Double/Debiased Machine Learning (DML)

Yonghan Jung

Purdue University

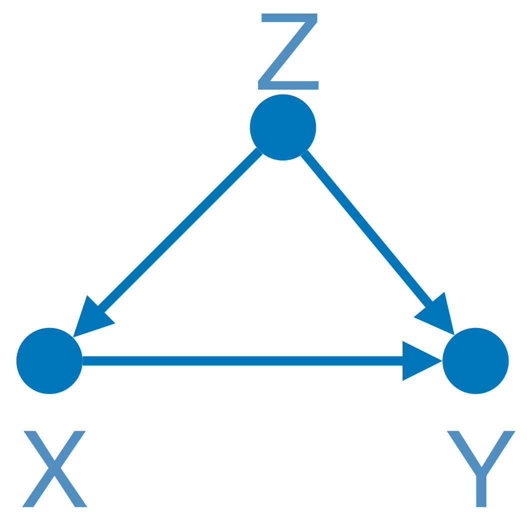
yonghanjung.me

Causal Effect Identification

Given {G, P, Q_0},

the causal effect identification (ID) task

finds the functional $C(P)$ s.t. $C(P) = Q_0$.



Causal graph (G)

$P(Z, X, Y)$

Distribution on G (P)

$Q_0 := \mathbb{E}[Y \mid do(x)]$

Causal Query (Q_0)

ID(G, P, Q_0)

ID algorithm

$C(P) = \sum_z \mathbb{E}[Y \mid x, z] P(z)$

Causal functional $C(P)$

s.t. $C(P) = Q_0$

Task of Causal Effect Estimation

Given {C(P), D},

the causal effect estimation (EST) task

finds the estimator T that estimates the query Q_0.

$C(P) = \sum_z \mathbb{E}[Y|x,z]P(z)$

Causal functional $C(P)$

s.t. $C(P) = Q_0$

$D \sim P(Z, X, Y)$

N samples ($N = |D|$**) drawn from P, where P is corresponding to G**

EST$(C(P), D)$

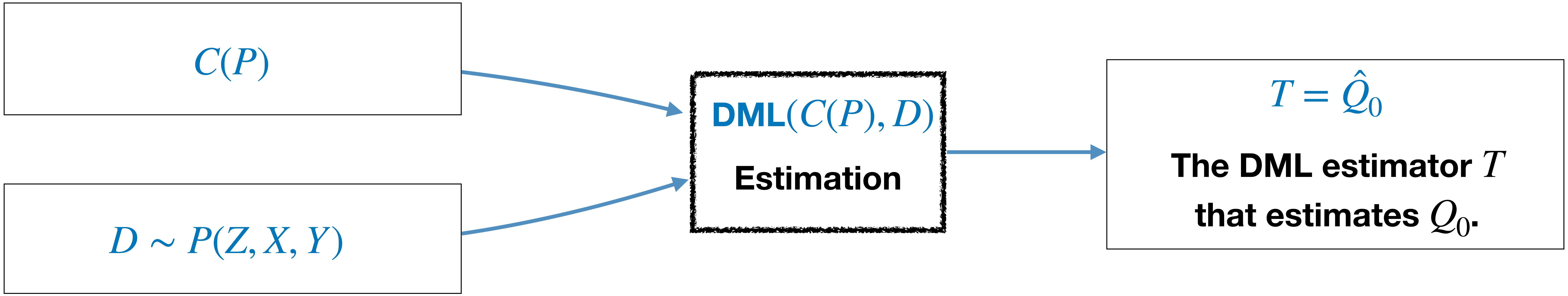
Estimation

$T = \hat{Q}_0$

The estimator T **that estimates** Q_0**.**

Toward Double/Debiased Machine Learning

Double/Debiased Machine Learning (DML) [Chernozhukov et al., 2018] is a framework of constructing the estimator T.



Goal of the talk

We will understand the mechanism of the DML estimator by constructing the estimator for

$$C(P) = \sum_z \mathbb{E}[Y \mid x, z] P(z)$$

We assume the followings in the lecture.

$X \in \{0,1\}$ a binary treatment variable.

$P(v) > 0$ for any v.

Y is 1-dimensional variable (continuous/discrete); Z can be multivariate (continuous/discrete)

Preliminary — Law of Expectation

Let $\mu(X, Z), A(X, Z)$ denote an arbitrary function of $\{X, Z\}$. Then,

Let $\mu_0(X, Z) := \mathbb{E}[Y | X, Z]$.

$$\begin{aligned}
\mathbb{E}[A(X,Z)\big\{Y - \mu(X,Z)\big\}] &= \sum_{x,y,z} A(x,z)\{y - \mu(x,z)\}P(y|x,z)P(x|z)P(z) \\
&= \sum_{x,z} A(x,z) \underbrace{\sum_{y} \{yP(y|x,z)\}}_{=\mu_0(X,Z)} - \mu(x,z)\}P(x,z) \\
&= \mathbb{E}[A(X,Z)\{\mu(X,Z) - \mu_0(X,Z)\}]
\end{aligned}$$

Preliminary — Law of Expectation

Let $\mu(X,Z), A(X,Z)$ denote an arbitrary function of $\{X,Z\}$. Let $\mu_0(X,Z) := \mathbb{E}[Y|X,Z]$.

$$\mathbb{E}[A(X,Z)\{Y-\mu(X,Z)\}] = \mathbb{E}[A(X,Z)\{\mu_0(X,Z)-\mu(X,Z)\}]$$

$$\begin{aligned}
\mathbb{E}[A(X,Z)\left\{Y-\mu(X,Z)\right\}] &= \sum_{x,y,z} A(x,z)\{y-\mu(x,z)\}P(y|x,z)P(x|z)P(z) \\
&= \sum_{x,z} A(x,z) \underbrace{\sum_{y} \{yP(y|x,z)\}}_{=\mu_0(X,Z)} - \mu(x,z)\}P(x,z) \\
&= \mathbb{E}[A(X,Z)\{\mu(X,Z)-\mu_0(X,Z)\}]
\end{aligned}$$

Preliminary — Law of Expectation

Let $A(X, Z)$ denote an arbitrary function of $\{X, Z\}$. Let $\pi_0(X \mid Z) := P(X \mid Z)$.

$$\mathbb{E}[A(X,Z)I_x(X)] = \mathbb{E}[A(x,Z)\pi_0(x \mid Z)]$$

$$\begin{aligned}
\mathbb{E}[A(X,Z)I_x(X)] &= \sum_{x',z} A(x,z)I_x(x')\underbrace{P(x' \mid z)}_{=\pi_0(x' \mid z)}P(z) \\
&= \sum_{z} A(x,z)\pi_0(x \mid z)P(z) \\
&= \mathbb{E}[A(x,Z)\pi_0(x \mid Z)]
\end{aligned}$$

Challenges in estimating $C(P)$

$$C(P) = \sum_z \mathbb{E}[Y | x, z] P(z)$$

Estimating $C(P)$ directly is challenging when Z is high-dimensional and a mixture of continuous/discrete variables …

… because estimating the density $P(z)$ is challenging, and

… computing the marginalization $\sum_z$ is hard.

Estimand — Alternative Representation for $C(P)$

$$C(P) = \sum_z \mathbb{E}[Y \mid x, z] P(z)$$

Instead of directly estimating $C(P)$, we find an *alternative representation* (***Estimand***) of $C(P)$, denoted $f(V, \eta)$, where…

$\mathbb{E}[f(V; \eta_0)] = C(P) = Q_0$ when $\eta = \eta_0$ for some η_0.

$V := \{Z, X, Y\}$ all variables; and $\eta := \eta(P)$ is some function of P called "*nuisance*".

Estimating the expectation will be easier than estimating $\sum_z$

Outcome-Regression-based Estimand (REG)

$$C(P) = \sum_z \mathbb{E}[Y|x,z]p(z)$$

$$= \mathbb{E}_Z\left[\mathbb{E}[Y|x,Z]\right]$$

Expectation over Z

$$C(P) = \sum_z \mathbb{E}[Y|x,z]P(z)$$

$$\mathbb{E}[f(V;\eta_0)] = C(P)$$

Let $\mu(X,Z)$ will be any arbitrary function of $\{X,Z\}$, and $\mu_0(X,Z) := \mathbb{E}[Y|X,Z]$.

$$f^{REG}(V;\eta := \mu) = \mu(x,Z)$$

Inverse Probability Weighting-based Estimand - 1

$$C(P) = \sum_{y,z} yP(y \mid x, z)P(z)$$

$$= \sum_{y,z} yP(y \mid x, z)\frac{P(x \mid z)}{P(x \mid z)}P(z)$$

Indicator s.t. 1 when x'=x

$$= \sum_{y,x',z} yI_x(x')P(y \mid x', z)\frac{P(x' \mid z)}{P(x' \mid z)}P(z)$$

$$= \sum_{y,x',z} \frac{I_x(x')}{P(x' \mid z)} yP(z, x', y) = \mathbb{E}\left[\frac{I_x(X)}{P(X \mid Z)}Y\right]$$

P(z,x,y) = P(y|x,z)P(x|z)P(z)

$$C(P) = \sum_{z} \mathbb{E}[Y \mid x, z]P(z)$$

$$\mathbb{E}[f(V;\eta_0)] = C(P)$$

Inverse Probability Weighting-based Estimand - 2

$$C(P) = \mathbb{E}\left[\frac{I_x(X)}{P(X \mid Z)} Y\right]$$

$$C(P) = \sum_z \mathbb{E}[Y \mid x, z] P(z)$$

$$\mathbb{E}[f(V; \eta_0)] = C(P)$$

Let $\pi(X \mid Z)$ is an arbitrary positive function and $\pi_0(X \mid Z) := P(X \mid Z)$.

Let $f^{IPW}(V; \eta := \pi) := \frac{I_x(X)}{\pi(X \mid Z)} Y$

Doubly Robust Estimand - 1

$$C(P) = \mathbb{E}\left[f^{IPW}(V;\pi_0)\right]$$ = C(P)

$$+\mathbb{E}[f^{REG}(V;\mu_0)]$$ = C(P)

$$-\mathbb{E}\left[\frac{I_x(X)}{\pi_0(X|Z)}\mu_0(X,Z)\right]$$

= C(P) is shown next

$$C(P) = \sum_z \mathbb{E}[Y|x,z]P(z)$$

$$\mathbb{E}[f(V;\eta_0)] = C(P)$$

$$f^{REG}(V;\eta_0 := \mu) := \mu(x,Z)$$

$$f^{IPW}(V;\eta := \pi) := \frac{I_x(X)}{\pi(X|Z)}Y$$

Doubly Robust Estimand - 2

$$\mathbb{E}\left[\frac{I_x(X)}{\pi_0(X|Z)}\mu_0(X,Z)\right] = \sum\nolimits_{x',z} \frac{I_x(x')}{\pi_0(x'|z)} \underbrace{\mu_0(x',z)}_{=\mathbb{E}[Y|x,z]} \underbrace{P(x'|z)}_{=\pi_0(x'|z)} P(z)$$

$$= \sum_z \mathbb{E}[Y|x,z]P(z) = C(P)$$

Doubly Robust Estimand - 3

$$C(P) = \mathbb{E}\left[f^{IPW}(V;\pi_0)\right]$$ = C(P)

$$+\mathbb{E}[f^{REG}(V;\mu_0)]$$ = C(P)

$$-\mathbb{E}\left[\frac{I_x(X)}{\pi_0(X|Z)}\mu_0(X,Z)\right]$$ = C(P)

$$\mathbb{E}[f(V;\eta_0)] = C(P)$$

$$f^{REG}(V;\eta_0 := \mu) := \mu(x,Z)$$

$$f^{IPW}(V;\eta := \pi) := \frac{I_x(X)}{\pi(X|Z)}Y$$

Doubly Robust Estimand - 3

$$C(P) = \mathbb{E}\left[f^{IPW}(V;\pi_0)\right] + \mathbb{E}[f^{REG}(V;\mu_0)] - \mathbb{E}\left[\frac{I_x(X)}{\pi_0(X|Z)}\mu_0(X,Z)\right]$$

$$= \mathbb{E}\left[f^{IPW}(V;\pi_0) + f^{REG}(V;\mu_0) - \frac{I_x(X)}{\pi_0(X|Z)}\mu_0(X,Z)\right]$$

$$= \mathbb{E}\left[\frac{I_x(X)}{\pi_0(X|Z)}\left\{Y - \mu_0(X,Z)\right\} + \mu_0(x,Z)\right]$$

$$f^{DR}(V;\eta = \{\pi,\mu\}) := \frac{I_x(X)}{\pi(X|Z)}\{Y - \mu(X,Z)\} + \mu(x,Z)$$

Doubly robustness of DR-Estimand - 1

If $\pi = \pi_0, \ldots$ (correctly estimated), for any μ,

$$A = \mathbb{E}[f^{IPW}(V; \pi_0)] = C(P)$$

$$\mathbb{E}[f^{DR}(V; \eta = \{\mu, \pi\})]$$
$$= \mathbb{E}\left[f^{IPW}(V; \pi)\right] \text{ —— A}$$
$$+\mathbb{E}[f^{REG}(V; \mu)] \text{ —— B}$$
$$-\mathbb{E}\left[\frac{I_x(X)}{\pi(X|Z)}\mu(X, Z)\right] \text{ —— C}$$

Doubly robustness of DR-Estimand - 2

If $\pi = \pi_0, \ldots$ (correctly estimated), for any μ,

$$C = \mathbb{E}\left[\frac{I_x(X)}{\pi_0(X|Z)}\mu(X,Z)\right]$$

$$= \sum_{z,x'} \frac{I_x(x')}{\pi_0(x'|z)}\mu(x',z)\underbrace{P(x'|z)}_{=\pi_0(x'|z)} P(z)$$

$$= \sum_{z} \mu(x,z)P(z) = \mathbb{E}[\mu(x,Z)] = \mathbb{E}[f^{REG}(V;\mu)] = B$$

$$\mathbb{E}[f^{DR}(V;\eta = \{\mu,\pi\})]$$
$$= \mathbb{E}\left[f^{IPW}(V;\pi)\right] \text{ — A}$$
$$+\mathbb{E}[f^{REG}(V;\mu)] \text{ — B}$$
$$-\mathbb{E}\left[\frac{I_x(X)}{\pi(X|Z)}\mu(X,Z)\right] \text{ — C}$$

$$\mathbb{E}[f^{DR}(V;\eta = \{\mu,\pi_0\})] := C(P) + B - B = C(P)$$

Doubly robustness of DR-Estimand - 3

If $\mu = \mu_0$,… (correctly estimated), for any positive π,

$$B = \mathbb{E}[f^{REG}(V;\mu_0)] = C(P)$$

$$\mathbb{E}[f^{DR}(V;\eta = \{\mu,\pi\})]$$

$$= \mathbb{E}\left[f^{IPW}(V;\pi)\right] \text{ ——— A}$$

$$+\mathbb{E}[f^{REG}(V;\mu)] \text{ ——— B}$$

$$-\mathbb{E}\left[\frac{I_x(X)}{\pi(X|Z)}\mu(X,Z)\right] \text{ ——— C}$$

Doubly robustness of DR-Estimand - 4

If $\mu = \mu_0, \ldots$ (correctly estimated), for any positive π,

$$C = \mathbb{E}\left[\frac{I_x(X)}{\pi(X|Z)}\mu_0(X,Z)\right]$$

$$= \sum_{z,x'} \frac{I_x(x')}{\pi(x'|z)} \underbrace{\mu_0(x',z)}_{=\sum_y yP(y|x',z)} P(x'|z)P(z)$$

$$= \sum_{z,x',y} \frac{I_x(x')}{\pi(x'|z)} yP(z,x',y) = \mathbb{E}\left[\frac{I_x(X)}{\pi(X|Z)}Y\right] = A$$

$$\mathbb{E}[f^{DR}(V;\eta = \{\mu,\pi\})]$$
$$= \mathbb{E}\left[f^{IPW}(V;\pi)\right] \text{ — A}$$
$$+\mathbb{E}[f^{REG}(V;\mu)] \text{ — B}$$
$$-\mathbb{E}\left[\frac{I_x(X)}{\pi(X|Z)}\mu(X,Z)\right] \text{ — C}$$

$$\mathbb{E}[f^{DR}(V;\eta = \{\mu,\pi_0\})] := C(P) + A - A = C(P)$$

Doubly robustness of DR-Estimand - 6

$$\mathbb{E}[f^{DR}(V; \eta = \{\mu, \pi\})] = C(P)$$

If $\mu = \mu_0$ either $\pi = \pi_0$

“**Doubly robustness!**”: Double chances for being correct!

Intermediate Summary - Estimands

$$f^{REG}(V; \eta := \mu) := \mu(x, Z)$$

$$f^{IPW}(V; \eta := \pi) := \frac{I_x(X)}{\pi(X|Z)} Y$$

$$f^{DR}(V; \eta = \{\pi, \mu\}) := \frac{I_x(X)}{\pi(X|Z)} \{Y - \mu(X, Z)\} + \mu(x, Z)$$

Given $\mathbb{E}[f(V; \eta_0)] = C(P)$,

1. Estimate η_0 (as $\hat{\eta}$) from data D
2. Take the empirical average

$$\mathbb{E}_D[f(V; \hat{\eta})] := \frac{1}{N} \sum_{i=1}^{N} f(V_i; \hat{\eta})$$

Which estimand should be chosen?

$f(V; \eta)$ s.t. $\mathbb{E}[f(V; \hat{\eta})]$ converge fast despite slow convergence of $\hat{\eta}$

Orthogonal Estimand - Rough Idea

Debiasedness: Even if $\hat{\eta}$ converges to η_0 slow, $\mathbb{E}[f(V;\hat{\eta})]$ converges to $\mathbb{E}[f(V;\eta_0)]$ fast.

If $\mathbb{E}[f(V;\eta)]$ is invariant to the small perturbation of η,

… even if the error of η is somewhat large,

… $\mathbb{E}[f(V;\eta)]$ will not be suffered by the error of η.

We will formalize this idea by considering the directional derivative of $\mathbb{E}[f(V;\eta)]$.

Orthogonal Estimand

Directional Derivative: For a function $g(\eta)$, its derivative at the direction h is given as

$$D_\eta g(\eta)\{h\} := \frac{\partial}{\partial t} g(\eta + th)\Big|_{t=0}$$

Orthogonal Estimand: $f(V;\eta)$ is an *orthogonal estimand* if

$$D_\eta \mathbb{E}[f(V;\eta_0)]\{\eta - \eta_0\} = 0$$

… which is equivalent to state $\mathbb{E}\left[(\eta - \eta_0) \cdot \frac{\partial}{\partial \eta} f(V;\eta)\Big|_{\eta=\eta_0}\right] = 0$

The estimand is invariant to the small perturbation of η (near η_0)

Orthogonal Estimand - 2

$f(V;\eta)$ is an *orthogonal estimand* if $D_\eta \mathbb{E}[f(V;\eta_0)]\{\eta - \eta_0\} = 0$

Then, by Taylor's expansion (up to the 2nd order),

$$\mathbb{E}[f(V;\eta)] - \mathbb{E}[f(V;\eta_0)] = \cancel{D_\eta \mathbb{E}[f(V;\eta_0)]\{\eta - \eta_0\}} + \frac{1}{2} D^2_\eta \mathbb{E}[f(V;\eta)]\{\eta - \eta_0\}^2$$

0

$O_P\left(\| \eta - \eta_0 \|^2_{L_2(P)}\right) := O_P\left(\mathbb{E}[(\eta - \eta_0)^2]\right)$, shortly, $O_P\left(\| \eta - \eta_0 \|^2\right)$.

For any η, $\mathbb{E}[f(V;\eta)] - C(P) = O_P\left(\| \eta - \eta_0 \|^2_2\right)$

Orthogonal Estimand - Two nuisances

$$\mathbb{E}[f(V;\{\eta^a,\eta^b\})] - C(P) = O_P\left(\|\eta^a-\eta_0^a\|^2\right) + O_P\left(\|\eta^b-\eta_0^b\|^2\right) + O_P\left(\|\eta^a-\eta_0^a\|\|\eta^b-\eta_0^b\|\right)$$

$$\mathbb{E}[f(V;\{\eta^a,\eta^b\})] - \mathbb{E}[f(V;\{\eta_0^a . \eta_0^b\})]$$

$$= D_{\eta^a}\mathbb{E}[f(V;\{\eta_0^a,\eta_0^b\})]\{\eta^a-\eta_0^a\}$$

$$+D_{\eta^b}\mathbb{E}[f(V;\{\eta_0^a,\eta_0^b\})]\{\eta^b-\eta_0^b\}$$

0

$$+\frac{1}{2}D^2_{\eta^a}\mathbb{E}[f(V;\{\eta^a,\eta^b\})]\{\eta-\eta_0\}^2 \quad = O_P\left(\|\eta^a-\eta_0^a\|^2\right)$$

$$+\frac{1}{2}D^2_{\eta^b}\mathbb{E}[f(V;\{\eta^a,\eta^b\})]\{\eta^b-\eta_0^b\}^2 \quad = O_P\left(\|\eta^b-\eta_0^b\|^2\right)$$

$$+D_{\eta_a}D_{\eta_b}\mathbb{E}[f(V;\{\eta^a,\eta^b\}]\{\eta^a-\eta_0^a,\eta^b-\eta_0^b\} \quad = O_P\left(\|\eta^a-\eta_0^a\|\|\eta^b-\eta_0^b\|\right)$$

Orthogonal Estimand - Two nuisances

$$\mathbb{E}[f(V;\{\eta^a,\eta^b\})] - C(P) = O_P\left(\| \eta^a - \eta_0^a \|^2\right) + O_P\left(\| \eta^b - \eta_0^b \|^2\right) + O_P\left(\| \eta^a - \eta_0^a \|\| \eta^b - \eta_0^b \|\right)$$

Whenever η^a and η^b converges to $N^{-1/4}$, $\mathbb{E}[f(V;\{\eta^a,\eta^b\}]$ converges to $C(P)$ at

$(N^{-1/4})^2 + (N^{-1/4})^2 + (N^{-1/4})(N^{-1/4}) = N^{-1/2}$ rate.

Orthogonal Estimand - Debiasedness

$$\mathbb{E}[f(V;\eta)] - C(P) = O_P\left(\| \eta - \eta_0 \|_2^2\right)$$

If η converges to η_0 at some rate, say $N^{-1/4}$

$\mathbb{E}[f(V;\eta)]$ converges to $C(P)$ at $(N^{-1/4})^2 = N^{-1/2}$ rate.

Debiasedness property of orthogonal estimands

$f(V;\eta)$ is an ***orthogonal estimand*** $\Rightarrow \mathbb{E}[f(V;\eta)]$ converges much faster than η

Is the REG estimand orthogonal?

$$D_\mu \mathbb{E}[f^{REG}(V;\mu_0)]\{\mu - \mu_0\} := \frac{\partial}{\partial t}\mathbb{E}\left[f^{REG}(V;\mu + t(\mu - \mu_0)\right] |_{t=0}$$

$$= \mathbb{E}\left[\frac{\partial}{\partial t} f^{REG}(V;\mu + t(\mu - \mu_0) |_{t=0}\right]$$

$$= \mathbb{E}\left[\frac{\partial}{\partial t}\{\mu + t(\mu - \mu_0)\} |_{t=0}\right]$$

$$= \mathbb{E}[\mu(x,Z) - \mu_0(x,Z)]$$

$$\neq 0$$

$f^{REG}(V;\mu)$ estimand is non-orthogonal.

Is the IPW estimand orthogonal?

$$D_\mu \mathbb{E}[f^{IPW}(V;\pi_0)]\{\pi - \pi_0\} := \frac{\partial}{\partial t}\mathbb{E}\left[f^{IPW}(V;\pi + t(\pi - \pi_0)\right]|_{t=0}$$

$$= \mathbb{E}\left[\frac{\partial}{\partial t} f^{IPW}(V;\pi + t(\pi - \pi_0)\,|_{t=0}\right]$$

$$= \mathbb{E}\left[(\pi - \pi_0)\frac{\partial}{\partial \pi} f^{IPW}(V;\pi)\,|_{\pi=\pi_0}\right]$$

$$= -\mathbb{E}\left[\{\pi - \pi_0\}\left\{\frac{I_x(X)}{\pi_0^2(X\,|\,Z)}Y\right\}\right]$$

$\neq 0$

$f^{IPW}(V;\pi)$ estimand is non-orthogonal.

Is the DR estimand orthogonal?

$$
\begin{aligned}
D_\mu \mathbb{E}[f^{DR}(V; \{\mu_0, \pi\})]\{\mu - \mu_0\} := \frac{\partial}{\partial t}\mathbb{E}\left[f^{DR}(V; \{\mu + t(\mu - \mu_0), \pi_0\}\right] |_{t=0} \\
= \mathbb{E}\left[\frac{\partial}{\partial t} f^{DR}(V; \{\mu + t(\mu - \mu_0, \pi_0\}) |_{t=0}\right] \\
= \mathbb{E}\left[(\mu - \mu_0)\frac{\partial}{\partial \mu} f^{DR}(V; \{\mu, \pi\}) |_{\mu=\mu_0}\right] \\
= \mathbb{E}\left[\{\mu - \mu_0\}\left\{-\frac{I_x(X)}{\pi_0(X|Z)}\mu_0(X, Z) + \mu_0(x, Z)\right\}\right] \\
= \mathbb{E}\left[\{\mu - \mu_0\}\{-\mu_0(x, Z) + \mu_0(x, Z)\}\right]
\end{aligned}
$$

0

Is the DR estimand orthogonal? - 2

$$D_\pi \mathbb{E}[f^{DR}(V;\{\mu_0,\pi\})]\{\pi-\pi_0\} := \frac{\partial}{\partial t}\mathbb{E}\left[f^{DR}(V;\{\mu_0,\pi+t(\pi-\pi_0)\}\right]|_{t=0}$$

$$= \mathbb{E}\left[\frac{\partial}{\partial t} f^{DR}(V;\{\mu_0,\pi+t(\pi-\pi_0)\})|_{t=0}\right]$$

$$= \mathbb{E}\left[(\pi-\pi_0)\frac{\partial}{\partial \pi} f^{DR}(V;\{\mu_0,\pi\})|_{\pi=\pi_0}\right]$$

$f^{DR}(V;\{\mu,\pi\})$ is an ***orthogonal estimand!***

Intermediate Summary - Orthogonal Estimands

Debiasedness: If $f(V;\eta)$ is orthogonal, $\mathbb{E}[f(V;\eta)] - C(P) = O_P\left(\| \eta - \eta_0 \|_2^2\right)$

$f^{DR}(V;\eta = \{\pi, \mu\}) := \dfrac{I_x(X)}{\pi(X|Z)}\{Y - \mu(X,Z)\} + \mu(x,Z)$ is an orthogonal estimand.

Therefore

$\mathbb{E}[f^{DR}(V;\{\pi,\mu\})] \to C(P)$ at $N^{-1/2}$ rate if π, μ converges to π_0, μ_0 at $N^{-1/4}$ rate.

Intermediate Summary - Orthogonal Estimands

$$\mathbb{E}[f^{DR}(V; \{\pi, \mu\})] - C(P) = O_P\left(\| \pi - \pi_0 \| \| \mu - \mu_0 \|\right)$$

$= N^{-1/2}$ if π, μ converges at $N^{-1/4}$
("*debiasedness*")

$= 0$ if either $\pi = \pi_0$ or $\mu = \mu_0$
("doubly-robustness")

Intermediate Summary - Orthogonal Estimands

$$\mathbb{E}[f^{DR}(V;\{\pi,\mu\})] - C(P) = O_P\left(\| \pi - \pi_0 \| \| \mu - \mu_0 \|\right)$$

$$\mathbb{E}[f^{DR}(V;\{\pi,\mu\})] - C(P) = \mathbb{E}[f^{DR}(V;\{\pi,\mu\}) - f^{DR}(V;\{\pi_0,\mu_0\})]$$

$$= \mathbb{E}\left[\frac{I_x(X)}{\pi(X|Z)}\{Y - \mu(X,Z)\} + \mu(x,Z) - \frac{I_x(X)}{\pi_0(X|Z)}\{Y - \mu_0(X,Z)\} - \mu_0(x,Z)\right]$$

$$= \mathbb{E}\left[\frac{I_x(X)}{\pi(X|Z)}\{Y - \mu(X,Z)\} + \mu(x,Z) - \mu_0(x,Z)\right]$$

$$= \mathbb{E}\left[\frac{I_x(X)}{\pi(X|Z)}\{\mu_0(X,Z) - \mu(X,Z)\} + \mu(x,Z) - \mu_0(x,Z)\right]$$

Intermediate Summary - Orthogonal Estimands

$$\mathbb{E}[f^{DR}(V; \{\pi, \mu\})] - C(P) = \mathbb{E}\left[\frac{I_x(X)}{\pi(X|Z)}\{\mu_0(X,Z) - \mu(X,Z)\} + \mu(x,Z) - \mu_0(x,Z)\right]$$

$$= \mathbb{E}\left[\frac{\pi_0(x|Z)}{\pi(x|Z)}\{\mu_0(x,Z) - \mu(x,Z)\} + \mu(x,Z) - \mu_0(x,Z)\right]$$

$$= \mathbb{E}\left[\left\{\frac{\pi_0(x|Z)}{\pi(x|Z)} - 1\right\}\{\mu_0(x,Z) - \mu(x,Z)\}\right]$$

$$= \mathbb{E}\left[\left\{\frac{\pi_0(x|Z)}{\pi(x|Z)} - 1\right\}\{\mu_0(x,Z) - \mu(x,Z)\}\right]$$

Intermediate Summary - Orthogonal Estimands

$$\mathbb{E}[f^{DR}(V;\{\pi,\mu\})] - C(P) = \mathbb{E}\left[\left\{\frac{\pi_0(x|Z)}{\pi(x|Z)} - 1\right\}\{\mu_0(x,Z) - \mu(x,Z)\}\right]$$

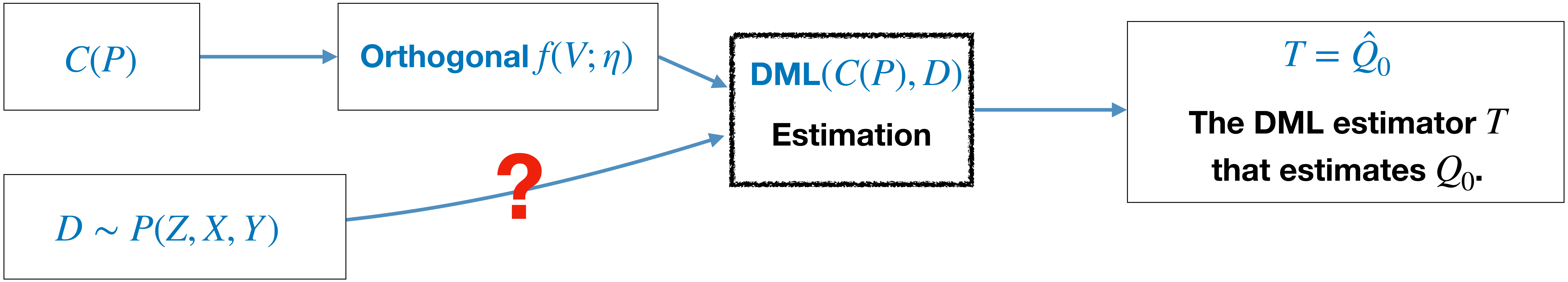
$$= \mathbb{E}\left[\frac{1}{\pi(x|Z)}\left\{\pi_0(x|Z) - \pi(x|Z)\right\}\left\{\mu_0(x,Z) - \mu(x,Z)\right\}\right]$$

$$= O_P\left(\| \pi_0 - \pi \| \| \mu - \mu_0 \|\right)$$

Estimating with finite samples

So far, we study the power of the orthogonal estimand.

Now, we connect the estimand to the estimation task using finite samples.



Estimating with Finite Samples

If $f(V;\eta)$ is orthogonal, $\mathbb{E}[f(V;\eta)] - C(P) = O\left(\| \eta - \eta_0 \|_2^2\right)$

Recall the notation $\mathbb{E}_D\left[f(V;\hat{\eta})\right] := \frac{1}{N}\sum_{i=1}^{N} f(V_i;\hat{\eta})$, where $\hat{\eta}$ denotes estimated nuisance.

Given samples D, $\mathbb{E}_D[f(V;\hat{\eta})]$ is our estimator for $C(P)$. Then, we are interested in the error

$$\mathbb{E}_D[f(V;\hat{\eta})] - C(P) \quad = \mathbb{E}_D[f(V;\hat{\eta})] - \mathbb{E}_P[f(V;\hat{\eta})] + \mathbb{E}_P[f(V;\hat{\eta})] - C(P)$$

We will focus on analyzing the remaining term: $\mathbb{E}_D[f(V;\hat{\eta})] - \mathbb{E}_P[f(V;\hat{\eta})]$.

Law of Large Numbers (LLN)

Remaining term: $\mathbb{E}_D[f(V;\hat{\eta})] - \mathbb{E}_P[f(V;\hat{\eta})] = \frac{1}{N}\sum_{i=1}^{N} f(V_i;\hat{\eta}) - \mathbb{E}_P[f(V;\hat{\eta})]$.

Law of Large Numbers (LLN)

For any fixed η_*, $\mathbb{E}_D[f(V;\eta_*)]$ converges to $\mathbb{E}_P[f(V;\eta_*)]$.

Concentration inequalities (e.g., Hoeffding's inequality)

For any fixed η_*, if $f(V;\eta_*)$ is bounded, $\mathbb{E}_D[f(V;\eta_*)]$ converges to $\mathbb{E}_P[f(V;\eta_*)]$ at $N^{-1/2}$ rate.

Challenges in LLN

For any fixed η_*, (if $f(V;\eta_*)$ is bounded), $\mathbb{E}_D[f(V;\eta_*)] - \mathbb{E}_P[f(V;\eta_*)] \to 0$ at $N^{-1/2}$ rate.

Consider $\mathbb{E}_D[f(V;\hat{\eta})] - \mathbb{E}_P[f(V;\hat{\eta})]$, where $\hat{\eta} := \hat{\eta}(D)$ is an estimate using samples D.

… Equivalently, consider $\frac{1}{N}\sum_{i=1}^{N} f(V_i, \hat{\eta}(N)) - \mathbb{E}[f(V;\hat{\eta}(N))]$

The LLN is not applicable since $\hat{\eta}$ is not fixed w.r.t. D (and N).

… Without any special treatises, $\mathbb{E}_D[f(V;\hat{\eta})] - \mathbb{E}_P[f(V;\hat{\eta})]$ is *not necessarily converging to 0.*

Uniform Convergence

To guarantee $\mathbb{E}_D[f(V;\hat{\eta})] - \mathbb{E}_P[f(V;\hat{\eta})] \to 0$, we should have

Uniform convergence: $\sup_{\eta \in H} \left(\mathbb{E}_D[f(V;\eta)] - \mathbb{E}_P[f(V;\eta)]\right) \to 0$

Then, $\mathbb{E}_D[f(V;\hat{\eta})] - \mathbb{E}_P[f(V;\hat{\eta})] \to 0$ obviously holds.

Donsker Class

For some function class $H(\ni \eta)$, the uniform convergence holds.

Donsker class: A class H s.t. $\sup_{\eta \in H} \left(\mathbb{E}_D[f(V;\eta)] - \mathbb{E}_P[f(V;\eta)] \right) \to 0$ at $N^{-1/2}$ rate

Example: A function class with bounded VC-dimension (called VC-class).

: differentiable functions

Error analysis under Donsker

If a nuisance function class H is "Donsker", $\mathbb{E}_D[f(V;\hat{\eta})] - \mathbb{E}_P[f(V;\hat{\eta})] \to 0$ at $N^{-1/2}$ rate.

$$\mathbb{E}_D[f(V;\hat{\eta})] - C(P) \quad = \mathbb{E}_D[f(V;\hat{\eta})] - \mathbb{E}_P[f(V;\hat{\eta})] + \mathbb{E}_P[f(V;\hat{\eta})] - C(P)$$

$\to 0$ at $N^{-1/2}$ rate

$= O\left(\| \hat{\eta} - \eta_0 \|_2^2\right)$

Then, the estimator $\mathbb{E}_D[f(V;\hat{\eta})]$ *converges to* $C(P)$ *fast even if* $\hat{\eta}$ *converges slow*

Limitation of Donsker

Donsker class: A class H s.t. $\sup_{\eta \in H} \left(\mathbb{E}_D[f(V;\eta)] - \mathbb{E}_P[f(V;\eta)] \right) \to 0$ at $N^{-1/2}$ rate

Example: A function class with bounded VC-dimension (called VC-class).

: differentiable functions

However, confining on the Donsker class is restrictive in the modern ML era.

***There is no guarantee that deep and complicated neural networks fall into the Donsker*.**

Releasing Donsker Assumption

Recall the Law of Large Numbers:

For any fixed η_*, $\mathbb{E}_D[f(V;\eta_*)] - \mathbb{E}_P[f(V;\eta_*)] \to 0$ at $N^{-1/2}$ rate.

This can be rewritten as a following principle ([Robins et al., 1997, Kennedy et al., 2019], etc.)

For any η s.t. independent to samples D, $\mathbb{E}_D[f(V;\eta)] - \mathbb{E}_P[f(V;\eta)] \to 0$ at $N^{-1/2}$ rate.

This doesn't require the Donsker class assumption!

Suppose $\hat{\eta}$ is estimated from a separate dataset D' that is independent to D. Then,

$$\mathbb{E}_D[f(V;\hat{\eta})] - \mathbb{E}_P[f(V;\hat{\eta})] \to 0 \text{ at } N^{-1/2} \text{ rate.}$$

Sample Splitting

Sample splitting procedure

Split D into two random halves D_0, D_1.

For $k \in \{0,1\}$,

Let $\hat{\eta}_k$ denote the estimated nuisance using D_k.

Let $T_k := \mathbb{E}_{D_{1-k}}\left[f(V;\hat{\eta}_k)\right]$

Let $T := (T_0 + T_1)/2$.

Then, $T - \mathbb{E}\left[f(V;\hat{\eta})\right] \to 0$ at $N^{-1/2}$ rate

Donsker class assumption is dropped.

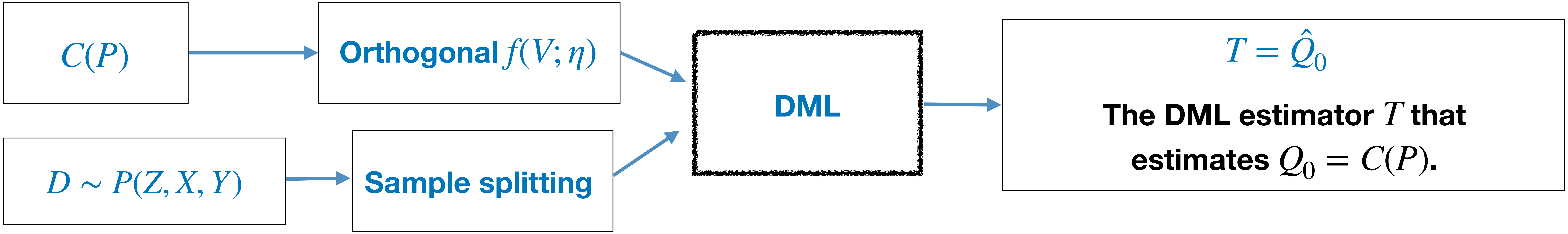
Any ML models can be employed!

DML Definition (Intermediate version)

Double/Debiased Machine Learning (DML)

Given a target quantity $C(P)$ and the data D, a DML estimator T is an estimator derived from

1. an orthogonal estimand $f(V;\eta)$, and
2. the sample-splitting procedure.



Toward Score-based Definition

The original DML definition by [Chernozhukov et al., 2018] is stated somewhat different, but share the crux of the idea.

e.g., $\mathbb{E}[Y \mid do(x)]$

Score: For a nuisance η and a target Q (where η_0, Q_0 denote true {nuisance, target}),

$g(V; \eta, Q)$ is a *score function* if $\mathbb{E}[g(V; \eta_0, Q_0)] = 0$

Example: $g(V; \eta, Q) = f(V; \eta) - Q$, where $f(V; \eta)$ is an *estimand* s.t. $\mathbb{E}[f(V; \eta_0)] = C(P) = Q_0$

$\Rightarrow \mathbb{E}[g(V; \eta_0, Q_0)] = \mathbb{E}[f(V; \eta_0)] - Q_0 = \mathbb{E}[f(V; \eta_0)] - C(P) = 0$ by def. of the estimand.

$\Rightarrow$ Therefore, $f(V; \eta) - Q$ is a valid score function.

Score-based estimation

Score-based estimation: Given data D and $\hat{\eta}$, find $\hat{Q}$ satisfying

$$\mathbb{E}_D[g(V; \hat{\eta}, \hat{Q})] = 0$$

Sample-splitting is applicable: We can use dataset D_0 for training $\hat{\eta}$ and find $\hat{Q}$ using D_1 by

$$\mathbb{E}_{D_1}[g(V; \hat{\eta}, \hat{Q})] = 0$$

Consider $f^{DR}(V; \eta = \{\mu, \pi\})$, and let $g(V; \eta, Q) := f^{DR}(V; \eta) - Q$.

Then, $\mathbb{E}_D[g(V; \hat{\eta}, \hat{Q})] = \mathbb{E}_D[f^{DR}(V; \hat{\eta})] - \hat{Q}$.

Therefore, the score-based estimation gives $\hat{Q} = \mathbb{E}_D[f^{DR}(V; \hat{\eta})]$

Orthogonal score

The original DML definition is stated somewhat different, but share the crux of the idea.

Orthogonal score: A score $g(V;\eta,Q)$ s.t. $\quad D_\eta g(V;\eta,Q_0)\{\eta-\eta_0\}=0$

Example: $g(V;\eta,Q)=f(V;\eta)-Q$ where $f(V;\eta)$ is an orthogonal estimand. Then,

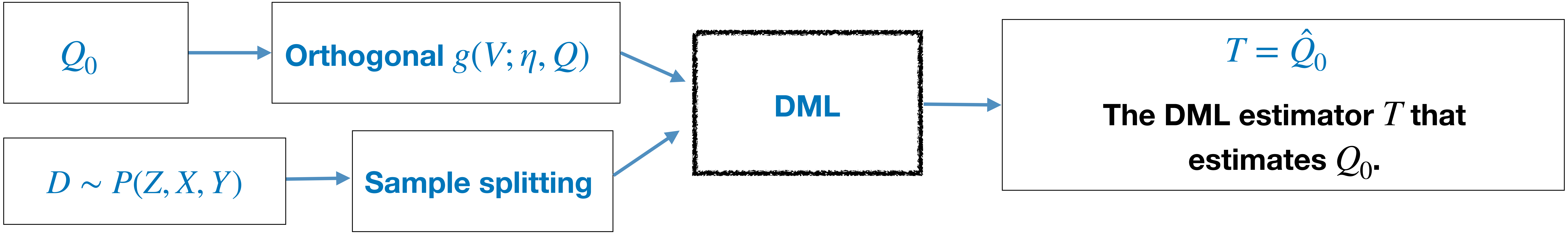
$$\Rightarrow D_\eta g(V;\eta,Q_0)\{\eta-\eta_0\}=D_\eta f(V;\eta)\{\eta-\eta_0\}=0$$

DML Definition

Double/Debiased Machine Learning (DML)

For a target quantity Q and the data D, a DML estimator T is an estimator derived from

1. an orthogonal score $g(V; \eta, Q)$, and
2. the sample-splitting procedure.



Debiasedness property

Double/Debiased Machine Learning (DML)

For a target quantity Q, a DML estimator T is an estimator derived from

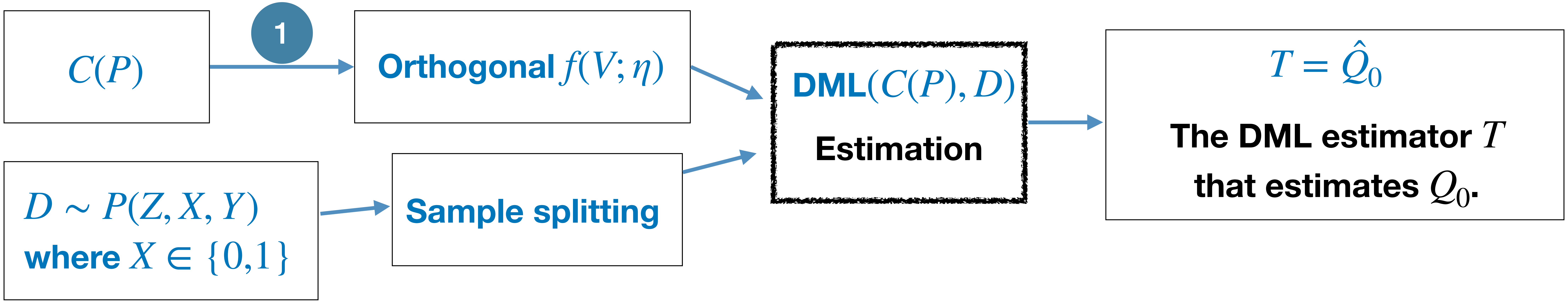
1. an orthogonal score $g(V; \eta, Q)$, and
2. the sample-splitting procedure.

$$T - Q_0 = O(N^{-1/2}) + O\left(\| \hat{\eta} - \eta_0 \|^2\right)$$

A DML estimator T converges to $C(P)$ at a $N^{-1/2}$ even if $\hat{\eta}$ converges $N^{-1/4}$ rate…

… without any function class assumption! (Any ML models can be used for $\hat{\eta}$)

Uncovered subjects - 1



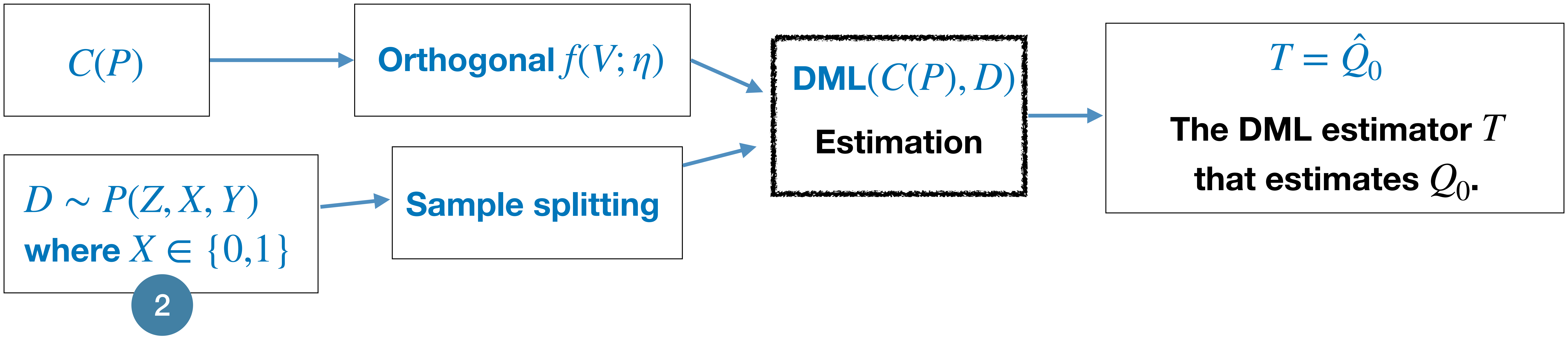
1 How to derive the orthogonal estimand $f(V;\eta)$ from an identified causal functional $C(P)$?

… The orthogonal estimand for the back-door adjustment and the truncated factorization (a.k.a. sequential back-door (SBD) or g-functional) are known.

… [Jung et al., 2021] showed that any ID functional can be represented as a function of SBDs.

… [Jung et al., 2021] proposed an algorithm for deriving the ortho. functional.

Uncovered subjects - 2



2 If X is continuous or $Q_0 := p(y \mid do(x))$, then what happens?

… The orthogonal functional may not exist, because the indicator $I_x(X)$ or $I_y(Y)$ are not well-defined for X.

… Special treatises to smooth out $I_x(X)$ (e.g., use a smoothing kernel density instead of $I_x(X)$) should be applied.

… [Jung et al., 2021] propose an estimator for $p(y \mid do(x))$ for the instruments setting.

Any Questions ❓