April 29, 2021

Influence functions and functional derivatives

Helene Rytgaard and Anders Munch

I think it is important to point out that this is background material. They are not supposed to understand everything fully and, particularly, they are not supposed to write like this in their reports (maybe only be able to grasp it intutively?).

I think it will also be very useful to write a small overview of the note here :-) Perhaps point the reader the most important – and maybe most practical – results. In a way, it gets most practical in Section and then the sections before serve as the (crucial) background story.

Moreover, I think for Section we should focus on the ATE? As part of that particularly present the efficient influence function for the ATE.

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.

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 components 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 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 $O_1, \ldots, O_n$, $n \in \mathbb{N}$ of a random variable $O \in \mathcal{O}$ distributed according to an unknown distribution function P_0

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 $\Psi_1 \colon \mathcal{P} \to \mathbb{R}$ on distributions $P \in \mathcal{P}$, for our purposes defined as

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

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. In (1) we abuse notation and consider Ψ_1 *both* as a map defined on the full model $\mathcal{P}$ *and* as a map defined on the product space containing μ_X and f. To avoid confusion, we always specify on what domain Ψ_1 is defined in the following. •

ref

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 $(\mathcal{P}, \Psi)$ taken 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 $\Psi_1(\mu_X, f) = \Psi_2(\mu, \pi)$ and $\Psi_1(\mu_X, f) = \Psi_3(\mu, f, \pi)$, where

$$\Psi_2(\mu, \pi) := \int \left\{ \frac{a\, y}{\pi(x)} - \frac{(1 - a)\, y}{1 - \pi(x)} \right\} \mathrm{d}\mu(y, a, x),$$

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

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.

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

(1) Estimate the nuisance parameters

(2) Plug the estimates into the expression for target parameter Ψ

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

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

which yields an estimator for the target parameter.

... Does it really matter which of these parametrizations we pick? And could we not just pick 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. 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

$$\hat{\theta}_n := \int_{-\infty}^{x} \hat{f}_n(z)\,\mathrm{d}z, \quad \text{and} \quad \hat{f}_n(z) = \hat{\mathbb{P}}_n[k_h(z, \cdot)], \tag{2}$$

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 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.

Add explicit bias variance decomposition?

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 can be problematic. With this example in mind, we should not be too confident about our ATE estimator. In fact, with a bit more work it can be showed that similar phenomenon happens if we use the first estimator suggested above.

todo

The issue from the toy example can be understood more generally by considering the decomposition

$$\begin{aligned}
\sqrt{n}\left(\hat{\theta}_n - \theta\right) &= \sqrt{n}\left(\Psi(\hat{\mathbb{P}}_n, \hat{\nu}_n) - \Psi(\mathrm{P}, \nu)\right) \\
&= \mathbb{G}_n[\varphi(\cdot, \hat{\nu}_n)] + \sqrt{n}\left\{\Psi(\mathrm{P}, \hat{\nu}_n) - \Psi(\mathrm{P}, \nu)\right\} \\
&= \mathbb{G}_n[\varphi(\cdot, \hat{\nu}_n)] + \sqrt{n}\left\{\mathrm{D}_\nu \Psi(\hat{\nu}_n - \nu) + \mathcal{O}_{\mathrm{P}}(\|\hat{\nu}_n - \nu\|^2)\right\}.
\end{aligned}$$

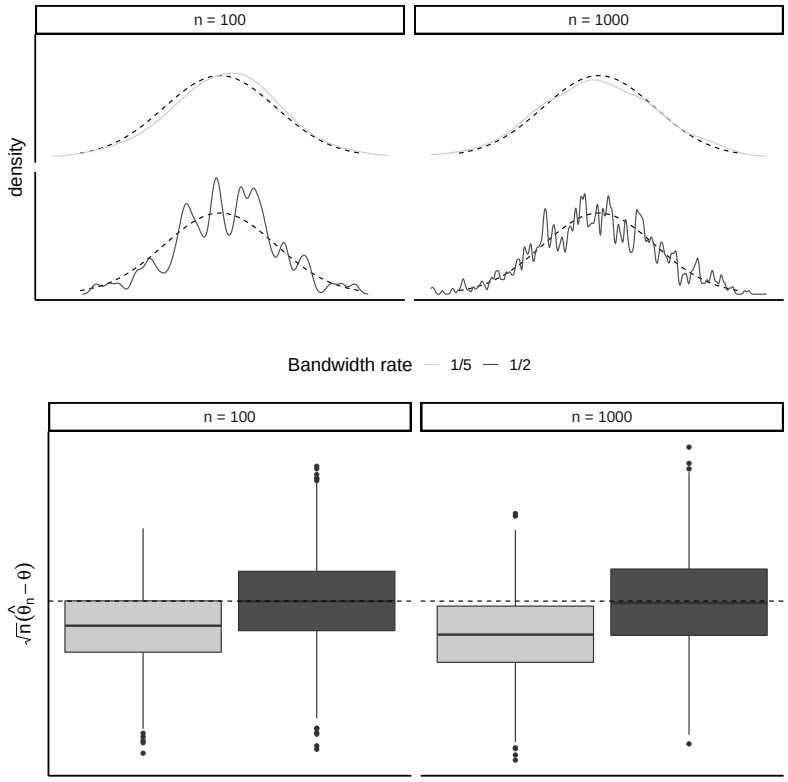

Figure 1: Simulation illustrating the problem for the plug-in approach. The top 2 rows give representative kernel density estimates for two different bandwidths scaling with n; the dashed line is the density of the distribution used to generate the data. The last row gives the distribution of the centralized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

Here we use $\mathrm{D}_\nu\Psi$ to denote some kind of derivative (so be defined shortly) of the map $\nu \mapsto \Psi(\mathrm{P}, \nu)$, where P is held fixed, and $\|\cdot\|$ is a norm defined on the space in which the nuisance estimator $\hat{\nu}_n$ takes it values – typically a function space. Using empirical process theory or sample spitting, it can in many cases be shown that $\mathbb{G}_n[\varphi(\cdot, \hat{\nu}_n)] = \mathbb{G}_n[\varphi(\cdot, \nu)] + \mathcal{o}_\mathrm{P}(1)$ [van der Vaart and Wellner, 1996, van der Vaart, 2000, Chernozhukov et al., 2018]. Furthermore, if we assume that $\hat{\nu}_n$ can be estimated at rate $n^{-1/4}$, meaning that $\|\hat{\nu}_n - \nu\| = \mathcal{o}_\mathrm{P}(n^{-1/4})$, the decomposition becomes

can we make this more practical and give an example?

$\mathbb{G}_n$ needs to be defined

$\mathcal{o}_\mathrm{P}$ needs to be defined

$$\sqrt{n}\left(\hat{\theta}_n - \theta\right) = \mathbb{G}_n[\varphi(\cdot, \nu)] + \mathrm{D}_\nu\Psi\left(\sqrt{n}(\hat{\nu}_n - \nu)\right) + \mathcal{o}_\mathrm{P}(1), \tag{3}$$

where we use that a derivative $\dot{\Psi}_\nu$ should be linear. The first term of this expression is controlled by reference to the central limit theorem, but unless we are able to estimate ν at the improved parametric rate of $n^{-1/2}$, we cannot hope to control the second term. Indeed, the fact that the default kernel estimator from the example converges at rate $n^{-2/5}$ is what ruins the asymptotic behavior of the target estimator. In the example we showed that the target estimator could be improved by using a suitably undersmoothed density estimator. However, we would not in general know how to choose a properly undersmoothed nuisance estimator, so this strategy is not applicable in practice (though see ...). If we want to use flexible machine learning estimators (which cannot be expected to converge at parametric rate) a better strategy is to design the map to Ψ such that its partial derivative $\mathrm{D}_\nu\Psi$ equals 0 at ν. This would make the second term in (3) vanish, and we would be able to estimate

check

ref to undersmoothed HAL

the target parameter θ at parametric rate, while still using a nuisance estimator converging only at rate $n^{-1/4}$.

In the next section we formally consider how to define a functional derivative, and then indicate how this can be used to design Ψ with partial derivative equal to 0. In section 4 we show how these results in addition allow us to analyze and construct *efficient* estimators of θ, that is, estimators with lowest possible asymptotic variance.

jeg er dum, det forstår jeg ikke? -> parametrization

3 Functional derivatives and the canonical gradient

In this section...

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 $h \in \mathcal{M}$

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

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

$$\left\| \frac{\Psi(x + \varepsilon h) - \Psi(x)}{\varepsilon} - \dot{\Psi}_x(h) \right\|_{\mathcal{Y}} \longrightarrow 0, \quad \text{when} \quad \varepsilon \longrightarrow 0. \tag{4}$$

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.,

$$\dot{\Psi}_x(h) = \partial_0 \, \Psi(x + \varepsilon h) := \frac{\partial}{\partial \varepsilon}\bigg|_{\varepsilon = 0} \Psi(x + \varepsilon h). \tag{5}$$

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 (5). Gâteaux differentiability is used by Chernozhukov et al. [2018] to define *Neyman orthogonality*: A function $\varphi\colon \mathcal{Z} \times \mathcal{V} \to \mathbb{R}$ fulfills the Neyman orthogonality condition (wrt. $\mathcal{V}$) if the map (as defined in section 2)

hely would like to change this notation
define the sample space $\mathcal{Z}$
proper def?

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

is Gâteaux differentiable at ν with vanishing derivative, i.e., $\dot{F}_\nu(h) = 0$ for all $h \in \mathcal{V}$.

Neyman orthogonality ensures that the second component of the decomposition in (3) vanishes, and as the condition can be checked using (5) it is rather straightforward to verify that a given function fulfills the condition. Still, Gâteaux is a 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. Similarly, to get a richer theory in the functional setting 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 (4) 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 a *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 (5) 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 $\mathbb{R}^3$, see picture).

Make the picture

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 $\dot{\mathcal{P}}_x$.

Finally, 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

define this earlier

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

3.2 Tangent spaces and gradients for statistical problems

And, start with motivation? Very interested in tangent space + canonical gradient + information bound, etc.

bla bla

Now that we properly defined differentiability of the statistical estimation problem, we are now ready to introduce so-called *tangent spaces* for the model $\mathcal{P}$, and particularly, a specific element of closure of the linear space of the tangent space called the *canonical gradient*, or, as we will refer to it in Section 4, the *efficient influence function*. The latter is central component in our analysis of the statistical estimation problem, representing the optimal asymptotic variance (Lemma ??) + debiasing? –> jeg skal lige forstå slutningen dette afsnit.

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,

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

should there be added some note about this – always sensible thing, or need to assume something?

The definition of a tangent space for a collection of probability measures simple 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*.

first time canonical gradient is mentioned, I think we need motivation, c.f., above :)

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. Restricting attention to such submodels for which the function

interval...

$$\dot{\ell}_0 = \partial_0 \log(p_\varepsilon)$$

exists as an element in $\mathcal{L}^2_\mathrm{P}$, we call $\dot{\ell}_0$ the *score* of the submodel p_ε. This reflects the terminology used in likelihood inference for ordinary parametric models.

Proposition 3.6. *We denote by $\Gamma_\mathrm{P} := \overline{\operatorname{span}}\{\dot{\ell}_0\} \subset \mathcal{L}^2_\mathrm{P}$ the closed linear span of the tangent space $\dot{\mathcal{P}}_\mathrm{P}$ of all score functions as defined above, considered as a subset of $\mathcal{L}^2_\mathrm{P}$.*

todo/finish. Pretty sure this results is correct...

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

Proof. Something like this: We can also represent $\mathcal{P}$ as a subset of $\mathcal{L}^2_\mu$ through the map $\mathrm{P} \mapsto \sqrt{p}$, and the topology on $\mathcal{P}$ induced in this way is the same as that induced by $\|\cdot\|_{\mathcal{M}_\mu}$ [Bickel et al., 1993]. This implies that convergence of $(\mathrm{P}_{\varepsilon h} - \mathrm{P}_0)\varepsilon^{-1}$ in $\mathcal{M}_\mu$ is equivalent to convergence of convergence of $(\sqrt{p_{\varepsilon h}} - \sqrt{p_0})\varepsilon^{-1}$ in $\mathcal{L}^2_\mu$. Then Gâteaux + map with $/p_0$... □

check – also the generalized versions on the whole space?

finish

Proposition 3.6 in essence states that we can think of the tangent space as the (closed linear span of) the collection of score functions for all parametric submodels $\mathcal{P}_\varepsilon$ passing through P; hence we shall often simply identify $\dot{\mathcal{P}}_\mathrm{P}$ with Γ_P. This representation in turn implies the following useful representation of the canonical gradient.

Proposition 3.7. *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*

$$\partial_0 \Psi(\mathrm{P}_\varepsilon) = \langle \varphi_\mathrm{P}, \dot{\ell}_0 \rangle_\mathrm{P} \tag{6}$$

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

canonical gradient not used in the statement this proposition? perhaps add extra equality to (6)?

Proof. The chain rule for Hadamard derivative implies that $\partial_0 \Psi(\mathrm{P}_\varepsilon) = \dot{\Psi}_\mathrm{P}(\partial_0 \mathrm{P}_\varepsilon)$, and then the representation given by proposition 3.6 implies that this expression equals $\Phi_\mathrm{P}(\dot{\ell}_0)$ for some continuous linear functional $\Phi_\mathrm{P} \colon \Gamma_\mathrm{P} \to \mathbb{R}$. 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 $\dot{\ell}_0 \in \Gamma_\mathrm{P}$. □

Combine this and next remark into a proof sketch

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 3.7 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].

I think this is correct

As with the identification $\dot{\mathcal{P}}_\mathrm{P} = \Gamma_\mathrm{P}$ we also identify the function φ_P with the canonical gradient $\dot{\Psi}_\mathrm{P}$. Note that if Γ_P is a proper subset of $\mathcal{L}^2_\mathrm{P}$ there can be several functions $\tilde{\varphi} \in \mathcal{L}^2_\mathrm{P}$ fulfilling condition (6); 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 representation of the tangent space and the gradient is very useful.

3.3 Properties of the canonical gradient

Seems useful to distinguish this part from the remainder of Section 3, but I may not be right about this. I am not sure I understood this result yet, but seems nice :)

wait, is the decomposition needed for this result, or can we wait with returning to the decomposition till talking about estimation?

Returning to the decomposition in (3) (Section 2) we have the following results.

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 $\varphi(\cdot, \nu(\mathrm{P}))$ is the canonical gradient of $(\mathcal{P}, \Psi)$ then, under regularity conditions, φ fulfills the Neyman orthogonality condition (wrt. $\dot{\nu}_\mathrm{P}(\dot{\mathcal{P}}_\mathrm{P})$), i.e., $\partial_0 \mathrm{P}[\varphi(\cdot, \nu(\mathrm{P}) + \varepsilon h)] = 0$ for all $h \in \dot{\nu}_\mathrm{P}(\dot{\mathcal{P}}_\mathrm{P})$.*

can these be made precise?

ok? Or should it actually be $\nu(\mathcal{P})$?

some unaligned defintions here?

Proof. Under regularity conditions that allow us to interchange integration and differentation in the following, we have for any differentiable path $\{\mathrm{P}_\varepsilon\}_\varepsilon \subset \mathcal{P}$ with $\mathrm{P}_0 = \mathrm{P}$ that

$$\begin{aligned}\partial_0 \Psi(\mathrm{P}_\varepsilon) &= \partial_0 \int p_\varepsilon(z)\varphi(z, \nu(\mathrm{P}_\varepsilon))\,\mathrm{d}\mu(z) \\ &= \int p\{\partial_0\varphi(z, \nu(\mathrm{P}_\varepsilon))\} + \{\partial_0 p_\varepsilon(z)\}\varphi(z, \nu(\mathrm{P}))\,\mathrm{d}\mu(z) \\ &= \partial_0 \mathrm{P}[\varphi(\cdot, \nu(\mathrm{P}_\varepsilon))] + \langle \dot{\ell}_0, \varphi(\cdot, \nu(\mathrm{P}))\rangle_{\mathrm{P}},\end{aligned} \tag{7}$$

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 3.7 and (7) imply that $\partial_0 \mathrm{P}[\varphi(\cdot, \nu(\mathrm{P}_\varepsilon))] = 0$. Assuming Hadamard differentiability of $\mathrm{P} \mapsto \nu(\mathrm{P})$ and $\nu \mapsto F(\nu) = \mathrm{P}[\varphi(\cdot, \nu)]$ we can write $\partial_0 \mathrm{P}[\varphi(\cdot, \nu(\mathrm{P}) + \varepsilon h)] = \partial_0 \dot{F}_{\nu(\mathrm{P})}(h)$, for $h \in \dot{\nu}_{\mathrm{P}}(\dot{\mathcal{P}}_{\mathrm{P}})$; hence (by definition of $\dot{\mathcal{P}}_{\mathrm{P}}$) we can find path P_ε such that $\dot{\nu}_{\mathrm{P}}(\partial_0 \mathrm{P}_\varepsilon) = h$, and then picking this path in the expression above we get $\partial_0 \dot{F}_{\nu(\mathrm{P})}(h) = 0$. □

Maybe simplify this simply to the statement that the EIF fulfills the N-O condition and then have the N-O as a proper def earlier?

finish the argument with something like this

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 above only used that $\varphi(\cdot, \nu(\mathrm{P}))$ satisfies (6), and not that $\varphi(\cdot, \nu(\mathrm{P})) \in \Gamma_{\mathrm{P}}$; hence, any gradient for the statistical problem will fulfill the Neyman orthogonality condition (under regularity conditions). •

I think this must be correct

Debiasing property of an estimator based on the canonical gradient. The result shows that estimators based on the canonical gradient has a "build-in" so-called *debiasing* mechanism because the first order bias (the second expression in (3)), due to estimation of the nuisance parameter $\hat{\nu}_n$, vanishes. This debiasing mechanism is crucial for $n^{-1/2}$-rate inference of a target parameter in a statistical model with nuisance parameters that are not estimable at this rate themselves.

I sort of feel like moving this to the next section? It is not until here we need the decomposition of (3), am I right?

Proposition 3.10. *I just want this to be van der Vaart Lemma 25.19, i.e., variance of the canonical gradient is the Cramér-Rao lower bound.*

Proof. Cauchy-Schwartz. □

What is the intuitive explanation of this result? In standard calculus, the derivatives are orthogonal – can this intuition explain the results?

4 Estimation based on the canonical gradient

proposed change to section title here

If you agree on the above (moving the debiasing remark down), I think we should return to the decomposition in (3). The first nice property to have is now exactly debiasing. Then we can move on to define asymptotically linear estimators and constructing an asymptotically linear estimator based on the decomposition in (3) which has influence function equal to the efficient influence function. If we have the result that all influence functions are gradients, then Proposition 3.10 gives asymptotic efficiency if the remainder is asymptotic negligible.

The last result of the previous section demonstrates that estimators based on the canonical gradient provide target parameter estimators with no first order bias. In this section we show that they also provide *efficient* estimators.

briefly to this

- RAL estimators and IFs
- Connection to canonical gradient

5 Further directions

- *Semi*-parametric inference.
- Undersmoothing
- Empirical process theory (or/versus sample splitting)
- Nuisance estimators (and their rate of convergence)
- Other/further topic on functional derivatives / derivatives of operators:
  - Functional delta method
  - Analysis of estimator
- Concrete strategies for finding estimators:
  - Debiasing / solving the (efficient) score equation
  - TMLE
  - ...?

References

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.

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.

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