Influence functions and functional derivatives

Anders Munch

May 11, 2021

Outline

Disclaimer about the note

- The note is work in progress, and we have not used it before – you are very welcome to comment on weird/unclear passages.

Disclaimer about the note

- The note is work in progress, and we have not used it before – you are very welcome to comment on weird/unclear passages.
- You should see it as a service – some exact mathematical statements are collected there if you care about it, but the important part is the intuition which we talk about today.

Disclaimer about the note

- The note is work in progress, and we have not used it before – you are very welcome to comment on weird/unclear passages.
- You should see it as a service – some exact mathematical statements are collected there if you care about it, but the important part is the intuition which we talk about today.
- Do NOT write like this in your report!

A statistical problem

We call a collection of probability measures $\mathcal{P}$ together with a functional $\Psi\colon \mathcal{P} \to \mathbb{R}$ *a statistical problem*.

A statistical problem

We call a collection of probability measures $\mathcal{P}$ together with a functional $\Psi \colon \mathcal{P} \to \mathbb{R}$ *a statistical problem*.

Example (Average treatment effect)

We are given n iid. sample of $O \sim \mathrm{P}$, with $\mathrm{P} \in \mathcal{P}$ and where $O = (X, A, Y)$, with $X \in \mathbb{R}^d$, $A \in \{0, 1\}$, and $Y \in \{0, 1\}$. We want to estimate the average treatment effect

$$\mathbb{E}_{\mathrm{P}}\left[f(1, X) - f(0, X)\right],$$

with $f(a, x) := \mathbb{E}_{\mathrm{P}}\left[Y \mid A = a, X = x\right]$. The target parameter is

$$\Psi(\mathrm{P}) = \mathbb{E}_{\mathrm{P}}\left[f_{\mathrm{P}}(1, X) - f_{\mathrm{P}}(0, X)\right].$$

A statistical problem

We call a collection of probability measures $\mathcal{P}$ together with a functional $\Psi\colon \mathcal{P} \to \mathbb{R}$ *a statistical problem*.

Example (Average treatment effect)

We are given n iid. sample of $O \sim \mathrm{P}$, with $\mathrm{P} \in \mathcal{P}$ and where $O = (X, A, Y)$, with $X \in \mathbb{R}^d$, $A \in \{0,1\}$, and $Y \in \{0,1\}$. We want to estimate the average treatment effect

$$\mathbb{E}_{\mathrm{P}}\left[f(1, X) - f(0, X)\right],$$

with $f(a, x) := \mathbb{E}_{\mathrm{P}}\left[Y \mid A = a, X = x\right]$. The target parameter is

$$\Psi(\mathrm{P}) = \mathbb{E}_{\mathrm{P}}\left[f_{\mathrm{P}}(1, X) - f_{\mathrm{P}}(0, X)\right].$$

Target and nuisance parameters

Target and nuisance parameters

Target parameter Low-dimensional, scientifically meaningful.

Target and nuisance parameters

Target parameter Low-dimensional, scientifically meaningful.

Nuisance parameters Needed to express the target parameter.

Target and nuisance parameters

Target parameter Low-dimensional, scientifically meaningful.

Nuisance parameters Needed to express the target parameter.

Example (ATE)

The ATE can be written as $\Psi(\mathrm{P}) = \mathrm{P}[\varphi_1] = \mathrm{P}[\varphi_2] = \mathrm{P}[\varphi_3]$, for

$$\varphi_1(o; f) := f(1, x) - f(0, x),$$

$$\varphi_2(o; \pi) := \frac{a\,y}{\pi(x)} - \frac{(1-a)\,y}{1-\pi(x)},$$

$$\varphi_3(o; f, \pi) := \varphi_1(o; f) + \varphi_2(o; \pi) - \frac{a\,f(1, x)}{\pi(x)} + \frac{(1-a)\,f(0, x)}{1-\pi(x)},$$

with $f(a, x) := \mathbb{E}_{\mathrm{P}}\left[Y \mid A = a, X = x\right]$, $\pi(x) := \mathrm{P}(A = 1 \mid X = x)$.

$\mathrm{P}[\varphi]$ means

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

Infinite-dimensional nuisance parameters

A parametric setting means that $\mathcal{P}$ is finite-dimensional. We are interested in *nonparametric* or *semiparametric* settings which mean that $\mathcal{P}$ is infinite-dimensional.

Infinite-dimensional nuisance parameters

A parametric setting means that $\mathcal{P}$ is finite-dimensional. We are interested in *nonparametric* or *semiparametric* settings which mean that $\mathcal{P}$ is infinite-dimensional.

Having our data set and scientific question in mind, why would it be of interest to use infinite-dimensional nuisance parameters?

Infinite-dimensional nuisance parameters

A parametric setting means that $\mathcal{P}$ is finite-dimensional. We are interested in *nonparametric* or *semiparametric* settings which mean that $\mathcal{P}$ is infinite-dimensional.

Having our data set and scientific question in mind, why would it be of interest to use infinite-dimensional nuisance parameters?

Trying to control for confounding $\implies$ nice to have:

- flexible model
- many covariates

Toy example: Integrated kernel density

$\mathcal{P}$ consist all probability measures with continuous Lebesgue-density (this is an infinite-dimensional space). We want to estimate $F(x) = \mathrm{P}(X \leq x)$ for unknown $\mathrm{P} \in \mathcal{P}$.

Toy example: Integrated kernel density

$\mathcal{P}$ consist all probability measures with continuous Lebesgue-density (this is an infinite-dimensional space). We want to estimate $F(x) = \mathrm{P}(X \leq x)$ for unknown $\mathrm{P} \in \mathcal{P}$. Our target parameter is then $\theta = \Psi(\mathrm{P}) = F_{\mathrm{P}}(x)$ which we can express as

$$\Psi(\mathrm{P}) = \Psi_0(f) := \int_{-\infty}^{x} f(z)\,\mathrm{d}z, \quad \text{for} \quad \mathrm{P} = f \cdot \lambda,$$

because of our assumption about $\mathcal{P}$.

Toy example: Integrated kernel density

$\mathcal{P}$ consist all probability measures with continuous Lebesgue-density (this is an infinite-dimensional space). We want to estimate $F(x) = \mathrm{P}(X \leq x)$ for unknown $\mathrm{P} \in \mathcal{P}$. Our target parameter is then $\theta = \Psi(\mathrm{P}) = F_{\mathrm{P}}(x)$ which we can express as

$$\Psi(\mathrm{P}) = \Psi_0(f) := \int_{-\infty}^{x} f(z)\,\mathrm{d}z, \quad \text{for} \quad \mathrm{P} = f \cdot \lambda,$$

because of our assumption about $\mathcal{P}$. We want to use **machine learning** (!) for this problem, so use a kernel estimator, i.e.,

$$\hat{f}_n(x) = \hat{\mathbb{P}}_n[k_h(X, x)] = \frac{1}{n}\sum_{i=1}^{n} k_h(X_i, x),$$

where k_h is, e.g, $k_h(x, y) = \frac{1}{h} k\left(\frac{x-y}{h}\right)$, with k the density for the standard Gaussian distribution, and the bandwidth h is chosen using cross-validation.

Toy example: Integrated kernel density

$\mathcal{P}$ consist all probability measures with continuous Lebesgue-density (this is an infinite-dimensional space). We want to estimate $F(x) = \mathrm{P}(X \leq x)$ for unknown $\mathrm{P} \in \mathcal{P}$. Our target parameter is then $\theta = \Psi(\mathrm{P}) = F_{\mathrm{P}}(x)$ which we can express as

$$\Psi(\mathrm{P}) = \Psi_0(f) := \int_{-\infty}^{x} f(z)\,\mathrm{d}z, \quad \text{for} \quad \mathrm{P} = f \cdot \lambda,$$

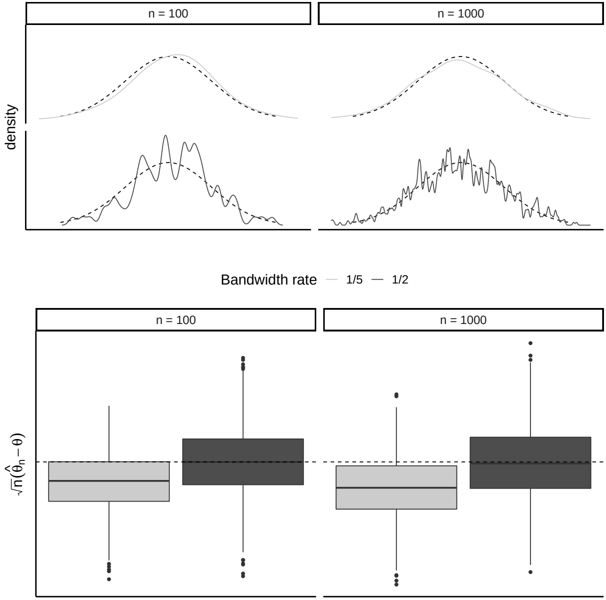
because of our assumption about $\mathcal{P}$. We want to use **machine learning** (!) for this problem, so use a kernel estimator, i.e.,

$$\hat{f}_n(x) = \hat{\mathbb{P}}_n[k_h(X, x)] = \frac{1}{n} \sum_{i=1}^{n} k_h(X_i, x),$$

where k_h is, e.g, $k_h(x, y) = \frac{1}{h} k\left(\frac{x-y}{h}\right)$, with k the density for the standard Gaussian distribution, and the bandwidth h is chosen using cross-validation. We then obtain the target estimator $\hat{\theta}_n = \Psi_0(\hat{f}_n)$.

How does this work in practice?

How does this work in practice?



What happened?

What happened?

Consider a general problem $(\mathcal{P}, \Psi)$ for which we can write $\Psi(\mathrm{P}) = \Psi_0(\mathrm{P}, \nu) = \mathrm{P}[\varphi(O, \nu)]$.

What happened?

Consider a general problem $(\mathcal{P}, \Psi)$ for which we can write $\Psi(\mathrm{P}) = \Psi_0(\mathrm{P}, \nu) = \mathrm{P}[\varphi(O, \nu)]$. We have

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

with $\mathbb{G}_n := \sqrt{n}(\hat{\mathbb{P}}_n - \mathrm{P})$ the empirical process.

What happened?

Consider a general problem $(\mathcal{P}, \Psi)$ for which we can write $\Psi(\mathrm{P}) = \Psi_0(\mathrm{P}, \nu) = \mathrm{P}[\varphi(O, \nu)]$. We have

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

with $\mathbb{G}_n := \sqrt{n}(\hat{\mathbb{P}}_n - \mathrm{P})$ the empirical process.

$\mathbb{G}_n[\varphi(O, \hat{\nu}_n)]$ determines the (main) variance

What happened?

Consider a general problem $(\mathcal{P}, \Psi)$ for which we can write $\Psi(\mathrm{P}) = \Psi_0(\mathrm{P}, \nu) = \mathrm{P}[\varphi(O, \nu)]$. We have

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

with $\mathbb{G}_n := \sqrt{n}(\hat{\mathbb{P}}_n - \mathrm{P})$ the empirical process.

$\mathbb{G}_n[\varphi(O, \hat{\nu}_n)]$ determines the (main) variance
$\Psi_0(\mathrm{P}, \hat{\nu}_n) - \Psi_0(\mathrm{P}, \nu)$ is bias!

What to do? – Taylor expansion

What to do? – Taylor expansion

Assume we could make a Taylor expansion of $\nu \mapsto \Psi_0(\mathrm{P}, \nu)$, so that

$$\Psi_0(\mathrm{P}, \hat{\nu}_n) - \Psi_0(\mathrm{P}, \nu) = \mathrm{D}_\nu \Psi_0[\hat{\nu}_n - \nu] + \mathcal{O}_{\mathrm{P}}(\|\hat{\nu}_n - \nu\|^2_{\mathcal{V}}).$$

What to do? – Taylor expansion

Assume we could make a Taylor expansion of $\nu \mapsto \Psi_0(\mathrm{P}, \nu)$, so that

$$\Psi_0(\mathrm{P}, \hat{\nu}_n) - \Psi_0(\mathrm{P}, \nu) = \mathrm{D}_\nu \Psi_0[\hat{\nu}_n - \nu] + \mathcal{O}_{\mathrm{P}}(\|\hat{\nu}_n - \nu\|_{\mathcal{V}}^2).$$

The decomposition then becomes

$$\begin{aligned}
\sqrt{n}\left(\hat{\theta}_n - \theta\right) &= \mathbb{G}_n[\varphi(O, \hat{\nu}_n)] && (1)\\
&\quad + \mathrm{D}_\nu \Psi_0\left[\sqrt{n}(\hat{\nu}_n - \nu)\right] && (2)\\
&\quad + \mathcal{O}_{\mathrm{P}}(\sqrt{n}\|\hat{\nu}_n - \nu\|_{\mathcal{V}}^2). && (3)
\end{aligned}$$

What to do? – Taylor expansion

Assume we could make a Taylor expansion of $\nu \mapsto \Psi_0(\mathrm{P}, \nu)$, so that

$$\Psi_0(\mathrm{P}, \hat{\nu}_n) - \Psi_0(\mathrm{P}, \nu) = \mathrm{D}_\nu \Psi_0[\hat{\nu}_n - \nu] + \mathcal{O}_{\mathrm{P}}(\|\hat{\nu}_n - \nu\|_{\mathcal{V}}^2).$$

The decomposition then becomes

$$\begin{aligned}
\sqrt{n}\left(\hat{\theta}_n - \theta\right) &= \mathbb{G}_n[\varphi(O, \hat{\nu}_n)] && (1)\\
&\quad + \mathrm{D}_\nu \Psi_0\left[\sqrt{n}(\hat{\nu}_n - \nu)\right] && (2)\\
&\quad + \mathcal{O}_{\mathrm{P}}(\sqrt{n}\|\hat{\nu}_n - \nu\|_{\mathcal{V}}^2). && (3)
\end{aligned}$$

(1) can be handled by empirical process theory or sample splitting

What to do? – Taylor expansion

Assume we could make a Taylor expansion of $\nu \mapsto \Psi_0(\mathrm{P}, \nu)$, so that

$$\Psi_0(\mathrm{P}, \hat{\nu}_n) - \Psi_0(\mathrm{P}, \nu) = \mathrm{D}_\nu \Psi_0[\hat{\nu}_n - \nu] + \mathcal{O}_{\mathrm{P}}(\|\hat{\nu}_n - \nu\|_{\mathcal{V}}^2).$$

The decomposition then becomes

$$\begin{aligned} \sqrt{n}\left(\hat{\theta}_n - \theta\right) &= \mathbb{G}_n[\varphi(O, \hat{\nu}_n)] && (1) \\ &\quad + \mathrm{D}_\nu \Psi_0\left[\sqrt{n}(\hat{\nu}_n - \nu)\right] && (2) \\ &\quad + \mathcal{O}_{\mathrm{P}}(\sqrt{n}\|\hat{\nu}_n - \nu\|_{\mathcal{V}}^2). && (3) \end{aligned}$$

(1) can be handled by empirical process theory or sample splitting

(2) is our focus! $\rightarrow$ make sense of this

What to do? – Taylor expansion

Assume we could make a Taylor expansion of $\nu \mapsto \Psi_0(\mathrm{P}, \nu)$, so that

$$\Psi_0(\mathrm{P}, \hat{\nu}_n) - \Psi_0(\mathrm{P}, \nu) = \mathrm{D}_\nu \Psi_0[\hat{\nu}_n - \nu] + \mathcal{O}_\mathrm{P}(\|\hat{\nu}_n - \nu\|^2_{\mathcal{V}}).$$

The decomposition then becomes

$$\begin{aligned}
\sqrt{n}\left(\hat{\theta}_n - \theta\right) = \mathbb{G}_n[\varphi(O, \hat{\nu}_n)] \qquad (1)\\
+ \mathrm{D}_\nu \Psi_0\left[\sqrt{n}(\hat{\nu}_n - \nu)\right] \qquad (2)\\
+ \mathcal{O}_\mathrm{P}(\sqrt{n}\|\hat{\nu}_n - \nu\|^2_{\mathcal{V}}). \qquad (3)
\end{aligned}$$

(1) can be handled by empirical process theory or sample splitting

(2) is our focus! $\rightarrow$ make sense of this

(3) is specific to the functional Ψ, but importantly the rate $\sqrt{n}\|\hat{\nu}_n - \nu\|_{\mathcal{V}} = o_\mathrm{P}(n^{-1/4})$ is sufficient; whether this holds then depends on the specific nuisance estimator.

What to do? – Taylor expansion

Assume we could make a Taylor expansion of $\nu \mapsto \Psi_0(\mathrm{P}, \nu)$, so that

$$\Psi_0(\mathrm{P}, \hat{\nu}_n) - \Psi_0(\mathrm{P}, \nu) = \mathrm{D}_\nu \Psi_0[\hat{\nu}_n - \nu] + \mathcal{O}_{\mathrm{P}}(\|\hat{\nu}_n - \nu\|_{\mathcal{V}}^2).$$

The decomposition then becomes

$$\begin{aligned}
\sqrt{n}\left(\hat{\theta}_n - \theta\right) = {} & \mathbb{G}_n[\varphi(O, \hat{\nu}_n)] && (1)\\
& + \mathrm{D}_\nu \Psi_0\left[\sqrt{n}(\hat{\nu}_n - \nu)\right] && (2)\\
& + \mathcal{O}_{\mathrm{P}}(\sqrt{n}\|\hat{\nu}_n - \nu\|_{\mathcal{V}}^2). && (3)
\end{aligned}$$

(1) can be handled by empirical process theory or sample splitting

(2) is our focus! $\rightarrow$ make sense of this

(3) is specific to the functional Ψ, but importantly the rate $\sqrt{n}\|\hat{\nu}_n - \nu\|_{\mathcal{V}} = o_{\mathrm{P}}(n^{-1/4})$ is sufficient; whether this holds then depends on the specific nuisance estimator.

Defining a functional derivative

What is a derivative?

Defining a functional derivative

What is a derivative?

A linear approximation $\dot{\Psi}_x$ to the map Ψ at $x \in \mathcal{M}$, i.e.,

$$\left\| \Psi(x + \varepsilon_n h_n) - \Psi(x) - \dot{\Psi}_x(\varepsilon_n h_n) \right\| = \mathcal{O}(\varepsilon_n),$$

when $\varepsilon_n \to 0$.

Defining a functional derivative

What is a derivative?

A linear approximation $\dot{\Psi}_x$ to the map Ψ at $x \in \mathcal{M}$, i.e.,

$$\left\| \Psi(x + \varepsilon_n h_n) - \Psi(x) - \dot{\Psi}_x(\varepsilon_n h_n) \right\| = \mathcal{O}(\varepsilon_n),$$

when $\varepsilon_n \to 0$.

This expression also makes sense for functionals (or operators) Ψ.

Defining a functional derivative

What is a derivative?

A linear approximation $\dot{\Psi}_x$ to the map Ψ at $x \in \mathcal{M}$, i.e.,

$$\left\| \Psi(x + \varepsilon_n h_n) - \Psi(x) - \dot{\Psi}_x(\varepsilon_n h_n) \right\| = \mathcal{O}(\varepsilon_n),$$

when $\varepsilon_n \to 0$.

This expression also makes sense for functionals (or operators) Ψ.

- For which h_n should this hold? Along "lines", "paths", or "uniformly" (h_n fixed, converging, or bounded)?

Defining a functional derivative

What is a derivative?

A linear approximation $\dot{\Psi}_x$ to the map Ψ at $x \in \mathcal{M}$, i.e.,

$$\left\| \Psi(x + \varepsilon_n h_n) - \Psi(x) - \dot{\Psi}_x(\varepsilon_n h_n) \right\| = \mathcal{O}(\varepsilon_n),$$

when $\varepsilon_n \to 0$.

This expression also makes sense for functionals (or operators) Ψ.

- For which h_n should this hold? Along "lines", "paths", or "uniformly" (h_n fixed, converging, or bounded)?
- Which norm on $\mathcal{M}$ should we use?

Defining a functional derivative

What is a derivative?

A linear approximation $\dot{\Psi}_x$ to the map Ψ at $x \in \mathcal{M}$, i.e.,

$$\left\| \Psi(x + \varepsilon_n h_n) - \Psi(x) - \dot{\Psi}_x(\varepsilon_n h_n) \right\| = \mathcal{O}(\varepsilon_n),$$

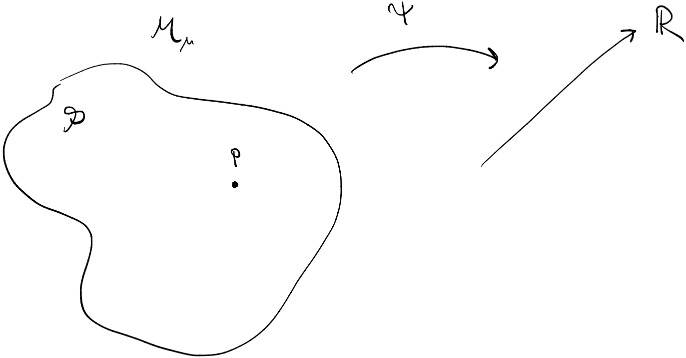
when $\varepsilon_n \to 0$.

This expression also makes sense for functionals (or operators) Ψ.

- For which h_n should this hold? Along "lines", "paths", or "uniformly" (h_n fixed, converging, or bounded)?
- Which norm on $\mathcal{M}$ should we use?
- In which space should we represent $\mathcal{P}$?

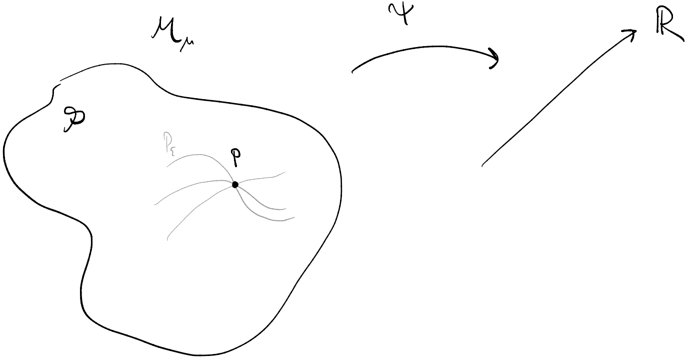
Pathwise Hadamard differentiability

Think of the gradient of a function defined on a manifold (surface).



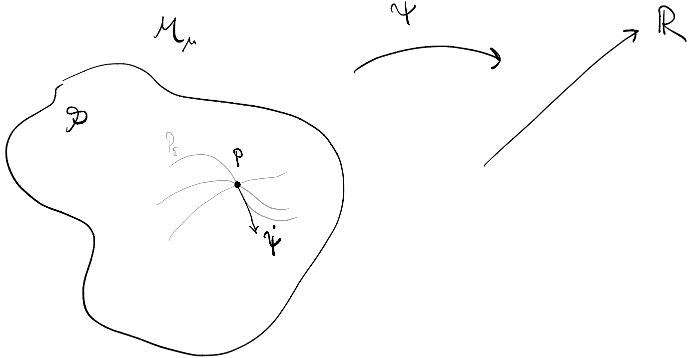
Pathwise Hadamard differentiability

Think of the gradient of a function defined on a manifold (surface).



Pathwise Hadamard differentiability

Think of the gradient of a function defined on a manifold (surface).



Canonical gradient

Definition (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*.

Canonical gradient

Definition (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*.

Characterizing property

With $\Gamma_{\mathrm{P}} := \overline{\mathrm{span}}\{\dot{\ell}_0\} \subset \mathcal{L}^2_{\mathrm{P}}$, where $\dot{\ell}_0 = \partial_0 \log(p_\varepsilon)$ is the score function of the sub-model P_ε, there exists a unique element $\varphi_{\mathrm{P}} \in \Gamma_{\mathrm{P}}$ such that

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

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

Canonical gradient for the ATE

Example (ATE)

When we make no assumptions about $\mathcal{P}$, the canonical gradient for the ATE problem

$$
\begin{aligned}
\varphi_{\mathrm{P}}(o; f, \pi) := f(1,x) - f(0,x) \\
&+ \frac{a\,y}{\pi(x)} - \frac{(1-a)\,y}{1-\pi(x)} \\
&- \frac{a\,f(1,x)}{\pi(x)} + \frac{(1-a)\,f(0,x)}{1-\pi(x)} \\
&- \Psi(\mathrm{P})
\end{aligned}
$$

Canonical gradient for the ATE

Example (ATE)

When we make no assumptions about $\mathcal{P}$, the canonical gradient for the ATE problem

$$\begin{aligned}\varphi_{\mathrm{P}}(o; f, \pi) := f(1,x) - f(0,x) \\ &+ \frac{a\,y}{\pi(x)} - \frac{(1-a)\,y}{1-\pi(x)} \\ &- \frac{a\,f(1,x)}{\pi(x)} + \frac{(1-a)\,f(0,x)}{1-\pi(x)} \\ &- \Psi(\mathrm{P})\end{aligned}$$

One way to show this is to first show that the tangent space Γ_{P} is the full subset $\mathbb{H}_0 \subset \mathcal{L}^2_{\mathrm{P}}$ of zero-mean functions, and then show that $\partial_0 \Psi(\mathrm{P}_\varepsilon) = \langle \varphi_{\mathrm{P}}, \dot{\ell}_0 \rangle_{\mathrm{P}}$ for all P_ε (see for instance Kennedy [2016]).

Neyman orthogonality

Theorem (Neyman orthogonality)

If $\Psi(\mathrm{P}) = \Psi_0(\mathrm{P}, \nu) = \mathrm{P}[\varphi(O, \nu(\mathrm{P}))]$ *and* $\varphi(\cdot, \nu) - \mathrm{P}[\varphi(O, \nu)]$ *is the canonical gradient of* $(\mathcal{P}, \Psi)$ *then* $\mathrm{D}_\nu \Psi_0 = 0$*.*

Neyman orthogonality

Theorem (Neyman orthogonality)

If $\Psi(\mathrm{P}) = \Psi_0(\mathrm{P}, \nu) = \mathrm{P}[\varphi(O, \nu(\mathrm{P}))]$ *and* $\varphi(\cdot, \nu) - \mathrm{P}[\varphi(O, \nu)]$ *is the canonical gradient of* $(\mathcal{P}, \Psi)$ *then* $\mathrm{D}_\nu \Psi_0 = 0$*.*

Debiasing

The *first order* bias, coming from $\Psi_0(\mathrm{P}, \hat{\nu}_n) - \Psi_0(\mathrm{P}, \nu)$, is removed.

Efficiency

Efficiency

Definition (RAL estimators)

An estimator $\hat{\theta}_n$ of the parameter $\theta = \Psi(\mathrm{P})$ under the model $\mathcal{P}$, is called *asymptotically linear* with *influence function* $\mathrm{IF}(\cdot, \mathrm{P}) \in \mathcal{L}^2_{\mathrm{P}}$, if $\mathrm{P}[\mathrm{IF}(O, \mathrm{P})] = 0$ for all $\mathrm{P} \in \mathcal{P}$, and

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

Efficiency

Definition (RAL estimators)

An estimator $\hat{\theta}_n$ of the parameter $\theta = \Psi(\mathrm{P})$ under the model $\mathcal{P}$, is called *asymptotically linear* with *influence function* $\mathrm{IF}(\cdot, \mathrm{P}) \in \mathcal{L}^2_{\mathrm{P}}$, if $\mathrm{P}[\mathrm{IF}(O, \mathrm{P})] = 0$ for all $\mathrm{P} \in \mathcal{P}$, and

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

Theorem (Efficient influence function)

The RAL estimator with lowest possible asymptotic variance has the canonical gradient as its influence function.

Constructing estimators 1: Solve the efficient score equation

Constructing estimators 1: Solve the efficient score equation

Find a parametrization $\Psi(\mathrm{P}) = \mathrm{P}[\varphi(O, \nu)]$ such that φ is the (canonical) gradient.

Constructing estimators 1: Solve the efficient score equation

Find a parametrization $\Psi(\mathrm{P}) = \mathrm{P}[\varphi(O, \nu)]$ such that φ is the (canonical) gradient. Then by Neyman orthogonality and assumptions we can write

$$\begin{aligned}
\sqrt{n}\left(\hat{\theta}_n - \theta\right) = \mathbb{G}_n[\varphi(O, \hat{\nu}_n)] & \\
+ \mathrm{D}_\nu \Psi_0\left[\sqrt{n}(\hat{\nu}_n - \nu)\right] & \qquad {\color{red}= 0} \\
+ \mathcal{O}_{\mathrm{P}}(\sqrt{n}\|\hat{\nu}_n - \nu\|_{\mathcal{V}}^2) &
\end{aligned}$$

Constructing estimators 1: Solve the efficient score equation

Find a parametrization $\Psi(\mathrm{P}) = \mathrm{P}[\varphi(O, \nu)]$ such that φ is the (canonical) gradient. Then by Neyman orthogonality and assumptions we can write

$$
\begin{aligned}
\sqrt{n}\left(\hat{\theta}_n - \theta\right) = \mathbb{G}_n[\varphi(O, \hat{\nu}_n)] \qquad & = \mathbb{G}_n[\varphi(O, \nu)] \\
+ \mathrm{D}_\nu \Psi_0\left[\sqrt{n}(\hat{\nu}_n - \nu)\right] \qquad & = 0 \\
+ \mathcal{O}_{\mathrm{P}}(\sqrt{n}\|\hat{\nu}_n - \nu\|_{\mathcal{V}}^2) \qquad & = o_{\mathrm{P}}(1)
\end{aligned}
$$

Constructing estimators 1: Solve the efficient score equation

Find a parametrization $\Psi(\mathrm{P}) = \mathrm{P}[\varphi(O, \nu)]$ such that φ is the (canonical) gradient. Then by Neyman orthogonality and assumptions we can write

$$
\begin{aligned}
\sqrt{n}\left(\hat{\theta}_n - \theta\right) &= \mathbb{G}_n[\varphi(O, \hat{\nu}_n)] && = \mathbb{G}_n[\varphi(O, \nu)] \\
&\quad + \mathrm{D}_\nu \Psi_0\left[\sqrt{n}(\hat{\nu}_n - \nu)\right] && = 0 \\
&\quad + \mathcal{O}_{\mathrm{P}}(\sqrt{n}\|\hat{\nu}_n - \nu\|_{\mathcal{V}}^2) && = o_{\mathrm{P}}(1) \\
&= \mathbb{G}_n[\varphi(O, \nu)] + o_{\mathrm{P}}(1).
\end{aligned}
$$

Constructing estimators 1: Solve the efficient score equation

Find a parametrization $\Psi(\mathrm{P}) = \mathrm{P}[\varphi(O, \nu)]$ such that φ is the (canonical) gradient. Then by Neyman orthogonality and assumptions we can write

$$
\begin{aligned}
\sqrt{n}\left(\hat{\theta}_n - \theta\right) &= \mathbb{G}_n[\varphi(O, \hat{\nu}_n)] & &= \mathbb{G}_n[\varphi(O, \nu)] \\
&\quad + \mathrm{D}_\nu \Psi_0\left[\sqrt{n}(\hat{\nu}_n - \nu)\right] & &= 0 \\
&\quad + \mathcal{O}_{\mathrm{P}}(\sqrt{n}\|\hat{\nu}_n - \nu\|^2_{\mathcal{V}}) & &= \mathcal{o}_{\mathrm{P}}(1) \\
&= \mathbb{G}_n[\varphi(O, \nu)] + \mathcal{o}_{\mathrm{P}}(1).
\end{aligned}
$$

Hence $\hat{\theta}_n$ is a RAL estimator, and if $\varphi - \mathrm{P}[\varphi]$ is the canonical gradient it will be *asymptotically efficient*.

Constructing estimators 1: Solve the efficient score equation

Find a parametrization $\Psi(\mathrm{P}) = \mathrm{P}[\varphi(O, \nu)]$ such that φ is the (canonical) gradient. Then by Neyman orthogonality and assumptions we can write

$$
\begin{aligned}
\sqrt{n}\left(\hat{\theta}_n - \theta\right) &= \mathbb{G}_n[\varphi(O, \hat{\nu}_n)] && = \mathbb{G}_n[\varphi(O, \nu)] \\
&\quad + \mathrm{D}_\nu \Psi_0\left[\sqrt{n}(\hat{\nu}_n - \nu)\right] && = 0 \\
&\quad + \mathcal{O}_{\mathrm{P}}(\sqrt{n}\|\hat{\nu}_n - \nu\|_{\mathcal{V}}^2) && = o_{\mathrm{P}}(1) \\
&= \mathbb{G}_n[\varphi(O, \nu)] + o_{\mathrm{P}}(1).
\end{aligned}
$$

Hence $\hat{\theta}_n$ is a RAL estimator, and if $\varphi - \mathrm{P}[\varphi]$ is the canonical gradient it will be *asymptotically efficient*.

This is the approach taken in Chernozhukov et al. [2018]. See also Example 4.1 of the note.

Constructing estimators 2: TMLE

References

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.

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