that, given a statistical model $\mathcal{P}$ and a target parameter $\theta = \Psi(\mathrm{P})$ defined through the functional $\Psi \colon \mathcal{P} \to \mathbb{R}$, it is possible to define the derivative of Ψ. This derivative is referred to as the *canonical gradient* and has two important properties (propositions 3.8 and 3.10), which in turn imply that estimators based on the canonical gradient will asymptotically have *vanishing first order bias* and be *efficient*. This is demonstrated in section 4 where we also present the canonical gradient for the ATE problem. In section 5 we very briefly mention further directions and related topics that are neglected in this note.

Notation We use notation from empirical process theory and write $\mathrm{P}[f]$ for a (probability) measure P and a measurable function f defined on the sample space $\mathcal{O}$ to mean

$$\mathrm{P}[f] := \mathbb{E}_{\mathrm{P}}[f(O)] = \int f(o)\,\mathrm{d}\mathrm{P}(o).$$

This allows us to write the empirical average as

$$\hat{\mathbb{P}}_n[f] = \frac{1}{n}\sum_{i=1}^{n} f(O_i).$$

We also use stochastic O-notation, $\mathcal{o}_{\mathrm{P}}$ and $\mathcal{O}_{\mathrm{P}}$. $X_n = \mathcal{O}_{\mathrm{P}}(1)$ means that the sequence X_n is bounded in probability, and more generally

$$X_n = \mathcal{O}_{\mathrm{P}}(R_n) \quad \text{means} \quad X_n = Y_n R_n \quad \text{for some} \quad Y_n = \mathcal{O}_{\mathrm{P}}(1),$$

and

$$X_n = \mathcal{o}_{\mathrm{P}}(R_n) \quad \text{means} \quad X_n = Y_n R_n \quad \text{for some} \quad Y_n \xrightarrow{\mathrm{P}} 0.$$

For example, $X_n = \mathcal{o}_{\mathrm{P}}(n^{-1/2})$ means that $\sqrt{n}X_n \xrightarrow{\mathrm{P}} 0$, and $X_n = \mathcal{o}_{\mathrm{P}}(1)$ simply means that $X_n \xrightarrow{\mathrm{P}} 0$. This follows the notation used in, e.g., van der Vaart [2000], see for instance chp. 2.

1 The average treatment effect as a statistical problem

Many statistical problems are naturally formulated using one or more so-called *nuisance parameters*. A nuisance parameter is a component we need to introduce in our the statistical model, which is not of interest in itself, but is nevertheless needed to model the question of interest. First of all let us note that by a *statistical problem* we formally mean the tuple $(\mathcal{P}, \Psi)$, where $\mathcal{P}$ is a collection of probability measures and $\Psi \colon \mathcal{P} \to \mathbb{R}$ is a functional defined on this collection of probability measures. The definition of $\mathcal{P}$ determines what kind of assumptions we make, while Ψ is determined by our scientific question of interest.

We are particularly interested in statistical problems with infinite-dimensional nuisance parameter. A good example of this is the average treatment effect (ATE) which will be our running example throughout.

Example 1.1 (ATE)
Suppose we observed independent and identically distributed (iid) samples $O_1, \dots, O_n$, $n \in \mathbb{N}$ of a random variable $O \in \mathcal{O}$ distributed according to an unknown distribution function P belonging to a statistical model $\mathcal{P}$. Each observation consists of $O = (X, A, Y)$ where $X \in \mathbb{R}^d$ are covariates, $A \in \{0, 1\}$ is a binary exposure and $Y \in \{0, 1\}$ is a binary outcome variable. The target parameter can be written as the functional $\Psi \colon \mathcal{P} \to \mathbb{R}$ on distributions $\mathrm{P} \in \mathcal{P}$, for our purposes defined as

$$\Psi(\mathrm{P}) = \Psi_1(\mu_X, f) = \int \big(f(1, x) - f(0, x)\big) d\mu_X(x), \tag{1}$$

where $f(a, x) = \mathbb{E}_{\mathrm{P}}[Y \mid A = a, X = x]$ and μ_X is the marginal distribution of X, where we suppress the dependence on P. Here we have used a subscript to differentiate between the functional Ψ considered as a map defined on $\mathcal{P}$ and the functional Ψ_1 considered as a map defined on a product space containing (μ_X, f). •

The ATE can be interpreted causally under structural assumptions. In this note, we do not consider these important issues, but merely consider the estimation of the ATE as a statistical problem as introduced above, i.e., a tuple $(\mathcal{P}, \Psi)$ taking the concrete form given in (1). In this case, the nuisance parameters are the conditional expectation of the outcome Y given covariates $X = x$ and treatment $A = a$, which we denoted by f, and the marginal distribution of X; we are not really interested in these functions, but we need them to be able to express the ATE.

We note here that we could also have expressed or "parametrized" the target parameter differently: Using iterated expectations it is straightforward to show both that $\Psi(\mathrm{P}) = \Psi_2(\mu, \pi)$ and $\Psi(\mathrm{P}) = \Psi_3(\mu, f, \pi)$, where

$$\begin{aligned}
\Psi_2(\mu, \pi) &:= \int \left\{ \frac{a\, y}{\pi(x)} - \frac{(1-a)\, y}{1 - \pi(x)} \right\} \mathrm{d}\mu(y, a, x), \\
\Psi_3(\mu, f, \pi) &:= \int \left\{ \frac{a\, (y - f(x, 1))}{\pi(x)} + f(x, 1) - \frac{(1-a)\, (y - f(x, 0))}{1 - \pi(x)} - f(x, 0) \right\} \mathrm{d}\mu(y, a, x),
\end{aligned}$$

with $\pi(x) := \mathrm{P}(A = 1 \mid X = x)$ denoting the conditional probability of treatment given covariate status. Hence, using Ψ_2 the nuisance parameters would instead be the treatment mechanism π and the full measure μ, while using Ψ_3 the nuisance parameters would be f, π, and μ. For later reference we define

$$\begin{aligned}
&\varphi_1(x; f) := f(1, x) - f(0, x), \quad \varphi_2(y, a, x; \pi) := \frac{a\, y}{\pi(x)} - \frac{(1-a)\, y}{1 - \pi(x)}, \quad \text{and} \\
&\varphi_3(y, a, x; f, \pi) := \varphi_2(y, a, x; \pi) + \varphi_1(x; f) - \frac{a\, f(1, x)}{\pi(x)} + \frac{(1-a)\, f(0, x)}{1 - \pi(x)},
\end{aligned} \tag{2}$$

such that we can write $\Psi(\mathrm{P}) = \mathrm{P}[\varphi_1(O, f)] = \mathrm{P}[\varphi_2(O, \pi)] = \mathrm{P}[\varphi_3(O, f, \pi)]$.

Statistical problems involving nuisance parameters lead to an obvious two-step estimation strategy:

(1) Estimate the nuisance parameters.

(2) Plug the estimates into the target parameter functional Ψ.

For instance, when estimating the ATE and using the parametrization in (1), we would (1) estimate the conditional outcome $f(x, y) = \mathbb{E}[Y \mid A = a, X = x]$ and the marginal distribution μ_X with estimators $\hat{f}_n$ and $\hat{\mu}_n$, and then (2) plug these into Ψ_1. Estimation of μ_X is straightforward using the empirical measure $\hat{\mathbb{P}}_n$, which gives the estimator

$$\hat{\theta}_n = \Psi_1(\hat{\mathbb{P}}_n, \hat{f}_n) = \frac{1}{n}\sum_{i=1}^{n}\left\{\hat{f}_n(1, X_i) - \hat{f}_n(0, X_i)\right\}, \tag{3}$$

where $\hat{f}_n$ is some estimated regression function, for instance obtained by linear regression. Using instead the parametrization given by Ψ_2 would demand estimation of π in step (1), giving the estimator $\Psi_2(\hat{\mathbb{P}}_n, \hat{\pi}_n)$, while using Ψ_3 would demand estimation of both f and π, giving the estimator $\Psi_3(\hat{\mathbb{P}}_n, \hat{f}_n, \hat{\pi}_n)$.

Now, does it really matter which of these parametrizations we pick? And could we not just choose any of them and then instead focus on picking a good nuisance estimator? Note that both estimation of f and π are well-studied problems: The first is a regression problem while the second is a classification problem, so we have a whole zoo of possible estimators, containing everything from linear regression models to random forests, neural networks, etc. Should we not simply focus on finding the best possible estimator of, say, f, and then just plug that into Ψ_1? The example in the following section demonstrates that some additional thought might be needed.

2 Motivation

To motivate our theoretical considerations on estimation of the ATE in this note, consider the following simple toy example.

Example 2.1 (Integrated kernel)
Given n samples $X_i \in \mathbb{R}$ from some unknown distribution with cumulative distribution function F, we want to estimate $F(x) = \mathrm{P}(X \leq x)$. Let us say that we are willing to assume that F has a continuous Lebesgue-density f. Then one estimation strategy would be to first use a kernel density estimator to estimate f, and then plug this into the integral operator

$$f \longmapsto \int_{-\infty}^{x} f(z)\,\mathrm{d}z,$$

to obtain an estimate of the cumulative distribution function $F(x)$ at the fixed point $x \in \mathbb{R}$. In this setting our nuisance parameter is f, and our target parameter is $\theta = \Psi(f)$ with

$$\Psi \colon \mathcal{F} \to \mathbb{R}, \quad \Psi(f) = \int_{-\infty}^{x} f(z)\,\mathrm{d}z,$$

where $\mathcal{F}$ is some suitable function space, for instance the collection of continuous functions. This procedure results in the target and nuisance estimators given as

$$\hat{\theta}_n := \int_{-\infty}^{x} \hat{f}_n(z)\,\mathrm{d}z, \quad \text{and} \quad \hat{f}_n(z) = \hat{\mathbb{P}}_n[k_h(z, \cdot)], \tag{4}$$

for some kernel function k_h with bandwidth h_n. It is well-known that the optimal choice of bandwidth h_n is $h_n \propto n^{-1/5}$ [Wasserman, 2006], so this would also be the natural choice in

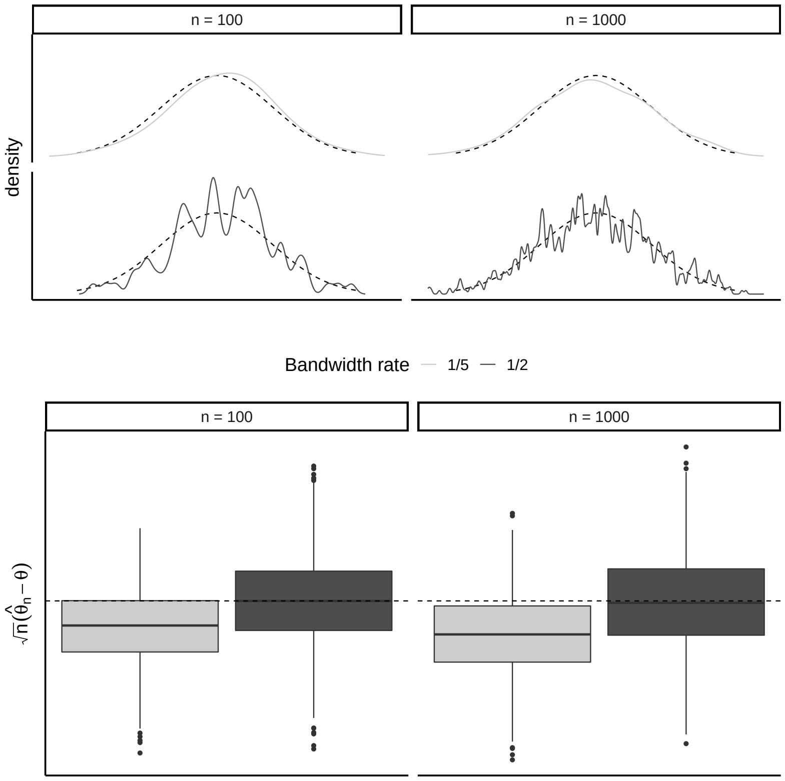

Figure 1: Simulation illustrating the problem for the plug-in approach. The top 2 rows give representative kernel density estimates for two different bandwidths scaling with n; the dashed line is the density of the distribution used to generate the data. The last row gives the distribution of the centralized and $\sqrt{n}$-scaled corresponding plug-in estimates based on 1000 Monte Carlo samples.

our case; indeed, the upper panel of figure 1 demonstrates how this choice of bandwidth is superior to the choice $h_n \propto n^{-1/2}$, which instead results in a very rough or undersmoothed estimate. Surprisingly, however, the lower panel shows that for estimation of the *target parameter*, plugging the undersmoothed estimate into Ψ is superior to using the default, optimal bandwidth estimator. This example is in fact simply enough to allow an exact analytic calculation of the bias and variance of the target parameter, and hence it is fairly straightforward to mathematically prove the behavior suggested by figure 1. •

Though the above example is somewhat silly because we already have an obvious estimator of the parameter of interest in this situation, it clearly illustrates that the modus operandi outlined in the two-step estimation procedure in the previous section can be problematic. With this example in mind, we should maybe not be too confident about the first ATE estimator we considered in (3). In fact, with a bit more work it can be showed that the same phenomenon as illustrated in figure 1 appears if we use a kernel-based regression estimator in (3).

The issue from the toy example can be understood more generally by considering a target parameter Ψ that can be parametrized as $\Psi(\mathrm{P}) = \Psi_0(\mathrm{P}, \nu) = \mathrm{P}[\varphi(O, \nu)]$ for some function $\varphi(\cdot, \nu)$ defined on the sample space $\mathcal{O}$ that depend on the nuisance parameter ν. (Many target parameters can be written on this form; in particular, as shown above, the ATE can be written on this form in at least three different ways.) Plugging the nuisance estimators $\hat{\mathbb{P}}_n$

and $\hat{\nu}_n$ into this expression gives the estimator $\hat{\theta}_n = \Psi_0(\hat{\mathbb{P}}_n, \hat{\nu}_n)$. To analyze the (asymptotic) behavior of this estimator we would typically consider the expression

$$\sqrt{n}\left(\hat{\theta}_n - \theta\right) = \sqrt{n}\left(\Psi_0(\hat{\mathbb{P}}_n, \hat{\nu}_n) - \Psi_0(\mathrm{P}, \nu)\right) = \sqrt{n}\left(\hat{\mathbb{P}}_n[\varphi(O, \hat{\nu}_n)] - \mathrm{P}[\varphi(O, \nu)]\right),$$

which can be further expanded as

$$\sqrt{n}\left(\hat{\theta}_n - \theta\right) = \mathbb{G}_n[\varphi(O, \hat{\nu}_n)] + \sqrt{n}\left\{\Psi_0(\mathrm{P}, \hat{\nu}_n) - \Psi_0(\mathrm{P}, \nu)\right\},$$

where $\mathbb{G}_n$ denotes the empirical process defined as $\mathbb{G}_n := \sqrt{n}(\hat{\mathbb{P}}_n - \mathrm{P})$. Let us now consider some heuristic arguments: If we could somehow make sense of a Taylor expansion of the function $\nu \mapsto \Psi_0(\mathrm{P}, \nu)$ we could further expand the above as

$$\sqrt{n}\left(\hat{\theta}_n - \theta\right) = \mathbb{G}_n[\varphi(O, \hat{\nu}_n)] + \mathrm{D}_\nu \Psi_0\left[\sqrt{n}(\hat{\nu}_n - \nu)\right] + \mathcal{O}_{\mathrm{P}}(\sqrt{n}\|\hat{\nu}_n - \nu\|_{\mathcal{V}}^2), \tag{5}$$

where $\|\cdot\|_{\mathcal{V}}$ is some norm defined on the space where the nuisance parameter ν takes it values; this would typically be a function space, so the norm could, for instance, be the $\mathcal{L}^2$-norm or the supremum-norm. The asymptotic analysis of each of the three components in (5) requires some fairly advanced mathematical tools, and in this note we only focus on the second part. For the other two components we simply note that in many cases empirical process theory or sample splitting can be used to show that $\mathbb{G}_n[\varphi(O, \hat{\nu}_n)] = \mathbb{G}_n[\varphi(O, \nu)] + o_{\mathrm{P}}(\sqrt{n})$, and that $\|\hat{\nu}_n - \nu\|_{\mathcal{V}} = o_{\mathrm{P}}(n^{-1/4})$ has been established for several machine learning estimators under the right conditions (see section 5 for a few additional comments on this). Assuming these two equations to hold, (5) simplifies to

$$\sqrt{n}\left(\hat{\theta}_n - \theta\right) = \mathbb{G}_n[\varphi(O, \nu)] + \mathrm{D}_\nu \Psi_0\left[\sqrt{n}(\hat{\nu}_n - \nu)\right] + o_{\mathrm{P}}(1).$$

The first component is now straightforward to analyze using the central limit theorem, as it is a sum of iid. zero-mean variables. This term contributes the main part of the variance of the asymptotic behavior of $\hat{\theta}_n$. The second component can contribute with additional variance, if the $\hat{\nu}_n$ converges at $n^{-1/2}$-rate; however, if $\hat{\nu}_n$ converges at a slower rate, the main issue with the second component becomes bias. Indeed, as the default kernel estimator converges at rate $n^{-2/5}$, this component is exactly what ruined the target estimator in example 2.1. If we want to use flexible, non-parametric nuisance estimators we can in general not hope to be able to have these converge at the parametric $n^{-1/2}$-rate (for instance, the $n^{-2/5}$-rate is the best possible for estimation of twice times continuously differentiable densities [van der Vaart, 2000, chp. 24]). Hence we cannot hope to achieve $n^{-1/2}$-rate inference about θ by optimizing the nuisance parameter, and thus we instead turn to the derivative $\mathrm{D}_\nu \Psi_0$. In particular, if we can design Ψ_0 in such a way that $\mathrm{D}_\nu \Psi_0 = 0$, we would get rid of this problematic bias term altogether.

This goal is our motivation for studying functional derivatives in the next section. As mentioned, we will see that this study will also allow us to analyze how efficiently a given statistical problem can be estimated.

3 Functional derivatives and the canonical gradient

For infinite-dimensional spaces, unlike for $\mathbb{R}^d$, there are several sensible and non-equivalent ways of defining a derivative. In addition, again unlike for $\mathbb{R}^d$, for such spaces there are typically several, non-equivalent norms, which in turn determine which kind of functionals and operators are differentiable according to what definition. We do not consider these issues in much detail but only consider two types of derivatives, namely Gâteaux and Hadamard. We then consider some slight generalizations of the latter, and then, in subsection 3.2, describe in greater detail how Hadamard derivatives for a statistical problem can be represented; this allows us to derive some useful properties in subsection 3.3.

3.1 Functional derivatives

The most straightforward form of functional differentiabiliy is the generalization of the *directional derivative* of multivariate calculus. This is known as Gâteaux differentiability and defined as follows.

Definition 3.1 (Gâteaux derivative)**.** Let $\mathcal{M}$ and $\mathcal{Y}$ be a normed real vector spaces, $\Psi\colon \mathcal{M} \to \mathcal{Y}$ a map, and $x \in \mathcal{M}$ a point in the domain. If there exists a linear, continuous operator $\dot{\Psi}_x\colon \mathcal{M} \to \mathcal{Y}$ such that for all $h \in \mathcal{M}$

$$\left\| \Psi(x + \varepsilon h) - \Psi(x) - \dot{\Psi}_x(\varepsilon h) \right\|_{\mathcal{Y}} = \mathcal{o}(\varepsilon), \quad \text{when} \quad \varepsilon \longrightarrow 0 \in \mathbb{R},$$

we say that Ψ is Gâteaux differentiable at x. We call the operator $\dot{\Psi}_x$ the Gâteaux derivative of Ψ at x.

If the map Ψ is Gâteaux differentiable at x then it follows from the definition that

$$\left\| \frac{\Psi(x + \varepsilon h) - \Psi(x)}{\varepsilon} - \dot{\Psi}_x(h) \right\|_{\mathcal{Y}} \longrightarrow 0, \quad \text{when} \quad \varepsilon \longrightarrow 0. \tag{6}$$

In particular, when $\mathcal{Y} = \mathbb{R}$ the Gâteaux derivative at x in the direction h can be derived as the ordinary derivative of the real-valued function $\varepsilon \mapsto \Psi(x + \varepsilon h)$ evaluated at 0, i.e.,

$$\dot{\Psi}_x(h) = \partial_0 \Psi(x + \varepsilon h) := \frac{\partial}{\partial \varepsilon}\bigg|_{\varepsilon=0} \Psi(x + \varepsilon h). \tag{7}$$

To simplify notation in the following we will use the notation $\partial_0 f(\varepsilon)$ for maps f with a real domain as just defined in (7). Gâteaux differentiability is used by Chernozhukov et al. [2018] to define *Neyman orthogonality*: A function $\varphi\colon \mathcal{O} \times \mathcal{V} \to \mathbb{R}$, defined on the sample space $\mathcal{O}$ and depending on a nuisance parameter $\nu \in \mathcal{V}$, fulfills the Neyman orthogonality condition (wrt. $\mathcal{V}$) if the map

$$F\colon \mathcal{V} \longrightarrow \mathbb{R}, \quad \nu \longmapsto F(\nu) := \Psi_0(\mathrm{P}, \nu) = \mathrm{P}[\varphi(O, \nu)]$$

is Gâteaux differentiable at ν with vanishing derivative, i.e., $\dot{F}_\nu(h) = 0$ for all $h \in \mathcal{V}$. We encountered the map above in section 2, and the point of Neyman orthogonality is exactly to ensure that the second component of the decomposition in (5) vanishes. Given a function φ it is straightforward to use (7) to verify that a given function fulfills the condition. Indeed, one can examine φ_1, φ_2, and φ_3 defined in (2) and see that φ_3 fulfills the Neyman orthogonality condition while φ_1 and φ_2 do not.

This shows that Gâteaux is already a useful concept. Still, it is a rather weak form of differentiability; for instance, as it is equivalent to the directional derivative for ordinary multivariate functions, Gâteaux differentiability is not enough to guarantee (ordinary) differentiability of such functions. To get a richer theory, a stronger notion of differentiability is needed. For our setting, a particular useful concept is *Hadamard* differentiability.

Definition 3.2 (Hadamard derivative)**.** Let $\mathcal{M}$ and $\mathcal{Y}$ be a normed real vector spaces, $\Psi\colon \mathcal{M} \to \mathcal{Y}$ a map, and $x \in \mathcal{M}$ a point in the domain. If there exists a linear, continuous operator $\dot{\Psi}_x\colon \mathcal{M} \to \mathcal{Y}$ such that

$$\left\| \frac{\Psi(x + \varepsilon_n h_n) - \Psi(x)}{\varepsilon_n} - \dot{\Psi}_x(h) \right\|_{\mathcal{Y}} \longrightarrow 0,$$

for any $\varepsilon_n \to 0$ and $\{h_n\}_{n\in\mathbb{N}} \subset \mathcal{M}$ with $h_n \to h \in \mathcal{M}$, we say that Ψ is Hadamard differentiable at x and call the operator $\dot{\Psi}_x$ the Hadamard derivative of Ψ at x.

Comparing this definition with (6) we see that the only difference is that the linear approximation provided by $\dot{\Psi}_x$ should hold along any *converging sequence* h_n and not merely in a fixed direction h. For this reason Hadamard differentiability is also known as *path-wise* differentiability. A still stronger condition is to demand that the approximation should hold for any *bounded sequence* h_n; this gives the concept of *Fréchet* differentiability, which we will not use in this text. One can show that *if* a map is Hadamard (or Fréchet) differentiable, then the Hadamard (or Fréchet) derivative is equal to the Gâteaux derivative. Hence, to find the Hadamard derivative of a given functional Ψ, the common strategy is to first use high school math tools to calculate (7) and then verify that the obtained candidate fulfills the requirements of definition 3.2.

So far we have assumed the domain to be a *linear* space. When working with probability measures, this is often not the case, and hence we need one final definition, which allows the operator Ψ and its derivative $\dot{\Psi}$ to be defined only on subsets of the normed vector space $\mathcal{M}$. We should think of the following as mirroring differentiation of a multivariate function defined on a manifold embedded in a higher dimensional Euclidean space (for instance, a surface embedded in $\mathbb{R}^3$).

Definition 3.3 (Tangential Hadamard derivative)**.** Let $\mathcal{P}$ and $\dot{\mathcal{P}}_x$ be subsets of the normed real vector space $\mathcal{M}$, with $x \in \mathcal{P} \subset \mathcal{M}$. For a map $\Psi\colon \mathcal{P} \to \mathcal{Y}$, we say that Ψ is *Hadamard differentiable (at x) tangential to $\dot{\mathcal{P}}_x$* if there exists a continuous, linear operator $\dot{\Psi}_x\colon \dot{\mathcal{P}}_x \to \mathcal{Y}$ such that

$$\left\| \frac{\Psi(x+\varepsilon_n h_n) - \Psi(x)}{\varepsilon_n} - \dot{\Psi}_x(h) \right\|_{\mathcal{Y}} \longrightarrow 0,$$

for any $\{h_n\} \subset \mathcal{M}$ and $\{\varepsilon_n\} \subset \mathbb{R}$ with $h_n \to h \in \dot{\mathcal{P}}_x$, $\varepsilon_n \to 0$, and $x + \varepsilon_n h_n \in \mathcal{P}$.

The only change from the previous definition is that the "path" $x + \varepsilon_n h_n$ is restricted to lie in the subset $\mathcal{P}$, and that the "direction" is h is restricted to lie in $\dot{\mathcal{P}}_x$.

3.2 Tangent spaces and gradients for statistical problems

Using the tools introduced so far, we can now define the so-called *tangent space for the model* $\mathcal{P}$ and then formally define the derivative of the map $\Psi\colon \mathcal{P} \to \mathbb{R}$ from the statistical problem $(\mathcal{P}, \Psi)$. We refer to this derivative as the *canonical gradient*, or later in section 4 the *efficient influence function*. The canonical gradient is a central component in our analysis of the statistical problem as it represents the optimal asymptotic variance (proposition 3.10); in addition it allows us to construct estimators with vanishing first order bias (proposition 3.8 and section 4).

As shown in the previous subsection, to be able to talk about differentiability of the statistical problem $(\mathcal{P}, \Psi)$, we need to embed the model $\mathcal{P}$ into a suitable normed vector space. To do so we now assume for simplicity that the family $\mathcal{P}$ is dominated by a single σ-finite measure μ (for our running example of the ATE this measure would be a product of Lebesgue and counting measures), and then think of $\mathcal{P}$ as lying inside the Banach space $\mathcal{M}_\mu$ of finite signed measures dominated by μ equipped with the variational norm

$$\|M\|_{\mathcal{M}_\mu} = \int |m| \, \mathrm{d}\mu, \qquad \text{for} \qquad M = m \cdot \mu.$$

Definition 3.4 (Tangent space for $\mathcal{P}$)**.** Let $\mathcal{P} \subset \mathcal{M}_\mu$ be a collection of probability measures and let $\mathrm{P} \in \mathcal{P}$. For any one-dimensional path $\varepsilon \mapsto \mathrm{P}_\varepsilon \in \mathcal{P}$ with $\mathrm{P}_0 = \mathrm{P}$, which is Hadamard[1] differentiable at 0, let $\partial_0 \mathrm{P}_\varepsilon$ be the derivative, and let $\{\partial_0 \mathrm{P}_\varepsilon\}$ be the collection of derivatives of all such paths. We call the closed linear span of this collection the *tangent space of* $\mathcal{P}$ *at* P and denote it by $\dot{\mathcal{P}}_\mathrm{P}$. Formally,

$$\dot{\mathcal{P}}_\mathrm{P} := \overline{\mathrm{span}}\,\{\partial_0 \mathrm{P}_\varepsilon \mid \varepsilon \mapsto \mathrm{P}_\varepsilon \text{ is Hadamard differentiable and } \mathrm{P}_0 = \mathrm{P}\}\,.$$

The definition of a tangent space for a collection of probability measures simply mimics the definition of a tangent space for a surface embedded in $\mathbb{R}^3$: We move along differentiable paths through a given point on the surface, and the tangent space is then the span of the derivatives of all such paths. With a differentiable structure on the collection $\mathcal{P}$ we can talk about a *gradient* of a functional defined on this set.

Definition 3.5 (Canonical gradient)**.** Let $(\mathcal{P}, \Psi)$ be a statistical problem, with $\mathcal{P} \subset \mathcal{M}_\mu$, and $\dot{\mathcal{P}}_\mathrm{P}$ the tangent space of $\mathcal{P}$ at $\mathrm{P} \in \mathcal{P}$. If $\Psi \colon \mathcal{P} \to \mathbb{R}$ is Hadamard differentiable at P tangential to $\dot{\mathcal{P}}_\mathrm{P}$, we refer to the Hadamard derivative $\dot{\Psi}_\mathrm{P}$ as the *canonical gradient of the statistical problem.*

The definitions of the tangent space $\dot{\mathcal{P}}_\mathrm{P}$ and the gradient $\dot{\Psi}_\mathrm{P}$ above capture the intuitive meaning of these concepts as generalizations of well-known concepts from multivariate calculus: The statistical model $\mathcal{P}$ is viewed as a subset of the unit sphere in $\mathcal{M}_\mu$ and the canonical gradient $\dot{\Psi}_\mathrm{P}$ is simply the infinite-dimensional version of the gradient of a map defined on a surface in Euclidean space. We shall see in a moment why this object $\dot{\Psi}_\mathrm{P}$ is of interest for the statistician, but firstly we consider a different representation of $\mathcal{P}$ which allows us to represent $\dot{\mathcal{P}}_\mathrm{P}$ as a subset of $\mathcal{L}^2_\mathrm{P}$; this is particularly useful as we then have the rich Hilbert space structure of $\mathcal{L}^2_\mathrm{P}$ at our disposal.

For a fixed element $\mathrm{P} \in \mathcal{P}$ with μ-density p, consider a one-dimensional parametric submodel of μ-densities p_ε with $p_0 = p$ and $p_\varepsilon \cdot \mu \in \mathcal{P}$ for all ε in some small interval. We restrict attention to such submodels for which the score function $\dot{\ell}_0 := \partial_0 \log(p_\varepsilon)$ exists as an element in $\mathcal{L}^2_\mathrm{P}$; here we use the terminology from ordinary maximum likelihood inference where the derivative of the likelihood for a parametric model p_ε is referred to as the score function. We denote the closed linear span of all such score functions by Γ_P, i.e., $\Gamma_\mathrm{P} := \overline{\mathrm{span}}\{\dot{\ell}_0\} \subset \mathcal{L}^2_\mathrm{P}$, where $\{\dot{\ell}_0\}$ is read as the collection of all possible score functions. It is argued by Bickel et al. [1993] that, for all relevant purposes, Γ_P is equivalent to the tangent space $\dot{\mathcal{P}}_\mathrm{P}$ from definition 3.5, and hence Γ_P is also referred to as the tangent space. It can be shown that all elements of Γ_P have mean zero (under P), i.e.,

$$\Gamma_\mathrm{P} \subset \mathbb{H}_0 := \left\{h \in \mathcal{L}^2_\mathrm{P} \mid \mathrm{P}[h] = 0\right\},$$

and that if we make no assumptions about the model $\mathcal{P}$ then $\Gamma_\mathrm{P} = \mathbb{H}_0$ [Bickel et al., 1993].

Due to the well-known structure of $\mathcal{L}^2_\mathrm{P}$, is often easier to analyze the tangent space as the subspace $\Gamma_\mathrm{P} \subset \mathcal{L}^2_\mathrm{P}$; for instance, one useful result is the following proposition, which provides an alternative characterization of the canonical gradient.

Proposition 3.6. *Let* $(\mathcal{P}, \Psi)$ *be a statistical problem with canonical gradient* $\dot{\Psi}_\mathrm{P}$ *at* $\mathrm{P} \in \mathcal{P}$. *There exists a unique element* $\varphi_\mathrm{P} \in \Gamma_\mathrm{P}$ *such that*

$$\partial_0 \Psi(\mathrm{P}_\varepsilon) = \dot{\Psi}_\mathrm{P}(\partial_0 \mathrm{P}_\varepsilon) = \langle \varphi_\mathrm{P}, \dot{\ell}_0 \rangle_\mathrm{P} \tag{8}$$

holds for any differentiable submodel P_ε *with score function* $\dot{\ell}_0$.

[1] When the domain is $\mathbb{R}$, all the previously considered types of differentiability are equivalent, so any type could be used here.

Sketch of proof. The first equality follows from the chain rule for Hadamard derivatives [van der Vaart, 2000, chp. 20], while the second follows from the arguments given above: If the tangent space can be represented as Γ_{P}, the linear map $\dot{\Psi}_{\mathrm{P}}$ on $\dot{\mathcal{P}}_{\mathrm{P}}$ can also be thought of as a linear map on Γ_{P}. As Γ_{P} is a closed subspace of a Hilbert space it is itself a Hilbert space, and hence Riesz representation theorem (for Hilbert spaces) gives the existence of a unique element $\varphi_{\mathrm{P}} \in \Gamma_{\mathrm{P}}$ such that $\Phi_{\mathrm{P}}(\dot{\ell}_0) = \langle \varphi_{\mathrm{P}}, \dot{\ell}_0 \rangle$ for all elements $\dot{\ell}_0 \in \Gamma_{\mathrm{P}}$. □

As with the identification $\dot{\mathcal{P}}_{\mathrm{P}} = \Gamma_{\mathrm{P}}$ we also identify the function φ_{P} with the canonical gradient $\dot{\Psi}_{\mathrm{P}}$. By the chain rule of ordinary multivariate calculus we have for a differentiable function $f\colon \mathbb{R}^d \to \mathbb{R}$ and any smooth curve $s\colon \mathbb{R} \to \mathbb{R}^d$ with $s(0) = x_0 \in \mathbb{R}^d$ that

$$\partial_0(f \circ s) = \nabla f(s(0)) \cdot (\partial_0 s)^\top = \langle \nabla f(x_0), \partial_t s \rangle ,$$

and as $\mathbb{R}^d$ can be spanned by smooth 1-dimensional curves, this property characterizes the gradient locally. Proposition 3.6 above states that this characterization extends to the Hilbert space setting with Hadamard differentiability, and in fact this characterization can be used to define Hadamard differentiability of the map Ψ tangential to Γ_{P} [Bickel et al., 1993, A.5].

Remark 3.7. Note that it is part of the defining property of the unique function φ_{P} that it lies in the tangent space Γ_{P}. If Γ_{P} is a proper subset of $\mathbb{H}_0$ there can be several functions $\tilde{\varphi} \in \mathbb{H}_0$ fulfilling condition (8); we refer to such functions as *gradients*. It follows from standard Hilbert space theory that the unique canonical gradient φ_{P} can be derived from any gradient $\tilde{\varphi}$ as the projection onto Γ_{P}, i.e., $\varphi_{\mathrm{P}} = \Pi(\tilde{\varphi} \mid \Gamma_{\mathrm{P}})$. This geometric picture of the tangent space and the canonical gradient can be very useful. •

3.3 Properties of the canonical gradient

In this subsection we derive two important properties of the canonical gradient. The first result states that the canonical gradient fulfills the Neyman orthogonality condition, which is the same as saying that the mean of the canonical gradient is insensitive to small (first order) perturbations of the nuisance parameters. The second results states that the canonical gradient in addition provides us with a theoretical bound on how efficient the target parameter can be estimated.

Proposition 3.8. *Let $(\mathcal{P}, \Psi)$ be a statistical problem with $\mathcal{P}$ dominated by μ and such that $\Psi(\mathrm{P}) = \mathrm{P}[\varphi(Z, \nu(\mathrm{P}))]$ for some function $\varphi\colon \mathcal{Z} \times \mathcal{V} \to \mathbb{R}$. If $\varphi(\cdot, \nu(\mathrm{P})) - \mathrm{P}[\varphi(O, \nu(\mathrm{P}))]$ is the canonical gradient of $(\mathcal{P}, \Psi)$ at P then, under regularity conditions, φ fulfills the Neyman orthogonality condition (wrt. $\nu(\mathcal{P})$).*

Sketch of proof. Under suitable regularity conditions, it holds for any differentiable path $\{\mathrm{P}_\varepsilon\}_\varepsilon \subset \mathcal{P}$ with $\mathrm{P}_0 = \mathrm{P}$ that

$$\begin{aligned}
\partial_0 \Psi(\mathrm{P}_\varepsilon) &= \partial_0 \int p_\varepsilon(z) \varphi(z, \nu(\mathrm{P}_\varepsilon)) \,\mathrm{d}\mu(z) \\
&= \int p(z)\{\partial_0 \varphi(z, \nu(\mathrm{P}_\varepsilon))\} + \{\partial_0 p_\varepsilon(z)\} \varphi(z, \nu(\mathrm{P})) \,\mathrm{d}\mu(z) \\
&= \partial_0 \mathrm{P}[\varphi(\cdot, \nu(\mathrm{P}_\varepsilon))] + \langle \dot{\ell}_0, \varphi(\cdot, \nu(\mathrm{P})) \rangle_{\mathrm{P}},
\end{aligned} \tag{9}$$

where $\dot{\ell}_0$ is the score function of the parametric submodel P_ε and we used the relation $\partial_0 \log p_\varepsilon = (\partial_0 p_\varepsilon) p_0^{-1}$. As $\varphi(\cdot, \nu(\mathrm{P}))$ is the canonical gradient at P, proposition 3.6 and (9) imply that $\partial_0 \mathrm{P}[\varphi(\cdot, \nu(\mathrm{P}_\varepsilon))] = 0$. □

Remark 3.9. Note that the fact that a function φ fulfills the Neyman orthogonality condition does not necessarily imply that φ is the canonical gradient (see, for instance, Chernozhukov et al. [2016] for an example). This can be seen from the fact that we in the proof sketch above only used that $\varphi(\cdot, \nu(\mathrm{P}))$ satisfies (8), and not that $\varphi(\cdot, \nu(\mathrm{P})) \in \Gamma_{\mathrm{P}}$ (see also remark 3.7); hence, any gradient for the statistical problem will fulfill the Neyman orthogonality condition (under regularity conditions). •

The reason to be particularly interested in the *canonical* gradient, instead of simply any gradient, comes from the following results. First, recall that for any statistical problem with a *finite-dimensional* model $\mathcal{P}$, the Fisher information provides a measure of the information obtainable about the parameters of the model; this measure is justified by the Cramer-Rao bound, which states that the least possible asymptotic variance of any estimator of the parameters of the model is given by the inverse of the Fisher information. In a more general form, the Cramer-Rao bound states that, for a one-dimensional family indexed by $t \in \mathbb{R}$ with parameter of interest $\Psi(\mathrm{P}_t)$, the optimal asymptotic variance is

$$\frac{(\partial_0 \Psi(\mathrm{P}_t))^2}{\mathrm{P}[\dot{\ell}_0^2]}, \tag{10}$$

where $\dot{\ell}_0$ is the score of the model, i.e., $\mathrm{P}[\dot{\ell}_0^2]$ is the Fisher information. Hence, the information bound for the parameter $\theta = \Psi(\mathrm{P}_t)$ in this model is defined as the inverse of (10). Now, a statistical problem with an infinite-dimensional model $\mathcal{P}$ contains all the one-dimensional submodels $\{\mathrm{P}_\varepsilon\} \subset \mathcal{P}$, first introduced in definition 3.5; each of these models have information bound wrt. $\Psi(\mathrm{P}_\varepsilon)$ given by the inverse of (10), and hence it is natural to define the information bound for the whole model as the information bound for the *least informative submodel*. Formally, this means we define the information bound of the statistical problem $(\mathcal{P}, \Psi)$ as

$$\mathcal{I}(\mathcal{P}, \Psi) := \inf \left\{ \frac{\mathrm{P}[\dot{\ell}_0^2]}{(\partial_0 \Psi(\mathrm{P}_\varepsilon))^2} \right\},$$

where the infimum is taken over all submodels $\{\mathcal{P}_\varepsilon\}$ with score functions $\dot{\ell}_0$. We then have the following relation.

Proposition 3.10. *For a statistical problem $(\mathcal{P}, \Psi)$ with canonical gradient $\dot{\Psi}_{\mathrm{P}} = \varphi_{\mathrm{P}}$ it holds that*

$$\mathcal{I}(\mathcal{P}, \Psi)^{-1} = \mathrm{P}[\varphi_{\mathrm{P}}^2].$$

Proof. Inverting the expression changes the infimum to a supremum, and then using proposition 3.6 we get

$$\mathcal{I}(\mathcal{P}, \Psi)^{-1} = \sup_{\dot{\ell} \in \Gamma_{\mathrm{P}}} \frac{\langle \varphi_{\mathrm{P}}, \dot{\ell} \rangle_{\mathrm{P}}^2}{\mathrm{P}[\dot{\ell}^2]}.$$

Using the Cauchy-Schwarz inequality gives that $\langle \varphi_{\mathrm{P}}, \dot{\ell} \rangle_{\mathrm{P}} \leq \mathrm{P}[\varphi_{\mathrm{P}}]\mathrm{P}[\dot{\ell}]$, and as $\varphi_{\mathrm{P}} \in \Gamma_{\mathrm{P}}$, the supremum is achieved at $\dot{\ell} = \varphi_{\mathrm{P}}$. □

4 Estimation based on the canonical gradient

Let us now return to estimation of the target parameter of the statistical problem $(\mathcal{P}, \Psi)$ and the decomposition in (5). Recall that we assume

A1: $\mathbb{G}_n[\varphi(O, \hat{\nu}_n)] = \mathbb{G}_n[\varphi(O, \nu)] + o_{\mathrm{P}}(1)$

A2: $\|\hat{\nu}_n - \nu\| = o_{\mathrm{P}}(n^{-1/4})$

where $\nu = \nu(\mathrm{P})$ is the "true" value of the nuisance parameter. Now, if $\varphi - \mathrm{P}[\varphi]$ is the canonical gradient, proposition 3.8 guarantees that the second component in (5) vanishes, and hence, under assumptions A1 and A2, the expression simplifies to

$$\sqrt{n}\left(\hat{\theta}_n - \theta\right) = \mathbb{G}_n[\varphi(O,\nu)] + o_{\mathrm{P}}(1)$$

By the central limit theorem, the dominating term on the right hand side converges to a centered Gaussian distribution with variance equal to $\mathrm{Var}[\varphi(O,\nu)]$. By proposition 3.10, this is the inverse of the information bound, and hence the estimator $\hat{\theta}_n$ based on the canonical gradient achieves the information bound for $(\mathcal{P}, \Psi)$.

We can now connect our discussion about functional derivatives to the notion of *influence functions* for *regular asymptotically linear* estimators. An estimator $\hat{\theta}_n$ of the parameter $\theta = \Psi(\mathrm{P})$ under the model $\mathcal{P}$, is called asymptotically linear if there exists a function $I \colon \mathcal{O} \times \mathcal{P} \to \mathbb{R}$ with $I(\cdot, \mathrm{P}) \in \mathcal{L}^2_{\mathrm{P}}$ and $\mathrm{P}[I(O,\mathrm{P})]$ for all $\mathrm{P} \in \mathcal{P}$, such that

$$\hat{\theta}_n - \theta = \hat{\mathbb{P}}_n[I(O,\mathrm{P})] + o_{\mathrm{P}}(n^{-1/2}).$$

to avoid pathological counter-examples, an additional technical regularity condition is imposed, which gives the notion of *regular* asymptotically linear (RAL) estimators. The function I is referred to as the *influence function* of the RAL estimator. As an example, the arguments above showed that the estimator $\hat{\theta}_n$ is a RAL estimator with influence function $I(O,\mathrm{P}) = \varphi(O,\nu) - \mathrm{P}[\varphi(O,\nu)]$.

By the central limit theorem, every RAL estimator is asymptotically normally distributed with variance given by the squared expectation of its influence function, and hence every RAL estimator is uniquely characterized by its influence function. All RAL estimators for a given statistical problem can thus be compared in terms of their asymptotic variance, and so it is natural to search for the RAL estimator with lowest possible variance, as this will provide the most efficient way of estimating the parameter of interest.

Now, it can be shown that there is a one-to-one correspondence between influence functions of RAL estimators and gradients of a statistical problem (see remark 3.7): Every RAL estimator will have a gradient as its influence function, and for every gradient there exists a RAL estimator with that gradient as its influence function. Furthermore, and in line with proposition 3.10, the RAL estimator with minimal asymptotic variance will have the canonical gradient as its influence function.

Example 4.1 (Efficient estimation of the ATE)
Returning to the problem of estimating the average treatment effect, we already argued in section 3.1 that the function φ_3 defined in (2) fulfills the Neyman orthogonality condition. In fact, it can be shown $\varphi_3 - \mathrm{P}[\varphi_3]$ indeed is the canonical gradient of the statistical problem $(\mathcal{P}, \Psi)$ when we impose no assumptions on the model $\mathcal{P}$. Hence, assuming A1 and A2, the discussion above shows that the estimator $\Psi_3(\hat{\mathbb{P}}_n, \hat{f}_n, \hat{\pi}_n)$ will converge to the true value $\theta = \Psi(\mathrm{P})$ at $n^{-1/2}$-rate, and that this estimator is efficient, meaning that it is asymptotically unbiased with the lowest possible asymptotic variance. In particular, with an estimate of the asymptotic variance, we can construct a Wald confidence interval of the parameter θ, which will (asymptotically) be the smallest possible. •

Remark 4.2. When we say that we impose "no assumptions" on the model $\mathcal{P}$, this is, strictly speaking, not completely true. For instance, as a minimum, we would need to impose that the function π is uniformly bounded away from 0 and 1 for the function φ_3 to be well-defined. In addition, though we do not want to impose parametric assumptions on f and π, for estimation to be possible at all we would need to impose some minimum amount of regularity, as for instance, continuity or a uniformly bounded total variation norm. Fortunately, imposing such regularity conditions does not affect the conclusion obtained in the example above. However, more restrictive assumptions, such as, for instance, randomization of the treatment *can* change the tangent space and thereby potentially also the canonical gradient and the information bound (see section 5 for a few additional comments). •

5 Further directions and topics

Semi-parametric inference In this note, our main example has been the ATE with no assumption on the model $\mathcal{P}$, which is called a fully non-parametric setting. We could change this by imposing assumptions on $\mathcal{P}$; for instance, we might assume randomization of the treatment, meaning that $A \perp\!\!\!\perp X$, or we might impose a particular functional form on f or π, e.g., that π is a linear function. Such assumptions restrict the model space $\mathcal{P}$ and can impact how efficiently we can estimate our parameter of interest. In a fully non-parametric setting, the calculation of the canonical gradient is somewhat simpler compared to the "truly" semi-parametric setting, where we impose additional assumptions on $\mathcal{P}$. This is because, in the non-parametric setting, the tangent space Γ_{P} is the whole space $\mathbb{H}_0$ of mean zero functions, and hence we simply need to find a function φ that fulfills (8); for semi-parametric models, the tangent space will be a proper subset of $\mathbb{H}_0$ and hence we also need to make sure that φ lies in this smaller space, and this can turn out to be a challenging task. The geometric tools from Hilbert space theory are particular useful for this job; Tsiatis [2007] and Bickel et al. [1993] treat this in great detail.

Empirical process theory Assumption A1, which essentially allows us to "remove the hat" from the nuisance estimator in φ, can be verified in a broad range of cases using empirical process theory. This theory provides extremely useful results, but the details of it demands some attention to measure theoretic subtleties. It is a mathematically interesting study which is covered thoroughly in, e.g., van der Vaart and Wellner [1996] and more briefly in van der Vaart [2000]. Alternatively, similar results can be obtained using **sample splitting**, which relies on much simpler mathematical tools (see for instance the appendix of Chernozhukov et al. [2018]).

Construction of estimators In this note we have mostly been concerned with the mathematical theory involved in defining the canonical gradient and showing that estimator that have this gradient as their influence functions are attractive. On the other hand, we have not been so much concerned with the actual *construction* of such estimator, except for the short example 4.1 above. There, we constructed an estimator by solving the efficient score equation; a similar strategy is taken and considered in much more detail in Chernozhukov et al. [2018]. A different approach is taken in the targeted minimum loss-based estimator (TMLE) framework [van der Laan and Rubin, 2006, van der Laan and Rose, 2011]; good introductory sources for the framework are van der Laan et al. [2014] and Rosenblum and van der Laan [2011].

Nuisance estimators In this note we have assumed we were given a nuisance estimator, such that A2 holds. However, this of course needs to be verified for the specific estimator of choice. For random forest some recent results in this direction can be found in Wager and Walther [2015], Duroux and Scornet [2016], Athey et al. [2019].

Doubly robustness A more careful analysis of the remainder term $\mathcal{O}_{\mathrm{P}}(\|\hat{\nu}_n - \nu\|_{\mathcal{V}}^2)$ from (5) will in some cases show that this has the special form of a *product* of two nuisance parameter estimation errors – in particular, this holds for the ATE setting, where the remainder term becomes the product $\|\hat{\pi}_n - \pi_0\| \times \|\hat{f}_n - f_0\|$. This implies that the estimator will be consistent if just *one* of the nuisance estimators is correctly specified, and hence it is *robust* again (some) model mis-specification. This topic has been analyzed in many articles, see for instance Glynn and Quinn [2010] and Bang and Robins [2005].

Undersmoothing TODO

Functional derivatives TODO: Functional delta method and linearization.

References

S. Athey, J. Tibshirani, S. Wager, et al. Generalized random forests. *Annals of Statistics*, 47(2):1148–1178, 2019.

H. Bang and J. M. Robins. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973, 2005.

P. J. Bickel, C. A. Klaassen, P. J. Bickel, Y. Ritov, J. Klaassen, J. A. Wellner, and Y. Ritov. *Efficient and adaptive estimation for semiparametric models*, volume 4. Johns Hopkins University Press Baltimore, 1993.

V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, and W. K. Newey. Double machine learning for treatment and causal parameters. Technical report, cemmap working paper, 2016.

V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins. Double/debiased machine learning for treatment and structural parameters, 2018.

R. Duroux and E. Scornet. Impact of subsampling and pruning on random forests. *arXiv preprint arXiv:1603.04261*, 2016.

A. N. Glynn and K. M. Quinn. An introduction to the augmented inverse propensity weighted estimator. *Political analysis*, 18(1):36–56, 2010.

M. Rosenblum and M. J. van der Laan. Simple examples of estimating causal effects using targeted maximum likelihood estimation. Technical report, Berkeley Division of Biostatistics Working Paper Series, 2011.

A. Tsiatis. *Semiparametric theory and missing data*. Springer Science & Business Media, 2007.

M. J. van der Laan and S. Rose. *Targeted learning: causal inference for observational and experimental data*. Springer Science & Business Media, 2011.

M. J. van der Laan and D. Rubin. Targeted maximum likelihood learning. *The international journal of biostatistics*, 2(1), 2006.

M. J. van der Laan, D. Benkeser, and O. Sofrygin. Targeted minimum loss-based estimation. *Wiley StatsRef: Statistics Reference Online*, pages 1–8, 2014.

A. van der Vaart and J. Wellner. *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer Science & Business Media, 1996.

A. W. van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

S. Wager and G. Walther. Adaptive concentration of regression trees, with application to random forests. *arXiv preprint arXiv:1503.06388*, 2015.

L. Wasserman. *All of nonparametric statistics*. Springer Science & Business Media, 2006.