LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

LMU MUNICH SCHOOL OF MANAGEMENT

INSTITUTE OF ARTIFICIAL INTELLIGENCE (AI) IN MANAGEMENT

mcml Munich Center for Machine Learning

AISTATS

Orthogonal Representation Learning for Estimating Causal Quantities

Valentyn Melnychuk, Dennis Frauen, Jonas Schweisthal, Stefan Feuerriegel

LMU Munich & Munich Center for Machine Learning (MCML), Munich, Germany

AISTATS 2026, Oral

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

LMU MUNICH SCHOOL OF MANAGEMENT

INSTITUTE OF ARTIFICIAL INTELLIGENCE (AI) IN MANAGEMENT

Munich Center for Machine Learning

AISTATS

Agenda

Introduction

CAPOs/CATE estimation

OR-learners

Research question 1

Research question 2

Takeaways

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN | LMU MUNICH SCHOOL OF MANAGEMENT | INSTITUTE OF ARTIFICIAL INTELLIGENCE (AI) IN MANAGEMENT

Munich Center for Machine Learning

AISTATS

Agenda

Introduction

CAPOs/CATE estimation

OR-learners

Research question 1

Research question 2

Takeaways

Introduction: Estimation of individualized causal quantities

Why is this important?

- Estimating conditional average potential outcomes (**CAPOs**) and conditional average treatment effect (**CATE**) from observational data is **one of the core challenges** in causal ML

- Existing end-to-end **representation learning methods**
  - work well in practice (based on numerous semi-synthetic benchmarks)
  - but lack asymptotic optimality

- Two-stage **Neyman-orthogonal learners**
  - offer such asymptotic optimality (e.g., quasi-oracle efficiency/double robustness)
  - but do not explicitly benefit from representation learning

Introduction: Estimation of individualized causal quantities

- Estimating conditional average potential outcomes (**CAPOs**) and conditional average treatment effect (**CATE**) from observational data is **one of the core challenges** in causal ML

Why is this important?

- Existing end-to-end **representation learning methods**
  - work well in practice (based on numerous semi-synthetic benchmarks)
  - but lack asymptotic optimality

- Two-stage **Neyman-orthogonal learners**
  - offer such asymptotic optimality (e.g., quasi-oracle efficiency/double robustness)
  - but do not explicitly benefit from representation learning

Central tension in our paper

Introduction: Estimation of individualized causal quantities

Problem formulation: CAPOs & CATE estimation

Given i.i.d. observational dataset $\mathcal{D} = \{X_i, A_i, Y_i\}_{i=1}^n \sim \mathbb{P}(X, A, Y)$ with

- X covariates
- A binary treatments
- Y continuous (factual) outcomes

we aim to estimate covariate-level causal quantities:

- conditional average potential outcomes (CAPOs):
  $\xi_a^x(x) = \mathbb{E}(Y[a] \mid X = x)$
- conditional average treatment effect (CATE):
  $\tau^x(x) = \mathbb{E}(Y[1] - Y[0] \mid X = x)$

However, we **never** observe both potential (counterfactual) outcomes!

Patient	Covariates X	Treatment A	Outcome $Y = Y(0)$	Outcome $Y = Y(1)$
👩🏼		0	-1.0	
👨🏽		1		2.3
👨🏻		1		0.3
...	...	...	...	...

Patient	Covariates X	Potential outcomes $Y(0)$	Potential outcomes $Y(1)$	Treatment effect $Y(1) - Y(0)$
🧕🏽		?	?	?
👨🏿		?	?	?
...	...	...	...	...

Introduction: Research gap – Our contributions

Research gap

- From a representation learning perspective:
  - multiple works simply suggested **specific** representation learning models
  - they often apply a **balancing constraint** as a tool to reduce estimation variance
  - yet, they do not (explicitly) offer assymptic optimality properties
- From a perspective of Neyman-orthogonal learners:
  - no rigorous study on **structural assumptions** and **the usage of representation learning** for CAPOs/CATE estimation (only for APOs/ATE (Schulte et al., 2025[1]))

Our contributions

- We are the first to **unify** representation learning methods and Neyman-orthogonal learners into a joint framework of orthogonal representation learners (**OR-learners**)
- At the same time, we provide answers to two main research questions (**RQ 1** & **RQ 2**):

1) Rickmer Schulte, David Rügamer, and Thomas Nagler. Adjustment for confounding using pre-trained representations. In International Conference on Machine Learning, 2025.

Introduction: Research gap – Our contributions

Research gap

- From a representation learning perspective:
  - multiple works simply suggested **specific** representation learning models
  - they often apply a **balancing constraint** as a tool to reduce estimation variance
  - yet, they do not (explicitly) offer assymptic optimality properties
- From a perspective of Neyman-orthogonal learners:
  - no rigorous study on **structural assumptions** and **the usage of representation learning** for CAPOs/CATE estimation (only for APOs/ATE (Schulte et al., 2025[1]))

Our contributions

- We are the first to **unify** representation learning methods and Neyman-orthogonal learners into a joint framework of orthogonal representation learners (**OR-learners**)
- At the same time, we provide answers to two main research questions (**RQ 1** & **RQ 2**):

RQ ①. When do representations strengthen the existing Neyman-orthogonal learners?

1) Rickmer Schulte, David Rügamer, and Thomas Nagler. Adjustment for confounding using pre-trained representations. In International Conference on Machine Learning, 2025.

Introduction: Research gap – Our contributions

Research gap

- From a representation learning perspective:
  - multiple works simply suggested **specific** representation learning models
  - they often apply a **balancing constraint** as a tool to reduce estimation variance
  - yet, they do not (explicitly) offer assymptic optimality properties
- From a perspective of Neyman-orthogonal learners:
  - no rigorous study on **structural assumptions** and **the usage of representation learning** for CAPOs/CATE estimation (only for APOs/ATE (Schulte et al., 2025[1]))

Our contributions

- We are the first to **unify** representation learning methods and Neyman-orthogonal learners into a joint framework of orthogonal representation learners (**OR-learners**)
- At the same time, we provide answers to two main research questions (**RQ 1** & **RQ 2**):

RQ ①. When do representations strengthen the existing Neyman-orthogonal learners?

RQ ②. When can the balancing constraint improve the efficiency of learning similarly to Neyman-orthogonality?

1) Rickmer Schulte, David Rügamer, and Thomas Nagler. Adjustment for confounding using pre-trained representations. In International Conference on Machine Learning, 2025.

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

LMU MUNICH SCHOOL OF MANAGEMENT

INSTITUTE OF ARTIFICIAL INTELLIGENCE (AI) IN MANAGEMENT

Munich Center for Machine Learning

Agenda

Introduction

CAPOs/CATE estimation

OR-learners

Research question 1

Research question 2

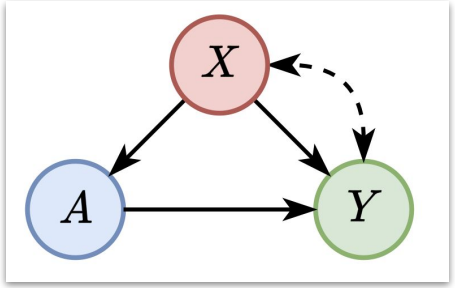
Takeaways

CAPOs/CATE estimation: Assumptions

Identifiability assumptions

- Potential outcomes framework (Neyman-Rubin):

  1. Consistency. If $A = a$ is a treatment for some patient, then $Y = Y[a]$
  2. Strong overlap. There is always a non-zero probability of receiving treatment, conditioning on the covariates: $\epsilon > 0, \mathbb{P}(1 - \epsilon \geq \pi_a(X) \geq \epsilon) = 1$
  3. Unconfoundedness. Current treatment is independent of the potential outcome, conditioning on the covariates: $A \perp\!\!\!\perp Y[a] \mid X$ for all a.



The causal diagram of a DGP that satisfies Assumptions (1) - (3)

- Under assumptions (1) - (3) CAPOs/CATE are identifiable as

$$\xi_a^x(x) = \mu_a^x(x) \qquad \tau^x(x) = \mu_1^x(x) - \mu_0^x(x)$$

where $\mu_a^x(x) = \mathbb{E}(Y \mid X = x, A = a)$ is an expected covariate-level outcome

Estimability assumptions

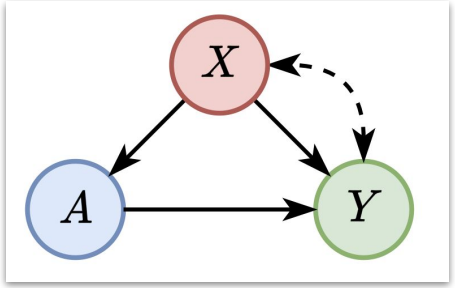
- To consistently estimate causal quantities, we assume that:
  - Covariate space $\mathcal{X}$ is compact
  - Ground-truth causal quantities (CATE/CAPOs) and nuisance functions (expected covariate-level outcome & propensity score) are **s-Hölder smooth** (for s > 0) with the corresponding **Hölder norms**

CAPOs/CATE estimation: Assumptions

Identifiability assumptions

- Potential outcomes framework (Neyman-Rubin):

  1. Consistency. If $A = a$ is a treatment for some patient, then $Y = Y[a]$
  2. Strong overlap. There is always a non-zero probability of receiving treatment, conditioning on the covariates: $\epsilon > 0, \mathbb{P}(1 - \epsilon \geq \pi_a(X) \geq \epsilon) = 1$
  3. Unconfoundedness. Current treatment is independent of the potential outcome, conditioning on the covariates: $A \perp\!\!\!\perp Y[a] \mid X$ for all a.



The causal diagram of a DGP that satisfies
ns (1) - (3)

- Under assumptions (1) - (3) CAPOs/CATE are identifiable as

$\xi_a^x(x) = \mu_a^x(x)$ $\tau^x(x) = \mu_1^x(x$

where $\mu_a^x(x) = \mathbb{E}(Y \mid X = x, A = a)$ i

ne

All the partial derivatives up to ⌊s⌋ exist and ⌊s⌋-th partial derivatives are Hölder, namely:

$$|D^m f(x) - D^m f(x')| \leq L \|x - x'\|_2^{s - \lfloor s \rfloor}$$

with $|m| = \sum_j m_j = \lfloor s \rfloor$

(i) Upper-bound on all the partial derivatives < ⌊s⌋
+ (ii) Hölder semi-norm for ⌊s⌋-th partial derivatives

$$L = \|f\|_{C^s(\mathcal{X})} := \sum_{|m| \leq \lfloor s \rfloor} \sup_{x \in \mathcal{X}} |D^m f(x)| + \sum_{|m| = \lfloor s \rfloor} \sup_{x, x' \in \mathcal{X}, x \neq x'} \frac{|D^m f(x) - D^m f(x')|}{\|x - x'\|_2^{s - \lfloor s \rfloor}}$$

es, we

TE/CAPOs) and nuisance functions (expected covariate-level outcome & propensity score) are **s-Hölder smooth** (for s > 0) with the corresponding **Hölder norms**

CAPOs/CATE estimation: Existing approaches

Existing approaches: Representation learning methods

- End-to-end representation learning aim to minimize a factual MSE:

$$\hat{\mathcal{L}}_{\Phi}(h_a \circ \Phi) = \mathbb{P}_n\{(Y - h_A(\Phi(X)))^2\}$$

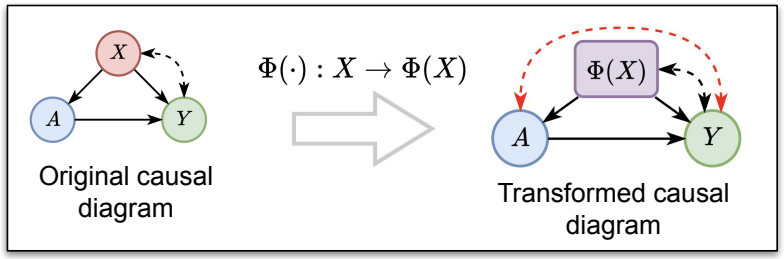
where Φ(.) is a representation subnetwork and $h_A(.)$ is an outcome subnetwork

- Often, **balancing constraint** is enforced to reduce the estimation variance:

$$\hat{\mathcal{L}}_{\mathrm{Bal}(\Phi)}(h_a \circ \Phi) = \hat{\mathcal{L}}_{\Phi}(h_a \circ \Phi) + \alpha\, \hat{\mathcal{L}}_{\mathrm{Bal}}(\Phi)$$

where the last term is a distributional distance: $\widehat{\mathrm{dist}}(\mathbb{P}(\Phi(X) \mid A = 0), \mathbb{P}(\Phi(X) \mid A = 1))$

- However, as discovered in (Melnychuk et al., 2024)[1], non-ivertible representations can induce confounding bias (RICB):



- Numerous specific neural models were proposed for CAPOs/CATE estimation

1) Valentyn Melnychuk, Dennis Frauen, and Stefan Feuerriegel. Bounds on representation-induced confounding bias for treatment effect estimation. In International Conference on Learning Representations, 2024.

CAPOs/CATE estimation: Existing approaches

Existing approaches: Representation learning methods

- Enc
- whe
- Ofte
- whe
- Hov
- indu
- Nur

Method	Learner type	Balancing constraint	Invertibility	Consistency of estimation	Neyman-orthogonality	
					CAPOs	CATE
TARNet (Shalit et al., 2017; Johansson et al., 2022)	PI	–	–	✓	✗	✗
BNN (Johansson et al., 2016); CFR (Shalit et al., 2017; Johansson et al., 2022); ESCFR (Wang et al., 2024); ORIC (Yan et al., 2025)	PI	IPM	(any) / –	✗ [✓: invertible]	✗	✗
RCFR (Johansson et al., 2018, 2022)	WPI	IPM + LW	(any) / –	✗ [✓: invertible]	✗	✗
DACPOL (Atan et al., 2018); CRN (Bica et al., 2020); ABCEI (Du et al., 2021); CT (Melnychuk et al., 2022); MitNet (Guo et al., 2023); BNCDE (Hess et al., 2024)	PI	JSD	–	✗	✗	✗
SITE (Yao et al., 2018)	PI	LS	MPD	✗ [✓: invertible]	✗	✗
DragonNet (Shi et al., 2019)	PI / (DR)	–	–	✓	(✓$^{\text{DR}^{\text{K}}}$)	(✓$^{\text{DR}^{\text{K}}}$)
PM (Schwab et al., 2018); StableCFR (Wu et al., 2023)	WPI	IPM + UVM	–	✓	✗	✗
CFR-ISW (Hassanpour and Greiner, 2019a);	IPTW	IPM + RP	–	✗	✗	✗
DR-CFR (Hassanpour and Greiner, 2019b); DeR-CFR (Wu et al., 2022)	IPTW	IPM + CP	–	✓	✗	✗
DKLITE (Zhang et al., 2020)	PI	CV	RL	✗ [✓: invertible]	✗	✗
BWCFR (Assaad et al., 2021)	IPTW	IPM + CP	–	✓	✗	✗
SNet (Curth and van der Schaar, 2021b; Chauhan et al., 2023)	DR	–	–	✓	(✓$^{\text{DR}^{\text{K}}}$)	✓$^{\text{DR}^{\text{K}}}$
GWIB (Yang et al., 2024)	PI	MI	–	✗	✗	✗
CausalEGM (Liu et al., 2024)	PI	–	GAN	✓	✗	✗

$= 1))$

ons can

1) Valentyn Melnychuk, Dennis Frauen, and Stefan Feuerriegel. Bounds on representation-induced confounding bias for treatment effect estimation. In International Conference on Learning Representations, 2024.

CAPOs/CATE estimation: Existing approaches

Existing approaches: Neyman-orthogonal learners

- Meta-learners for CAPOs/CATE estimation are **model-agnostic methods** that proceed in two stages: (i) nuisance functions estimation and (ii) target model fit
- They aim to minimize a weighted MSE in a target model class $\mathcal{G} = \{g(\cdot) : \mathcal{V} \subseteq \mathcal{X} \to \mathcal{Y}\}$

$$\mathcal{L}_{\mathcal{G}}(g, \eta) = \mathbb{E}\left[w(\pi_a^x(X))\left(\chi^x(X, \eta) - g(V)\right)^2\right]$$

where $\chi^x(.)$ is a causal quantity (CAPOs/CATE)

- Neyman-orthogonal learners use **debiased MSE** (so that the gradient of the MSE wrt. g(.) is first-order insensitive to the nuisance functions):

$$\hat{\mathcal{L}}_{\mathcal{G}}(g, \hat{\eta}) = \mathbb{P}_n\left\{\rho(A, \hat{\pi}_a^x(X))\left(\phi(Z, \hat{\eta}) - g(V)\right)^2\right\}$$

where φ(.) is a pseudo-outcome (matches to a causal quantity in expectation)

- They possess **quasi-oracle efficiency** and (often) **double-robustness:**

$$\|\hat{g} - g^*\|_{L_2}^2 \lesssim \text{Rate}_{\mathcal{D}}(\mathcal{G}; \hat{g}, \hat{\eta}) + \underbrace{\|\hat{\pi}_1^x - \pi_1^x\|_{L_4}^2 \|\hat{\mu}_a^x - \mu_a^x\|_{L_4}^2}_{R_2(\eta, \hat{\eta})}$$

- By choosing different weight functions, we can **reduce the variance of estimation** (yet, when MSE is regularized, we might add bias)

CAPOs/CATE estimation: Existing approaches

Existing approaches: Neyman-orthogonal learners

- Meta-learners for CAPOs/CATE estimation are **model-agnostic methods** that proceed in two stages: (i) nuisance functions estimation and (ii) target model fit
- They aim to minimize a weighted MSE in a target model class $\mathcal{G} = \{g(\cdot) : \mathcal{V} \subseteq \mathcal{X} \to \mathcal{Y}\}$

$$\mathcal{L}_{\mathcal{G}}(g, \eta) = \mathbb{E}\left[w(\pi_a^x(X))\left(\chi^x(X, \eta) - g(V)\right)^2\right]$$

where $\chi^x(.)$ is a causal quantity (CAPOs/CATE)

- Neyman-orthogonal **MSE** (so that the gradient of the MSE wrt. g(.) is first-order insensitive to the nuisance functions):

$$\hat{\mathcal{L}}_{\mathcal{G}}(g, \hat{\eta}) = \mathbb{P}_n\left\{\rho(A, \hat{\pi}_a^x(X))\left(\phi(Z, \hat{\eta}) - g(V)\right)^2\right\}$$

- [illegible] pseudo-outcome (matches to a causal quantity in expectation)
- They possess **quasi-oracle efficiency** and (often) **double-robustness:**

$$\|\hat{g} - g^*\|_{L_2}^2 \lesssim \mathrm{Rate}_{\mathcal{D}}(\mathcal{G}; \hat{g}, \hat{\eta}) + \underbrace{\|\hat{\pi}_1^x - \pi_1^x\|_{L_4}^2 \|\hat{\mu}_a^x - \mu_a^x\|_{L_4}^2}_{R_2(\eta, \hat{\eta})}$$

- By [illegible] we can **reduce the variance of estimation** (yet, when MSE is regularized, we might add bias

Best in-class projection

Empirical minimizer

Finite-sample second-stage error

Second-order remainder

CAPOs/CATE estimation: Existing approaches

- Meta-learners for CAPOs/CATE estimation are **model-agnostic methods** that proceed in two stages: (i) nuisance functions estimation and (ii) target model fit
- They aim to minimize a weighted MSE in a target model class $\mathcal{G} = \{g(\cdot) : \mathcal{V} \subseteq \mathcal{X} \to \mathcal{Y}\}$

Ex ap Ne or le

Causal quantity	Meta-learner	Neyman-orthogonal loss, $\hat{\mathcal{L}}_{\mathcal{G}}(g, \hat{\eta})$	Second-order remainder, $R_2(\eta, \hat{\eta})$
CAPOs	DR_a^{K}	$\mathbb{P}_n\left\{\left(\frac{\mathbb{1}\{A=a\}}{\hat{\pi}_a^x(X)}(Y-\hat{\mu}_a^x(X))+\hat{\mu}_a^x(X)-g(V)\right)^2\right\}$	$\|\hat{\pi}_1^x-\pi_1^x\|_{L_4}^2\,\|\hat{\mu}_a^x-\mu_a^x\|_{L_4}^2$
	DR_a^{FS}	$\mathbb{P}_n\left\{\frac{\mathbb{1}\{A=a\}}{\hat{\pi}_a^x(X)}(Y-g(V))^2+\left(1-\frac{\mathbb{1}\{A=a\}}{\hat{\pi}_a^x(X)}\right)(\hat{\mu}_a^x(X)-g(V))^2\right\}$	$\|\hat{\pi}_1^x-\pi_1^x\|_{L_4}^2\,\|\hat{\mu}_a^x-\mu_a^x\|_{L_4}^2$
CATE	DR^{K}	$\mathbb{P}_n\left\{\left(\frac{A-\hat{\pi}_1^x(X)}{\hat{\pi}_0^x(X)\,\hat{\pi}_1^x(X)}(Y-\hat{\mu}_A^x(X))+\hat{\mu}_1^x(X)-\hat{\mu}_0^x(X)-g(V)\right)^2\right\}$	$\sum_{a\in\{0,1\}}\|\hat{\pi}_1^x-\pi_1^x\|_{L_4}^2\,\|\hat{\mu}_a^x-\mu_a^x\|_{L_4}^2$
	R	$\mathbb{P}_n\left\{(A-\hat{\pi}_1^x(X))^2\left(\frac{Y-\hat{\mu}^x(X)}{A-\hat{\pi}_1^x(X)}-g(V)\right)^2\right\}$	$\sum_{a\in\{0,1\}}\|\hat{\pi}_1^x-\pi_1^x\|_{L_4}^2\,\|\hat{\mu}_a^x-\mu_a^x\|_{L_4}^2+\|\hat{\pi}_1^x-\pi_1^x\|_{L_4}^4$
	IVW	$\mathbb{P}_n\left\{(A-\hat{\pi}_1^x(X))^2\left(\frac{A-\hat{\pi}_1^x(X)}{\hat{\pi}_0^x(X)\,\hat{\pi}_1^x(X)}(Y-\hat{\mu}_A^x(X))+\hat{\mu}_1^x(X)-\hat{\mu}_0^x(X)-g(V)\right)^2\right\}$	$\sum_{a\in\{0,1\}}\|\hat{\pi}_1^x-\pi_1^x\|_{L_4}^2\,\|\hat{\mu}_a^x-\mu_a^x\|_{L_4}^2+\|\hat{\pi}_1^x-\pi_1^x\|_{L_4}^4$

References:
DR_a^{K} (Kennedy, 2023); DR_a^{FS} (Foster and Syrgkanis, 2023); DR^{K} (Kennedy, 2023); R (Nie and Wager, 2021); IVW (Fisher, 2024)

$R_2(\eta,\hat{\eta})$

Finite-sample second-stage error

- By ve can **reduce the variance of estimation** (yet, when MSE is regularized, we might add bias

Second-order remainder

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

LMU MUNICH SCHOOL OF MANAGEMENT

INSTITUTE OF ARTIFICIAL INTELLIGENCE (AI) IN MANAGEMENT

Munich Center for Machine Learning

Agenda

Introduction

CAPOs/CATE estimation

OR-learners

Research question 1

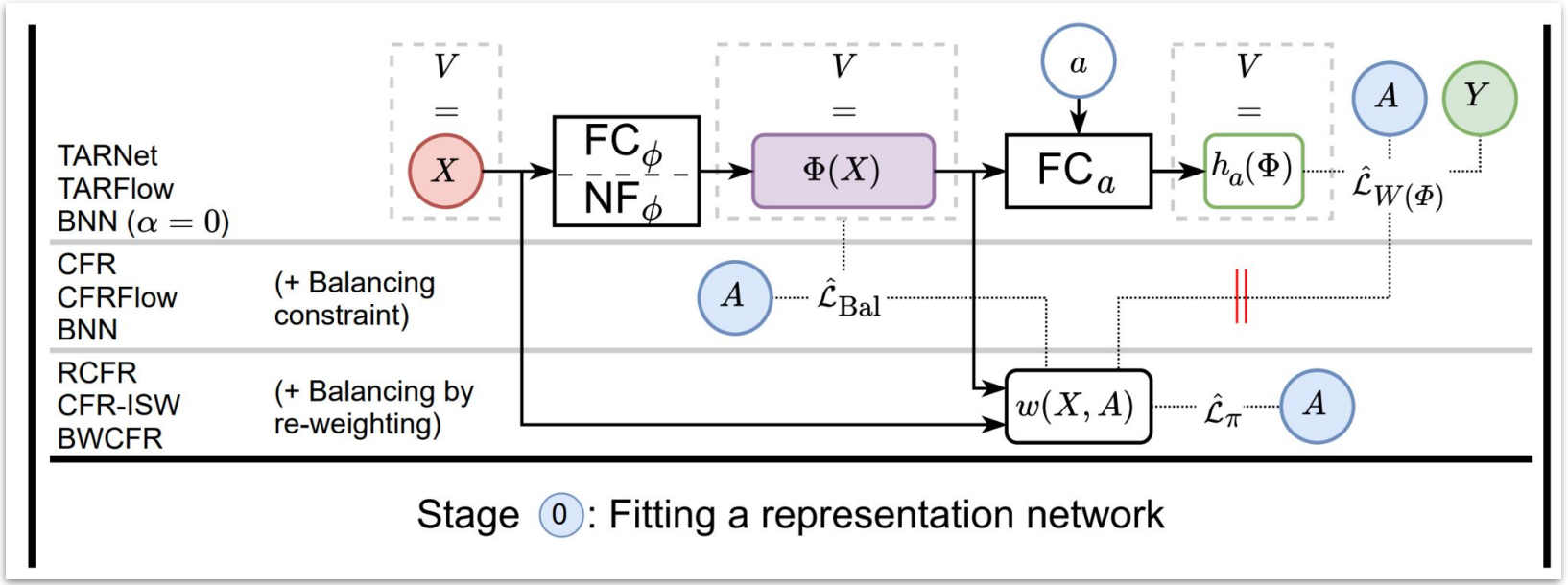
Research question 2

Takeaways

OR-learners: Unified framework

- To answer **RQ 1** & **RQ 2,** we propose a unified framework of (i) representation learning methods and (ii) Neyman-orthogonal learners, called orthogonal representation learners (**OR-learners**)
- OR-learners combine “best of both worlds”: (i) representation learning capabilities and (ii) favorable asymptotic properties of Neyman-orthogonality
- They proceed in three stages:

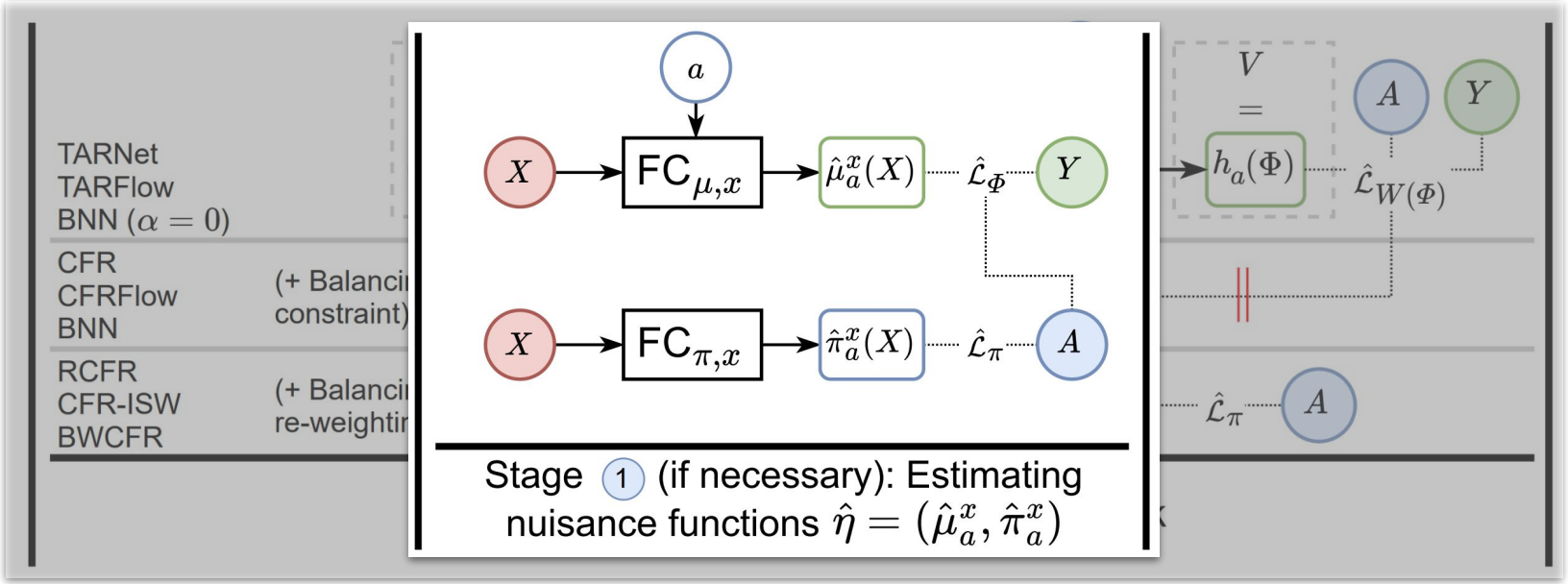
Unified framework



OR-learners: Unified framework

- To answer **RQ 1** & **RQ 2,** we propose a unified framework of (i) representation learning methods and (ii) Neyman-orthogonal learners, called orthogonal representation learners (**OR-learners**)
- OR-learners combine “best of both worlds”: (i) representation learning capabilities and (ii) favorable asymptotic properties of Neyman-orthogonality
- They proceed in three stages:

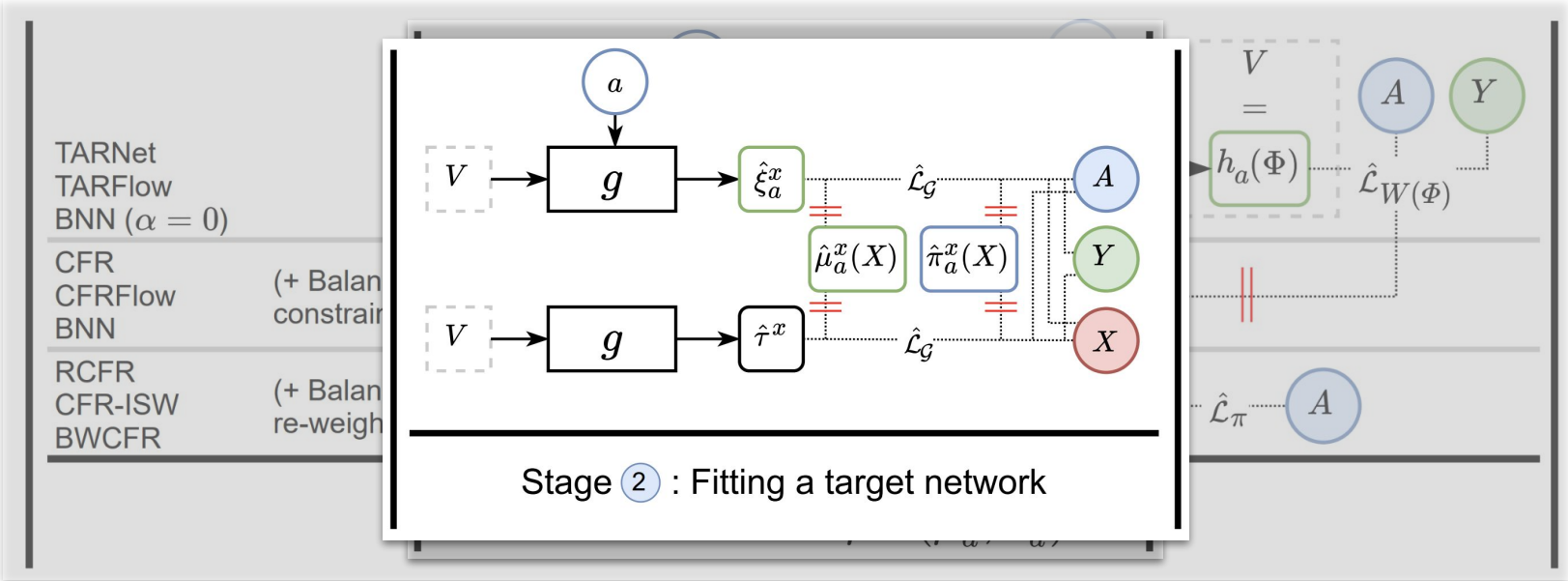
Unified framework



OR-learners: Unified framework

- To answer **RQ 1** & **RQ 2,** we propose a unified framework of (i) representation learning methods and (ii) Neyman-orthogonal learners, called orthogonal representation learners (**OR-learners**)
- OR-learners combine “best of both worlds”: (i) representation learning capabilities and (ii) favorable asymptotic properties of Neyman-orthogonality
- They proceed in three stages:

Unified framework



LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN | LMU MUNICH SCHOOL OF MANAGEMENT | INSTITUTE OF ARTIFICIAL INTELLIGENCE (AI) IN MANAGEMENT

Munich Center for Machine Learning

Agenda

Introduction

CAPOs/CATE estimation

OR-learners

Research question 1

Research question 2

Takeaways

RQ1: Error bounds under smoothness – Heterogeneity trade-off

RQ ①. When do representations strengthen the existing Neyman-orthogonal learners?

RQ 1: Error bounds under smoothness

- Under Hölder smoothness, the following **error bound** holds assymptotically for non-parametric models (e.g. local polynomial regression (Kennedy, 2023[1])):

$$\|g^{*v} - \hat{g}^v\|_{L_2}^2 \lesssim (L^v)^{\frac{2d_v}{(2s^v+d_v)}} n^{-\frac{2s^v}{(2s^v+d_v)}} + R_2$$

  where g^{*v} is the best projection, s^v - Hölder smooth with Hölder norm L^v, and $R_2 = R_2(\eta, \hat{\eta})$ is a second-order remainder that depends on the smoothness of nuisance functions $\eta = (\mu_0^x, \mu_1^x, \pi_1^x)$
- Similar error bounds can be found for **neural networks** (Schulte et al., 2025[2])

RQ 1: Heterogeneity trade-off

- While the R_2 term is irreducible, we can **reduce the first term** by choosing different **V**:
  - V = X: no asymptotic bias between g^{*v} and ground truth CAPOs/CATE, but **the first term** is the **largest**
  - V = ∅: first term becomes a parametric rate (1/n), but we **lose all the heterogeneity** of CAPOs/CATE (= asymptotic bias)

1) Edward H. Kennedy. Towards optimal doubly robust estimation of heterogeneous causal effects. Electronic Journal of Statistics, 17(2):3008–3049, 2023.
2) Rickmer Schulte, David Rügamer, and Thomas Nagler. Adjustment for confounding using pre-trained representations. In International Conference on Machine Learning, 2025.

RQ1: Error bounds under smoothness – Heterogeneity trade-off

RQ ①. When do representations strengthen the existing Neyman-orthogonal learners?

RQ 1: Error bounds under smoothness

- Under Hölder smoothness, the following **error bound** holds assymptotically for non-parametric models (e.g. local polynomial regression (Kennedy, 2023[1])):

$$\|g^{*v} - \hat{g}^v\|_{L_2}^2 \lesssim \left(L^v\right)^{\frac{2d_v}{(2s^v+d_v)}} n^{-\frac{2s^v}{(2s^v+d_v)}} + R_2$$

  where g^{*v} is the best projection, s^v - Hölder smooth with Hölder norm L^v, and $R_2 = R_2(\eta, \hat{\eta})$ is a second-order remainder that depends on the smoothness of nuisance functions $\eta = (\mu_0^x, \mu_1^x, \pi_1^x)$
- Similar error bounds can be found for **neural networks** (Schulte et al., 2025[2])

RQ 1: Heterogeneity trade-off

- While the R_2 term is irreducible, we can **reduce the first term** by choosing different **V**:
  - V = X: no asymptotic bias between g^{*v} and ground truth CAPOs/CATE, but **the first term** is the **largest**
  - V = ∅: first term becomes a parametric rate (1/n), but we **lose all the heterogeneity** of CAPOs/CATE (= asymptotic bias)

1) Edward H. Kennedy. Towards optimal doubly robust estimation of heterogeneous causal effects. Electronic Journal of Statistics, 17(2):3008–3049, 2023.
2) Rickmer Schulte, David Rügamer, and Thomas Nagler. Adjustment for confounding using pre-trained representations. In International Conference on Machine Learning, 2025.

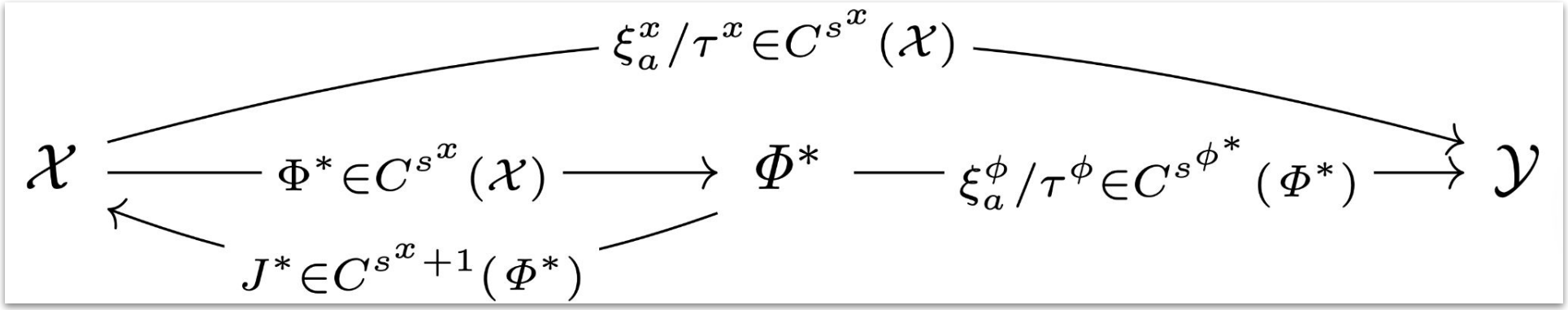
RQ1: Low-dimensional manifold hypothesis

RQ ①. When do representations strengthen the existing Neyman-orthogonal learners?

- To resolve the heterogeneity trade-off (= bias-variance trade-off for a target model), we assume a low-dimensional manifold hypothesis:

RQ 1: Low-dimensional manifold hypothesis

Assumption 1 (informal). (i) CAPOs/CATE are supported on a low-dimensional, compact, smooth manifold (representation space) $\Phi^* \subseteq \mathbb{R}^{d_{\phi^*}}, d_{\phi^*} \ll d_x$ such that $\xi_a^x(x) = \xi_a^{\phi^*}(\Phi^*(x))$ *and* $\tau^x(x) = \tau^{\phi^*}(\Phi^*(x))$ and (ii) there exists a sufficiently smooth (at least once differentaible) pullback map J^*



RQ1: Theoretical results

RQ ①. When do representations strengthen the existing Neyman-orthogonal learners?

RQ 1: Theoretical results

- Under Assumption 1, the following holds:

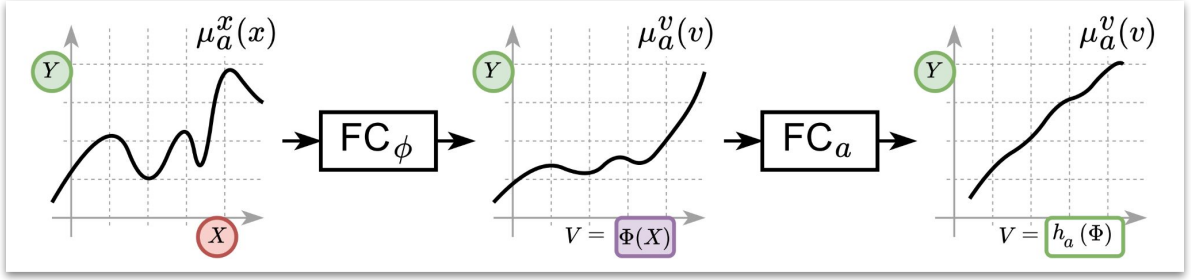
Proposition 1 (informal). Under Assumption 1 and when J^* is a contractive map, representation-level CAPOs/CATE are easier to learn:

$$||g^{*\phi^*} - \hat{g}^{\phi^*}||_{L_2}^2 \lesssim ||g^{*x} - \hat{g}^{x}||_{L_2}^2$$

- Can we learn the ideal representation from Assumption 1 in practice?

Proposition 2 (informal). Under mild conditions on the representation network trained with the factual MSE $\hat{\mathcal{L}}_{\Phi}(h_a \circ \Phi)$, there exists a hidden layer $V = \hat{f}(X)$ where the regression target becomes smoother

- Hence, the outputs of hidden layers can serve as **substitutes** for the ideal representation from Assumption 1 (= OR-learners)



RQ1: Theoretical results

RQ 1: Theoretical results

RQ ①. When do representations strengthen the existing Neyman-orthogonal learners?

- Under Assumption 1, the following holds:

Proposition 1 (informal). Under Assumption 1 and when J^* is a contractive map, representation-level CAPOs/CATE are easier to learn:

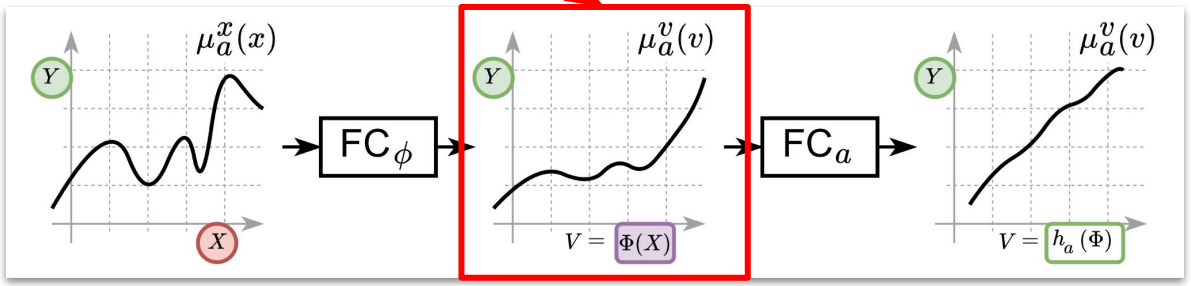
$$||g^{*\phi^*} - \hat{g}^{\phi^*}||_{L_2}^2 \lesssim ||g^{*x} - \hat{g}^{x}||_{L_2}^2$$

- Can we learn

Empirically, we saw that using **middle layers** is the best trade-off between (a) full re-training of the representation by a target model and (b) using only outputs that contain plug-in bias

Proposition ork trained with the factual MSE $\mathcal{L}_\Phi(h_a \circ \Phi)$, there exists a hidden layer $V = \hat{f}(X)$ where the regression target becomes smoother

- Hence, the outputs of hidden layers can serve as **substitutes** for the ideal representation from Assumption 1 (= OR-learners)



RQ1: Empirical results

RQ ①. When do representations strengthen the existing Neyman-orthogonal learners?

- In a numerous (semi-)synthetic benchmarks, our **OR-learners with V = Φ(X)** achieve the best performance **when the low-dimensional manifold hypothesis holds** (compared to standard Neyman-orthogonal learners and other variants):

RQ 1: Empirical results

		DR_0^K	DR_0^{FS}	DR_1^K	DR_1^{FS}	DR^K	R	IVW
TARNet	$V = (\hat{\mu}_0^x, \hat{\mu}_1^x)$	22.3%	20.9%	27.6%	25.5%	37.4%	37.1%	37.4%
	$V = X$	25.0%	20.4%	23.5%	13.2%	19.3%	6.8%	15.3%
	$V = X^*$	27.0%	28.7%	26.0%	23.4%	13.2%	6.2%	10.8%
	$V = \hat{\Phi}(X)$	64.7%	60.3%	69.0%	57.9%	68.6%	69.1%	67.4%
BNN ($\alpha = 0.0$)	$V = (\hat{\mu}_0^x, \hat{\mu}_1^x)$	40.9%	41.1%	40.7%	42.1%	45.4%	45.8%	44.6%
	$V = X$	38.2%	37.6%	33.5%	29.6%	24.4%	8.7%	19.6%
	$V = X^*$	40.5%	50.0%	34.6%	39.6%	13.8%	7.7%	10.9%
	$V = \hat{\Phi}(X)$	70.6%	70.6%	68.6%	73.4%	84.2%	79.4%	82.5%

Higher = better. Improvement over the baseline in more than 50% of runs marked in green

ACIC 2016 datasets collection

		DR_0^K	DR_0^{FS}	DR_1^K	DR_1^{FS}	DR^K	R	IVW
TARNet	$V = (\hat{\mu}_0^x, \hat{\mu}_1^x)$	+0.549 ± 0.006	+0.564 ± 0.006	+0.589 ± 0.003	+0.589 ± 0.003	+0.509 ± 0.004	+0.510 ± 0.004	+0.509 ± 0.004
	$V = X$	+0.011 ± 0.006	+0.082 ± 0.065	+0.017 ± 0.005	+0.011 ± 0.005	+0.002 ± 0.007	+0.215 ± 0.247	+0.004 ± 0.008
	$V = X^*$	+0.033 ± 0.009	−0.001 ± 0.007	+0.052 ± 0.014	−0.017 ± 0.003	+0.063 ± 0.012	+0.129 ± 0.179	+0.052 ± 0.005
	$V = \hat{\Phi}(X)$	−0.011 ± 0.004	+0.007 ± 0.053	−0.014 ± 0.002	−0.014 ± 0.006	−0.017 ± 0.005	−0.014 ± 0.020	−0.016 ± 0.005
BNN ($\alpha = 0.0$)	$V = (\hat{\mu}_0^x, \hat{\mu}_1^x)$	−0.004 ± 0.015	±0.000 ± 0.017	−0.013 ± 0.014	−0.014 ± 0.011	+0.001 ± 0.010	−0.002 ± 0.008	−0.002 ± 0.009
	$V = X$	+0.013 ± 0.028	+0.054 ± 0.043	+0.005 ± 0.021	−0.012 ± 0.025	+0.021 ± 0.025	+0.121 ± 0.102	+0.025 ± 0.031
	$V = X^*$	+0.040 ± 0.056	−0.006 ± 0.037	+0.048 ± 0.043	−0.039 ± 0.022	+0.087 ± 0.032	+0.075 ± 0.056	+0.096 ± 0.040
	$V = \hat{\Phi}(X)$	−0.019 ± 0.019	−0.029 ± 0.022	−0.034 ± 0.019	−0.040 ± 0.023	−0.020 ± 0.020	−0.027 ± 0.020	−0.022 ± 0.021

Lower = better. Significant improvement over the baseline in green, significant worsening of the baseline in red

HC-MNIST dataset

RQ1: Empirical results

RQ ①. When do representations strengthen the existing Neyman-orthogonal learners?

- In a numerous (semi-)synthetic benchmarks, our **OR-learners with V = Φ(X)** achieve the best performance **when the low-dimensional manifold hypothesis holds** (compared to standard Neyman-orthogonal learners and other variants):

RQ 1: Empirical results

		DR_0^K	DR_0^{FS}	DR_1^K	DR_1^{FS}	DR^K	R	IVW
TARNet	$V = (\hat{\mu}_0^x, \hat{\mu}_1^x)$	22.3%	20.9%	27.6%	25.5%	37.4%	37.1%	37.4%
	$V = X$	25.0%	20.4%	23.5%	13.2%	19.3%	6.8%	15.3%
	$V = X^*$	27.0%	28.7%	26.0%	23.4%	13.2%	6.2%	10.8%
	$V = \hat{\Phi}(X)$	64.7%	60.3%	69.0%	57.9%	68.6%	69.1%	67.4%
BNN ($\alpha = 0.0$)	$V = (\hat{\mu}_0^x, \hat{\mu}_1^x)$	40.9%	41.1%	40.7%	42.1%	45.4%	45.8%	44.6%
	$V = X$	38.2%	37.6%	33.5%	29.6%	24.4%	8.7%	19.6%
	$V = X^*$	40.5%	50.0%	34.6%	39.6%	13.8%	7.7%	10.9%
	$V = \hat{\Phi}(X)$	70.6%	70.6%	68.6%	73.4%	84.2%	79.4%	82.5%

Higher = better. Improvement over the baseline in more than 50% of runs marked in green

ACIC 2016 datasets collection

		DR_0^K	DR_0^{FS}	DR_1^K	DR_1^{FS}	DR^K	R	IVW
TARNet	$V = (\hat{\mu}_0^x, \hat{\mu}_1^x)$	+0.549 ± 0.006	+0.564 ± 0.006	+0.589 ± 0.003	+0.589 ± 0.003	+0.509 ± 0.004	+0.510 ± 0.004	+0.509 ± 0.004
	$V = X$	+0.011 ± 0.006	+0.082 ± 0.065	+0.017 ± 0.005	+0.011 ± 0.005	+0.002 ± 0.007	+0.215 ± 0.247	+0.004 ± 0.008
	$V = X^*$	+0.033 ± 0.009	−0.001 ± 0.007	+0.052 ± 0.014	−0.017 ± 0.003	+0.063 ± 0.012	+0.129 ± 0.179	+0.052 ± 0.005
	$V = \hat{\Phi}(X)$	−0.011 ± 0.004	+0.007 ± 0.053	−0.014 ± 0.002	−0.014 ± 0.006	−0.017 ± 0.005	−0.014 ± 0.020	−0.016 ± 0.005
BNN ($\alpha = 0.0$)	$V = (\hat{\mu}_0^x, \hat{\mu}_1^x)$	−0.004 ± 0.015	±0.000 ± 0.017	−0.013 ± 0.014	−0.014 ± 0.011	+0.001 ± 0.010	−0.002 ± 0.008	−0.002 ± 0.009
	$V = X$	+0.013 ± 0.028	+0.054 ± 0.043	+0.005 ± 0.021	−0.012 ± 0.025	+0.021 ± 0.025	+0.121 ± 0.102	+0.025 ± 0.031
	$V = X^*$	+0.040 ± 0.056	−0.006 ± 0.037	+0.048 ± 0.043	−0.039 ± 0.022	+0.087 ± 0.032	+0.075 ± 0.056	+0.096 ± 0.040
	$V = \hat{\Phi}(X)$	−0.019 ± 0.019	−0.029 ± 0.022	−0.034 ± 0.019	−0.040 ± 0.023	−0.020 ± 0.020	−0.027 ± 0.020	−0.022 ± 0.021

Lower = better. Significant improvement over the baseline in green, significant worsening of the baseline in red

HC-MNIST dataset

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

LMU MUNICH SCHOOL OF MANAGEMENT

INSTITUTE OF ARTIFICIAL INTELLIGENCE (AI) IN MANAGEMENT

Munich Center for Machine Learning

Agenda

Introduction

CAPOs/CATE estimation

OR-learners

Research question 1

Research question 2

Takeaways

RQ2: Balancing constraint & representation-induced bias

RQ ②. When can the balancing constraint improve the efficiency of learning similarly to Neyman-orthogonality?

RQ 2: Balancing constraint & representa-tion-induced bias

- As mentioned previously, balancing constraint can induce an asymptotic bias (representation induced confounding bias, RICB), when applied to the existing representation learning methods:

$$\xi_a^{\hat{\phi}}(\phi) \neq \mu_a^{\hat{\phi}}(\phi) \text{ and } \tau^{\hat{\phi}}(\phi) \neq \mu_1^{\hat{\phi}}(\phi) - \mu_0^{\hat{\phi}}(\phi)$$

- The RICB can be circumvented with:
  - **invertibility** of representations (yet, this compromises the main purpose of low-dimensional representations)
  - by using our **OR-learners** (as we do not put any constraints while estimating the nuisance functions)
  - by enfocing balancing constraint for the **target model only** (an example of overlap-adaptive regularization (Melnychuk et al., 2026[1]))

1) Valentyn Melnychuk, Dennis Frauen, Jonas Schweisthal, and Stefan Feuerriegel. Overlap-Adaptive Regularization for Conditional Average Treatment Effect Estimation. In International Conference on Learning Representations, 2026.

RQ2: Theoretical results

RQ ②. When can the balancing constraint improve the efficiency of learning similarly to Neyman-orthogonality?

- However, in general, balancing constraint can be **inherently detrimental!**

Proposition 3 (informal). Under (relatively) mild conditions, factual MSE minimization $\hat{\mathcal{L}}_{\Phi}(h_a \circ \Phi)$ makes the representation an **expanding map** (= Lipschitz constant >= 1)

RQ 2: Theoretical results

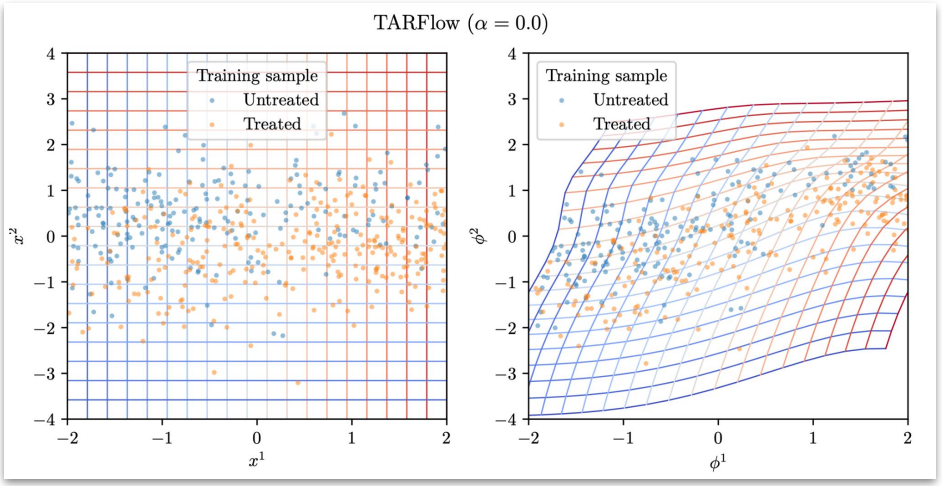
Proposition 4 (informal). Balancing constraints (i.e., Wasserstein distance & maximum mean discrepancy) force the representation to be a **contractive map** (Lipschitz constant <= 1)

- For both to work together, the balancing constraint requires an **inductive bias**: the low-overlap regions of the covariate space coincide with the low CAPOs/CATE heterogeneity (e.g., instrumental variables)
- In general, balancing **cannot recover the lack of Neyman-orthogonality** and is asymptotically detrimental

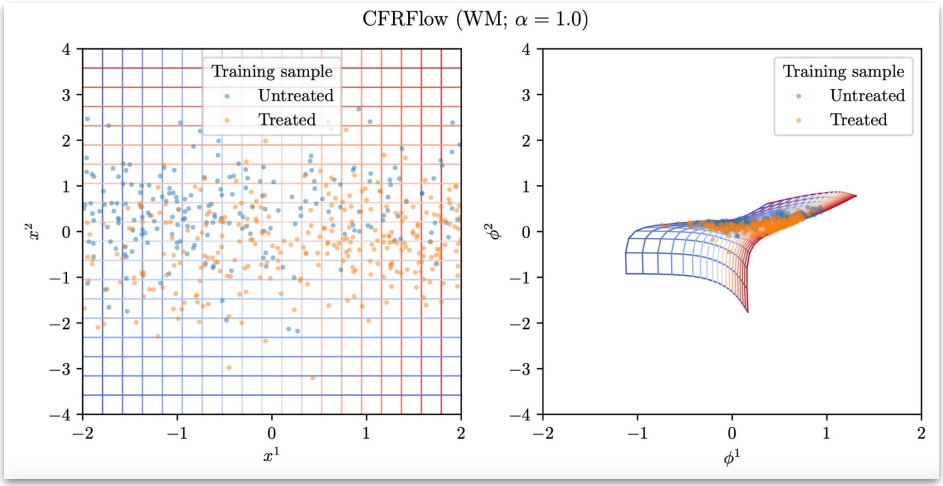
RQ2: Empirical results

RQ ②. When can the balancing constraint improve the efficiency of learning similarly to Neyman-orthogonality?

RQ 2: Empirical results



Invertible representation network w/o balancing

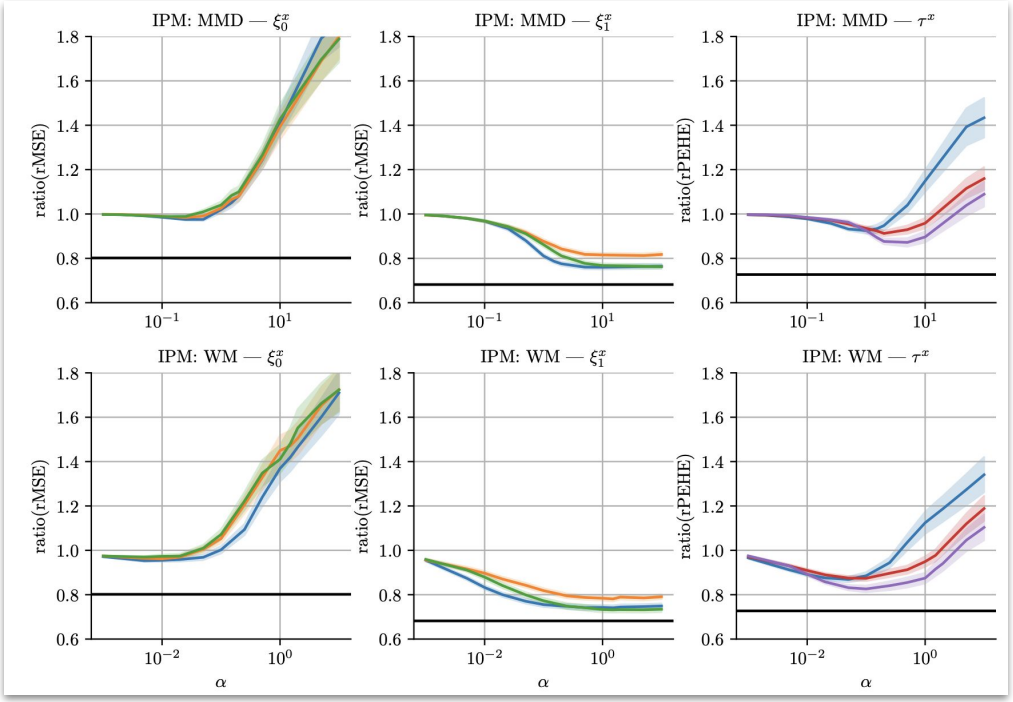


Invertible representation network w/ balancing

RQ2: Empirical results

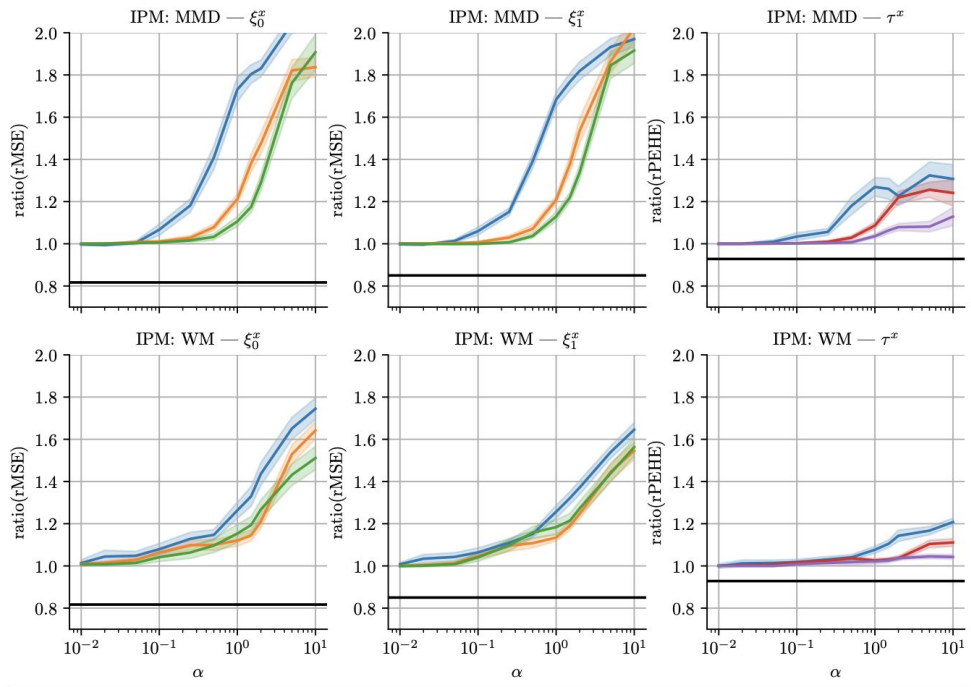
RQ ②. When can the balancing constraint improve the efficiency of learning similarly to Neyman-orthogonality?

RQ 2: Empirical results



Balancing constraint is beneficial (in blue):

- when the inductive bias is present
- only in low-sample regime
- hyperparameter α needs to be tuned



Balancing is detrimental (in blue):

- majority of the datasets

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

LMU MUNICH SCHOOL OF MANAGEMENT

INSTITUTE OF ARTIFICIAL INTELLIGENCE (AI) IN MANAGEMENT

Munich Center for Machine Learning

AISTATS

Agenda

Introduction

CAPOs/CATE estimation

OR-learners

Research question 1

Research question 2

Takeaways

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

LMU MUNICH SCHOOL OF MANAGEMENT

INSTITUTE OF ARTIFICIAL INTELLIGENCE (AI) IN MANAGEMENT

Munich Center for Machine Learning

AISTATS

Takeaways

Given that there is **no nuisance-free** way to do CATE/CAPOs model selection based solely on the observational data, we showed that:

RQ 1: Given the **low-dimensional manifold assumption**, one can simplify the task of CAPOs/CATE estimation and use the proposed framework of OR-learners

RQ 2: We advise **against the balancing constraint**, unless one can assume the underlying inductive bias and uses it correctly (so that it does not induce a confounding bias)

ArXiv: arxiv.org/abs/2502.04274

See you at the Poster session 3, # 171

RQ1: Conclusion

RQ ①. When do representations strengthen the existing Neyman-orthogonal learners?

RQ 1: Conclusion

Guidelines from RQ ①. We suggest using the *OR-learners* as they instrumentalize the core Assumption 1: Under it, the representation-based Neyman-orthogonal learners outperform the standard Neyman-orthogonal learners. Furthermore, *OR-learners* offer a middle-ground solution between **(a)** the full re-training of the representation network at the second-stage and **(b)** debiasing only the representation network outputs.

RQ2: Conclusion

RQ ②. When can the balancing constraint improve the efficiency of learning similarly to Neyman-orthogonality?

RQ 2: Empirical results

Guidelines from RQ ②. The balancing constraint relies on the strong inductive bias that the low-overlap regions of the covariate space coincide with the low CAPOs/CATE heterogeneity. The *OR-learners*, on the other hand, do not make such an assumption and provide general asymptotic optimality guarantees.