Orthogonal Representation Learning for Estimating Causal Quantities

Valentyn Melnychuk **Dennis Frauen** **Jonas Schweisthal** **Stefan Feuerriegel**
LMU Munich & Munich Center for Machine Learning
Munich, Germany
melnychuk@lmu.de

Abstract

End-to-end representation learning has become a powerful tool for estimating causal quantities from high-dimensional observational data, but its efficiency remained unclear. Here, we face a central tension: End-to-end representation learning methods often work well in practice but lack asymptotic optimality in the form of the quasi-oracle efficiency. In contrast, two-stage Neyman-orthogonal learners provide such a theoretical optimality property but do not explicitly benefit from the strengths of representation learning. In this work, we step back and ask two research questions: (1) *When do representations strengthen existing Neyman-orthogonal learners?* and (2) *Can a balancing constraint – a commonly proposed technique in the representation learning literature – provide improvements to Neyman-orthogonality?* We address these two questions through our theoretical and empirical analysis, where we introduce a unifying framework that connects representation learning with Neyman-orthogonal learners (namely, *OR-learners*). In particular, we show that, under the low-dimensional manifold hypothesis, the OR-learners can strictly *improve* the estimation error of the standard Neyman-orthogonal learners. At the same time, we find that the balancing constraint requires an additional inductive bias and *cannot* generally compensate for the lack of Neyman-orthogonality of the end-to-end approaches. Building on these insights, we offer guidelines for how users can effectively combine representation learning with the classical Neyman-orthogonal learners to achieve both practical performance and theoretical guarantees.

Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300. Copyright 2026 by the author(s).

1 INTRODUCTION

Estimating causal quantities has many applications in medicine (Feuerriegel et al., 2024), policy-making (Kuzmanovic et al., 2024), marketing (Varian, 2016), and economics (Basu et al., 2011). In this work, we focus on the *individualized causal quantities*, namely, the conditional average treatment effect (CATE) and the conditional average potential outcomes (CAPOs).

Recently, **end-to-end representation learning methods** have gained wide popularity in estimating causal quantities from observational data (e.g., Johansson et al., 2016; Shalit et al., 2017; Hassanpour and Greiner, 2019a,b; Zhang et al., 2020; Assaad et al., 2021; Johansson et al., 2022). The key benefit of the representation learning methods is that they utilize a key *low-dimensional manifold hypothesis* (Fefferman et al., 2016), namely, that the nuisance functions and the ground-truth causal quantity are both defined on some low-dimensional manifold of the covariate space.

A different literature stream seeks to estimate causal quantities through **two-stage Neyman-orthogonal learners** (Morzywolek et al., 2023; Vansteelandt and Morzywołek, 2025). Prominent examples are the DR-learner (Kennedy, 2023), R-learner (Nie and Wager, 2021), and IVW-learner (Fisher, 2024). Unlike end-to-end representation learning, Neyman-orthogonal learners offer several *favorable asymptotic properties*, namely quasi-oracle efficiency and double robustness (Chernozhukov et al., 2017; Foster and Syrgkanis, 2023).

However, we arrive at a central tension: On the one hand, end-to-end representation learning methods might substantially help with the high-dimensional data; while, on the other, Neyman-orthogonal meta-learners possess asymptotic optimality properties. In this work, we aim to address this tension by answering *two core research questions* (**RQ**).

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

RQ ① has not yet been fully addressed. For example, the prior work (Schulte et al., 2025) studied how the representations can facilitate a semi-parametric average-treatment effect (ATE) estimation (analogous to quasi-oracle efficiency when the causal quantity is finitely-dimensional). (Curth and van der Schaar, 2021b) provided an empirical study of CATE meta-learners with representation learning used for the estimation of the nuisance functions, but the learned representations have not been used as the target model inputs. Hence, strategies to reconcile the advantages of representation learning and Neyman-orthogonal learners in a principled manner are unclear.

As an alternative strategy to reduce the variance of the estimation, some end-to-end representation learning methods suggested a *balancing constraint* for the representations (Johansson et al., 2016; Shalit et al., 2017). This motivates our second question.

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

To the best of our knowledge, **RQ** ② has not yet been studied. As discovered in (Melnychuk et al., 2024), the balancing constraint can lead to a *representation-induced confounding bias* (RICB), and it is only guaranteed to omit the RICB (and perform a consistent estimation) for *invertible* representations (Johansson et al., 2019).

We analyze the **RQ** ① and **RQ** ② by introducing a unifying framework of Neyman-orthogonal representation learners for CAPOs/CATE, namely *OR-learners*. The *OR-learners* employ representation learning at *both stages of training*: (a) we use representation learning to jointly estimate the nuisance functions, and (b) we use the learned representations as the inputs for the target models. Based on the OR-learners, we then answer both research questions:

To answer **RQ** ①, we provide *sufficient theoretical guarantees*, under which the *OR-learners* are guaranteed to outperform the existing Neyman-orthogonal learners. Here, unlike (Curth and van der Schaar, 2021b), we use the learned representations at both stages of Neyman-orthogonal learners.

In response to **RQ** ②, we discover that the balancing constraint heavily relies on an additional inductive bias that the *low-overlap regions of the covariate space also have the low heterogeneity of the ground-truth CAPOs/CATE*. Therefore, in general, the balancing constraint (no matter whether invertibility is enforced) is *not* a substitute for Neyman-orthogonality and is, thus, asymptotically inferior to the *OR-learners*. One of the important consequences of **RQ** ② is that the *OR-learners*, due to their Neyman-orthogonality, *asymptotically outperform both categories of end-to-end methods: (a) with unconstrained and (b) with balanced representations*.

In sum, **our main contribution** is the following: By answering our **RQ** ① and **RQ** ②, we propose *guidelines* on how a causal ML practitioner can effectively combine the representation learning with the Neyman-orthogonal meta-learners. Specifically, we provide the conditions under which the *OR-learners*, the combination of both (i) the representation learning and (ii) Neyman-orthogonal meta-learners, perform better than each approach (i) and (ii) separately.

2 RELATED WORK

Our work aims to unify two streams of work, namely, representation learning methods and Neyman-orthogonal learners. We briefly review both in the following (a detailed overview is in Appendix A).

End-to-end representation learning. Several methods have been previously introduced for *end-to-end* representation learning of CAPOs/CATE (see, in particular, the seminal works (Johansson et al., 2016; Shalit et al., 2017; Johansson et al., 2022)). A large number of works later suggested different extensions to these. Existing methods fall into three main streams: (1) One can fit an *unconstrained shared representation* to directly estimate both potential outcome surfaces (e.g., **TARNet** (Shalit et al., 2017)). (2) Some methods additionally enforce a *balancing constraint* based on empirical probability metrics, so that the distributions of the treated and untreated representations become similar (e.g., **CFR** and **BNN** (Johansson et al., 2016; Shalit et al., 2017)). Importantly, balancing based on empirical probability metrics is only guaranteed to perform a consistent estimation for *invertible* representations since, otherwise, balancing leads to a *representation-induced confounding bias* (RICB) (Johansson et al., 2019; Melnychuk et al., 2024). (3) One can additionally perform *balancing by re-weighting* the loss and the distributions of the representations with learnable weights (e.g., **RCFR** (Johansson et al., 2022)). We later adopt the methods from (1)–(3) as baselines.

Neyman-orthogonal learners. Causal quantities can be estimated using model-agnostic methods, so-called *meta-learners* (Künzel et al., 2019). Prominent examples are the DR-learner (Kennedy, 2023; Curth et al., 2020), R-learner (Nie and Wager, 2021), and IVW-learner (Fisher, 2024). Meta-learners are model-agnostic in the sense that any base model (e.g., neural network) can be used for estimation. Also, meta-learners have several practical advantages (Morzywolek

et al., 2023): (i) they oftentimes offer favorable theoretical guarantees such as Neyman-orthogonality (Chernozhukov et al., 2017; Foster and Syrgkanis, 2023); (ii) they can address the causal inductive bias that the CATE is "simpler" than CAPOs (Curth and van der Schaar, 2021a), and (iii) the target model obtains a clear interpretation as a projection of the ground-truth CAPOs/CATE on the target model class. (Curth and van der Schaar, 2021b; Frauen et al., 2025) provided a comparison of meta-learners implemented via neural networks with different representations, yet with the target model based on the original covariates (the representations were only used as an interim tool to estimate nuisance functions). In contrast, here, we study the learned representations as primary inputs to the target model.

Representation learning and efficient estimation. Perhaps the closest to ours is (Schulte et al., 2025). Therein, the authors provided theoretical and empirical evidence on how representation learning might help with semi-parametric efficient estimation, but only for the ATE. Yet, a comprehensive study for CAPOs/CATE about the combination of representation learning and Neyman-orthogonal learners (= an analogue of semi-parametric estimators for the infinitely-dimensional estimands) is *missing*.

Research gap. From a *representation learning perspective*, multiple works simply suggested specific representation learning models, yet they did not properly explain where the gain of their performance comes from. On the other hand, from a perspective of *Neyman-orthogonal learners*, rigorous studies of structural assumptions and representation learning are very recent and have not been done for CAPOs/CATE. Our work is the first to unify representation learning methods and Neyman-orthogonal learners and give definitive answers to **RQ** ① and **RQ** ②. As a result, we develop guidelines for how practitioners can effectively combine representation learning with the classical Neyman-orthogonal learners.

3 PRELIMINARIES

Notation. We denote random variables with capital letters Z, their realizations with small letters z, and their domains with calligraphic letters $\mathcal{Z}$. Let $\mathbb{P}(Z)$, $\mathbb{P}(Z = z)$, $\mathbb{E}(Z)$ be the distribution, probability mass function/density, and expectation of Z, respectively. Let $\mathbb{P}_n\{f(Z)\} = \frac{1}{n}\sum_{i=1}^{n} f(z_i)$ be the sample average of $f(Z)$. We also denote the L_p norm as $\|f\|_{L_p} = (\mathbb{E}(|f(Z)|^p))^{1/p}$. We define the following nuisance functions: $\pi_a^x(x) = \mathbb{P}(A = a \mid X = x)$ is the *covariate propensity score* for the treatment A, and $\mu_a^x(x) = \mathbb{E}(Y \mid X = x, A = a)$ is the *expected covariate-level outcome* for the outcome Y. We also denote $\mu^x(x) = \pi_0^x(x)\mu_0^x(x) + \pi_1^x(x)\mu_1^x(x)$. Similarly, we define $\pi_a^\phi(\phi) = \mathbb{P}(A = a \mid \Phi(X) = \phi)$ and $\mu_a^\phi(\phi) = \mathbb{E}(Y \mid \Phi(X) = \phi, A = a)$ as the *representation propensity score* and the *expected representation-level outcome* for a representation $\Phi(x) = \phi$, respectively. Importantly, the superscripts in $\pi_a^x, \mu_a^x, \pi_a^\phi, \mu_a^\phi$ indicate whether the corresponding nuisance functions depend on the covariates x or on the representation ϕ.

Problem setup. To estimate the causal quantities, we make use of an observational dataset $\mathcal{D}$ that contains high-dimensional covariates $X \in \mathcal{X} \subseteq \mathbb{R}^{d_x}$, a binary treatment $A \in \{0, 1\}$, and a continuous outcome $Y \in \mathcal{Y} \subseteq \mathbb{R}$. For example, a common setting is an anti-cancer therapy, where the outcome is the tumor growth, the treatment is whether chemotherapy is administered, and covariates are patient information such as age and sex. The dataset $\mathcal{D} = \{z_i = (x_i, a_i, y_i)\}_{i=1}^n$ is assumed to be sampled i.i.d. from a joint distribution $\mathbb{P}(Z) = \mathbb{P}(X, A, Y)$ with dataset size n.

Causal quantities. We are interested in the estimation of two important causal quantities at the covariate level of heterogeneity: • *conditional average potential outcomes (CAPOs)* given by $\xi_a^x(x)$, and • the *conditional average treatment effect (CATE)* given by $\tau^x(x)$, with $\xi_a^x(x) = \mathbb{E}(Y[a] \mid X = x)$ and $\tau^x(x) = \mathbb{E}(Y[1] - Y[0] \mid X = x) = \xi_1^x(x) - \xi_0^x(x)$. If we could directly observe the true outcomes $Y[a]$ under both treatments (or the corresponding treatment effect $Y[1] - Y[0]$), then the consistent estimation of CAPOs and CATE, respectively, would reduce to a standard regression problem, but this is impossible due to the fundamental problem of causal inference.

Assumptions. To consistently estimate the causal quantities given only the observational data $\mathcal{D}$, we need to make standard assumptions (Rubin, 1974; Curth and van der Schaar, 2021b; Kennedy, 2023). **(1)** For **identifiability**, we assume (i) *consistency*: if $A = a$, then $Y[a] = Y$; (ii) *overlap*: $\mathbb{P}(0 < \pi_a^x(X) < 1) = 1$; and (iii) *unconfoundedness*: $(Y[0], Y[1]) \perp\!\!\!\perp A \mid X$. Given the assumptions (i)–(iii), both CAPOs and CATE are identifiable from $\mathbb{P}(Z)$ as expected covariate-level outcomes, $\xi_a^x(x) = \mu_a^x(x)$, or as the difference of expected covariate-level outcomes, $\tau^x(x) = \mu_1^x(x) - \mu_0^x(x)$, respectively. **(2)** For **estimability** of the CAPOs/CATE from the finite-sized dataset $\mathcal{D}$ (e. g., with neural networks), we assume that (i) $\mathcal{X}$ is compact; and (ii) the Hölder smoothness of the ground-truth causal quantities and the nuisance functions (see Appendix B for definitions). Specifically, we assume the CAPOs, $\xi_a^x(\cdot) = \mu_a^x(\cdot)$, to be $s_{\mu_a}^x$-smooth; the CATE $\tau^x(\cdot)$ to be s_τ^x-smooth; and the ground-truth propensity score $\pi_a^x(\cdot)$ to be s_π^x-smooth ($s_{\mu_a}^x, s_\tau^x, s_\pi^x > 0$). We also denote the corresponding Hölder norms $\|\cdot\|_{C^{s^x}(\mathcal{X})}$ as $L_{\mu_a}^x, L_\tau^x, L_\pi^x > 0$; and Lipschitz constants as $[\cdot]_{C^1(\mathcal{X})} = \mathrm{Lip}(\cdot)$, where

$[\cdot]_{C^s(\mathcal{X})}$ is a Hölder semi-norm.

End-to-end representation learning. Representation learning methods make use of the identifiability formulas for CAPOs/CATE to fit a *representation network*. Hence, the majority of the methods (e.g., Johansson et al., 2016; Shalit et al., 2017; Hassanpour and Greiner, 2019a,b; Zhang et al., 2020; Assaad et al., 2021; Johansson et al., 2022) aim to (jointly) learn both expected covariate-level outcomes μ_a^x as a composition of a (a) representation subnetwork $\Phi(x) : \mathcal{X} \to \Phi$ and an (b) outcome subnetwork $h_a(\phi) : \Phi \times \mathcal{A} \to \mathcal{Y}$. Both (a) and (b) then aim to minimize a factual mean squared error (MSE) risk:

$$\hat{\mathcal{L}}_{\Phi}(h_a \circ \Phi) = \mathbb{P}_n\big\{(Y - h_A(\Phi(X)))^2\big\}. \tag{1}$$

Some approaches further apply the covariate or representation propensity weights in Eq. (1). As a result, all end-to-end representation learning methods are either plug-in or inverse probability of treatment weighted (IPTW) learners (Curth and van der Schaar, 2021b), which both generally suffer from the *first-order error* of the model misspecification (Kennedy, 2023; Morzywolek et al., 2023).

Neyman-orthogonal learners. Here, we focus on *two-stage Neyman-orthogonal learners* due to their theoretical advantages (Chernozhukov et al., 2017; Foster and Syrgkanis, 2023). Formally, two-stage learners aim to find the best projection of CAPOs/CATE onto a *target model class* $\mathcal{G} = \{g(\cdot) : \mathcal{V} \subseteq \mathcal{X} \to \mathcal{Y}\}$ by minimizing different *target risks* wrt. $g(V)$ ($V \subseteq X$ is a conditioning set and the input for the target model). Then, the target risk is chosen as a (weighted) MSE:

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

where $\eta = (\mu_0^x, \mu_1^x, \pi_1^x)$ are the nuisance functions; $w(\cdot) > 0$ is a weighting function; and $\chi^x(\cdot)$ is the target causal quantity (i. e., $\chi^x(x, \eta) = \mu_a^x(x)$ for CAPOs and $\chi^x(x, \eta) = \mu_1^x(x) - \mu_0^x(x)$ for CATE). The Neyman-orthogonal learners proceed in two stages: (i) the nuisance functions are learned, $\hat{\eta}$, and (ii) *debiased estimators* of the target risks $\mathcal{L}_{\mathcal{G}}(g, \eta)$ are fitted wrt. g:

$$\hat{\mathcal{L}}_{\mathcal{G}}(g, \hat{\eta}) = \mathbb{P}_n\big\{\rho(A, \hat{\pi}_a^x(X))\,(\phi(Z, \hat{\eta}) - g(V))^2\big\}, \tag{3}$$

where $\rho(\cdot)$ and $\phi(\cdot)$ are a learner-specific function and pseudo-outcome, respectively. For example, the DR-learner in the style[1] of Kennedy (2023) is given by $w = \rho = 1$, $\phi_{\xi_a}(Z, \eta) = \mathbb{1}\{A = a\}(Y - \mu_a^x(X))/\pi_a^x(X) + \mu_a^x(X)$ for CAPOs estimation, and $\phi_\tau(Z, \eta) = (A - \pi_1^x(X))(Y - \mu_A^x(X))/(\pi_0^x(X)\,\pi_1^x(X)) + \mu_1^x(X) - \mu_0^x(X)$ for CATE estimation. The R-learner (Nie and Wager, 2021) is given by $w = \pi_0^x(X)\pi_1^x(X)$, $\rho = (A - \pi_1^x(X))^2$,

[1]We introduce an alternative DR-learner in the style of Foster and Syrgkanis (2023) in Appendix B.

$$\begin{tikzcd}
\mathcal{X} \arrow[r, "\Phi^* \in C^{s^x}(\mathcal{X})"] \arrow[rr, bend left, "\xi_a^x/\tau^x \in C^{s^x}(\mathcal{X})"] & \Phi^* \arrow[l, bend left, "J^* \in C^{s^x+1}(\Phi^*)"] \arrow[r, "\xi_a^\phi/\tau^\phi \in C^{s^{\phi^*}}(\Phi^*)"] & \mathcal{Y}
\end{tikzcd}$$

Figure 1: Visual summary of the relationships between the representation map Φ^*, its pullback J^*, and target CAPOs/CATE from Assumption 1.

and $\phi_\tau(Z, \eta) = (Y - \mu^x(X))/(A - \pi_1^x(X))$. Similarly, the IVW-learner (Fisher, 2024) is defined by the combination of the weighting functions of the R-learner, w and ρ, and the pseudo-outcome of the DR-learner, ϕ_τ. Because the Neyman-orthogonal learners use the debiased target risks, they possess a *quasi-oracle efficiency* and *double robustness* (= asymptotic optimality properties). We provide a more detailed overview of meta-learners in Appendix B.

End-to-end Neyman-orthogonality. In a special case, the end-to-end IPTW-learner for CAPOs, (e.g., Assaad et al., 2021) can possess Neyman-orthogonality. Yet, this IPTW-learner is a *special case* of a general DR-learner (Foster and Syrgkanis, 2023), where the target model and the nuisance models coincide (see Appendix B for details).

4 ORTHOGONAL REPRESENTATION LEARNING

To answer our two core research questions **RQ** ① and **RQ** ②, we first introduce our unified framework of representation learning methods and Neyman-orthogonal learners, called *OR-learners*. We provide proofs of theoretical statements in Appendix C.

OR-learners. The *OR-learners* proceed in three stages.[2] In stage ⓪, we fit a representation function $\hat{\Phi}$ that minimizes the plug-in/IPTW MSE from Eq. (1). Then, in stage ①, we fit additional nuisance functions $\hat{\eta}$ ($\hat{\pi}_a^x$ and, optionally, $\hat{\mu}_a^x$). Finally, in stage ②, we use the learned representations from stage ⓪ as inputs to the target model $\hat{g}(\phi)$ that minimizes a Neyman-orthogonal loss from Eq. (3). For **implementation details**, we refer to Appendix F.

"Best of both worlds". The OR-learners combine the benefits of two streams of literature. On the one hand, (i) the OR-learners are Neyman-orthogonal learners and thus inherit the same favorable asymptotic properties such as quasi-oracle efficiency and double robustness. On the other hand, they use (ii) representation learning twice (unlike (Curth and van der Schaar, 2021b)), namely, once for nuisance-function estimation and once for target-model fitting. As we will show later, under reasonable assumptions, the *OR-learners* outperform both (i) standard Neyman-orthogonal learners with the

[2]Code is available at `https://github.com/Valentyn1997/OR-learners`.

original covariates as the inputs to the target models and (ii) end-to-end representation learning.

4.1 RQ ①: When Can Representations Improve Neyman-orthogonality?

To compare the standard Neyman-orthogonal learners with $V = X$ and the *OR-learners* with $V = \Phi(X)$, we make use of the following lemma:

Lemma 1 (Quasi-oracle efficiency of a non-parametric model)**.** *Assume a non-parametric*[3] *target model* $g(v), v \in \mathcal{V} \subseteq \mathbb{R}^{d_v}$. *Then, the error between* $g^{*v} = \arg\min_{g \in \mathcal{G}} \mathcal{L}_{\mathcal{G}}(g, \eta)$ *and* $\hat{g}^v = \arg\min_{g \in \mathcal{G}} \hat{\mathcal{L}}_{\mathcal{G}}(g, \hat{\eta})$ *can be upper-bounded as:*

$$\|g^{*v} - \hat{g}^v\|_{L_2}^2 \lesssim (L^v)^{\frac{2d_v}{(2s^v+d_v)}} n^{-\frac{2s^v}{(2s^v+d_v)}} + R_2, \quad (4)$$

where g^{*v} *is* 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* n*,* $s^x_{\mu_a}, s^x_{\pi}$*, and* d_x*.*

We immediately see that we can reduce the error between $g^{*v}(v)$ and $\hat{g}^v(v)$ by (1) decreasing the dimensionality of the conditioning set d_v, (2) decreasing the Hölder norm L^v of the $g^{*v}(v)$, and (3) increasing the Hölder smoothness s^v.

Heterogeneity trade-off. In principle, all (1)-(3) can be optimized by choosing $V = \emptyset$, and, in this case, we recover a well-known *semi-parametric efficient* estimator of APOs/ATE. However, we then lose all the *heterogeneity* of the potential outcomes/treatment effect: Although the error $\|g^{*v} - \hat{g}^v\|_{L_2}^2$ gets smaller, the error between the ground-truth causal quantity and $\hat{g}$, $\|\xi^x_a - \hat{g}^v\|_{L_2}^2$ or $\|\tau^x - \hat{g}^v\|_{L_2}^2$, gets larger. Thus, we get a trade-off between *(i) reducing the error of the second-stage estimation* and *(ii) reducing the error with the ground-truth.*

Structure-agnostic learner. If we only care about the *(ii) reducing the error with the ground-truth*, the learner with $V = X$ is generally the best choice. That is, without any assumptions on the structure of covariates, $\hat{g}^v(v)$ provides a min-max optimal estimator of CAPOs/CATE (Balakrishnan et al., 2023; Jin and Syrgkanis, 2025). Specifically, in this case, we implicitly rely on the "worst-case" scenario that a ground-truth causal quantity *densely depends on the full covariate set* X. Yet, at the same time, we struggle with the curse of dimensionality.

Low-dimensional manifold hypothesis. The CAPOs/CATE estimation can be greatly improved if we depart from the "worst-case" scenario and assume a low-dimensional manifold hypothesis.

[3]Similar convergence rate can be shown for neural networks, see Sec. 5.2 in (Schulte et al., 2025).

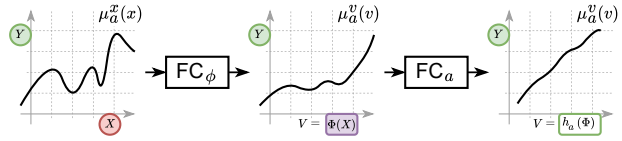

Figure 2: Hidden layers of the representation network induce spaces where the regression task is simpler.

Assumption 1 (Manifold hypothesis)**.** *We assume (i) the ground-truth causal quantities are supported on the low-dimensional compact smooth manifold (representation space)* $\Phi^* \subseteq \mathbb{R}^{d_{\phi^*}}, d_{\phi^*} \ll d_x$ *embedded into the covariate space:*

$$\xi^x_a(x) = \xi^{\phi^*}_a(\Phi^*(x)) \text{ and } \tau^x(x) = \tau^{\phi^*}(\Phi^*(x)), \quad (5)$$

where $\Phi^*(\cdot) : \mathcal{X} \to \Phi^*$ *is an* s^x*-Hölder smooth surjective embedding.*[4] *Furthermore, (ii) there exists a (not necessarily unique) pullback map* $J^* : \Phi^* \to \mathcal{X}$*, such that* $\Phi^* \circ J^* = \mathrm{id}_{\Phi^*}$*, and* J^* *is* $(s^x + 1)$*-Hölder smooth with a Hölder norm* L^{J^*}*.*

Assumption 1 (see a summary in Fig. 1) helps to improve the error bound in Eq. (4). Specifically, if we assume that the CAPOs/CATE $\tau^{\phi^*}/\xi^{\phi^*}_a \in \mathcal{G}$ (w.l.o.g.), then $g^{*\phi^*} \circ \Phi^* = g^{*x} = \xi^x_a/\tau^x$. However, the error bound gets lower when we learn $g^{*\phi^*}$ compared to g^{*x}.

Proposition 1. *Under Assumption 1, the following holds: (1)* $d_{\phi^*} \ll d_x$*, (2)* $g^{*\phi^*}$ *is an* s^{ϕ^*}*-Hölder smooth function with Hölder norm* L^{ϕ^*} *such that* $s^{\phi^*} \geq s^x$ *and* $L^{\phi^*} \leq c(L^{J^*}) \cdot L^x$ *with non-decreasing* $c(\cdot)$*. Also, when* L^{J^*} *is sufficiently small (i. e., a contractive map),*

$$||g^{*\phi^*} - \hat{g}^{\phi^*}||_{L_2}^2 \lesssim \|g^{*x} - \hat{g}^x\|_{L_2}^2 . \quad (6)$$

Note that in both cases ($V = X$ and $V = \Phi^*(X)$) the second-order remainder $R_2(\eta, \hat{\eta})$ is the same and depends on the dimensionality and the smoothness of the nuisance functions in the original space $\mathcal{X}$. Proposition 1 then answers our main **RQ** ①: Under Assumption 1, the *OR-learners* with the known representation function Φ^* outperform the standard Neyman-orthogonal learners that are based on $V = X$. Now, we ask two follow-up questions: (i) *Is Assumption 1 reasonable?* and (ii) *Can we learn* Φ^**?*

(i) Is Assumption 1 reasonable? This assumption has two main parts: (i) Eq. (5) and (ii) Hölder smoothness of the pullback J^*. (i) Eq. (5) describes so-called *valid representations* (Melnychuk et al., 2024) or, alternatively, *outcome mean sufficient representations* (Christgau and Hansen, 2024). It means that the low-dimensional representation Φ^* has to contain all the sufficient information to model the ground-truth CAPOs/CATE. A trivial example of a valid representation is $\Phi^* = (\mu^x_0(x), \mu^x_1(x)) \in \mathbb{R}^2$. The condition

[4]Note that we *do not assume* that the propensity score is supported on the same manifold.

(ii) additionally requires that the representation pullback J^* is Hölder smooth and, when L^{J^*} is sufficiently small, it contracts the representation space. For example, by properties of Hölder smooth functions (see Appendix B), when $L^{J^*} \leq 1/\sqrt{d_{\phi^*}}$, $\text{Lip}(J^*) \leq 1$. Alternatively, Assumption 1 states that, (i) when $d_{\phi^*} \ll d_x$, the representation projects or averages irrelevant dimensions of $\mathcal{X}$; and, (ii) when L^{J^*} is sufficiently small, the representation smoothens/expands the original covariate space $\mathcal{X}$ (i. e., when $s^x \geq 1$ and $L^{J^*} \leq 1/\sqrt{d_{\phi^*}}$, then $\text{Lip}(\Phi^*) \geq \text{Lip}(J^*)^{-1} \geq 1$).

We argue that the manifold hypothesis (Assumption 1) is a very flexible and, often, the most realistic assumption for CATE/CAPOs estimation among *different forms of structural knowledge* (such as additivity, sparsity, or linearity (Schulte et al., 2025)). For example, in the context of anti-cancer therapy, we might consider X-ray scans as high-dimensional covariates X. For them, it can be reasonable to assume that the ground-truth CATE/CAPOs lie in some low-dimensional manifold of the whole image space (e. g., a tumor might be fully characterized by several high-level covariates like shape, size, density, etc.).

(ii) Can we learn Φ^*? A natural question arises is whether we can learn the representation $\Phi^*(\cdot)$ from observational data $\mathcal{D}$. Interestingly, a result similar to Proposition 1 can be obtained by using a *neural representation* $\hat{\Phi}$ that minimizes a plug-in MSE in Eq. (1).

Proposition 2 (Smoothness of the hidden layers)**.** *We denote the trained representation network as $\hat{\mu}_a^x = \hat{h}_a \circ \hat{\Phi} = \arg\min \hat{\mathcal{L}}_{\Phi}(h_a \circ \Phi)$. Then, under mild conditions on the representation network, there exists a hidden layer $V = \hat{f}(X)$ where the regression target becomes smoother: $s^v_{\mu_a} \geq s^x_{\mu_a}$ and $L^v_{\mu_a} \leq L^x_{\mu_a}$.*

We illustrate Proposition 2 in Fig. 2. Importantly, Proposition 2 ensures that, with the plug-in loss, we can obtain a representation $\hat{\Phi}$ that simplifies learning μ_a^x (analogously to how Proposition 1 follows from Assumption 1) and, thus, *can serve as a substitute for the ideal* Φ^*. Yet, as mentioned previously, the plug-in loss is *not* Neyman-orthogonal and is thus sub-optimal for learning the causal quantities ξ_a^x/τ^x. Hence, by trying to debias the prediction based on $\hat{\Phi}$, we yield the *OR-learners*. In this way, we provide efficient learners of the representation-level causal quantities, namely $\tau^{\hat{\phi}}/\xi_a^{\hat{\phi}}$.

Alternative debiasing strategies. We might wonder whether we can use the Neyman-orthogonal losses differently. For example, we can **(a)** learn $\hat{\Phi}$ better (e. g., by using a Neyman-orthogonal loss in Eq. (3) with $g = h_a \circ \Phi$ and $V = X$), or **(b)** use the output of the representation network (the smoothest hidden layer according to Proposition 2) as the input for the target model ($V = (\hat{h}_0(\hat{\Phi}(X)), \hat{h}_1(\hat{\Phi}(X)))$).

(a) Re-learning the representation. As we will see later in our experiments, using the learned representation $\hat{\Phi}$ (as suggested by the *OR-learners*) *is more effective than learning the representation from scratch by using the Neyman-orthogonal losses* from Eq. (3). A possible reason for this is that Neyman-orthogonal losses (e. g., DR-learners) have larger variance and, thus, may fail to learn a low-dimensional representation well.

(b) Usage of the outputs. As another extreme, we can use the outputs of the representation network $V = (\hat{h}_0(\hat{\Phi}(X)), \hat{h}_1(\hat{\Phi}(X))) \in \mathbb{R}^2$ to fit the target model. Yet, in this case, the debiasing with the Neyman-orthogonal losses only *calibrates* the outputs of the representation network $\hat{h}_a \circ \hat{\Phi}$ and cannot compensate for larger errors in learning μ_a^x.

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.

4.2 RQ ②: Can a Balancing Constraint Substitute Neyman-orthogonality?

Balancing constraint. A balancing constraint was introduced in (Johansson et al., 2016; Shalit et al., 2017; Johansson et al., 2022) to reduce finite-sample estimation variance for the end-to-end representation learning methods. It then modifies the plug-in loss of Eq. (1):

$$\hat{\mathcal{L}}_{\text{Bal}(\Phi)}(h_a \circ \Phi) = \hat{\mathcal{L}}_{\Phi}(h_a \circ \Phi) + \alpha\, \hat{\mathcal{L}}_{\text{Bal}}(\Phi), \quad (7)$$

where $\alpha \geq 0$ is a balancing strength, and $\hat{\mathcal{L}}_{\text{Bal}}(\Phi) = \widehat{\text{dist}}(\mathbb{P}(\Phi(X) \mid A = 0), \mathbb{P}(\Phi(X) \mid A = 1))$ is an empirical probability metric (e. g., Wasserstein metric (WM) or maximum mean discrepancy (MMD)). The main intuition behind balancing is that it tries to construct a representation space $\hat{\Phi}$ in which both treatments are equally probable, namely, $\pi_0^{\hat{\phi}}(\phi) = \pi_1^{\hat{\phi}}(\phi)$. Yet, as we will show in the following, this strategy generally harms the CAPOs/CATE estimation.

Representation-induced confounding bias (RICB). As discovered in (Johansson et al., 2019; Melnychuk et al., 2024), setting α too high might lead to the RICB. In this case, the learned representation *stops being asymptotically valid* (= it does not contain the sufficient information to adjust for the covariates X):

$$\xi_a^{\hat{\phi}}(\phi) \neq \mu_a^{\hat{\phi}}(\phi) \text{ and } \tau^{\hat{\phi}}(\phi) \neq \mu_1^{\hat{\phi}}(\phi) - \mu_0^{\hat{\phi}}(\phi), \quad (8)$$

where $\hat{\Phi}$ is learned with the population version of Eq. 7. As a simple demonstration, we consider $\hat{\Phi}(x) = \text{const}$ that minimizes the loss in Eq. (7) when $\alpha \to \infty$. In this case, $\xi_a^{\hat{\phi}}(\phi)$ is the average potential outcome (APO), while $\mu_a^{\hat{\phi}}(\phi)$ is the mean outcome $\mathbb{E}(Y \mid A = a)$ (analogously $\tau^{\hat{\phi}}$ is the ATE, and $\mu_1^{\hat{\phi}}(\phi) - \mu_0^{\hat{\phi}}(\phi)$ is the difference in means).

Addressing RICB with OR-learners. Even if the learned representation $\hat{\Phi}$ contains the RICB, the *OR-learners* still yield the quasi-oracle efficient estimator of the representation-level causal quantities (see Lemma 1), as they have access to the *unconstrained* nuisance function estimators. Specifically, if we use DR-learners, we yield the augmented IPTW (A-IPTW) estimators of the APOs/ATE; and if we employ R-/IVW-learners, we get the A-IPTW estimators of the overlap-weighted ATE.

Omitting RICB with invertibility. As a remedy for the RICB, originally suggested by (Johansson et al., 2022), one can use *invertible representations* (e. g., $\Phi(X)$ can be implemented as a normalizing flow (Rezende and Mohamed, 2015)). In this case, we face a trade-off: we do not allow for the RICB, but also cannot benefit from Assumption 1.

RQ ②. A central question arises: Did it make sense to use the balancing constraint in the first place? In the following, we demonstrate that the balancing constraint relies on the *additional inductive bias*: low-overlap regions of the covariate space exhibit low CAPOs/CATE heterogeneity. To see that, we state two propositions.

First, as a consequence of Proposition 2, the representation subnetwork acts as an expanding mapping.

Proposition 3 (Smoothing via expanding mapping)**.** *Assume that the trained representation network $\hat{\Phi}$ minimizes $\hat{\mathcal{L}}_{\Phi}(h_a \circ \Phi)$ and is $s^{\hat{\Phi}}$-Hölder smooth ($s^{\hat{\Phi}} \geq 1$). Then, under mild conditions on $\hat{\mu}_a^x = \hat{h}_a \circ \hat{\Phi}$ and μ_a^x, (1) $\hat{\Phi}$ is an expanding mapping, namely,* $\text{Lip}(\hat{\Phi}) \geq 1$.

On the other hand, by trying to enforce the balancing constraint, we actually fit a *contracting mapping.*

Proposition 4 (Balancing via contracting mapping)**.** *Assume that the trained representation network $\hat{\Phi}$ minimizes $\hat{\mathcal{L}}_{\text{Bal}}(\Phi)$ with WM / MMD and is $s^{\hat{\Phi}}$-Hölder smooth ($s^{\hat{\Phi}} \geq 1$). Then, under mild conditions on $\hat{\Phi}$, (1) $\hat{\Phi}$ is a contracting mapping, namely* $\text{Lip}(\hat{\Phi}) \leq 1$. *Furthermore, if an analogue of Assumption 1 holds for $\hat{\Phi}$ with a pullback $\hat{J}$ (e. g., $\hat{\Phi}$ is smoothly invertible), (2) the pullback map is expanding, namely,* $\text{Lip}(\hat{J}) \geq 1$.

Interpretation. Hence, by minimizing the joint loss from Eq. (7), two things happen simultaneously. On the one hand, the plug-in loss $\hat{\mathcal{L}}_{\Phi}(h_a \circ \Phi)$ aims to expand the regions of the covariate space where μ_a^x (and thus CAPOs/CATE) are heterogeneous (to make the regression surface smoother). On the other hand, the balancing loss $\hat{\mathcal{L}}_{\text{Bal}}(\Phi)$ contracts the low-overlap regions of the covariate space (to minimize an empirical probability metric). Those considerations bring us to the following inductive bias.

"Low overlap – low heterogeneity" inductive bias. For the joint loss in Eq. (7) to perform well, we implicitly require that the regions of the covariate space with low CAPOs/CATE heterogeneity to *coincide* with the low-overlap regions. For example, instrumental variables $X_I \subseteq X$ induce no heterogeneity in CAPOs/CATE and, at the same time, create the low-overlap regions.

Inductive bias and OR-learners. The *OR-learners*, on the other hand, *do not require such an inductive bias to perform well*: they consider the low-overlap regions as *inherently uncertain*. In contrast, the DR-learners scale up the MSE risk in the low-overlap regions (as they rely on the IPTW weights); and R-/IVW-learners de-emphasize those regions and only fit CATE well in the overlapping parts of $\mathcal{X}$. Therefore, only *in special finite-sample cases* when the inductive bias is true, the end-to-end methods with the balancing constraint might yield a better estimator of CAPOs/CATE than the *OR-learners*. Yet, *in general*, the *OR-learners* are *asymptotically optimal* due to the Neyman-orthogonality. We sum up our findings in Fig. 5 of Appendix D.

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.

5 EXPERIMENTS

The primary aim of our numerical experiments is ***not*** standard benchmarking but *to validate the insights from* **RQ ①** and **RQ ②**, that is, when to use the OR-learners instead of the standard Neyman-orthogonal learners or instead of representations with balancing constraint.

Setup. We follow prior literature (Curth and van der Schaar, 2021b; Melnychuk et al., 2024) and use several (semi-)synthetic datasets where both counterfactual outcomes $Y[0]$ and $Y[1]$ and ground-truth covariate-level CAPOs / CATE are available. We perform experiments in two settings that correspond to each research question. • In **Setting ①**, we compare different *OR-learners* based on different target model inputs (i. e., original covariates, pre-trained representations, or the

Table 1: **Results for 77 ACIC 2016 datasets in Setting ①.** Reported: the % of runs, where the *OR-learners* improve over plug-in representation networks wrt. out-of-sample rMSE / rPEHE. Here, $d_{\hat{\phi}} = 8$.

		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

outputs of the pre-trained representation network). • In **Setting ②**, we show when the *OR-learners* improve the representation networks trained with the balancing constraint.

Datasets. We used three standard datasets for benchmarking in causal inference: (1) a fully-synthetic dataset ($d_x = 2$) (Kallus et al., 2019; Melnychuk et al., 2024); (2) the IHDP dataset ($n = 747; d_x = 25$) (Hill, 2011; Shalit et al., 2017); (3) a collection of 77 ACIC 2016 datasets ($n = 4802, d_x = 82$) (Dorie et al., 2019); and (4) a high-dimensional HC-MNIST dataset (Jesson et al., 2021) ($n = 70000, d_x = 785$). Details are in Appendix E. All the datasets (1)-(4) then help us to empirically study **RQ ①** and **RQ ②**. Specifically, for **RQ ①**, we know a priori that the manifold hypothesis *does not hold* for (1) the fully-synthetic dataset; *is believed to hold* for (3) ACIC 2016 datasets; and *definitely holds* for (2) the IHDP dataset (i. e., CAPOs/CATE are defined on the linear combinations of the covariates) and for (4) the HC-MNIST dataset (i. e., CAPOs/CATE depend on a two-dimensional latent manifold that encodes an image's mean pixel intensity and digit label). Furthermore, for **RQ ②**, we know that the "low overlap – low heterogeneity" inductive bias can only be assumed for (2) the IHDP dataset.

Performance metrics. We report (i) the out-of-sample root mean squared error (rMSE) and (ii) the root precision in estimating heterogeneous effect (rPEHE) for CAPOs and CATE, respectively. Recall that, in **RQ ①** and **RQ ②**, we are primarily interested in how the *OR-learners* improve the existing methods, and, therefore, we report the difference in the performance between the baseline representation learning method and the *OR-learners*. Formally, we compute Δ(rMSE) for CAPOs and Δ(rPEHE) for CATE for different variants of the *OR-learners*: DR-learner in the style of Kennedy (2023) ($\mathbf{DR}_a^{\mathbf{K}}$) and DR-learner in the style of Foster and Syrgkanis (2023) ($\mathbf{DR}_a^{\mathbf{FS}}$) for CAPOs; and DR-learner (Kennedy, 2023) ($\mathbf{DR}^{\mathbf{K}}$) / R-learner (Nie and Wager, 2021) (**R**) / IVW-learner (Fisher, 2024) (**IVW**) for CATE. Furthermore, we followed best benchmarking practices for CAPOs/CATE estimation (Curth et al., 2021). Namely, we often report robust performance metrics (i. e., medians/percentage of best runs) and always compare baselines with a similar structure of the nuisance functions (e. g., S-learners vs. S-learner-based DR/R-learners).

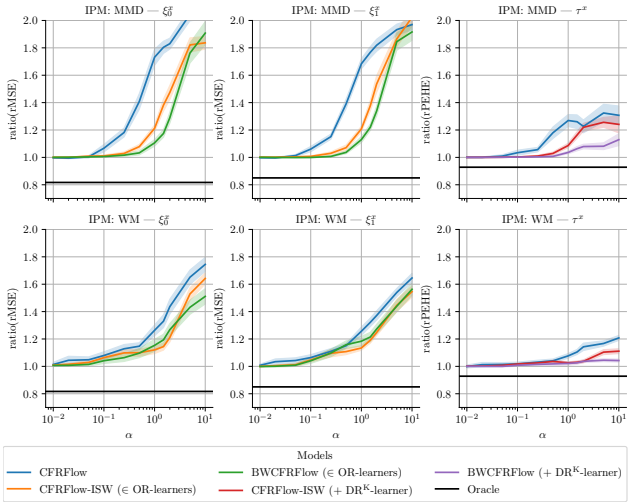

Figure 3: **Results for synthetic data in Setting ②.** Reported: ratio between the performance of TARFlow (CFRFlow with $\alpha = 0$) and invertible representation networks with varying α; mean ± SE over 15 runs. Lower is better. Here: $n_{\text{train}} = 500$, $d_{\hat{\phi}} = 2$.

Baselines. We implemented various state-of-the-art representation learning methods and combine each baseline with the *OR-learners* (see Appendix F): **TARNet** (Shalit et al., 2017); several variants of **BNN** (Johansson et al., 2016) (w/ or w/o balancing); several variants of **CFR** (Shalit et al., 2017; Johansson et al., 2022) (w/ balancing, non-/ invertible); several variants of **RCFR** (Johansson et al., 2018, 2022) (different types of balancing); several variants of **CFR-ISW** (Hassanpour and Greiner, 2019a) (w/ or w/o balancing, non-/ invertible); and **BWCFR** (Assaad et al., 2021) (w/ or w/o balancing, non-/invertible).

■ **Setting ①.** In Setting ①, we want to confirm our theoretical insights for the manifold hypothesis (Assumption 1) by comparing the performance of vanilla representation networks (i. e., TARNet and BNN ($\alpha = 0.0$)) versus the *OR-learners* applied on top of the learned unconstrained representations, where the latter is denoted $V = \hat{\Phi}(X)$. We compare two further variants of the *OR-learners*, where the target network has different inputs: (a) $V = (\hat{h}_0(\hat{\Phi}(X)), \hat{h}_1(\hat{\Phi}(X))) = (\hat{\mu}_0^x, \hat{\mu}_1^x)$, and (b) $V = X$, yet the same depth of one hidden layer. We also compare the *OR-learners* with (c) the target network, which matches the depth of the original representation network $V = X^*$. Therefore, (b) and (c) both provide a fair comparison of the *OR-learners* and the standard Neyman-orthogonal learners with $V = X$.

Results. Tables 1 and 2 show the results for the ACIC 2016 datasets and the HC-MNIST dataset, where, due to high-dimensionality, it is reasonable to assume the low-dimensional manifold hypothesis. We find that the *OR-learners* outperform the baseline representation learning networks due to their Neyman-orthogonality.

Table 2: **Results for HC-MNIST experiments in Setting ①.** Reported: improvements of the *OR-learners* over plug-in representation networks wrt. out-of-sample rMSE / rPEHE; mean ± std over 10 runs. Here, $d_{\hat{\phi}} = 78$.

		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

Table 3: **Results for IHDP experiments in Setting ①.** Reported: out-of-sample rMSE / rPEHE for different causal quantities (ξ_a^x/τ^x, respectively), median ± std over 100 train/test splits. Here, $d_{\hat{\phi}} = 12$ for neural baselines.

		ξ_0^x	ξ_1^x	τ^x
XGBoost S	DR^{K}	0.496 ± 0.118	0.723 ± 0.241	0.826 ± 0.239
	R	—	—	1.631 ± 0.260
	IVW	—	—	0.820 ± 0.247
XGBoost T	DR^{K}	0.501 ± 0.117	0.531 ± 0.283	0.754 ± 0.240
	R	—	—	1.762 ± 0.258
	IVW	—	—	0.749 ± 0.242
TARNet	Plug-in	0.367 ± 0.160	**0.379 ± 0.226**	**0.518 ± 0.270**
	DR^{K}	<u>0.366 ± 0.168</u>	0.390 ± 0.228	<u>0.523 ± 0.280</u>
	DR^{FS}	**0.364 ± 0.169**	0.421 ± 0.247	—
	R	—	—	0.563 ± 0.295
	IVW	—	—	0.530 ± 0.285
BNN ($\alpha = 0.0$)	Plug-in	0.386 ± 0.157	0.478 ± 0.140	0.595 ± 0.168
	DR^{K}	0.379 ± 0.168	0.465 ± 0.160	0.568 ± 0.180
	DR^{FS}	0.376 ± 0.169	<u>0.382 ± 0.213</u>	—
	R	—	—	0.543 ± 0.198
	IVW	—	—	0.568 ± 0.186
Oracle		0.303	0.315	0.434

Lower = better. Best in **bold**, second best <u>underlined</u>

Furthermore, the *OR-learners* with $V = \hat{\Phi}(X)$ outperform the standard Neyman-orthogonal learners (namely, with $V = X/X^*$) for CAPOs/CATE estimation. This confirms that *using the pre-trained representation for the target model input $V = \hat{\Phi}(X)$ is more effective than training the representation network from scratch at the second stage*. Furthermore, Table 3 shows the results for the IHDP dataset. Here, we report the absolute performance of different methods: non-neural Neyman-orthogonal learners instantiated with XGBoost (Chen et al., 2015) (with S-/T-learners for the first-stage models); plug-in representation learning methods; and the OR-learners used with the pre-trained representations $V = \hat{\Phi}(X)$. We see that the OR-learners improve over other non-neural Neyman-orthogonal learners: This was expected as *the potential outcomes for the IHDP dataset are defined via the low-dimensional manifold of the covariate space*. In Appendix G, we also provide **additional results** for (i) the synthetic dataset (where the low-dimensional manifold hypothesis cannot be assumed) and (ii) the HC-MNIST dataset (where we compare the *OR-learners* with the non-neural Neyman-orthogonal learners). This confirms our theory: (i) as expected, both $V = \hat{\Phi}(X)$ and $V = X/X^*$ perform similarly, and (ii) *OR-learners* outperform non-neural learners.

■ **Setting ②.** Here, we want to verify our finding that the balancing constraint only helps when the "low overlap – low heterogeneity" inductive bias can be assumed. For that, we study how the *OR-learners* compare with the representation networks trained with the balancing constraint and varying amounts of balancing strength α. We consider both invertible (TARFlow, CFRFlow, CFRFlow-ISW, and BWCFRFlow) and non-invertible (CFR, BNN, RCFR, CFR-ISW, and BWCFR) representation networks.

Table 4: **Results for 77 semi-synthetic ACIC 2016 experiments in Setting ②.** Reported: the % of runs, where the *OR-learners* improve over non-invertible plug-in/IPTW representation networks wrt. out-of-sample rMSE / rPEHE. Here, $d_{\hat{\phi}} = 8$.

	DR_0^{K}	DR_0^{FS}	DR_1^{K}	DR_1^{FS}	DR^{K}	R	IVW
CFR (MMD; $\alpha = 0.1$)	67.6%	60.3%	67.1%	62.9%	72.5%	66.8%	69.0%
CFR (WM; $\alpha = 0.1$)	73.2%	68.3%	72.8%	69.0%	74.6%	74.5%	75.2%
BNN (MMD; $\alpha = 0.1$)	74.0%	79.0%	66.3%	66.4%	70.6%	70.3%	68.7%
BNN (WM; $\alpha = 0.1$)	74.6%	80.5%	70.4%	73.2%	75.8%	77.5%	76.6%
RCFR (MMD; $\alpha = 0.1$)	78.8%	71.6%	75.1%	71.7%	72.4%	70.7%	74.5%
RCFR (WM; $\alpha = 0.1$)	80.2%	75.7%	77.5%	76.3%	71.0%	73.5%	75.3%
CFR-ISW (MMD; $\alpha = 0.1$)	70.8%	63.4%	69.2%	65.7%	67.2%	64.2%	69.8%
CFR-ISW (WM; $\alpha = 0.1$)	76.5%	69.6%	71.6%	71.3%	70.7%	73.7%	77.0%
BWCFR (MMD; $\alpha = 0.1$)	69.5%	66.7%	66.3%	63.9%	70.5%	68.4%	68.7%
BWCFR (WM; $\alpha = 0.1$)	73.4%	73.9%	73.2%	72.7%	71.5%	72.1%	71.7%

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

<u>Results.</u> Fig. 3 and Table 4 show the results for the synthetic data and the ACIC 2016 datasets. In both cases, the *OR-learners* manage to improve the representation networks that use balancing constraints (as, in general, the "low overlap – low heterogeneity" inductive bias cannot be assumed for these datasets). We also refer to Appendix G for **additional results** for (i) the synthetic and (ii) IHDP datasets. For (i), by varying the size of data $n_{\text{train}} \in \{250, 1000\}$, we show a similar pattern to Fig. 5 of Appendix D. Also, we visualize the expanding/contracting mappings, suggested by Propositions 3-4. For (ii), the balancing constraint helps the CAPOs/CATE estimation, as the underlying inductive bias is present in the IHDP dataset.

Limitations. Applications of our *OR-learners* should follow a cautious approach: We rely on a crucial manifold hypothesis (Assumption 1) that has to be supported by some *background knowledge* about the ground-truth causal quantity.

Takeaways. Our experiments confirm our theoretical findings for **RQ ①** and **RQ ②**. In general, there is no nuisance-free way to do CATE/CAPOs model selection based solely on the observational data (Curth and van der Schaar, 2023). However, ① given Assumption 1, one can *simplify* the task of CAPOs/CATE estimation and use the suggested framework of *OR-learners*. Similarly, ② we advise *against* the balancing constraint, unless one can assume the underlying inductive bias.

Acknowledgments. This paper is supported by the DAAD program "Konrad Zuse Schools of Excellence in Artificial Intelligence", sponsored by the Federal Ministry of Education and Research. S.F. acknowledges funding via Swiss National Science Foundation Grant 186932. This work has been supported by the German Federal Ministry of Education and Research (Grant: 01IS24082).

References

Joseph Antonelli, Matthew Cefalu, Nathan Palmer, and Denis Agniel. Doubly robust matching estimators for high dimensional confounding adjustment. *Biometrics*, 74(4):1171–1179, 2018.

Serge Assaad, Shuxi Zeng, Chenyang Tao, Shounak Datta, Nikhil Mehta, Ricardo Henao, Fan Li, and Lawrence Carin. Counterfactual representation learning with balancing weights. In *International Conference on Artificial Intelligence and Statistics*, 2021.

Onur Atan, William R. Zame, and Mihaela van der Schaar. Counterfactual policy optimization using domain-adversarial neural networks. 2018.

Sivaraman Balakrishnan, Edward Kennedy, and Larry Wasserman. The fundamental limits of structure-agnostic functional estimation. In *International Conference on Statistics and Data Science*, 2023.

Anirban Basu, Daniel Polsky, and Willard G. Manning. Estimating treatment effects on healthcare costs under exogeneity: is there a 'magic bullet'? *Health Services and Outcomes Research Methodology*, 11: 1–26, 2011.

Ioana Bica, Ahmed M. Alaa, James Jordon, and Mihaela van der Schaar. Estimating counterfactual treatment outcomes over time through adversarially balanced representations. In *International Conference on Learning Representations*, 2020.

Vinod K. Chauhan, Soheila Molaei, Marzia Hoque Tania, Anshul Thakur, Tingting Zhu, and David A. Clifton. Adversarial de-confounding in individualised treatment effects estimation. In *International Conference on Artificial Intelligence and Statistics*, 2023.

Ricky T.Q. Chen, Jens Behrmann, David K. Duvenaud, and Jörn-Henrik Jacobsen. Residual flows for invertible generative modeling. In *Advances in Neural Information Processing Systems*, 2019.

Tianqi Chen, Tong He, Michael Benesty, Vadim Khotilovich, Yuan Tang, Hyunsu Cho, Kailong Chen, Rory Mitchell, Ignacio Cano, Tianyi Zhou, et al. XGBoost: Extreme gradient boosting. *R package version 0.4-2*, 1(4):1–4, 2015.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, and Whitney Newey. Double/debiased/Neyman machine learning of treatment effects. *American Economic Review*, 107(5):261–265, 2017.

Alexander Mangulad Christgau and Niels Richard Hansen. Efficient adjustment for complex covariates: Gaining efficiency with DOPE. *arXiv preprint arXiv:2402.12980*, 2024.

Amanda Coston, Edward Kennedy, and Alexandra Chouldechova. Counterfactual predictions under runtime confounding. *Advances in Neural Information Processing Systems*, 2020.

Daniel Csillag, Claudio Jose Struchiner, and Guilherme Tegoni Goedert. Generalization bounds for causal regression: Insights, guarantees and sensitivity analysis. In *International Conference on Machine Learning*, 2024.

Alicia Curth and Mihaela van der Schaar. On inductive biases for heterogeneous treatment effect estimation. *Advances in Neural Information Processing Systems*, 2021a.

Alicia Curth and Mihaela van der Schaar. Nonparametric estimation of heterogeneous treatment effects: From theory to learning algorithms. In *International Conference on Artificial Intelligence and Statistics*, 2021b.

Alicia Curth and Mihaela van der Schaar. In search of insights, not magic bullets: Towards demystification of the model selection dilemma in heterogeneous treatment effect estimation. In *International Conference on Machine Learning*, 2023.

Alicia Curth, Ahmed M. Alaa, and Mihaela van der Schaar. Estimating structural target functions using machine learning and influence functions. *arXiv preprint arXiv:2008.06461*, 2020.

Alicia Curth, David Svensson, Jim Weatherall, and Mihaela van der Schaar. Really doing great at estimating CATE? A critical look at ML benchmarking practices in treatment effect estimation. *arXiv preprint arXiv:2107.13346*, 2021.

Alexander D'Amour and Alexander Franks. Deconfounding scores: Feature representations for causal effect estimation with weak overlap. *arXiv preprint arXiv:2104.05762*, 2021.

Rafael De La Llave and Rafael Obaya. Regularity of the composition operator in spaces of Hölder functions. *Discrete and Continuous Dynamical Systems*, 5:157–184, 1999.

Vincent Dorie, Jennifer Hill, Uri Shalit, Marc Scott, and Dan Cervone. Automated versus do-it-yourself methods for causal inference: Lessons learned from a data analysis competition. *Statistical Science*, 34 (1):43–68, 2019.

Xin Du, Lei Sun, Wouter Duivesteijn, Alexander Nikolaev, and Mykola Pechenizkiy. Adversarial balancing-based representation learning for causal effect inference with observational data. *Data Mining and Knowledge Discovery*, 35(4):1713–1738, 2021.

Charles Fefferman, Sanjoy Mitter, and Hariharan Narayanan. Testing the manifold hypothesis. *Journal of the American Mathematical Society*, 29(4): 983–1049, 2016.

Stefan Feuerriegel, Dennis Frauen, Valentyn Melnychuk, Jonas Schweisthal, Konstantin Hess, Alicia Curth, Stefan Bauer, Niki Kilbertus, Isaac S. Kohane, and Mihaela van der Schaar. Causal machine learning for predicting treatment outcomes. *Nature Medicine*, 2024.

Aaron Fisher. Inverse-variance weighting for estimation of heterogeneous treatment effects. In *International Conference on Machine Learning*, 2024.

Dylan J. Foster and Vasilis Syrgkanis. Orthogonal statistical learning. *The Annals of Statistics*, 51(3): 879–908, 2023.

Dennis Frauen, Valentyn Melnychuk, and Stefan Feuerriegel. Fair off-policy learning from observational data. In *International Conference on Machine Learning*, 2024.

Dennis Frauen, Konstantin Hess, and Stefan Feuerriegel. Model-agnostic meta-learners for estimating heterogeneous treatment effects over time. In *International Conference on Learning Representations*, 2025.

Xingzhuo Guo, Yuchen Zhang, Jianmin Wang, and Mingsheng Long. Estimating heterogeneous treatment effects: Mutual information bounds and learning algorithms. In *International Conference on Machine Learning*, 2023.

Ben B. Hansen. The prognostic analogue of the propensity score. *Biometrika*, 95(2):481–488, 2008.

Negar Hassanpour and Russell Greiner. CounterFactual regression with importance sampling weights. In *International Joint Conference on Artificial Intelligence*, 2019a.

Negar Hassanpour and Russell Greiner. Learning disentangled representations for counterfactual regression. In *International Conference on Learning Representations*, 2019b.

Konstantin Hess, Valentyn Melnychuk, Dennis Frauen, and Stefan Feuerriegel. Bayesian neural controlled differential equations for treatment effect estimation. In *International Conference on Learning Representations*, 2024.

Jennifer L. Hill. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240, 2011.

Ming-Yueh Huang and Kwun Chuen Gary Chan. Joint sufficient dimension reduction and estimation of conditional and average treatment effects. *Biometrika*, 104(3):583–596, 2017.

Yiyan Huang, Cheuk Hang Leung, Siyi Wang, Yijun Li, and Qi Wu. Unveiling the potential of robustness in evaluating causal inference models. In *Advances in Neural Information Processing Systems*, 2024.

Andrew Jesson, Sören Mindermann, Yarin Gal, and Uri Shalit. Quantifying ignorance in individual-level causal-effect estimates under hidden confounding. In *International Conference on Machine Learning*, 2021.

Jikai Jin and Vasilis Syrgkanis. Structure-agnostic optimality of doubly robust learning for treatment effect estimation. In *Annual Conference on Learning Theory*, 2025.

Fredrik D. Johansson, Uri Shalit, and David Sontag. Learning representations for counterfactual inference. In *International Conference on Machine Learning*, 2016.

Fredrik D. Johansson, Nathan Kallus, Uri Shalit, and David Sontag. Learning weighted representations for generalization across designs. *arXiv preprint arXiv:1802.08598*, 2018.

Fredrik D. Johansson, David Sontag, and Rajesh Ranganath. Support and invertibility in domain-invariant representations. In *International Conference on Artificial Intelligence and Statistics*, 2019.

Fredrik D. Johansson, Uri Shalit, Nathan Kallus, and David Sontag. Generalization bounds and representation learning for estimation of potential outcomes and causal effects. *Journal of Machine Learning Research*, 23:7489–7538, 2022.

Nathan Kallus, Xiaojie Mao, and Angela Zhou. Interval estimation of individual-level causal effects under unobserved confounding. In *International Conference on Artificial Intelligence and Statistics*, 2019.

Edward H. Kennedy. Towards optimal doubly robust estimation of heterogeneous causal effects. *Electronic Journal of Statistics*, 17(2):3008–3049, 2023.

Kwangho Kim and José R Zubizarreta. Fair and robust estimation of heterogeneous treatment effects for policy learning. In *International Conference on Machine Learning*, 2023.

Sören R. Künzel, Jasjeet S. Sekhon, Peter J. Bickel, and Bin Yu. Meta-learners for estimating heterogeneous treatment effects using machine learning. *Proceedings*

of the National Academy of Sciences, 116(10):4156–4165, 2019.

Milan Kuzmanovic, Dennis Frauen, Tobias Hatt, and Stefan Feuerriegel. Causal machine learning for cost-effective allocation of development aid. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2024.

Yann LeCun. The MNIST database of handwritten digits. *http://yann.lecun.com/exdb/mnist/*, 1998.

Qiao Liu, Zhongren Chen, and Wing Hung Wong. An encoding generative modeling approach to dimension reduction and covariate adjustment in causal inference with observational studies. *Proceedings of the National Academy of Sciences*, 121(23):e2322376121, 2024.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Wei Luo and Yeying Zhu. Matching using sufficient dimension reduction for causal inference. *Journal of Business & Economic Statistics*, 38(4):888–900, 2020.

David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. Learning adversarially fair and transferable representations. In *International Conference on Machine Learning*, 2018.

Valentyn Melnychuk, Dennis Frauen, and Stefan Feuerriegel. Causal transformer for estimating counterfactual outcomes. In *International Conference on Machine Learning*, 2022.

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

Pawel Morzywolek, Johan Decruyenaere, and Stijn Vansteelandt. On a general class of orthogonal learners for the estimation of heterogeneous treatment effects. *arXiv preprint arXiv:2303.12687*, 2023.

Xinkun Nie and Stefan Wager. Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika*, 108: 299–319, 2021.

Kenneth R. Niswander. The collaborative perinatal study of the National Institute of Neurological Diseases and Stroke. *The Woman and Their Pregnancies*, 1972.

Boris T. Polyak and Anatoli B. Juditsky. Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, 30(4):838–855, 1992.

Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *International Conference on Machine Learning*, 2015.

James M. Robins and Andrea Rotnitzky. Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association*, 90(429):122–129, 1995.

Paul R. Rosenbaum and Donald B. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

Donald B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688, 1974.

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

Patrick Schwab, Lorenz Linhardt, and Walter Karlen. Perfect match: A simple method for learning representations for counterfactual inference with neural networks. *arXiv preprint arXiv:1810.00656*, 2018.

Uri Shalit, Fredrik D. Johansson, and David Sontag. Estimating individual treatment effect: Generalization bounds and algorithms. In *International Conference on Machine Learning*, 2017.

Claudia Shi, David Blei, and Victor Veitch. Adapting neural networks for the estimation of treatment effects. *Advances in Neural Information Processing Systems*, 2019.

Charles J. Stone. Optimal global rates of convergence for nonparametric regression. *The Annals of Statistics*, pages 1040–1053, 1982.

Lars van der Laan, Marco Carone, and Alex Luedtke. Combining T-learning and DR-learning: a framework for oracle-efficient estimation of causal contrasts. *arXiv preprint arXiv:2402.01972*, 2024.

Mark J. van der Laan, Sherri Rose, et al. *Targeted learning: causal inference for observational and experimental data*, volume 4. Springer, 2011.

Stijn Vansteelandt and Paweł Morzywołek. Orthogonal prediction of counterfactual outcomes. *Journal of Causal Inference*, 13(1):20240051, 2025.

Hal R. Varian. Causal inference in economics and marketing. *Proceedings of the National Academy of Sciences*, 113(27):7310–7315, 2016.

Hao Wang, Jiajun Fan, Zhichao Chen, Haoxuan Li, Weiming Liu, Tianqiao Liu, Quanyu Dai, Yichao Wang, Zhenhua Dong, and Ruiming Tang. Optimal transport for treatment effect estimation. *Advances in Neural Information Processing Systems*, 2024.

Anpeng Wu, Junkun Yuan, Kun Kuang, Bo Li, Runze Wu, Qiang Zhu, Yueting Zhuang, and Fei Wu. Learning decomposed representations for treatment effect estimation. *IEEE Transactions on Knowledge and Data Engineering*, 35(5):4989–5001, 2022.

Anpeng Wu, Kun Kuang, Ruoxuan Xiong, Bo Li, and Fei Wu. Stable estimation of heterogeneous treatment effects. In *International Conference on Machine Learning*, 2023.

Yuguang Yan, Zongyu Li, Haolin Yang, Zeqin Yang, Hao Zhou, Ruichu Cai, and Zhifeng Hao. Reducing confounding bias without data splitting for causal inference via optimal transport. In *International Conference on Machine Learning*, 2025.

Hao Yang, Zexu Sun, Hongteng Xu, and Xu Chen. Revisiting counterfactual regression through the lens of Gromov-Wasserstein information bottleneck. *arXiv preprint arXiv:2405.15505*, 2024.

Liuyi Yao, Sheng Li, Yaliang Li, Mengdi Huai, Jing Gao, and Aidong Zhang. Representation learning for treatment effect estimation from observational data. *Advances in Neural Information Processing Systems*, 2018.

Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In *International Conference on Machine Learning*, 2013.

Yao Zhang, Alexis Bellot, and Mihaela van der Schaar. Learning overlapping representations for the estimation of individualized treatment effects. In *International Conference on Artificial Intelligence and Statistics*, 2020.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes] (see Sec. 3 and Appendix F)
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes] (see Appendix G)
   (c) (Optional) Source code, with specification of all dependencies, including external libraries. [Yes] (Code is available at `https://github.com/Valentyn1997/OR-learners`.)

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes] (See Sec. 3 and 4)
   (b) Complete proofs of all theoretical results. [Yes] (See Appendix C)
   (c) Clear explanations of any assumptions. [Yes] (See Sec. 4)

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes] (Code is available at `https://github.com/Valentyn1997/OR-learners`.)
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes] (See Appendix F)
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes] (In the caption of figures and tables)
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes] (see Appendix G)

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator, if your work uses existing assets. [Yes] (We used publicly available synthetic and semi-synthetic datasets).
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Orthogonal Representation Learning for Estimating Causal Quantities: Appendix

A EXTENDED RELATED WORK

Our work aims to unify two streams of work, namely, end-to-end representation learning methods (Sec. A.1) and two-stage meta-learners (Sec. A.2). We review both in the following and then discuss the implications for our work.

A.1 End-to-end Representation Learning Methods

Several methods have been previously introduced for *end-to-end* representation learning of CAPOs/CATE (see, in particular, the seminal works by Johansson et al., 2016; Shalit et al., 2017; Johansson et al., 2022). Existing methods fall into three main streams: (1) One can fit an *unconstrained shared representation* to directly estimate both potential outcome surfaces (e.g., **TARNet**; Shalit et al., 2017). (2) Some methods additionally enforce a *balancing constraint based on empirical probability metrics*, so that the distributions of the treated and untreated representations become similar (e.g., **CFR** and **BNN**; Johansson et al., 2016; Shalit et al., 2017). Importantly, the balancing constraint is only guaranteed to perform a consistent estimation for *invertible* representations since, otherwise, balancing leads to a *representation-induced confounding bias* (RICB) (Johansson et al., 2019; Melnychuk et al., 2024). Finally, (3) one can additionally perform *balancing by re-weighting* the loss and the distributions of the representations with learnable weights (e.g., **RCFR**; Johansson et al., 2022).

Table 5 provides a summary of the main representation learning methods for the estimation of causal quantities. Therein, we showed (1) how different constraints imposed on the representations relate to the consistency of estimation and (2) Neyman-orthogonality of the underlying methods. We highlight several important constrained representations below and discuss the implications for estimating causal quantities.

Balancing constraint and invertibility. Following CFR and BNN, several works proposed alternative strategies for implementing the balancing constraints, e. g., based on adversarial learning (Atan et al., 2018; Curth and van der Schaar, 2021a; Du et al., 2021; Melnychuk et al., 2022; Guo et al., 2023); metric learning (Yao et al., 2018); counterfactual variance minimization (Zhang et al., 2020); and empirical mutual information (Yang et al., 2024). To enforce *invertibility* (and, thus, consistency of estimation), several works suggested metric learning heuristics (Yao et al., 2018) or reconstruction loss (Zhang et al., 2020).

Balancing by re-weighting. Other methods extended *balancing by re-weighting*, as in RCFR but, for example, with weights based on matching (Schwab et al., 2018; Wu et al., 2023); or with inverse probability of treatment weights (IPTW) (Hassanpour and Greiner, 2019a,b; Assaad et al., 2021; Wu et al., 2022). Importantly, balancing by re-weighting on itself does not harm the consistency of the estimation and only changes the type of the underlying meta-learner (i. e., weighted plug-in or IPTW).

Validity of representations for consistent estimation. As mentioned previously, balancing representations with empirical probability metrics without strictly enforcing invertibility generally leads to *inconsistent estimation based on representations*. This issue was termed as a *representation-induced adaptation error* (Johansson et al., 2019) in the context of unsupervised domain adaptation and as a *representation-induced confounding bias (RICB)* (Melnychuk et al., 2024) in the context of estimation of causal quantities. More generally, the RICB can be recognized as a type of runtime confounding (Coston et al., 2020), i. e., when only a subset of covariates is available for the estimation of the causal quantities. Several works offered solutions to circumvent the RICB and achieve consistency. For example, Assaad et al. (2021) employed IPTW based on original covariates, and Melnychuk et al. (2024) used a sensitivity model to perform a partial identification. However, to the best of our knowledge, <u>no</u> Neyman-orthogonal method was proposed to resolve the RICB (see Fig. 4).

Balancing and finite-sample generalization error. In the original works on balancing representations

Table 5: Overview of representation learning methods for CAPOs/CATE estimation. Here, parentheses imply the possibility of an extension.

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	✓	✗	✗
OR-learners (our paper)	DR / R / IVW	(any)	NFs / –	✓	✓$^{\text{DR}^{\text{FS}}}$, ✓$^{\text{DR}^{\text{K}}}$	✓$^{\text{DR}^{\text{K}}}$, ✓^{R}, ✓^{IVW}

Legend:
- Learner type: plug-in (PI); weighted plug-in (WPI); inverse probability of treatment weighted (IPTW); doubly-robust (DR); Robinson's / residualized (R)
- Balancing: integral probability metric (IPM); learnable weights (LW); Jensen-Shannon divergence (JSD); local similarity (LS); upsampling via matching (UVM); representation propensity (RP); covariate propensity (CP); counterfactual variance (CV); mutual information (MI)
- Invertibility: middle point distance (MPD); reconstruction loss (RL); normalizing flows (NFs); GAN-based loss (GAN)
- Neyman-orthogonality: DR-learner in the style of Kennedy (2023) (DR^{K}); DR-learner in the style of Foster and Syrgkanis (2023) (DR^{FS})

(Shalit et al., 2017; Johansson et al., 2022), the authors provided finite-sample generalization error bounds for any estimator of CAPOs/CATE based on a factual estimation error and a distributional distance between treated and untreated populations. Therein, the authors employed integral probability metrics as the distributional distance. These bounds were further improved with other distributional distances, e. g., counterfactual variance (Zhang et al., 2020), χ^2-divergence (Csillag et al., 2024), and KL-divergence (Huang et al., 2024). However, the work by (Shalit et al., 2017; Johansson et al., 2022) suggests that the large distributional distance only *acknowledges the lack of overlap between treated and untreated covariates* (and, hence, the hardness of the estimation) but it *does not instruct how much balancing needs to be applied*. Moreover, the finite-sample generalization error bounds *do not instruct* how to design Neyman-orthogonal learners, and, thus, they are not relevant to our work.

Note on non-neural representations. Multiple works also explored the use of non-neural representations for the estimation of causal quantities, also known under the umbrella term of *scores*. Examples include propensity/balancing scores (Rosenbaum and Rubin, 1983; Antonelli et al., 2018), prognostic scores (Hansen, 2008; Huang and Chan, 2017; Luo and Zhu, 2020; Antonelli et al., 2018; D'Amour and Franks, 2021), and deconfounding scores (D'Amour and Franks, 2021). However, we want to highlight that these works focus on <u>different</u>, rather simpler than our settings:

- *Propensity, balancing, and deconfounding scores* (Rosenbaum and Rubin, 1983) were employed to estimate *average* causal quantities (Antonelli et al., 2018; D'Amour and Franks, 2021). Examples are average potential

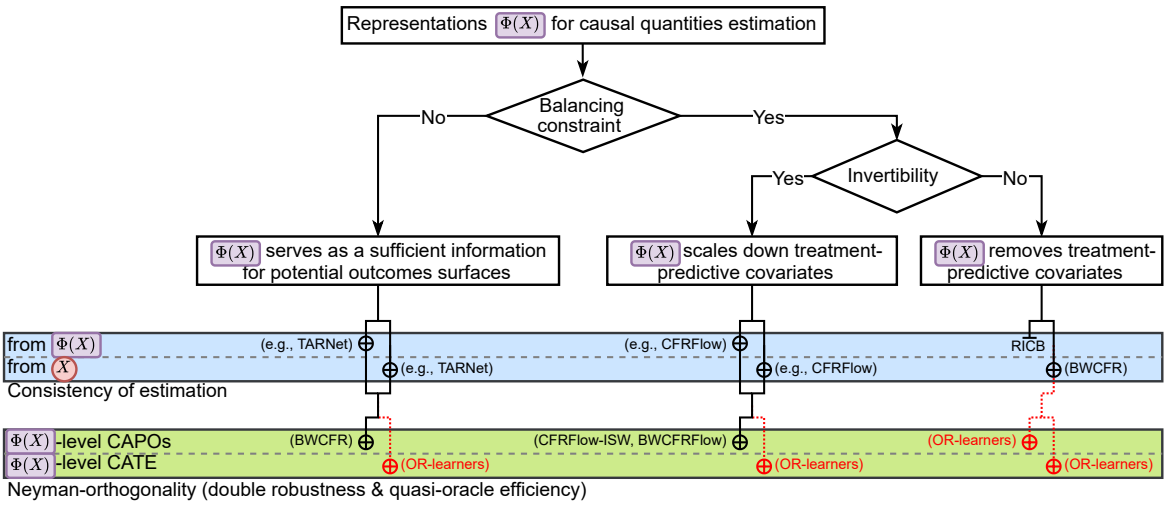

Figure 4: Flow chart of consistency and Neyman-orthogonality for representation learning methods. The *OR-learners* fill the gaps shown by red dotted lines.

outcomes (APOs) and average treatment effect (ATE). This is because they lose information about the heterogeneity of the potential outcomes/treatment effect. In our work, on the other hand, we study a general class of *heterogeneous* causal quantities, namely, representation-conditional CAPOs/CATE.

- *Prognostic scores* (Hansen, 2008) can be used for both averaged (Antonelli et al., 2018; Luo and Zhu, 2020; D'Amour and Franks, 2021) and heterogeneous causal quantities (Huang and Chan, 2017). In (Huang and Chan, 2017; Luo and Zhu, 2020), they are used in the context of a sufficient covariate dimensionality reduction. Yet, these works either (i) make simplifying strong assumptions (Antonelli et al., 2018; Luo and Zhu, 2020; D'Amour and Franks, 2021), so that the prognostic scores coincide with the expected covariate-conditional outcome; or (ii) consider only linear prognostic scores (Huang and Chan, 2017; Luo and Zhu, 2020). To the best of our knowledge, the first practical method for non-linear, learnable representations was proposed in (Johansson et al., 2016; Shalit et al., 2017; Johansson et al., 2022).

Hence, the above-mentioned works operate in much simpler settings and, therefore, are not relevant baselines for our work.

A.2 Two-stage Meta-learners

Meta-learners. Causal quantities can be estimated using model-agnostic methods, so-called *meta-learners* (Künzel et al., 2019). Meta-learners typically combine multiple models to perform two-stage learning, namely, (1) nuisance functions estimation and (2) target model fitting. As such, meta-learners must be instantiated with some machine learning model (e.g., a neural network) to perform (1) and (2). Notable examples include X- and U-learners (Künzel et al., 2019), R-learner (Nie and Wager, 2021), DR-learner (Kennedy, 2023; Curth et al., 2020), and IVW-learner (Fisher, 2024). Curth and van der Schaar (2021b) provided a comparison of meta-learners implemented via neural networks, where unconstrained representations are used solely to estimate (1) nuisance functions but not as inputs to the (2) target model (as we analyze in our work).

Neyman-orthogonal learners. Neyman-orthogonality (Foster and Syrgkanis, 2023), or double/debiased machine learning (Chernozhukov et al., 2017), directly extend the idea of semi-parametric efficiency to infinite-dimensional target estimands such as CAPOs and the CATE. Informally, Neyman-orthogonality means that the population loss of the target model is first-order insensitive to the misspecification of the nuisance functions. Examples of model-agnostic[5] Neyman-orthogonal learners are DR-learners for CAPOs (Vansteelandt and Morzywołek, 2025); and DR-, R-, IVW-learners for CATE (Morzywolek et al., 2023).

[5]Several works extended the theory of targeted maximum likelihood estimation (van der Laan et al., 2011) and proposed sieves-based Neyman-orthogonal learners (e. g., EP-learner for CATE (van der Laan et al., 2024) and i-learner for CAPOs (Vansteelandt and Morzywołek, 2025)). Yet, those methods are not fully model-agnostic (namely, they cannot be instantiated with neural networks) and are thus not considered in our work.

A.3 Estimation of Causal Quantities for General-purpose Learned Representations

Other constraints may be applied to the representations, for example, to achieve algorithmic fairness (Zemel et al., 2013; Madras et al., 2018). Some works combined Neyman-orthogonal learners and fairness constraints, but different from our setting. For example, (Kim and Zubizarreta, 2023) provided a DR-learner for fair CATE estimation based on the linear combination of the basis functions; and (Frauen et al., 2024) built fair representations for policy learning with DR-estimators of policy value. The latter work, nevertheless, can be seen as a special case of the general *OR-learners*.

B BACKGROUND MATERIALS

In this section, we provide the formal definitions of Neyman-orthogonality, Hölder smoothness, and integral probability metrics; we state the identifiability and smoothness assumptions; and we offer an overview of meta-learners for CAPOs/CATE estimation.

B.1 Neyman-orthogonality and Double Robustness

We use the following additional notation: $\|\cdot\|_{L_p}$ denotes the L_p-norm with $\|f\|_{L_p} = \mathbb{E}(|f(Z)|^p)^{1/p}$, $a \lesssim b$ means there exists $C \geq 0$ such that $a \leq C \cdot b$, and $X_n = o_{\mathbb{P}}(r_n)$ means $X_n/r_n \xrightarrow{p} 0$.

Definition 1 (Neyman-orthogonality (Foster and Syrgkanis, 2023; Morzywolek et al., 2023))**.** *A risk $\mathcal{L}$ is called* Neyman-orthogonal *if its pathwise cross-derivative equals zero, namely,*

$$D_\eta D_g \mathcal{L}(g^*, \eta)[g - g^*, \hat{\eta} - \eta] = 0 \quad \textit{for all } g \in \mathcal{G} \textit{ and } \hat{\eta} \in \mathcal{H}, \tag{9}$$

where $D_f F(f)[h] = \frac{\mathrm{d}}{\mathrm{d}t} F(f + th)|_{t=0}$ and $D_f^k F(f)[h_1, \ldots, h_k] = \frac{\partial^k}{\partial t_1 \ldots \partial t_k} F(f + t_1 h_1 + \cdots + t_k h_k)|_{t_1 = \cdots = t_k = 0}$ are pathwise derivatives (Foster and Syrgkanis, 2023); $g^* = \arg\min_{g \in \mathcal{G}} \mathcal{L}(g, \eta)$; and η is the ground-truth nuisance function.

Informally, this definition means that the risk is first-order insensitive wrt. the misspecification of the nuisance functions.

Definition 2 (Double robustness)**.** *An estimator $\hat{g}^* = \arg\min_{g \in \mathcal{G}} \mathcal{L}(g, \hat{\eta})$ of $g^* = \arg\min_{g \in \mathcal{G}} \mathcal{L}(g, \eta)$ is said to be* double robust *if, for any estimators $\hat{\mu}_a^x$ and $\hat{\pi}_1^x$ of the nuisance functions μ_a^x and π_1^x, it holds that*

$$\|\hat{g}^* - g^*\|_{L_2}^2 \lesssim \mathcal{L}(\hat{g}^*, \hat{\eta}) - \mathcal{L}(g^*, \hat{\eta}) + \underbrace{\|\hat{\pi}_1^x - \pi_1^x\|_{L_4}^2 \, \|\hat{\mu}_a^x - \mu_a^x\|_{L_4}^2}_{R_2(\eta, \hat{\eta})}, \tag{10}$$

where $\mathcal{L}(\hat{g}^*, \hat{\eta}) - \mathcal{L}(g^*, \hat{\eta})$ is the difference between the risks of the estimated target model and the optimal target model where the estimated nuisance functions are used, and $R_2(\eta, \hat{\eta})$ is a second-order remainder.

Definition 3 (Quasi-oracle efficiency)**.** *An estimator $\hat{g}^* = \arg\min_{g \in \mathcal{G}} \mathcal{L}(g, \hat{\eta})$ of $g^* = \arg\min_{g \in \mathcal{G}} \mathcal{L}(g, \eta)$ is said to be* quasi-oracle efficient *if the estimators $\hat{\mu}_a^x$ and $\hat{\pi}_1^x$ of the nuisance functions μ_a^x and π_1^x are allowed to have slow rates of convergence, $o_{\mathbb{P}}(n^{-1/4})$, and the following still holds asymptotically:*

$$\|\hat{g}^* - g^*\|_{L_2}^2 \lesssim \mathcal{L}(\hat{g}^*, \hat{\eta}) - \mathcal{L}(g^*, \hat{\eta}) + \underbrace{o_{\mathbb{P}}(n^{-1/2})}_{R_2(\eta, \hat{\eta})}, \tag{11}$$

where $\mathcal{L}(\hat{g}^*, \hat{\eta}) - \mathcal{L}(g^*, \hat{\eta})$ is the difference between the risks of the estimated target model and the optimal target model where the estimated nuisance functions are used, and $R_2(\eta, \hat{\eta})$ is a second-order remainder.

Furthermore, if the finite-sample estimator is used, namely $\hat{g} = \arg\min_{g \in \mathcal{G}} \hat{\mathcal{L}}(g, \hat{\eta})$, the error between $\hat{g}$ and g^* can be upper-bounded as

$$\|\hat{g} - g^*\|_{L_2}^2 \lesssim \mathrm{Rate}_{\mathcal{D}}(\mathcal{G}; \hat{g}, \hat{\eta}) + \underbrace{o_{\mathbb{P}}(n^{-1/2})}_{R_2(\eta, \hat{\eta})}, \tag{12}$$

where $\mathrm{Rate}_{\mathcal{D}}(\mathcal{G}; \hat{g}, \hat{\eta})$ is the convergence rate of the target model that satisfies $\mathcal{L}(\hat{g}, \hat{\eta}) - \mathcal{L}(g^*, \hat{\eta}) \leq \mathrm{Rate}_{\mathcal{D}}(\mathcal{G}; \hat{g}, \hat{\eta})$ for any $\hat{\eta} \in \mathcal{H}$ (Foster and Syrgkanis, 2023).

B.2 Hölder Smoothness

Definition 4 (Hölder smoothness)**.** *Let $s > 0, s - \lfloor s \rfloor \in (0, 1]$, and $\mathcal{X} \subseteq \mathbb{R}^{d_x}$. A function $f : \mathcal{X} \to \mathbb{R}$ is said to be* s-Hölder smooth *(i.e., belongs to the Hölder class $C^s(\mathcal{X})$) if it is $\lfloor s \rfloor$-times continuously differentiable, and, for any $x, x' \in \mathcal{X}$, there exists a constant $L > 0$ such that for every m:*

$$|D^m f(x) - D^m f(x')| \leq L \, \|x - x'\|_2^{s - \lfloor s \rfloor}, \tag{13}$$

where $m = (m_1, \ldots, m_{d_x})$ *such that* $|m| = \sum_j m_j = \lfloor s \rfloor$, $D^m = \frac{\partial^{\lfloor s \rfloor}}{\partial_{x_1}^{m_1} \ldots \partial_{x_{d_x}}^{m_{d_x}}}$, *and* $\|\cdot\|_2$ *is the Euclidean norm. In our work, we set the constant* L *to be a* Hölder norm $\|\cdot\|_{C^s(\mathcal{X})}$*:*

$$L = \|f\|_{C^s(\mathcal{X})} := \sum_{|m| \leq \lfloor s \rfloor} \sup_{x \in \mathcal{X}} |D^m f(x)| + \sum_{|m| = \lfloor s \rfloor} \sup_{x, x' \in \mathcal{X}, x \neq x'} \frac{|D^m f(x) - D^m f(x')|}{\|x - x'\|_2^{s - \lfloor s \rfloor}}, \tag{14}$$

where the second term is also called a Hölder semi-norm $[D^m f]_{C^{0, s - \lfloor s \rfloor}(\mathcal{X})}$.

Hölder smooth functions have a following useful property (De La Llave and Obaya, 1999): they can be well approximated by $\lfloor s \rfloor$-order Taylor approximations, namely,

$$f(x) = f(x') + \sum_{j=1}^{\lfloor s \rfloor} \frac{1}{j!} D^j f(x')[x - x', \ldots, x - x'] + R_{\lfloor s \rfloor}(x, x'), \tag{15}$$

where $D^j f(x')[x - x', \ldots, x - x']$ is a Taylor polynomial of degree j, and $R_{\lfloor s \rfloor}(x, x')$ is a remainder term for which $\left|R_{\lfloor s \rfloor}(x, x')\right| \leq (1/\lfloor s \rfloor!) \|f\|_{C^s(\mathcal{X})} \|x - x'\|_2^s$.

Another useful property connects the Hölder norm L with Lipschitz constant $\mathrm{Lip}(f)$ when $s \geq 1$:

$$\mathrm{Lip}(f) \leq \sup_{x \in \mathcal{X}} \|\nabla_x f(x)\| = \sup_{x \in \mathcal{X}} \left(\sum_{j=1}^{d_x} \left| D^j f(x) \right|^2 \right)^{1/2} \leq \left(\sum_{j=1}^{d_x} L^2 \right)^{1/2} = \sqrt{d_x}\, L, \tag{16}$$

where $\nabla_x f(x)$ is a Jacobian of f.

B.3 Integral Probability Metrics

Integral probability metrics (IPMs) are a broad class of distances between probability distributions, defined in terms of a family of functions $\mathcal{F}$. Given two probability distributions $\mathbb{P}(Z_1)$ and $\mathbb{P}(Z_2)$ over a domain $\mathcal{Z}$, an IPM measures the maximum difference in expectation over a class of functions $\mathcal{F}$:

$$\mathrm{IPM}(\mathbb{P}(Z_1), \mathbb{P}(Z_2)) = \sup_{f \in \mathcal{F}} \left| \mathbb{E}(f(Z_1)) - \mathbb{E}(f(Z_2)) \right|. \tag{17}$$

In this framework, $\mathcal{F}$ specifies the allowable ways in which the difference between the distributions can be measured. Depending on the choice of $\mathcal{F}$, different IPMs arise.

Wasserstein metric (WM). The Wasserstein metric is a specific IPM where the function class $\mathcal{F}$ is the set of 1-Lipschitz functions, which are functions where the absolute difference between outputs is bounded by the absolute difference between inputs:

$$W(\mathbb{P}(Z_1), \mathbb{P}(Z_2)) = \sup_{f \in \mathcal{F}_1} \left| \mathbb{E}(f(Z_1)) - \mathbb{E}(f(Z_2)) \right|. \tag{18}$$

This metric can be interpreted as the minimum cost required to transport probability mass from one distribution to another, where the cost is proportional to the distance moved.

Maximum mean discrepancy (MMD). Another popular example is the maximum mean discrepancy, where the function class $\mathcal{F}$ corresponds to functions in the unit ball of a reproducing kernel Hilbert space (RKHS), $\mathcal{F}_{\mathrm{RKHS},\,1} = \{f \in \mathcal{H} : \|f\|_{\mathcal{H}} \leq 1\}$:

$$\mathrm{MMD}(\mathbb{P}(Z_1), \mathbb{P}(Z_2)) = \sup_{f \in \mathcal{F}_{\mathrm{RKHS},1}} \left| \mathbb{E}(f(Z_1)) - \mathbb{E}(f(Z_2)) \right|. \tag{19}$$

The MMD is often used in hypothesis testing and in training generative models, particularly when the distributions are defined over high-dimensional data.

B.4 Meta-learners for CAPOs and CATE Estimation

Plug-in learners. A naïve way to estimate CAPOs and CATE is to simply estimate $\hat{\mu}_0^x(x)$ and $\hat{\mu}_1^x(x)$ and 'plug them into' the identification formulas for CAPOs and CATE. For example, an S-learner (S-Net) fits a single model with the treatment as an input, while a T-learner (T-Net) builds two models for each treatment (Künzel et al., 2019). Many end-to-end representation learning methods, such as TARNet (Shalit et al., 2017) and BNN without the balancing constraint (Johansson et al., 2016), can be seen as variants of the plug-in learner: In the end-to-end fashion, they build a representation of the covariates $\hat{\Phi}(x) \in \Phi \subseteq \mathbb{R}^{d_\phi}$ and then use $\hat{\Phi}$ to estimate $\hat{\mu}_a^x(x) = h_a(\Phi(x))$ with the S-Net (BNN w/o balancing) or the T-Net (TARNet).

Yet, plug-in learners have several major drawbacks (Morzywolek et al., 2023; Vansteelandt and Morzywołek, 2025). (a) They do not account for the selection bias, namely, that $\hat{\mu}_0^x$ is estimated better for the treated population and $\hat{\mu}_1^x$ for the untreated (this is also known as a plug-in bias (Kennedy, 2023)). (b) In the case of CATE estimation, the plug-in learners might additionally fail to address the causal inductive bias that the CATE is a "simpler" function than both CAPOs (Künzel et al., 2019; Curth and van der Schaar, 2021a), as it is impossible to add additional smoothing for the CATE model separately from CAPOs models. (c) It is also unclear how to consistently estimate the CAPOs/CATE depending on the subset of covariates $V \subseteq X$ with the aim of reducing the variance of estimation. For example, it is unclear how to estimate representation-level CAPOs, $\xi_a^\phi(\phi) = \mathbb{E}(Y[a] \mid \Phi(X) = \phi)$, and CATE, $\tau^\phi(\phi) = \mathbb{E}(Y[1] - Y[0] \mid \Phi(X) = \phi)$, especially when the representations are constrained.

Two-stage Neyman-orthogonal learners. To address the shortcomings of plug-in learners, two-stage meta-learners were proposed (see Appendix A.2). These proceed as follows:

- **(i)** First, one chooses a *target model class* $\mathcal{G} = \{g(\cdot) : \mathcal{V} \subseteq \mathcal{X} \to \mathcal{Y}\}$ such as, for example, neural networks. A target model takes a (possibly confounded) subset V of the original covariates X as an input and outputs the prediction of causal quantities conditioned on V, namely, CAPOs $\xi_a^v(v) = \mathbb{E}(Y[a] \mid V = v)$ or CATE $\tau^v(v) = \mathbb{E}(Y[1] - Y[0] \mid V = v)$.
- **(ii)** Then, two-stage meta-learners formulate one of the *target risks* for $g(v)$, where $v \in \mathcal{V}$:

$$\mathcal{L}_{\mathcal{G}}(g, \eta) = \mathbb{E}\big[w(\pi_a^x(X))\,(\chi^x(X, \eta) - g(V))^2\big], \tag{20}$$

  where $\eta = (\mu_0^x, \mu_1^x, \pi_1^x)$ are the nuisance functions, $w(\cdot) > 0$ is the weighting function, $\chi^x(x, \eta) = \mu_a^x(x)$ for CAPOs, and $\chi^x(x, \eta) = \mu_1^x(x) - \mu_0^x(x)$ for CATE. Based on $w(\cdot)$, there are multiple choices for choosing a target risk, each with different interpretations and implications for finite-sample two-stage estimation. For example, the DR-learners for CAPOs/CATE use $w(\pi_a^x(x)) = 1$, and the R-/IVW-learners for CATE use overlap weights $w(\pi_a^x(x)) = \pi_0^x(x)\,\pi_1^x(x)$. Furthermore, it is easy to see that the minimization of the target risks in Eq. (20) yields the best projection of the ground-truth V-level causal quantities: DR-learners yield the projection of CAPOs/CATE, $\xi_a^v(v)/\tau^v(v)$; and the R-/IVW-learners yield the projection of the overlap-weighted CATE, $\tau_{\pi_0\,\pi_1}^v(v) = \mathbb{E}[\pi_0^x(X)\,\pi_1^x(X)(Y[1] - Y[0]) \mid V = v]/\mathbb{E}[\pi_0^x(X)\,\pi_1^x(X) \mid V = v]$ (Morzywolek et al., 2023; Vansteelandt and Morzywołek, 2025).
- **(iii)** In the last step, two-stage meta-learners minimize an empirical version of the target risk $\hat{\mathcal{L}}_{\mathcal{G}}(g, \hat{\eta})$, which is estimated using observational data and the nuisance functions $\hat{\eta}$ estimated at the first stage. The latest step then yields so-called *Neyman-orthogonal learners* when the target risk is estimated with semi-parametric efficient estimators (Robins and Rotnitzky, 1995; Foster and Syrgkanis, 2023). We provide specific definitions of different Neyman-orthogonal learners in Table 6. Therein, we also specify the second-order remainder terms associated with Neyman-orthogonality $R_2(\eta, \hat{\eta})$ that relate to the *quasi-oracle efficiency* and *double robustness* (see Appendix B.1 for definitions).

End-to-end Neyman-orthogonality. As noted by Vansteelandt and Morzywołek (2025), under *certain conditions*, the end-to-end IPTW-learners for CAPOs might possess Neyman-orthogonality. Specifically, if we assume that the ground-truth CAPOs are contained in the target model class (i. e., $V = X$ and $\xi_x^a \in \mathcal{G}$ so that $g^* = \xi_x^a$), the target causal quantity ξ_x^a coincides with one of the nuisance functions μ_x^a. Therefore, by setting $g = \mu_x^a$, we can simplify the original DR-loss in the style of Foster and Syrgkanis (2023):

$$\hat{\mathcal{L}}_{\mathcal{G}}^{\mathrm{DR}_a^{\mathrm{FS}}}(g, \hat{\eta} = (\mu_a^x = g, \pi_a^x)) = \mathbb{P}_n\left\{\frac{\mathbb{1}\{A = a\}}{\hat{\pi}_a^x(X)}\big(Y - g(X)\big)^2\right\}, \tag{21}$$

Table 6: Overview of two-stage Neyman-orthogonal learners. Here, $\eta = (\mu_0^x, \mu_1^x, \pi_1^x)$ are the nuisance functions.

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}	$\mathbb{P}_n \left\{ \left(\frac{\mathbb{1}\{A=a\}}{\hat{\pi}_a^x(X)} \big(Y - \hat{\mu}_a^x(X)\big) + \hat{\mu}_a^x(X) - g(V) \right)^2 \right\}$	$\lVert \hat{\pi}_1^x - \pi_1^x \rVert_{L_4}^2 \lVert \hat{\mu}_a^x - \mu_a^x \rVert_{L_4}^2$
	DR_a^{FS}	$\mathbb{P}_n \left\{ \frac{\mathbb{1}\{A=a\}}{\hat{\pi}_a^x(X)} \big(Y - g(V)\big)^2 + \left(1 - \frac{\mathbb{1}\{A=a\}}{\hat{\pi}_a^x(X)}\right) \big(\hat{\mu}_a^x(X) - g(V)\big)^2 \right\}$	$\lVert \hat{\pi}_1^x - \pi_1^x \rVert_{L_4}^2 \lVert \hat{\mu}_a^x - \mu_a^x \rVert_{L_4}^2$
CATE	DR^{K}	$\mathbb{P}_n \left\{ \left(\frac{A - \hat{\pi}_1^x(X)}{\hat{\pi}_0^x(X)\, \hat{\pi}_1^x(X)} \big(Y - \hat{\mu}_A^x(X)\big) + \hat{\mu}_1^x(X) - \hat{\mu}_0^x(X) - g(V) \right)^2 \right\}$	$\sum_{a \in \{0,1\}} \lVert \hat{\pi}_1^x - \pi_1^x \rVert_{L_4}^2 \lVert \hat{\mu}_a^x - \mu_a^x \rVert_{L_4}^2$
	R	$\mathbb{P}_n \left\{ \big(A - \hat{\pi}_1^x(X)\big)^2 \left(\frac{Y - \hat{\mu}^x(X)}{A - \hat{\pi}_1^x(X)} - g(V) \right)^2 \right\}$	$\sum_{a \in \{0,1\}} \lVert \hat{\pi}_1^x - \pi_1^x \rVert_{L_4}^2 \lVert \hat{\mu}_a^x - \mu_a^x \rVert_{L_4}^2 + \lVert \hat{\pi}_1^x - \pi_1^x \rVert_{L_4}^4$
	IVW	$\mathbb{P}_n \left\{ \big(A - \hat{\pi}_1^x(X)\big)^2 \left(\frac{A - \hat{\pi}_1^x(X)}{\hat{\pi}_0^x(X)\, \hat{\pi}_1^x(X)} \big(Y - \hat{\mu}_A^x(X)\big) + \hat{\mu}_1^x(X) - \hat{\mu}_0^x(X) - g(V) \right)^2 \right\}$	$\sum_{a \in \{0,1\}} \lVert \hat{\pi}_1^x - \pi_1^x \rVert_{L_4}^2 \lVert \hat{\mu}_a^x - \mu_a^x \rVert_{L_4}^2 + \lVert \hat{\pi}_1^x - \pi_1^x \rVert_{L_4}^4$

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)

and original DR-loss in the style of Kennedy (2023):

$$\hat{\mathcal{L}}_{\mathcal{G}}^{\text{DR}_a^{\text{K}}}(g, \hat{\eta} = (\mu_a^x = g, \pi_a^x)) = \mathbb{P}_n \left\{ \frac{\mathbb{1}\{A = a\}}{(\hat{\pi}_a^x(X))^2} \big(Y - g(X)\big)^2 \right\}. \tag{22}$$

Both losses in Eq. (21)-(22) are examples of weighted plug-in (WPI) / IPTW learners, and it can be easily shown that they both possess Neyman-orthogonality and quasi-oracle efficiency. Thus, some of the end-to-end representation learning methods that use IPTW weighting (namely, CFR-ISW (Hassanpour and Greiner, 2019a) and BWCFR (Assaad et al., 2021)) are Neyman-orthogonal when no balancing constraint is enforced or the invertible representations are used (so that the above-mentioned conditions are met).

C THEORETICAL RESULTS

Lemma 1 (Quasi-oracle efficiency of a non-parametric model)**.** *Assume (i) a non-parametric target model $g(v), v \in \mathcal{V} \subseteq \mathbb{R}^{d_v}$ and (ii) that a convergence rate for the target model does not depend on whether estimated or the ground-truth nuisance functions are used. Then, the error between $g^{*v} = \arg\min_{g \in \mathcal{G}} \mathcal{L}_{\mathcal{G}}(g, \eta)$ and $\hat{g}^v = \arg\min_{g \in \mathcal{G}} \hat{\mathcal{L}}_{\mathcal{G}}(g, \hat{\eta})$ can be upper-bounded as:*

$$\|g^{*v} - \hat{g}^v\|_{L_2}^2 \lesssim (L^v)^{2d_v/(2s^v+d_v)} \cdot n^{-2s^v/(2s^v+d_v)} + R_2(\eta, \hat{\eta}), \tag{23}$$

where g^{*v} is an s^v-Hölder smooth function with the Hölder norm $L^v = \|g^{*v}\|_{C^{s^v}(\mathcal{V})}$, and $R_2(\eta, \hat{\eta})$ is a second-order remainder that depends on n, $s^x_{\mu_a}, s^x_\pi$, and d_x. For example, if the nuisance functions are estimated with non-parametric models, the second-order term has the following form:

$$\text{DR}^a : R_2(\eta, \hat{\eta}) \lesssim n^{-2(s^x_{\mu_a}/(s^x_{\mu_a}+d_x)+s^x_\pi/(s^x_\pi+d_x))}, \tag{24}$$

$$\text{DR} : R_2(\eta, \hat{\eta}) \lesssim n^{-2(\min_{a\in\{0,1\}} s^x_{\mu_a}/(s^x_{\mu_a}+d_x)+s^x_\pi/(s^x_\pi+d_x))}, \tag{25}$$

$$\text{R/IVW} : R_2(\eta, \hat{\eta}) \lesssim n^{-2(\min_{a\in\{0,1\}} s^x_{\mu_a}/(s^x_{\mu_a}+d_x)+s^x_\pi/(s^x_\pi+d_x))} + n^{-4s^x_\pi/(s^x_\pi+d_x))}, \tag{26}$$

where Hölder norms are omitted for clarity.

Proof. In the following, to simplify the notation, we drop an upper index v for both $g^{*v} = g^*$ and $\hat{g}^v = \hat{g}$. We adopt a classical convergence rate of a non-parametric regression (Stone, 1982). Specifically, we define a second-stage estimator $\hat{g}_p$ as a local polynomial/linear smoother estimator of order $p \geq \lfloor s \rfloor$ (we denote the class of such estimators as $\mathcal{G}$). Under the usual regularity assumptions (Kennedy, 2023) and given the Taylor expansion property of Hölder smooth functions (see Appendix B.2), there exists a constant $C > 0$ and a bandwidth $h > 0$ such that, for an oracle second-stage estimator g^*_p, the following holds:

$$\inf_{g_p \in \mathcal{G}} \|g_p - g^*\|_{L_2} = \left\|g^*_p - g^*\right\|_{L_2} \lesssim \sqrt{C} L^v h^{s^v}. \tag{27}$$

Let us denote $\hat{g}^*_p = \arg\min_{g_p \in \mathcal{G}} \hat{\mathcal{L}}_{\mathcal{G}}(g_p, \eta)$. Then, using the standard local Rademacher/VC arguments, the estimation error is as follows:

$$\left\|\hat{g}^*_p - g^*_p\right\|_{L_2}^2 \lesssim \frac{C}{nh^{d_v}}, \tag{28}$$

and, thus, the desired estimation error becomes

$$\left\|\hat{g}^*_p - g^*\right\|_{L_2}^2 \leq 2\left\|\hat{g}^*_p - g^*_p\right\|_{L_2}^2 + 2\left\|g^*_p - g^*\right\|_{L_2}^2 \lesssim \frac{C}{nh^{d_v}} + C(L^v)^2 h^{2s^v}. \tag{29}$$

Now, by choosing $h \asymp \left(\frac{1}{(L^v)^2 n}\right)^{1/(2s^v+d_v)}$, we recover the following bound (with the known nuisance functions):

$$\left\|\hat{g}^*_p - g^*\right\|_{L_2}^2 \lesssim C \underbrace{(L^v)^{2d_v/(2s^v+d_v)} \cdot n^{-2s^v/(2s^v+d_v)}}_{\text{Rate}_{\mathcal{D}}(\mathcal{G}; \hat{g}_p, \eta)}. \tag{30}$$

Finally, by using the error bound from the quasi-oracle efficiency of the Neyman-orthogonal learners (see Definition 3) and assuming that $\mathcal{L}_{\mathcal{G}}(\hat{g}_p, \hat{\eta}) - \mathcal{L}_{\mathcal{G}}(\hat{g}^*_p, \hat{\eta}) = \text{Rate}_{\mathcal{D}}(\mathcal{G}; \hat{g}_p, \hat{\eta}) \lesssim \text{Rate}_{\mathcal{D}}(\mathcal{G}; \hat{g}_p, \eta)$ (convergence rate for the target model does not depend on whether estimated or the ground-truth nuisance functions are used), we recover the desired inequality

$$\|\hat{g}_p - g^*\|_{L_2}^2 \leq 2\left\|\hat{g}_p - \hat{g}^*_p\right\|_{L_2}^2 + 2\left\|\hat{g}^*_p - g^*\right\|_{L_2}^2 \tag{31}$$

$$\lesssim \mathcal{L}_{\mathcal{G}}(\hat{g}_p, \hat{\eta}) - \mathcal{L}_{\mathcal{G}}(\hat{g}^*_p, \hat{\eta}) + R_2(\eta, \hat{\eta}) + \left\|\hat{g}^*_p - g^*\right\|_{L_2}^2 \tag{32}$$

$$\lesssim (L^v)^{2d_v/(2s^v+d_v)} \cdot n^{-2s^v/(2s^v+d_v)} + R_2(\eta, \hat{\eta}), \tag{33}$$

where $\hat{g}_p = \arg\min_{g_p \in \mathcal{G}} \hat{\mathcal{L}}_{\mathcal{G}}(g_p, \hat{\eta})$.

For the upper-bound on the second-order remainder, $R_2(\eta, \hat{\eta})$, we refer to existing results of (Curth and van der Schaar, 2021b; Kennedy, 2023; Schulte et al., 2025). □

Proposition 1. *Under Assumption 1, the following holds: (1) $d_{\phi^*} \ll d_x$, (2) $g^{*\phi^*}$ is an s^{ϕ^*}-Hölder smooth function with Hölder norm L^{ϕ^*} such that $s^{\phi^*} \geq s^x$ and $L^{\phi^*} \leq c(L^{J^*}) \cdot L^x$ with non-decreasing $c(\cdot)$. Also, when L^{J^*} is sufficiently small,*

$$\left\| g^{*\phi^*} - \hat{g}^{\phi^*} \right\|_{L_2}^2 \lesssim \| g^{*x} - \hat{g}^x \|_{L_2}^2 . \tag{34}$$

Proof. First, (1) $d_{\phi^*} \ll d_x$ is directly given by Assumption 1. Furthermore, (2) can be derived from the properties of the Hölder smooth functions. Specifically, under Assumption 1, the causal quantities in the representation space can be written as the composition of the functions:

$$\xi_a^{\phi^*}(\phi) = \xi_a^{\phi^*}(\Phi^*(J^*(\phi))) = \xi_a^x(J^*(\phi)) \quad \text{and} \quad \tau^{\phi^*}(\phi) = \tau^{\phi^*}(\Phi^*(J^*(\phi))) = \tau^x(J^*(\phi)). \tag{35}$$

Therefore, we can use the standard boundedness of the composition operator on Hölder spaces (see Theorem 4.3 in (De La Llave and Obaya, 1999)):

$$J^* \in C^{s^x+1}(\Phi^*),\ g^{x*} = \xi_a^x / \tau^x \in C^{s^x}(\mathcal{X}) \quad \Rightarrow \quad g^{*\phi^*} = g^{x*} \circ J^* \in C^{s^x}(\mathcal{X}), \tag{36}$$

and

$$\left\| g^{*\phi^*} \right\|_{C^{s^x}(\Phi^*)} = \| g^{x*} \circ J^* \|_{C^{s^x}(\Phi^*)} \leq K_{J^*} \| g^{x*} \|_{C^{s^x}(\mathcal{X}^*)} , \tag{37}$$

where $K_{J^*} = \sup_{g \neq 0} \frac{\| g \circ J^* \|_{C^{s^x}(\Phi^*)}}{\| g \|_{C^{s^x}(\mathcal{X})}}$ is an operator norm of the pullback J^*. Therefore, $s^{\phi^*} \geq s^x$ and $L^{\phi^*} \leq K_{J^*} \cdot L^x$.

Furthermore, it was demonstrated in Theorem 4.3 of De La Llave and Obaya (1999) that (1) if $0 < s^x \leq 1$, it can be shown that $K_{J^*} \leq \max\{1, \mathrm{Lip}(J^*)^{s^x}\} \leq \max\{1, (L^{J^*})^{s^x}\}$; and (2) if $s^x > 1$, then $K_{J^*} \leq P(s^x, L^{J^*})$, where P is a polynomial of degree $\lfloor s \rfloor$ with non-negative coefficients (follows from a Faa di Bruno expansion). Thus, in both cases (1)-(2), K_{J^*} can be upper-bounded by some non-decreasing function c depending on the Hölder norm L^{J^*}.

Finally, to compare the error bounds between $V = \Phi^*(X)$ and $V = X$, we refer to Lemma 1. That is, the error bounds are

$$V = \Phi^*(X): \quad \left\| g^{*\phi^*} - \hat{g}^{\phi^*} \right\|_{L_2}^2 \lesssim (L^{\phi^*})^{2d_{\phi^*}/(2s^{\phi^*}+d_{\phi^*})} \cdot n^{-2s^{\phi^*}/(2s^{\phi^*}+d_{\phi^*})} + R_2(\eta, \hat{\eta}), \tag{38}$$

$$V = X: \quad \| g^{*x} - \hat{g}^x \|_{L_2}^2 \lesssim (L^x)^{2d_x/(2s^x+d_x)} \cdot n^{-2s^x/(2s^x+d_x)} + R_2(\eta, \hat{\eta}). \tag{39}$$

Here, the second-order error terms are the same, yet the target model error terms differ:

$$(L^{\phi^*})^{2d_{\phi^*}/(2s^{\phi^*}+d_{\phi^*})} \leq [c(L^{J^*}) \cdot L^x]^{2d_x/(2s^x+d_x)} \quad \text{and} \quad n^{-2s^{\phi^*}/(2s^{\phi^*}+d_{\phi^*})} \leq n^{-2s^x/(2s^x+d_x)}. \tag{40}$$

Therefore, when L^{J^*} is sufficiently small such that $c(L^{J^*}) \leq 1$, the representation-level learners asymptotically achieve a lower error than the covariate-level learners. □

Proposition 2 (Smoothness of the hidden layers)**.** *We denote the trained representation network as $\hat{\mu}_a^x = \hat{h}_a \circ \hat{\Phi} = \arg\min \hat{\mathcal{L}}_\Phi(h_a \circ \Phi)$. Let $s \geq s_{\mu_a}^x$ and the trained representation network be factorized as*

$$\hat{\mu}_a^x = T_{L+1} \circ T_L \circ \cdots \circ T_1, \tag{41}$$

where $T_\ell : \mathcal{V}_{\ell-1} \to \mathcal{V}_\ell$, $\mathcal{V}_0 = \mathcal{X}$, and $T_{L+1} : \mathcal{V}_L \to \mathcal{Y}$. For each hidden layer $\ell = 1, \ldots, L$, define

$$\hat{f}^{(\ell)} := T_\ell \circ \cdots \circ T_1, \qquad \mathcal{V}_\ell := \hat{f}^{(\ell)}(\mathcal{X}), \tag{42}$$

and the corresponding tail network

$$\hat{h}_a^{(\ell)} := T_{L+1} \circ T_L \circ \cdots \circ T_{\ell+1} : \mathcal{V}_\ell \to \mathcal{Y}, \tag{43}$$

so that $\hat{\mu}_a^x(x) = \hat{h}_a^{(\ell)}(\hat{f}^{(\ell)}(x))$.

Assume:

(i) $\mathcal{X}$ *is compact;* $\mu_a^x \in C^{s_{\mu_a}^x}(\mathcal{X})$; $\|\mu_a^x\|_{C^{s_{\mu_a}^x}(\mathcal{X})} = L_{\mu_a}^x$.

(ii) There exists $B_a > 0$ *such that* $\|T_{L+1}\|_{C^s(\mathcal{V}_L)} \leq B_a$.

(iii) For each hidden layer $j = 2, \ldots, L$*, there exists* $K_{T_j} \in (0, 1]$ *such that, for every scalar-valued* $g \in C^s(\mathcal{V}_j)$*,* $\|g \circ T_j\|_{C^s(\mathcal{V}_{j-1})} \leq K_{T_j} \|g\|_{C^s(\mathcal{V}_j)}$.

(iv) For each $\ell = 1, \ldots, L$*, define the conditional residual* $\hat{r}_{\ell,n}(v) := \mathbb{E}[\mu_a^x(X) - \hat{h}_a^{(\ell)}(V_\ell) \mid V_\ell = v]$*. Assume that* $\hat{r}_{\ell,n} \in C^s(V_\ell)$ *and* $\|\hat{r}_{\ell,n}\|_{C^s(V_\ell)} \leq \varepsilon_n, \varepsilon_n \to 0$.

(v) There exists $\Delta > 0$ *such that* $B_a \prod_{j=2}^L K_{T_j} \leq L_{\mu_a}^x - \Delta, \varepsilon_n \leq \Delta$ *for all sufficiently large* n.

Then there exists a hidden layer $V_\ell = \hat{f}^{(\ell)}(\mathcal{X})$ *such that the representation-level regression target*

$$\mu_a^{v_\ell}(v) := \mathbb{E}[\mu_a^x(X) \mid V_\ell = v] \tag{44}$$

satisfies

$$\mu_a^{v_\ell} \in C^s(\mathcal{V}_\ell), \qquad \|\mu_a^{v_\ell}\|_{C^s(\mathcal{V}_\ell)} \leq L_{\mu_a}^x. \tag{45}$$

Hence,

$$s_{\mu_a}^{v_\ell} \geq s_{\mu_a}^x, \qquad L_{\mu_a}^{v_\ell} \leq L_{\mu_a}^x. \tag{46}$$

In fact, under assumption (v), one may take $\ell = 1$.

Proof. We choose $\ell = 1$. By repeated application of assumption (iii),

$$\|\hat{h}_a^{(1)}\|_{C^s(\mathcal{V}_1)} = \|T_{L+1} \circ T_L \circ \cdots \circ T_2\|_{C^s(\mathcal{V}_1)} \leq \Big(\prod_{j=2}^L K_{T_j}\Big) \|T_{L+1}\|_{C^s(\mathcal{V}_L)} \leq B_a \prod_{j=2}^L K_{T_j}. \tag{47}$$

Now define

$$\mu_a^{v_1}(v) := \mathbb{E}[\mu_a^x(X) \mid V_1 = v]. \tag{48}$$

By linearity of conditional expectation,

$$\mu_a^{v_1}(v) = \mathbb{E}\Big[\hat{h}_a^{(1)}(V_1) + \big(\mu_a^x(X) - \hat{h}_a^{(1)}(V_1)\big) \,\Big|\, V_1 = v\Big] = \hat{h}_a^{(1)}(v) + \hat{r}_{1,n}(v). \tag{49}$$

Since $\hat{h}_a^{(1)} \in C^s(\mathcal{V}_1)$ and $\hat{r}_{1,n} \in C^s(\mathcal{V}_1)$ by assumptions (ii)–(iv), it follows that $\mu_a^{v_1} \in C^s(\mathcal{V}_1)$. Using the triangle inequality for the full Hölder norm,

$$\|\mu_a^{v_1}\|_{C^s(\mathcal{V}_1)} \leq \|\hat{h}_a^{(1)}\|_{C^s(\mathcal{V}_1)} + \|\hat{r}_{1,n}\|_{C^s(\mathcal{V}_1)} \leq B_a \prod_{j=2}^L K_{T_j} + \varepsilon_n. \tag{50}$$

By assumption (v), for all sufficiently large n,

$$\|\mu_a^{v_1}\|_{C^s(\mathcal{V}_1)} \leq L_{\mu_a}^x. \tag{51}$$

Because $s \geq s_{\mu_a}^x$, we conclude that

$$s_{\mu_a}^{v_1} \geq s_{\mu_a}^x, \qquad L_{\mu_a}^{v_1} \leq L_{\mu_a}^x. \tag{52}$$

Thus, there exists a hidden layer with the desired property. □

Proposition 3 (Smoothing via expanding mapping)**.** *Assume that the trained representation network* $\hat{\Phi}$ *minimizes* $\hat{\mathcal{L}}_\Phi(h_a \circ \Phi)$ *and is* $s^{\hat{\Phi}}$*-Hölder smooth (*$s^{\hat{\Phi}} \geq 1$*). Then, under assumptions (i)-(v) of Proposition 2 and additional assumptions that:*

(vi) The ground-truth target is non-constant and satisfies $\mathrm{Lip}(\mu_a^x) > 0$.

(vii) The fitted plug-in predictor is accurate in Lipschitz semi-norm: $\mathrm{Lip}(\mu_a^x - \hat{\mu}_a^x) \leq \varepsilon_n, \varepsilon_n \to 0$.

(viii) There exists $\Delta > 0$ such that

$$\sqrt{d_{V_1}}\, B_a \prod_{j=2}^{L} K_{T_j} \leq \mathrm{Lip}(\mu_a^x) - \Delta, \qquad \varepsilon_n \leq \Delta \quad \textit{for all sufficiently large } n, \tag{53}$$

then $\hat{\Phi}$ is an expanding mapping, namely, $\mathrm{Lip}(\hat{\Phi}) \geq 1$.

Proof. By Proposition 2, for $\ell = 1$ and $\hat{h}_a := \hat{h}_a^{(1)}$,

$$\|\hat{h}_a\|_{C^s(\mathcal{V}_1)} \leq B_a \prod_{j=2}^{L} K_{T_j}. \tag{54}$$

Since $s^{\hat{\Phi}} \geq 1$, the standard relation between the Hölder norm and the Lipschitz constant yields

$$\mathrm{Lip}(\hat{h}_a) \leq \sqrt{d_{V_1}}\, \|\hat{h}_a\|_{C^s(\mathcal{V}_1)} \leq \sqrt{d_{V_1}}\, B_a \prod_{j=2}^{L} K_{T_j}. \tag{55}$$

Hence, by assumption (viii),

$$\mathrm{Lip}(\hat{h}_a) \leq \mathrm{Lip}(\mu_a^x) - \Delta. \tag{56}$$

Now, by the triangle inequality for the Lipschitz constant and assumption (vii),

$$\mathrm{Lip}(\mu_a^x) \leq \mathrm{Lip}(\hat{\mu}_a^x) + \mathrm{Lip}(\mu_a^x - \hat{\mu}_a^x) \leq \mathrm{Lip}(\hat{\mu}_a^x) + \varepsilon_n. \tag{57}$$

Since $\hat{\mu}_a^x = \hat{h}_a \circ \hat{\Phi}$, the composition rule for Lipschitz maps gives

$$\mathrm{Lip}(\hat{\mu}_a^x) \leq \mathrm{Lip}(\hat{h}_a)\ \mathrm{Lip}(\hat{\Phi}). \tag{58}$$

Therefore,

$$\mathrm{Lip}(\mu_a^x) \leq \mathrm{Lip}(\hat{h}_a)\ \mathrm{Lip}(\hat{\Phi}) + \varepsilon_n. \tag{59}$$

Rearranging,

$$\mathrm{Lip}(\hat{\Phi}) \geq \frac{\mathrm{Lip}(\mu_a^x) - \varepsilon_n}{\mathrm{Lip}(\hat{h}_a)}. \tag{60}$$

Using Eq. (56),

$$\mathrm{Lip}(\hat{\Phi}) \geq \frac{\mathrm{Lip}(\mu_a^x) - \varepsilon_n}{\mathrm{Lip}(\mu_a^x) - \Delta}. \tag{61}$$

For all sufficiently large n, assumption (viii) implies $\varepsilon_n \leq \Delta$, hence

$$\mathrm{Lip}(\hat{\Phi}) \geq 1. \tag{62}$$

This proves the claim. □

Proposition 4 (Balancing via contracting mapping)**.** *Assume that the trained representation network with $s^{\hat{\Phi}}$-Hölder smooth $\hat{\Phi}$ ($s^{\hat{\Phi}} \geq 1$) that minimizes $\hat{\mathcal{L}}_{\mathrm{Bal}}(\Phi)$ with WM / MMD. Then, assuming that $\hat{\Phi}$ is non-constant and the class of representation networks is closed under rescaling, (1) $\hat{\Phi}$ is a contracting mapping, namely* $\mathrm{Lip}(\hat{\Phi}) \leq 1$. *Furthermore, if Assumption 1 holds wrt. $\hat{\Phi}$ (e. g., $\hat{\Phi}$ is smoothly invertible), (2) then the pullback map is expanding, namely,* $\mathrm{Lip}(\hat{J}) \geq 1$.

Proof. (1) The proof proceeds separately for (a) WM-based balancing and (b) MMD-based balancing. Consider a scaling transformation applied to some s^{Φ}-Hölder smooth representation Φ (with $s^{\Phi} \geq 1$): $\beta \cdot \Phi$ with $\beta \in (0, 1)$. Without the loss of generality, we assume that the class of representation networks is closed under rescaling, so both Φ and $\beta \cdot \Phi$ belong to this class.

(a) WM. For any Φ and $\beta \in (0,1)$, we can use the pushforward property of the Wasserstein metric W:

$$W\Big((\beta \cdot \Phi)_{\#}\mathbb{P}(X \mid A=0), (\beta \cdot \Phi)_{\#}\mathbb{P}(X \mid A=1)\Big) = \beta\, W\big(\Phi_{\#}\mathbb{P}(X \mid A=0), \Phi_{\#}\mathbb{P}(X \mid A=1)\big) \tag{63}$$

$$\leq W\big(\Phi_{\#}\mathbb{P}(X \mid A=0), \Phi_{\#}\mathbb{P}(X \mid A=1)\big). \tag{64}$$

Hence, $\hat{\mathcal{L}}_{Bal}(\beta\,\Phi) \leq \beta \hat{\mathcal{L}}_{Bal}(\Phi)$ with strict inequality unless the class-conditional distributions exactly coincide after $\Phi(\cdot)$ (given $\hat{\mathcal{L}}_{Bal}$ is empirical, strict inequality holds almost always).

Therefore, the proof follows from contradiction: If for some empirical non-constant minimizer $\hat{\Phi}$, $\mathrm{Lip}(\hat{\Phi}) \geq 1$ ($\mathrm{Lip}(\cdot)$ is well-defined as $s^{\hat{\Phi}} \geq 1$), the rescaled $\tilde{\Phi} = \beta\,\hat{\Phi}$ with $\beta \leq 1/\,\mathrm{Lip}(\hat{\Phi})$ achieves almost always (strictly) better balancing loss and has

$$\mathrm{Lip}(\tilde{\Phi}) = \beta\ \mathrm{Lip}(\hat{\Phi}) \leq \frac{\mathrm{Lip}(\hat{\Phi})}{\mathrm{Lip}(\hat{\Phi})} = 1. \tag{65}$$

Thus, $\hat{\Phi}$ is not a proper minimizer of the empirical balancing loss, as we can select a strictly better minimizer with $\mathrm{Lip}(\tilde{\Phi}) \leq 1$.

(b) MMD. Consider RKHS induced by shift-invariant, Lipschitz kernels $k(z,z') = \kappa(\|z-z'\|_2)$ with κ monotonically decreasing. Then, scaling with $\beta \in (0,1)$ shrinks all pairwise distances, which reduces MMD:

$$\mathrm{MMD}\Big((\beta \cdot \Phi)_{\#}\mathbb{P}(X \mid A=0), (\beta \cdot \Phi)_{\#}\mathbb{P}(X \mid A=1)\Big) \leq \mathrm{MMD}\Big(\Phi_{\#}\mathbb{P}(X \mid A=0), \Phi_{\#}\mathbb{P}(X \mid A=1)\Big), \tag{66}$$

with strict decrease unless the two pushforwards already match. Therefore, by same rescaling argument as for (a) WM, we can always find such β so that the rescaled empirical minimizer $\tilde{\Phi} = \beta\hat{\Phi}$ has $\mathrm{Lip}(\tilde{\Phi}) \leq 1$.

(2) Finally, to show that $\mathrm{Lip}(\hat{J}) \geq 1$ under Assumption 1, we used the composition property of Lipschitz smooth functions:

$$\hat{\Phi} \circ \hat{J} = \mathrm{id}_{\hat{\Phi}} \quad \Rightarrow \quad 1 = \mathrm{Lip}(\mathrm{id}_{\hat{\Phi}}) \leq \mathrm{Lip}(\hat{\Phi}) \cdot \mathrm{Lip}(\hat{J}). \tag{67}$$

Hence, if $\mathrm{Lip}(\hat{\Phi}) \leq 1$, it is necessary that $\mathrm{Lip}(\hat{J}) \geq 1$.

□

D VISUAL SUMMARY OF THEORETICAL RESULTS

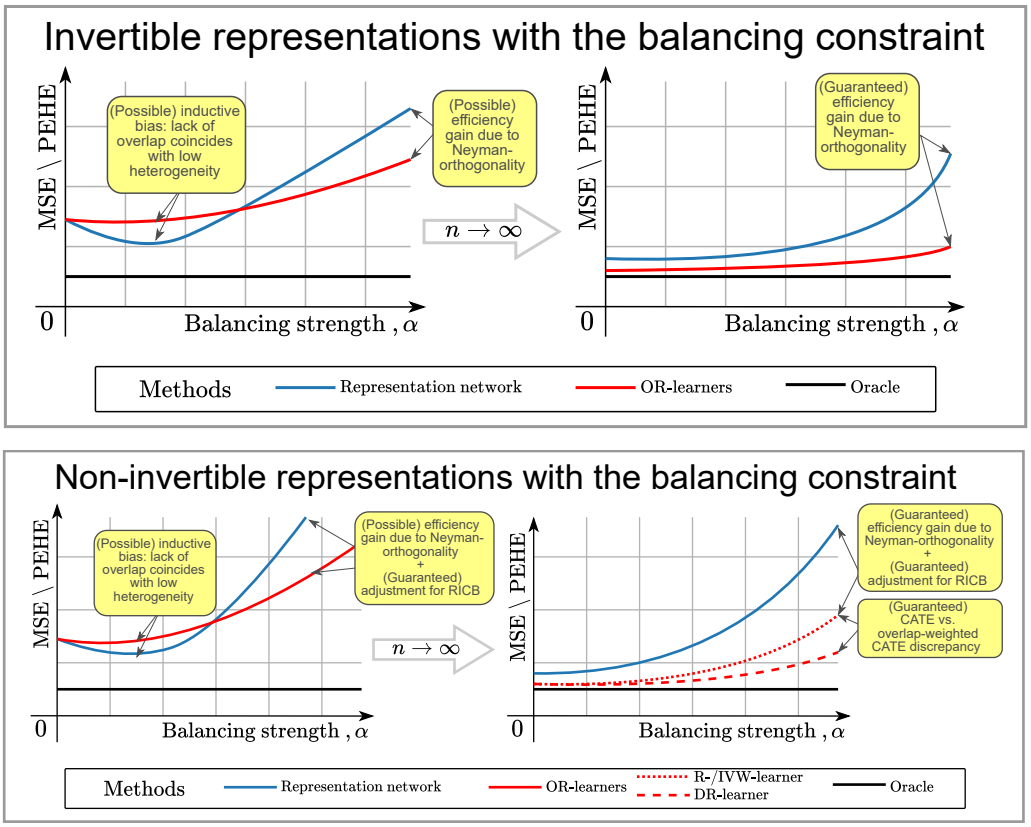

Figure 5: **Insights for RQ ②**. For both figures, we highlight in yellow boxes how the *OR-learners* (in red) can be beneficial in comparison with the end-to-end representation network (in blue). Specifically, we compare the generalization performance in terms of MSE / precision in estimating heterogeneous effect (PEHE) (lower is better), depending on the strength of balancing, α. In both cases, we show the behavior in a finite-sample vs. asymptotic regime ($n \to \infty$). The plots highlight the effectiveness of the *OR-learners* in the asymptotic regime, especially when too much balancing is applied.

E DATASET DETAILS

E.1 Synthetic Dataset

We use a synthetic benchmark dataset with hidden confounding as proposed by Kallus et al. (2019), but modify it by incorporating the confounder as the second observed covariate. Specifically, synthetic covariates X_1 and X_2 along with treatment A and outcome Y are generated by the following data-generating process:

$$\begin{cases} X_1 \sim \text{Unif}(-2, 2), \\ X_2 \sim N(0, 1), \\ A \sim \text{Bern}\left(\frac{1}{1+\exp(-(0.75\,X_1 - X_2 + 0.5))}\right) \\ Y \sim N\big((2\,A-1)\,X_1 + A - 2\,\sin(2\,(2\,A-1)\,X_1 + X_2) - 2\,X_2\,(1+0.5\,X_1), 1\big), \end{cases} \tag{68}$$

where X_1, X_2 are mutually independent.

E.2 IHDP Dataset

The Infant Health and Development Program (IHDP) dataset (Hill, 2011; Shalit et al., 2017) is a widely-used semi-synthetic benchmark for evaluating treatment effect estimation methods. It consists of 100 train/test splits, with $n_{\text{train}} = 672$, $n_{\text{test}} = 75$, and $d_x = 25$. However, this dataset suffers from significant overlap violations, leading to instability in methods that rely on inverse propensity weights (Curth and van der Schaar, 2021b; Curth et al., 2021).

E.3 ACIC 2016 Dataset Collection

The covariates for ACIC 2016 (Dorie et al., 2019) are derived from a large-scale study on developmental disorders (Niswander, 1972). The datasets in ACIC 2016 vary in the number of true confounders, the degree of overlap, and the structure of conditional outcome distributions. ACIC 2016 features 77 distinct data-generating mechanisms, each with 100 equal-sized samples ($n = 4802, d_x = 82$) after one-hot encoding the categorical covariates.

E.4 HC-MNIST Dataset

The HC-MNIST benchmark was introduced as a high-dimensional, semi-synthetic dataset (Jesson et al., 2021), derived from the original MNIST digit images (LeCun, 1998). It consists of $n_{\text{train}} = 60,000$ training images and $n_{\text{test}} = 10,000$ test images. Each image in HC-MNIST is compressed into a single latent coordinate, ϕ, such that the potential outcomes are non-linear functions of both the image's mean pixel intensity and its digit label. Treatment assignment then depends on this one-dimensional summary ϕ along with an additional latent (synthetic) confounder U, which is treated as an observed covariate. Specifically, the HC-MNIST dataset can be described by the following data-generating process:

$$\begin{cases} U \sim \text{Bern}(0.5), \\ X \sim \text{MNIST-image}(\cdot), \\ \phi := \left(\text{clip}\left(\frac{\mu_{N_x} - \mu_c}{\sigma_c}; -1.4, 1.4\right) - \text{Min}_c\right) \frac{\text{Max}_c - \text{Min}_c}{1.4 - (-1.4)}, \\ \alpha(\phi; \Gamma^*) := \frac{1}{\Gamma^* \,\text{sigmoid}(0.75\phi + 0.5)} + 1 - \frac{1}{\Gamma^*}, \quad \beta(\phi; \Gamma^*) := \frac{\Gamma^*}{\text{sigmoid}(0.75\phi + 0.5)} + 1 - \Gamma^*, \\ A \sim \text{Bern}\left(\frac{u}{\alpha(\phi;\Gamma^*)} + \frac{1-u}{\beta(\phi;\Gamma^*)}\right), \\ Y \sim N\big((2A-1)\phi + (2A-1) - 2\sin(2(2A-1)\phi) - 2(2U-1)(1+0.5\phi), 1\big), \end{cases} \tag{69}$$

where c is a label of the digit from the sampled image X; μ_{N_x} is the average intensity of the sampled image; μ_c and σ_c are the mean and standard deviation of the average intensities of the images with the label c; and $\text{Min}_c = -2 + \frac{4}{10}c, \text{Max}_c = -2 + \frac{4}{10}(c+1)$. The parameter Γ^* defines what factor influences the treatment assignment to a larger extent, i.e., the additional confounder or the one-dimensional summary. We set $\Gamma^* = \exp(1)$. For further details, we refer to Jesson et al. (2021).

F IMPLEMENTATION DETAILS AND HYPERPARAMETERS

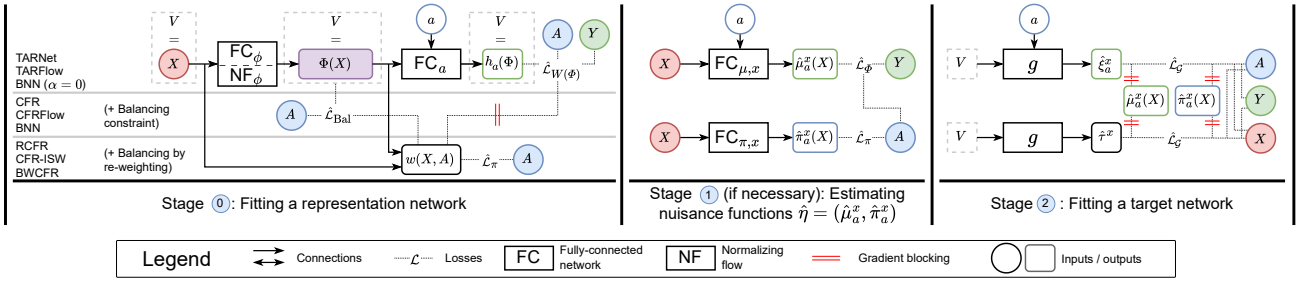

Figure 6: **An overview of the *OR-learners*.** The *OR-learners* proceed in three stages: 0 fitting a representation network, 1 estimation of the nuisance functions, and 2 fitting a target network. For stage 0, we also show different options for the target network input V. Depending on the choice of the input V, the second-stage model $g(V)$ obtains different interpretations: it either learns a new model from scratch or performs a calibration of the representation network outputs.

Overview. The *OR-learners* use neural networks to fit a target model g based on the learned representations $\hat{\Phi}(X)$. They proceed in three stages (see Fig. 6): 0 fitting a representation network; 1 estimating nuisance functions (if necessary); and 2 fitting a target network. The pseudocode is in Algorithm 1. Therein: In stage 0, the *representation network* consists of either (a) a fully-connected (FC_ϕ) or a normalizing flow (NF_ϕ) representation subnetwork, and (b) a fully-connected (FC_a) outcome subnetwork. Here, *any* representation learning method can be used, and, depending on the method, additional components might be added (e. g., a propensity subnetwork for CFR-ISW). Then, in stage 1, we might need to additionally fit nuisance functions (e. g., when the constrained representations were used in stage 0, so that $\hat{\mu}_a^\phi$ is inconsistent wrt. $\hat{\mu}_a^x$). Therein, we might optionally employ two additional networks, namely, a *propensity network* $\text{FC}_{\pi,x}$ and an *outcome network* $\text{FC}_{\mu,x}$. Finally, in stage 2, we utilize different DR- and R-losses, as presented in Sec. 3, to fit a fully-connected *target network* g and thus yield a final estimator of CAPOs/CATE.

Implementation. We implemented the *OR-learners* in PyTorch and Pyro. For better compatibility, the fully-connected subnetworks have one hidden layer with a tunable number of units and a RELU activation function. For the representation subnetworks involving normalizing flows, we employed residual normalizing flows (Chen et al., 2019) that have three hidden layers with a tunable synchronous number of units. All the networks for the *OR-learners* (see stages 0–2 in Fig. 6) are trained with AdamW (Loshchilov and Hutter, 2019). Each network was trained with $n_{\text{epoch}} = 200$ epochs for the synthetic dataset and $n_{\text{epoch}} = 50$ for the ACIC 2016 dataset collection. To further stabilize training of the target networks in stage 2, we (i) used exponential moving average (EMA) of model weights (Polyak and Juditsky, 1992) with a smoothing hyperparameter ($\lambda = 0.995$); and (ii) clipped too low propensity scores ($\hat{\pi}_a^x(X) < 0.05$).

Hyperparameters. We performed hyperparameter tuning of the *OR-learners* (at stages 0 and 1) and other non-neural Neyman-orthogonal learners (at stage 1) based on five-fold cross-validation using the training subset. At both stages, we did a random grid search with respect to different tuning criteria. For the final stage 2, on the other hand, we used fixed hyperparameters for all the experiments, as an exact hyperparameter search is not possible for target models solely with the observational data (Curth and van der Schaar, 2023). Table 7 provides all the details on hyperparameter tuning. For reproducibility, we made the tuned hyperparameters available in our GitHub.[6]

[6] `https://github.com/Valentyn1997/OR-learners`.

Algorithm 1 Pseudocode of the *OR-learners*

Input: Training dataset $\mathcal{D}$; strength of the balancing constraint $\alpha \geq 0$; dist $\in \{\text{WM}, \text{MMD}\}$
Stage 0: Fit a representation network $\in$ {TARNet/TARFlow, CFR/CFRFlow, RCFR/RCFRFlow, BNN/BNNFlow, CFR-ISW/CFRFlow-ISW, BWCFR/BWCFRFlow}
  if Representation network $\in$ {BWCFR/BWCFRFlow} **then**
    Fit a propensity network ($\text{FC}_{\pi,x}$) by minimizing a BCE loss $\hat{\mathcal{L}}_\pi$ and set $\hat{\pi}_a^x(X) \leftarrow \text{FC}_{\pi,x}(X)$
  end if
  for $i = 0$ **to** $\lceil n_{\text{epochs}} \cdot n/b_{\text{R}} \rceil$ **do**
    Draw a minibatch $\mathcal{B} = \{X, A, Y\}$ of size b_{R} from $\mathcal{D}$
    Initialize: $W \leftarrow \mathbb{1}_{b_R}$; $\hat{\mathcal{L}}_\pi \leftarrow 0$; $\hat{\mathcal{L}}_{\text{Bal}} \leftarrow 0$
    $\Phi \leftarrow \text{NF}_\phi$ / $\text{FC}_\phi(X)$
    $h_a(\Phi) \leftarrow \text{FC}_a(\Phi, a)$
    if Representation network $\in$ {CFR-ISW/CFRFlow-ISW} **then**
      $\hat{\pi}_a^\phi(\Phi) \leftarrow \text{FC}_{\pi,\phi}(\text{detach}(\Phi))$
      $\hat{\mathcal{L}}_\pi \leftarrow \text{BCE}(\hat{\pi}_A^\phi(\Phi), A)$
      $W \leftarrow \text{detach}\left(\mathbb{1}\{\hat{\pi}_A^\phi(\Phi) \geq 0.05\}/\hat{\pi}_A^\phi(\Phi)\right)$
    else if Representation network $\in$ {BWCFR/BWCFRFlow} **then**
      $W \leftarrow \mathbb{1}\{\hat{\pi}_A^x(X) \geq 0.05\}/\hat{\pi}_A^x(X)$
    else if Representation network $\in$ {RCFR/RCFRFlow} **then**
      $W \leftarrow \text{FC}_w(\text{detach}(\Phi))$
    end if
    $\hat{\mathcal{L}}_{W(\Phi)} \leftarrow \mathbb{P}_{b_R}\{W(Y - h_A(\Phi(X)))^2\}/\mathbb{P}_{b_R}\{W\}$
    if Representation network $\notin$ {TARNet/TARFlow} and $\alpha > 0$ **then**
      $\hat{\mathcal{L}}_{\text{Bal}} \leftarrow W$-weighted $\widehat{\text{dist}}(\mathbb{P}(\Phi(X) \mid A = 0), \mathbb{P}(\Phi(X) \mid A = 1))$
    end if
    Gradient update of the representation network wrt. $\hat{\mathcal{L}}_{W(\Phi)} + \alpha\hat{\mathcal{L}}_{\text{Bal}} + \hat{\mathcal{L}}_\pi$
  end for
  $V \leftarrow X \,/\, \hat{\Phi}(X) \,/\, (\hat{h}_0(\hat{\Phi}(X)), \hat{h}_1(\hat{\Phi}(X)))$
Stage 1: Estimate nuisance functions $\hat{\eta} = (\hat{\mu}_a^x, \hat{\pi}_a^x)$
  if Representation network $\notin$ {BWCFR/BWCFRFlow} **then**
    Fit a propensity network ($\text{FC}_{\pi,x}$) by minimizing a BCE loss $\hat{\mathcal{L}}_\pi$ and set $\hat{\pi}_a^x(X) \leftarrow \text{FC}_{\pi,x}(X)$
  end if
  if $\alpha > 0$ and FC_ϕ is used at stage 0 **then**
    Fit an outcome network ($\text{FC}_{\mu,x}$) by minimizing an unweighted MSE loss $\hat{\mathcal{L}}_\Phi$ and set $\hat{\mu}_a^x(X) \leftarrow \text{FC}_{\mu,x}(X, a)$
  else
    Set $\hat{\mu}_a^x(X) \leftarrow \hat{\mu}_a^\phi(\Phi(X))$
  end if
Stage 2: Fit a target network $\hat{g} = \arg\min \hat{\mathcal{L}}_\diamond(g, \hat{\eta})$
  for $i = 0$ **to** $\lceil n_{\text{epochs}} \cdot n/b_{\text{T}} \rceil$ **do**
    Draw a minibatch $\mathcal{B} = \{X, A, Y\}$ of size b_{T} from $\mathcal{D}$
    $\alpha_a(A, X) \leftarrow \mathbb{1}\{A = a\} \cdot \mathbb{1}\{\hat{\pi}_a^x(X) \geq 0.05\}/\hat{\pi}_a^x(X)$
    if Causal quantity == CAPO **then**
      $\hat{\mathcal{L}}_{\mathcal{G}}^{\text{DR}_a^{\text{K}}}(g, \hat{\eta}) \leftarrow \mathbb{P}_{b_{\text{T}}}\left\{\left(\alpha_a(A, X)\left(Y - \hat{\mu}_a^x(X)\right) + \hat{\mu}_a^x(X) - g(V)\right)^2\right\}$
      $\hat{\mathcal{L}}_{\mathcal{G}}^{\text{DR}_a^{\text{FS}}}(g, \hat{\eta}) \leftarrow \mathbb{P}_{b_{\text{T}}}\left\{\alpha_a(A, X)\left(Y - g(V)\right)^2 + \left(1 - \alpha_a(A, X)\right)\left(\hat{\mu}_a^x(X) - g(V)\right)^2\right\}$
    end if
    if Causal quantity == CATE **then**
      $\hat{\mathcal{L}}_{\mathcal{G}}^{\text{DR}^{\text{K}}}(g, \hat{\eta}) \leftarrow \mathbb{P}_{b_{\text{T}}}\left\{\left(\alpha_0(A, X)\left(Y - \hat{\mu}_0^x(X)\right) + \alpha_1(A, X)\left(Y - \hat{\mu}_1^x(X)\right) + \hat{\mu}_1^x(X) - \hat{\mu}_0^x(X) - g(V)\right)^2\right\}$
      $\hat{\mathcal{L}}_{\mathcal{G}}^{\text{R}}(g, \hat{\eta}) \leftarrow \mathbb{P}_{b_{\text{T}}}\left\{\left(\left(Y - \hat{\mu}^x(X)\right) - \left(A - \hat{\pi}_1^x(X)\right)g(V)\right)^2\right\}$
      $\hat{\mathcal{L}}_{\mathcal{G}}^{\text{IVW}}(g, \hat{\eta}) \leftarrow \mathbb{P}_{b_{\text{T}}}\left\{\left(A - \hat{\pi}_1^x(X)\right)^2\left(\left(\alpha_0(A, X)\left(Y - \hat{\mu}_0^x(X)\right) + \alpha_1(A, X)\left(Y - \hat{\mu}_1^x(X)\right) + \hat{\mu}_1^x(X) - \hat{\mu}_0^x(X) - g(V)\right)^2\right\}$
    end if
    Gradient & EMA update of the target network g wrt. $\hat{\mathcal{L}}_{\mathcal{G}}(g, \hat{\eta})$
  end for
Output: V-level estimator $\hat{g}$ for CAPOs/CATE

Table 7: Hyperparameter tuning for the *OR-learners* and other baselines.

Stage	Model	Hyperparameter	Range / Value
Stage ⓪	TARNet/TARFlow BNN/BNNFlow CFR/CFRFlow BWCFR/BWCFRFlow	Learning rate	0.001, 0.005, 0.01
		Minibatch size, b_R	32, 64, 128
		Weight decay	0.0, 0.001, 0.01, 0.1
		Hidden units in NF_ϕ / FC_ϕ	$R\,d_x$, 1.5 Rd_x, 2 Rd_x
		Hidden units in FC_a	$R\,d_\phi$, 1.5 Rd_ϕ, 2 Rd_ϕ
		Tuning strategy	random grid search with 50 runs
		Tuning criterion	factual MSE loss
		Optimizer	AdamW
	CFR-ISW/CFRFlow-ISW	Representation network learning rate	0.001, 0.005, 0.01
		Propensity network learning rate	0.001, 0.005, 0.01
		Minibatch size, b_R	32, 64, 128
		Representation network weight decay	0.0, 0.001, 0.01, 0.1
		Propensity network weight decay	0.0, 0.001, 0.01, 0.1
		Hidden units in NF_ϕ / FC_ϕ	$R\,d_x$, 1.5 Rd_x, 2 Rd_x
		Hidden units in FC_a	$R\,d_\phi$, 1.5 Rd_ϕ, 2 Rd_ϕ
		Hidden units in $\text{FC}_{\pi,\phi}$	$R\,d_\phi$, 1.5 Rd_ϕ, 2 Rd_ϕ
		Tuning strategy	random grid search with 50 runs
		Tuning criterion	factual MSE loss + factual BCE loss
		Optimizer	AdamW
	RCFR/RCFRFlow	Learning rate	0.001, 0.005, 0.01
		Minibatch size, b_R	32, 64, 128
		Weight decay	0.0, 0.001, 0.01, 0.1
		Hidden units in NF_ϕ / FC_ϕ	$R\,d_x$, 1.5 Rd_x, 2 Rd_x
		Hidden units in FC_a	$R\,d_\phi$, 1.5 Rd_ϕ, 2 Rd_ϕ
		Hidden units in FC_w	$R\,d_\phi$, 1.5 Rd_ϕ, 2 Rd_ϕ
		Tuning strategy	random grid search with 50 runs
		Tuning criterion	factual MSE loss
		Optimizer	AdamW
Stage ①	Propensity network	Learning rate	0.001, 0.005, 0.01
		Minibatch size, b_N	32, 64, 128
		Weight decay	0.0, 0.001, 0.01, 0.1
		Hidden units in $\text{FC}_{\pi,x}$	$R\,d_x$, 1.5 Rd_x, 2 Rd_x
		Tuning strategy	random grid search with 50 runs
		Tuning criterion	factual BCE loss
		Optimizer	AdamW
	Outcome network	Learning rate	0.001, 0.005, 0.01
		Minibatch size, b_N	32, 64, 128
		Hidden units in $\text{FC}_{\mu,x}$	$R\,d_x$, 1.5 Rd_x, 2 Rd_x
		Weight decay	0.0, 0.001, 0.01, 0.1
		Tuning strategy	random grid search with 50 runs
		Tuning criterion	factual MSE loss
		Optimizer	AdamW
	XGBoost	Number of estimators, $n_{e,N}$	50, 100, 150
		Maximum depth	3, 6, 9, 12
		L_1 regularization, α	40, 80, 120, 160
		Minimum sum of instance weight in a child	0, 3, 6, 9
		Minimum loss reduction needed for a split	1, 3, 5, 7, 9
		Tuning strategy	random grid search with 50 runs
		Tuning criterion	factual MSE/BCE loss
Stage ②	Target network	Learning rate	0.005
		Minibatch size, b_T	64
		EMA of model weights, λ	0.995
		Hidden units in g	Hidden units in FC_a
		Tuning strategy	no tuning
		Optimizer	AdamW
	XGBoost	Number of estimators, $n_{e,T}$	$n_{e,N}$
		Maximum depth	6
		L_1 regularization, α	0
		Minimum sum of instance weight in a child	1
		Minimum loss reduction needed for a split	0

$R = 2$ (synthetic data), $R = 1$ (IHDP dataset), $R = 0.25$ (ACIC 2016 datasets collection)

G ADDITIONAL EXPERIMENTS

G.1 Setting ①

(i) Synthetic data. Table 8 shows additional results for the synthetic dataset in Setting ①. Therein, we observe that the *OR-learners* with $V = \hat{\Phi}(X)$ are as effective as other variants (e. g., $V = X/X^*$). This was expected as, in the synthetic dataset (Melnychuk et al., 2024), the ground-truth CAPOs/CATE densely depend on the covariates X ($d_x = 2$), and, thus, the low-manifold hypothesis (e. g., with $d_\phi = 1$) can not be assumed. Hence, all four variants of the *OR-learners* perform similarly well, and all the improvements can be attributed to the Neyman-orthogonality of the *OR-learners*.

Table 8: **Results for synthetic experiments in Setting ①.** Reported: improvements of the *OR-learners* over plug-in representation networks wrt. out-of-sample rMSE / rPEHE; mean ± std over 15 runs. Here, $n_{\text{train}} = 500, d_{\hat{\phi}} = 2$.

		DR_0^{K}	DR_0^{FS}	DR_1^{K}	DR_1^{FS}	DR^{K}	R	IVW
TARNet	$V = (\hat{\mu}_0^x, \hat{\mu}_1^x)$	-0.002 ± 0.011	-0.002 ± 0.016	-0.004 ± 0.006	-0.004 ± 0.006	-0.006 ± 0.012	-0.009 ± 0.013	-0.009 ± 0.017
	$V = X$	$+0.064 \pm 0.034$	$+0.083 \pm 0.051$	$+0.078 \pm 0.053$	$+0.059 \pm 0.037$	-0.018 ± 0.012	-0.021 ± 0.018	-0.021 ± 0.015
	$V = X^*$	$+0.015 \pm 0.022$	$+0.023 \pm 0.020$	$+0.015 \pm 0.031$	$+0.004 \pm 0.016$	-0.013 ± 0.017	-0.017 ± 0.018	-0.017 ± 0.016
	$V = \hat{\Phi}(X)$	-0.002 ± 0.013	$\pm 0.000 \pm 0.017$	-0.004 ± 0.007	-0.003 ± 0.006	-0.011 ± 0.012	-0.012 ± 0.009	-0.012 ± 0.013
BNN ($\alpha = 0.0$)	$V = (\hat{\mu}_0^x, \hat{\mu}_1^x)$	-0.006 ± 0.014	$+0.001 \pm 0.022$	-0.009 ± 0.009	-0.009 ± 0.009	-0.007 ± 0.009	-0.006 ± 0.010	-0.009 ± 0.013
	$V = X$	$+0.067 \pm 0.033$	$+0.101 \pm 0.039$	$+0.045 \pm 0.037$	$+0.037 \pm 0.034$	-0.020 ± 0.018	-0.023 ± 0.016	-0.022 ± 0.019
	$V = X^*$	$+0.011 \pm 0.017$	$+0.023 \pm 0.033$	-0.005 ± 0.016	-0.008 ± 0.015	-0.010 ± 0.035	-0.017 ± 0.018	-0.016 ± 0.025
	$V = \hat{\Phi}(X)$	-0.008 ± 0.012	-0.002 ± 0.016	-0.010 ± 0.013	-0.011 ± 0.011	-0.012 ± 0.011	-0.012 ± 0.013	-0.015 ± 0.014

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

(ii) HC-MNIST dataset. Table 9 shows the results for the HC-MNIST dataset. Here, we again report the absolute performance of different methods: non-neural Neyman-orthogonal learners instantiated with XGBoost (Chen et al., 2015) (with S-/T-learners for the first-stage models); plug-in representation learning methods; and the OR-learners used with the pre-trained representations $V = \hat{\Phi}(X)$. We see that the OR-learners improve over the other, non-neural Neyman-orthogonal learners: This is again not surprising as the manifold hypothesis holds for the HC-MNIST dataset. Furthermore, the *OR-learners* outperform the baseline plug-in representation learning methods given *sufficient overlap* in the HC-MNIST data: This contrasts with the IHDP dataset results (see Table 3), as the IHDP dataset contains extreme propensity scores (Curth et al., 2021).

Table 9: **Results for HC-MNIST experiments in Setting ①.** Reported: out-of-sample rMSE / rPEHE for different causal quantities (ξ_a^x / τ^x, respectively), mean ± std over 10 runs. Here, $d_{\hat{\phi}} = 78$ for neural baselines.

		ξ_0^x	ξ_1^x	τ^x
XGBoost S	DR^{K}	0.610 ± 0.000	0.499 ± 0.000	0.766 ± 0.000
	R	—	—	0.702 ± 0.000
	IVW	—	—	0.723 ± 0.000
XGBoost T	DR^{K}	0.591 ± 0.000	0.498 ± 0.000	0.763 ± 0.000
	R	—	—	0.706 ± 0.000
	IVW	—	—	0.726 ± 0.000
TARNet	Plug-in	0.514 ± 0.006	0.485 ± 0.003	0.697 ± 0.004
	DR^{K}	$\underline{0.503 \pm 0.003}$	$\underline{0.471 \pm 0.003}$	0.681 ± 0.004
	DR^{FS}	0.520 ± 0.051	$\underline{0.471 \pm 0.006}$	—
	R	—	—	0.684 ± 0.021
	IVW	—	—	0.682 ± 0.006
BNN ($\alpha = 0.0$)	Plug-in	0.526 ± 0.021	0.508 ± 0.023	0.700 ± 0.020
	DR^{K}	0.507 ± 0.010	0.474 ± 0.006	0.680 ± 0.010
	DR^{FS}	$\mathbf{0.497 \pm 0.005}$	$\mathbf{0.468 \pm 0.004}$	—
	R	—	—	$\mathbf{0.673 \pm 0.008}$
	IVW	—	—	$\underline{0.678 \pm 0.009}$
Oracle		0.363	0.365	0.513

Lower = better. Best in **bold**, second best underlined

G.2 Setting ②

(i) Synthetic data. We report additional results for the synthetic dataset in Fig. 7 and Table 10 for invertible and non-invertible representations, respectively. Fig. 7 empirically demonstrates our intuition from Fig. 5 of Appendix D. Specifically, as the data size grows, the *OR-learners* manage to correct the RICB of the baseline representation learning methods more effectively (even for large values of the balancing strength α). Notably, for this dataset, the balancing constraint acts detrimentally as the heterogeneity of the CAPOs/CATE is high for both low and high overlap regions of the covariate space (thus, the underlying inductive bias "low overlap – low heterogeneity" cannot be assumed). Table 10 then provides similar evidence for the non-invertible representations. Here, the *OR-learners* achieve significant improvements in the majority of the cases (and no significant worsenings).

In Fig. 8, we additionally show how the learned normalizing flows transform the original space $\mathcal{X}$ (the models are the same as in Fig. 3). The rendered transformations match the theoretical results provided in Sec. 4.2. Specifically, TARFlow expands (scales up) the original space so that the regression task becomes easier in the representation space. At the same time, CRFFlows with different balancing hyperparameters α aim to contract (scale down) the space, thus achieving better balancing.

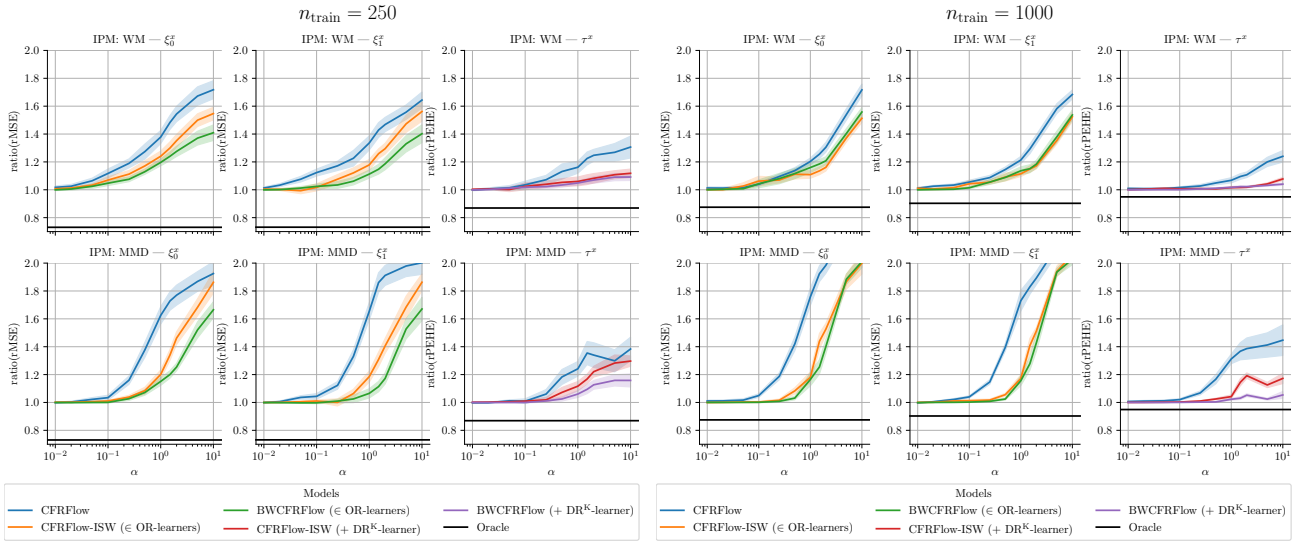

Figure 7: **Results for synthetic data in Setting ②.** Reported: ratio between the performance of TARFlow (CFRFlow with $\alpha = 0$) and invertible representation networks with varying α; mean ± SE over 15 runs. Lower is better. Here: $n_{\text{train}} \in \{250, 1000\}$, $d_{\hat{\phi}} = 2$.

Table 10: **Results for synthetic experiments in Setting ②.** Reported: improvements of the *OR-learners* over non-invertible plug-in / IPTW representation networks wrt. out-of-sample rMSE / rPEHE; mean ± std over 15 runs. Here, $n_{\text{train}} = 500, d_{\hat{\phi}} = 2$.

	DR_0^{K}	DR_0^{FS}	DR_1^{K}	DR_1^{FS}	DR^{K}	R	IVW
CFR (MMD; $\alpha = 0.1$)	-0.006 ± 0.024	-0.005 ± 0.026	-0.009 ± 0.026	-0.014 ± 0.022	-0.011 ± 0.039	-0.017 ± 0.032	-0.012 ± 0.042
CFR (WM; $\alpha = 0.1$)	-0.003 ± 0.016	-0.006 ± 0.014	-0.005 ± 0.010	-0.006 ± 0.005	-0.001 ± 0.023	-0.005 ± 0.023	-0.004 ± 0.027
BNN (MMD; $\alpha = 0.1$)	-0.058 ± 0.047	-0.051 ± 0.046	-0.011 ± 0.012	-0.006 ± 0.018	-0.048 ± 0.041	-0.038 ± 0.043	-0.039 ± 0.040
BNN (WM; $\alpha = 0.1$)	$+0.016 \pm 0.101$	-0.013 ± 0.035	-0.005 ± 0.037	$+0.007 \pm 0.043$	-0.026 ± 0.042	-0.026 ± 0.041	-0.025 ± 0.041
RCFR (MMD; $\alpha = 0.1$)	-0.010 ± 0.086	-0.032 ± 0.034	-0.012 ± 0.019	-0.012 ± 0.020	-0.040 ± 0.043	-0.028 ± 0.038	-0.034 ± 0.042
RCFR (WM; $\alpha = 0.1$)	-0.008 ± 0.020	-0.009 ± 0.019	-0.003 ± 0.015	-0.006 ± 0.015	-0.019 ± 0.021	-0.015 ± 0.022	-0.019 ± 0.022
CFR-ISW (MMD; $\alpha = 0.1$)	$+0.002 \pm 0.025$	-0.003 ± 0.016	-0.002 ± 0.009	-0.008 ± 0.008	$+0.001 \pm 0.023$	-0.002 ± 0.014	-0.001 ± 0.017
CFR-ISW (WM; $\alpha = 0.1$)	$+0.001 \pm 0.029$	-0.006 ± 0.017	-0.004 ± 0.018	-0.003 ± 0.029	-0.009 ± 0.017	-0.008 ± 0.014	-0.010 ± 0.016
BWCFR (MMD; $\alpha = 0.1$)	$+0.007 \pm 0.079$	-0.003 ± 0.018	-0.005 ± 0.014	-0.003 ± 0.012	-0.015 ± 0.016	-0.017 ± 0.012	-0.018 ± 0.011

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

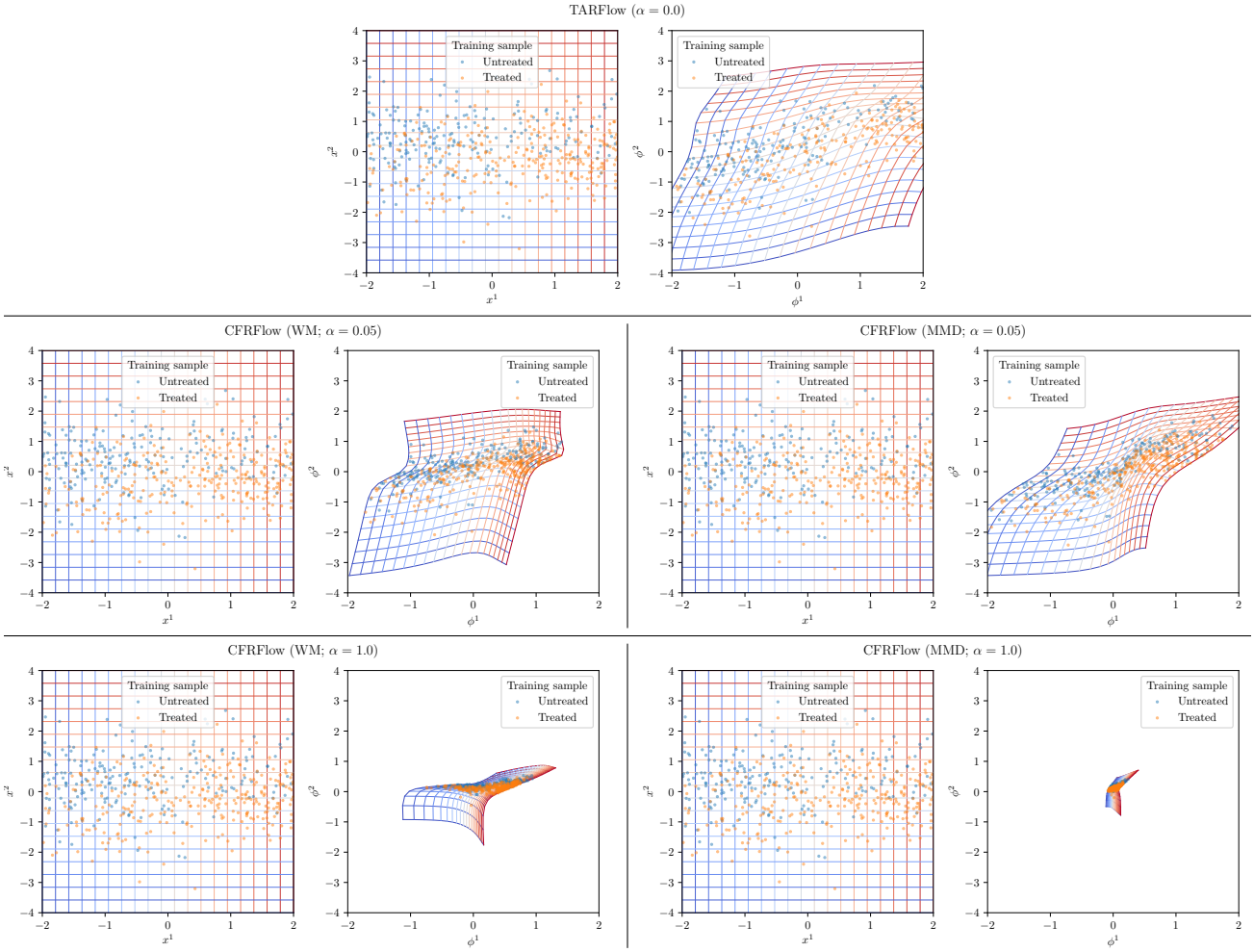

Figure 8: Visualization of the invertible transformations defined by the learned normalizing flow representation subnetworks for synthetic experiments in Setting ②. Here, $n_{\text{train}} = 500, d_{\hat{\phi}} = 2$. Specifically, we show how a grid in the original covariate space, $\mathcal{X} \subseteq \mathbb{R}^2$, gets transformed onto the representation space, $\hat{\Phi} \subseteq \mathbb{R}^2$. We vary the strength of balancing $\alpha \in \{0, 0.05, 1.0\}$ and the IPM $\in \{\text{WM}, \text{MMD}\}$. As suggested by the theory in Sec. 4.2, the covariate space gets *expanded* for $\alpha = 0$ and gets *contracted* for large values of α (e. g., $\alpha = 1$).

(ii) IHDP dataset. Fig. 9 shows the results for the IHDP dataset in Setting ② for invertible representations. Here, interestingly, balancing in CFRFlow seems to outperform the *OR-learners* for some values of α. This is not surprising, as the IHDP dataset contains strong overlap violations and one of the ground-truth CAPOs is linear (namely, ξ_1^x). Hence, the "low overlap – low heterogeneity" inductive bias partially holds for this dataset. Note, however, that the optimal α is different for both CAPOs and CATE, which renders balancing impractical (considering its value cannot be reliably tuned just with the observational data).

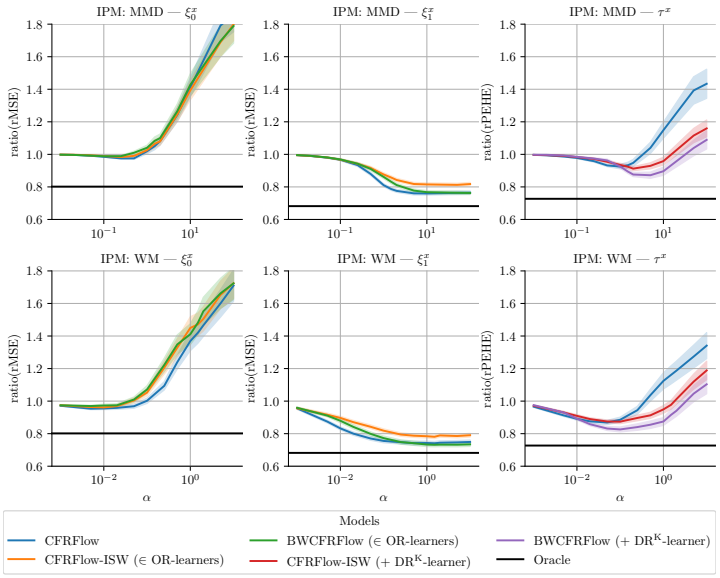

Figure 9: **Results for IHDP experiments in Setting ②.** Reported: ratio between the performance of TARFlow (CFRFlow with $\alpha = 0$) and invertible representation networks with varying α; mean ± SE over 100 train/test splits. Lower is better. Here: $d_{\hat{\phi}} = 12$.

(iii) HC-MNIST dataset. Finally, in Table 11 we showcase the results of the HC-MNIST experiments for non-invertible representations with balancing. Here, the *OR-learners* significantly improve over the baseline representation learning methods with the balancing constraint in the majority of the cases (and do not significantly worsen them). This was expected, as the ground-truth CAPOs for the HC-MNIST dataset are highly heterogeneous regardless of the overlap.

Table 11: **Results for HC-MNIST experiments in Setting ②.** Reported: improvements of the *OR-learners* over non-invertible plug-in / IPTW representation networks wrt. out-of-sample rMSE / rPEHE; mean ± std over 10 runs. Here, $d_{\hat{\phi}} = 78$.

	DR_0^{K}	DR_1^{K}	DR_0^{FS}	DR_1^{FS}	DR^{K}	IVW	R
CFR (MMD; $\alpha = 0.1$)	-0.025 ± 0.007	-0.011 ± 0.004	-0.015 ± 0.008	-0.014 ± 0.005	-0.023 ± 0.009	-0.022 ± 0.008	-0.016 ± 0.009
CFR (WM; $\alpha = 0.1$)	-0.062 ± 0.010	-0.021 ± 0.006	-0.049 ± 0.011	-0.018 ± 0.006	-0.035 ± 0.008	-0.027 ± 0.008	-0.024 ± 0.008
BNN (MMD; $\alpha = 0.1$)	-0.069 ± 0.174	-0.061 ± 0.132	-0.080 ± 0.175	-0.068 ± 0.130	-0.025 ± 0.059	-0.023 ± 0.069	-0.035 ± 0.055
BNN (WM; $\alpha = 0.1$)	-0.064 ± 0.012	-0.024 ± 0.004	-0.054 ± 0.011	-0.022 ± 0.003	-0.050 ± 0.018	-0.043 ± 0.016	-0.040 ± 0.018
RCFR (MMD; $\alpha = 0.1$)	-0.106 ± 0.039	-0.040 ± 0.018	$+0.044 \pm 0.387$	-0.042 ± 0.020	-0.090 ± 0.039	-0.056 ± 0.073	$+0.013 \pm 0.178$
RCFR (WM; $\alpha = 0.1$)	-0.405 ± 0.458	-0.178 ± 0.384	-0.406 ± 0.447	-0.148 ± 0.406	-0.233 ± 0.345	-0.225 ± 0.343	-0.210 ± 0.359
CFR-ISW (MMD; $\alpha = 0.1$)	-0.010 ± 0.007	-0.003 ± 0.004	-0.006 ± 0.006	-0.007 ± 0.002	-0.007 ± 0.008	-0.004 ± 0.010	-0.008 ± 0.008
CFR-ISW (WM; $\alpha = 0.1$)	-0.019 ± 0.009	-0.014 ± 0.007	-0.007 ± 0.017	-0.013 ± 0.007	-0.024 ± 0.005	-0.022 ± 0.006	-0.021 ± 0.005
BWCFR (MMD; $\alpha = 0.1$)	-0.008 ± 0.019	$+0.043 \pm 0.163$	-0.012 ± 0.006	-0.011 ± 0.004	$+0.005 \pm 0.063$	-0.013 ± 0.011	$+0.020 \pm 0.126$

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

G.3 Runtime

Table 12 provides the runtime comparison of different models from different stages of the *OR-learners*. Here, the *OR-learners* are well scalable.

Training stage	Model	Variant	Average duration of a training iteration (in ms)
Stage ⓪	TARNet/TARFlow	—	$\approx$ 9.3/85.4
	BNN/BNNFlow	MMD WM	$\approx$ 18.1/94.9 $\approx$ 15.6/166.1
	CFR/CFRFlow	MMD WM	$\approx$ 20.7/115.2 $\approx$ 21.7/174.0
	BWCFR/BWCFRFlow	MMD WM	$\approx$ 19.4/102.0 $\approx$ 19.7/101.1
	CFR-ISW/CFRFlow-ISW	MMD WM	$\approx$ 21.2/103.0 $\approx$ 23.1/116.0
	RCFR/RCFRFlow	MMD WM	$\approx$ 27.9/101.7 $\approx$ 27.1/127.7
Stage ①	Propensity network	—	$\approx$ 5.2
	Outcome network	—	$\approx$ 6.7
Stage ②	Target network	—	$\approx$ 6.5

Table 12: Total runtime (in milliseconds) for different models at stages ⓪-② of the *OR-learners*. Reported: average duration of a training iteration for the IHDP dataset (lower is better). Experiments were carried out on 2 GPUs (NVIDIA A100-PCIE-40GB) with IntelXeon Silver 4316 CPUs @ 2.30GHz.