Cost-Aware Optimized Front-Door Experimental Design

Leopold Mareis LEOPOLD.MAREIS@TUM.DE
Technical University of Munich, Germany

Mathias Drton MATHIAS.DRTON@TUM.DE
Technical University of Munich, Germany
Munich Center for Machine Learning, Munich, Germany

Editors: Bijan Mazaheri and Niels Richard Hansen

Abstract

Causal effect estimation often succeeds cost-constrained sequential data collection. This work considers multivariate linear front-door models with arbitrary unobserved confounding on treatment and response. We optimize the experimental design by balancing the statistical efficiency and measurement costs through partial data. The full-data efficient influence function for the causal effect is derived, together with the geometry of all observed-data influence functions. This characterization yields a closed-form optimal sampling policy and an estimator to minimize the asymptotic variance of regular asymptotically linear (RAL) estimators within a class of augmented full-data influence functions. The resulting design also covers back-door estimation. In simulations and applications to biological, medical, and industrial datasets, the optimized designs achieve substantial efficiency gains (5.3% to 31.9%) over naive full-sampling strategies.

Keywords: Causal inference, experimental design, front-door estimation, linear SEM, semiparametric estimation

1. Introduction

Estimating the linear causal effect between a treatment X_t and a response X_r from observational data under unobserved confounding is a central problem in causal inference, with applications ranging from pharmacology, medicine, economics, to social sciences. This paper addresses a problem that arises naturally in practice: how to allocate a limited measurement budget across staged sampling designs with intermediate costs so as to minimize the variance of a front-door causal effect estimator. We combine semiparametric efficiency theory and causal graphical modelling to derive RAL estimators and optimized experimental designs that achieve substantial efficiency gains over naive full-measurement strategies without sacrificing inferential validity.

A helpful way to view the setting is as a statistician planning a medical study. The goal is to estimate the effect of a medical procedure, but collecting additional biological measurements per patient is costly, risky or time intensive. As even partial data can be informative, the experimenter must decide which patients to measure more extensively and how to allocate limited measurement resources while maintaining statistical efficiency. In our framework, baseline covariate information allows the experimenter to make patient-specific probabilistic decisions about whether to measure mediators and, conditional on that, whether to measure the response. These decisions are based on information already available at the time of sampling. Importantly, the design operates at the population level: while every unit contributes information in expectation, it is not always optimal to collect all measurements for every individual.

We study these issues in a multivariate linear front-door model defined by structural equations

$$X = \beta X + \varepsilon, \quad X = (I - \beta)^{-1}\varepsilon,$$

where the lower-triangular parameter matrix β encodes linear causal relationships between the confounder, treatment, mediator, and response variables $X := X_{C,t,M,r}$ as visualized in Figure 1. To reflect the sequential data collection, we introduce coarsening levels $X_{C,t}$, $X_{C,t,M}$, $X_{C,t,M,r}$ with known propensities π, resulting in an observed-data distribution $\mathcal{P}_{\beta,\varepsilon,\pi}$.

In well-behaved statistical models all sensible consistent estimators, including the MLE, are regular and asymptotically linear (RAL). Our work builds on several strands of RAL literature. Robins et al. (1994) and Robins and Rotnitzky (1995) established semiparametric efficiency of parameter estimates in multivariate regression models under coarsening, but were limited to single-stage regressions. Over the years, this single-stage problem has been optimized with recent advances in automatic bias corrections using learned Riesz representers (Tan, 2010; Hines and Hines, 2025). Utilizing the graphical structure in acyclic directed mixed graphs (ADMGs), Bhattacharya et al. (2020) and Jung et al. (2021) developed efficient, doubly robust full-data estimators targeting interventional distributions (Hines et al., 2022; Chernozhukov et al., 2017; van der Vaart, 2014). In contrast, we impose a linear structure, further restricting the semiparametric tangent space and allowing for practical estimation without non-parametric density estimation. Unifying graphical models and missing data theory, Mohan and Pearl (2021) solved the distributional identification problem, allowing for reinforced causal effect estimation (Seitzer et al., 2021). Related work includes the search for back-door adjustment sets that minimize the asymptotic variance of linear regression estimators (Witte et al., 2020; Drton et al., 2011), as well as the optimization of efficient linear regression within instrumental-variable experimental designs (Mareis, 2025).

Contribution The primary contributions are the following.

- **Theorem 6**: We derive the efficient influence function $\varphi_\xi^{F,\text{eff}}$ for the causal effect ξ in the full-data model $\mathcal{M}_1$. This function characterizes the minimum achievable asymptotic variance among all RAL estimators when complete observations are available. It decomposes into two orthogonal components that separately and efficiently target the treatment-to-mediator parameter β_{Mt} and the mediator-to-response parameter β_{rM}.

- **Theorem 11:** Given $\varphi_\xi^{F,\text{eff}}$, we characterize the geometry of all observed-data influence functions in $\mathcal{M}_\pi$ and identify the optimal augmentation within the class $IF^\pi(\varphi^{F,\text{eff}})$; see Lemma 7. Theorem 11 yields a closed-form optimized propensity π^* that solves the constrained minimization of asymptotic variance over all RAL estimators within $IF^\pi(\varphi^{F,\text{eff}})$ and all sampling regimes under a fixed expected budget b_0. The optimal propensity balances the statistical efficiency of the intermediate estimators $\hat{\beta}_{Mt,n}$ and $\hat{\beta}_{rM,n}$ against measurement cost. The optimized design for back-door estimation follows as a corollary.

In a simulation study and in applications to biological, medical and industrial datasets, we demonstrate substantial efficiency gains of 5.3% to 31.9% compared to the naive full-sampling design in Section 5. To ensure reproducibility and support practical applications through step-by-step instructions, the accompanying code is available at https://doi.org/10.5281/zenodo.18960268.

2. Multivariate Front-Door Model

We begin by introducing the linear multivariate front-door model. Let $(\Omega, \mathcal{F}, \mathcal{P})$ be a probability space, and consider the Hilbert space $\mathcal{H} = L_0^2(\Omega; \mathbb{R}^d)$ consisting of all d dimensional mean-zero, finite variance functions $h : \Omega \to \mathbb{R}^d$, with its inner product $\mathrm{E}\big[h_1^\top h_2\big]$. Let $\mathcal{C}(A, B)$ denote the space of continuous functions from A to B.

Definition 1 *Fix integers $d_C, d_M \in \mathbb{N}$ and let $d := d_C + 1 + d_M + 1$. We partition the vector $x \in \mathbb{R}^d$ as $x = (x_C, x_t, x_M, x_r)$. The parameter set of admissible coefficient matrices is*

$$\mathcal{B} := \left\{ \beta \in \mathbb{R}^{d\times d} : \beta = \begin{pmatrix} 0 & 0 & 0 & 0 \\ \beta_{tC} & 0 & 0 & 0 \\ \beta_{MC} & \beta_{Mt} & 0 & 0 \\ \beta_{rC} & 0 & \beta_{rM} & 0 \end{pmatrix} \right\}.$$

The nuisance parameter space $\mathcal{E}$ of additive noise terms is restricted by a factorization of the induced measures P_ε. In particular, it is defined as

$$\mathcal{E} := \left\{ \varepsilon \in \mathcal{H} \,:\, \begin{array}{c} \varepsilon \text{ absolutely continuous,} \\ \mathcal{P}_\varepsilon = \mathcal{P}_{\varepsilon_C} \otimes \mathcal{P}_{\varepsilon_{t,r}} \otimes \mathcal{P}_{\varepsilon_M} \end{array} \right\}. \tag{1}$$

The linear multivariate front-door model is then defined as the set of pushforward measures

$$\mathcal{M}_1 = \{\mathcal{P}_{\beta,\varepsilon} = (I - \beta)^{-1}\mathcal{P}_\varepsilon \,:\, \beta \in \mathcal{B}, \varepsilon \in \mathcal{E}\},$$

which satisfy the linear structural equation $X = \beta X + \varepsilon$, or equivalently $X = (I - \beta)^{-1}\varepsilon$.

This formulation allows for confounded errors ε_t and ε_r, as well as arbitrary confounding and dependencies within the subvectors X_C and X_M. As motivated in the introduction, we further permit the data to be generated by a known coarsening or, more precisely, missingness process, affecting the mediating, and treatment variables (X_M, X_t) in two stages. The *full-data* model $\mathcal{M}_1$ is particularly important in the subsequent derivation of efficient estimators.

Definition 2 *Let X follow model $\mathcal{M}_1$. For given propensity functions $\pi_1 \in \mathcal{C}(\mathbb{R}^{d_C+1}, (0, 1])$ and $\pi_2 \in \mathcal{C}(\mathbb{R}^{d_C+1+d_M}, (0, 1])$, let the random variable $\Delta : \Omega \mapsto \{1, 2, \infty\}$, with probabilities*

$$\begin{aligned} \mathcal{P}_\pi(\Delta = 1 \,|\, X) &= 1 - \pi_1(X_{C,t}), \qquad & \mathcal{P}_\pi(\Delta = 2 \,|\, X) &= \pi_1(X_{C,t})(1 - \pi_2(X_{C,t,M})), \\ & & \mathcal{P}_\pi(\Delta = \infty \,|\, X) &= \pi_1(X_{C,t})\pi_2(X_{C,t,M}) \end{aligned}$$

indicate which subset of X is observed. The observed-data distribution is then denoted by $\mathcal{P}_{\theta,\pi}$, with $\theta = (\beta, \varepsilon)$, and generates realizations

$$\left\{ \left(\Delta^{(i)}, G_{\Delta^{(i)}}(X^{(i)}) \right) \right\}_{i=1,\dots,n}, \quad \textit{where} \quad G_{\Delta^{(i)}}(X^{(i)}) = \begin{cases} (X_C^{(i)}, X_t^{(i)}) & \textit{if } \Delta^{(i)} = 1 \\ (X_C^{(i)}, X_t^{(i)}, X_M^{(i)}) & \textit{if } \Delta^{(i)} = 2, \\ X^{(i)} & \textit{if } \Delta^{(i)} = \infty. \end{cases}$$

Model 3 *For each $\pi_1 \in \mathcal{C}(\mathbb{R}^{d_C+1}, (0, 1])$ and $\pi_2 \in \mathcal{C}(\mathbb{R}^{d_C+1+d_M}, (0, 1])$, define the observed-data multivariate front-door model as the collection of measures*

$$\mathcal{M}_\pi = \left\{ \mathcal{P}_{\theta,\pi} = \mathcal{L}(\Delta, G_\Delta(X)) : \begin{array}{c} \theta = (\beta, \varepsilon) \in \mathcal{B} \times \mathcal{E} = \Theta, \\ X \sim \mathcal{P}_{\beta,\varepsilon} \in \mathcal{M}_1, \\ \Delta | X \sim \mathcal{P}_\pi(\cdot \,|\, X) \end{array} \right\}.$$

A visualizations of the causal graph to which the model $\mathcal{M}_\pi$ is Markov is shown in Figure 1.

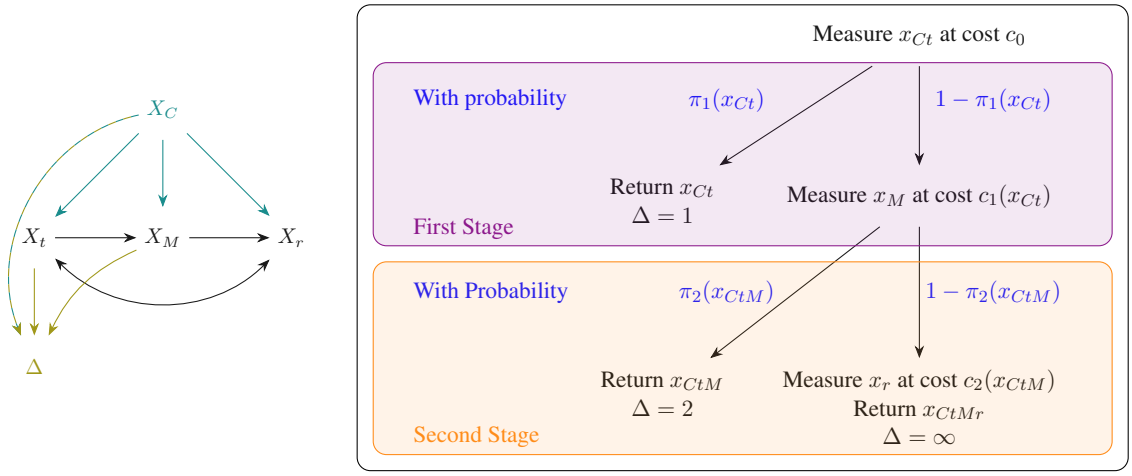

Figure 1: *Left:* ADMG for the multivariate linear front-door model. *Right:* Generation of a new sample. The first stage probabilistically decides whether to measure x_M. If so, the second stage decides probabilistically whether to additionally measure x_r.

2.1. Practical Workflow

Figure 1 illustrates the mechanism used to generate a new sample. We write '*Measure* x_i' to denote a measurement of $X_i = [(I-\beta)^{-1}\varepsilon]_i$. A dataset with four samples from model $\mathcal{M}_\pi$ can therefore take the form $\left\{(\Delta^{(1)} = \infty, x^{(1)}_{CtMr}),\ (2, x^{(2)}_{CtM}),\ (1, x^{(3)}_{Ct}),\ (1, x^{(4)}_{Ct})\right\}$, and no model can produce a sample on the components x_{Ctr}. The theoretically important full-data Model $\mathcal{M}_1$ with propensity $\pi \equiv (1,1)$ always yields datasets of the form $\{(\infty, x^{(i)}_{CtMr})\}_{i=1}^n$. A practical example is personalized staged measurement of patient features for medical diagnostics (von Kleist et al., 2025).

The experimenter adopts model $\mathcal{M}_\pi$ based on the unknown design matrix β and errors ε with the goal of estimating the causal effect ξ as cost-efficiently as possible. For each propensity $\pi = (\pi_1, \pi_2)$, the experimenter can generate datasets according to Figure 1 under Model $\mathcal{M}_\pi$. The experimenter specifies the base cost c_0 and cost functions c_1, c_2, based on real-world considerations, which together determine the cost of each sample. Varying π alters the proportion of X_{Ct}, X_{CtM} and X_{CtMr} occurrences and thus the associated cost. Since certain parameter configurations require relatively more information on X_{Ct} and less on X_{CtMr} for estimating ξ, the experimenter can exploit the staged cost and optimize (π_1^*, π_2^*) via Theorem 11 or Corollary 12. Data is then sampled exclusively from the induced model $\mathcal{M}_{\pi^*}$ to estimate the causal effect in a cost-efficient manner.

3. Background on Efficient Estimation

Our objective is to identify propensity functions π_1, π_2 that minimize the asymptotic variance of an estimator for the causal effect of the treatment X_t on the response X_r. This causal effect is defined as the derivative of the interventional expectation evaluated at the treatment level

$$\xi(x_t) = \frac{\partial}{\partial x_t^*} \mathrm{E}\left[X_r; do(X_t = x_t^*)\right]\Big|_{x_t^* = x_t}.$$

The $do(\cdot)$ operator modifies the structural equation $X = \beta X + \varepsilon$ in Definition 1, thereby altering the data-generating process (Maathuis et al., 2019). In the linear front-door model, the causal effect reduces to the sum over all mediating paths,

$$\xi = \phi(\beta_{Mt}, \beta_{rM}^\top) := \beta_{rM}\beta_{Mt} \in \mathbb{R}.$$

We restrict our attention to estimators of ξ that are regular and asymptotically linear (RAL) in $\mathcal{M}_\pi$, denoted by $\mathcal{E}_{\text{RAL}}(\xi, \mathcal{M}_\pi)$. Let $X^{(i)}$, $i \in [n]$, be i.i.d. draws from $\mathcal{P}_{\theta,\pi} \in \mathcal{M}_\pi$. An estimator $\hat{\xi}_n = (X^{(1)}, \ldots, X^{(n)})$ for the causal effect ξ is asymptotically linear if

$$\sqrt{n}(\hat{\xi}_n - \xi) = n^{-1/2} \sum_{i=1}^{n} \varphi(X^{(i)}) + o_{\mathcal{P}_{\theta;\pi}}(1),$$

for some influence function φ in the Hilbert space of centered, square-integrable functions. For any RAL estimator $\hat{\xi}_n$, the quantity $\sqrt{n}(\hat{\xi}_n - \xi)$ converges in distribution to $\mathcal{N}(0, \mathrm{E}_{\mathcal{P}_{\theta_0,\pi}}[\varphi\varphi^\top])$, and this convergence is stable under all local perturbations $\mathcal{P}_{\theta_0 + h/\sqrt{n}}$, $h \in \Theta$, of the data generating distribution (van der Vaart, 1998, Thm. 25.20). To ensure regularity, consider parametric submodels $\{\mathcal{P}_{(\beta,\varepsilon_0+\gamma h),\pi} : \gamma \in \mathbb{R}\} \subset \mathcal{M}_\pi$ passing through the truth $\mathcal{P}_{\beta_0,\varepsilon_0}$, for fixed $h \in \mathcal{E}$.

Assumption 4 (Newey (1990), A1, or van der Vaart (1998), Thm. 7.10) *For each parametric submodel, let $p_{\beta,\varepsilon_0+\gamma h}$ denote the density function. Assume that all mappings $(\beta, \gamma) \mapsto \sqrt{p_{\beta,\varepsilon_0+\gamma h}(x)}$ are almost surely continuously differentiable at θ_0, implying differentiability in quadratic mean. Further assume that the information matrix $\mathrm{E}\left[S_{\beta,\gamma} S_{\beta,\gamma}^\top\right]$ is finite and positive definite at θ_0.*

Influence functions must satisfy orthogonality conditions with respect to the semiparametric tangent space $\mathscr{T}$. The semiparametric tangent space is, by definition, the L^2 closure of submodel tangent spaces, where a submodel tangent space is the linear span of its score function $S_{\beta,\gamma}$ at θ_0. All influence functions are orthogonal to the nuisance tangent space Λ_ε, constructed as the component of $\mathscr{T}$ associated with the nuisance variation ε. This ensures that influence functions are sensitive only to perturbations of the parameter of interest. Consequently, the set of influence functions is characterized by the affine space $\varphi_0 + \mathscr{T}^\perp$, anchored at an arbitrary influence function φ_0 (Tsiatis, 2006, Thm. 4.3.). The efficient influence function minimizes the norm, or equivalently the asymptotic variance (with variances $V_1 \preceq V_2$ if $V_2 - V_1$ is positive semi-definite), and is characterized by the projection $\varphi^{\text{eff}} = \Pi(\varphi_0 | \mathscr{T})$. This leads to the following overarching optimization problem.

Problem 5 *Let $c_0 > 0$ denote the cost of measuring $X_{C,t}$, and let the measurable cost functions $c_1 \in \mathcal{C}(\mathbb{R}^{d_C+1}, \mathbb{R}_{>0})$, $c_2 \in \mathcal{C}(\mathbb{R}^{d_C+1+d_M}, \mathbb{R}_{>0})$ represent the additional cost of measuring X_M after $X_{C,t}$ is observed, and of measuring X_r after $X_{C,t,M}$ is observed. For an average per-sample budget $b_0 \in (c_0, c_0 + \mathrm{E}[c_1(X_{C,t}) + c_2(X_{C,t,M})]]$, the goal is to minimize the asymptotic variance $\mathrm{Var}_\infty(\hat{\xi}_n; \pi)$ over all RAL estimators $\hat{\xi}_n$ of ξ and all sampling regimes induced by $\mathcal{M}_{\pi=(\pi_1,\pi_2)}$:*

$$\begin{aligned} \underset{\substack{\pi_1 \in \mathcal{C}(\mathbb{R}^{d_C+1}) \\ \pi_2 \in \mathcal{C}(\mathbb{R}^{d_C+d_M+1}) \\ \hat{\xi}_n \in \mathcal{E}_{RAL}(\xi, \mathcal{M}_\pi)}}{\text{argmin}} \quad & \mathrm{Var}_\infty(\hat{\xi}_n; \pi) \quad \textit{s.t.} \quad 0 < \pi_1(X_{C,t}) \leq 1 \textit{ a.e.}, \\ & 0 < \pi_2(X_{C,t,M}) \leq 1 \textit{ a.e.}, \\ & \mathrm{E}[c_0 + (\pi_1 c_1)(X_{C,t}) + \pi_1(X_{C,t})(\pi_2 c_2)(X_{C,t,M})] = b_0. \end{aligned}$$

The optimal propensity functions (π_1^*, π_2^*) specify the experimental design. Since the problem cannot be solved in practice without prior knowledge, we require access to an initial estimate of $\mathcal{P}_{\theta,\pi}$.

4. Optimized Causal Effect Estimation

Since the asymptotic variance $\mathrm{Var}_\infty(\hat{\xi}_n;\pi) = \mathrm{E}[\varphi\varphi^\top] = \mathrm{Var}(\varphi)$ fully determines the limiting distribution of any RAL estimator, solving Problem 5 reduces to analyzing the influence function space and deriving the efficient influence function. A logical overview of the proceeding results is presented in Appendix A, clarifying the relationships between the full-data and observed-data influence functions established by the main theorems and lemmas.

4.1. Efficient Influence Function in $\mathcal{M}_1$ and its Optimal Augmentation in $\mathcal{M}_\pi$

As a first step towards solving Problem 5, we characterize the minimum-variance RAL estimator under the full-data model $\mathcal{M}_1$. All major proofs are presented in Appendix B.

Theorem 6 *Denote the residual error $\varepsilon_r - \mathrm{E}[\varepsilon_r|\varepsilon_t]$ by $\varepsilon_r^\perp$. In Model $\mathcal{M}_1$, the 0-indexed row and column of $\beta \in \mathcal{B}$ are structurally zero, and the efficient full-data influence function $\varphi^{F,eff}$ therefore only has non-zero components on the $(t,M,r)\times(C,t,M)$ block. This remaining block is given by*

$$\varphi^{F,eff}_{\beta_{(t,M,r)(C,t,M)}} = \begin{pmatrix} \varepsilon_t\varepsilon_C^\top \mathrm{Var}(\varepsilon_C)^{-1} & 0 & 0 \\ \varepsilon_M(\varepsilon_C^\top \mathrm{Var}(\varepsilon_C)^{-1} - \varepsilon_t \mathrm{Var}(\varepsilon_t)^{-1}\beta_{tC}) & \varepsilon_M\varepsilon_t \mathrm{Var}(\varepsilon_t)^{-1} & 0 \\ \varepsilon_r^\perp(\varepsilon_C^\top \mathrm{Var}(\varepsilon_C)^{-1} - \varepsilon_M^\top \mathrm{Var}(\varepsilon_M)^{-1}(\beta_{Mt}\beta_{tC}+\beta_{MC})) & 0 & \varepsilon_r^\perp\varepsilon_M^\top \mathrm{Var}(\varepsilon_M)^{-1} \end{pmatrix}.$$

Consequently, the full-data efficient influence function in $\mathcal{M}_1$ for the causal effect ξ is

$$\varphi^{F,eff}_\xi = \beta_{rM}\varepsilon_M\varepsilon_t \mathrm{Var}(\varepsilon_t)^{-1} + \varepsilon_r^\perp\varepsilon_M^\top \mathrm{Var}(\varepsilon_M)^{-1}\beta_{Mt}.$$

The proof decomposes the full-data efficient influence function $\varphi^{F,\mathrm{eff}}_\xi$ into two orthogonal components, efficiently and separately targeting β_{Mt} and β_{rM}. To transition from full-data to observed-data influence functions, we rely on an existing characterization of their geometry: Each observed-data influence function can be expressed as an augmentation of a full-data influence function.

Lemma 7 (Tsiatis (2006), Thm. 8.3.) *Denote the tangent space in Model $\mathcal{M}_\pi$ corresponding to variation in the propensity functions π by Λ_π. The observed-data augmentation space is*

$$\Lambda_{2,\pi} = \left\{ \sum_{\delta\in\{1,2\}} \frac{I(\Delta=\delta) - (1-\pi_\delta)I(\Delta\geq\delta)}{\Pi_{j=1}^\delta \pi_j}(h_\delta \circ G_\delta) \;:\; h_1 \in L_0^2(\mathbb{R}^{d_C+1},\mathbb{R}), h_2 \in L_0^2(\mathbb{R}^{d_C+1+d_M},\mathbb{R}) \right\}.$$

Each full-data influence function φ^F induces an functional family of augmented observed-data influence functions in Model $\mathcal{M}_\pi$, denoted by

$$IF^\pi(\varphi^F) = \left\{ \left[\frac{I(\Delta=\infty)\varphi^F}{\pi_1\pi_2} + h\right] - \Pi\left([\,\cdot\,]|\Lambda_\pi\right) \;:\; h \in \Lambda_{2,\pi} \right\},$$

The union of these sets coincides with the entire space of observed-data influence functions.

This characterization enables us to identify the optimal full-data influence function $\varphi^{\text{opt},\pi}$ within $IF^{\pi}(\varphi^{F,\text{eff}})$. It can be obtained by retaining the conditional mean term $\mathbb{E}\left[\varphi^{F,\text{eff}}|\varepsilon_C, \varepsilon_t\right]$ and weighting the remaining augmentation terms by indicators of the sampling stage variable Δ and the inverse propensity functions π.

Lemma 8 *Let $\pi_1 \in \mathcal{C}(\mathbb{R}^{d_C+1}, (0,1])$ and $\pi_2 \in \mathcal{C}(\mathbb{R}^{d_C+d_M+1}, (0,1])$ be given. Under the Model $\mathcal{M}_\pi$, the optimal observed-data influence-function augmentation of the full-data efficient influence function for the $(t, M, r) \times (C, t, M)$ block of the parameter matrix β and for the causal effect ξ are*

$$\varphi^{opt,\pi}_{\beta_{(t,M,r)(C,t,M)}}(X) = \begin{pmatrix} diag(1) & 0 & 0 \\ 0 & \frac{diag(I(\Delta\geq 2))}{\pi_1(X_{C,t})} & 0 \\ 0 & 0 & \frac{diag(I(\Delta=\infty))}{\pi_1(X_{C,t})\pi_2(X_{C,t,M})} \end{pmatrix} \varphi^{F,eff}_{\beta_{(t,M,r)(C,t,M)}}(X), \quad \textit{and}$$

$$\varphi^{opt,\pi}_{\xi}(X) = \frac{I(\Delta \geq 2)\beta_{rM}\varepsilon_M\varepsilon_t \text{Var}(\varepsilon_t)^{-1}}{\pi_1(X_{C,t})} + \frac{I(\Delta = \infty)\varepsilon_r^{\perp}\varepsilon_M^{\top}\text{Var}(\varepsilon_M)^{-1}\beta_{Mt}}{\pi_1(X_{C,t})\pi_2(X_{C,t,M})}.$$

The optimized influence function $\varphi^{\text{opt},\pi}$ might not be an efficient influence function in the observed-data Model $\mathcal{M}_\pi$. Recall that every full-data influence function lies in the affine linear space $\varphi^{F,\text{eff}} + \Lambda_\varepsilon^{\perp}$ as specified by Equation (2). By linearity of the operator $\mathcal{J}$, which maps a full-data influence function φ^F to its optimum in $IF^{\pi}(\varphi^F)$, the efficient observed-data influence function is given as $\varphi^{F,\text{eff}} + u$, where the orthogonal nuisance tangent space element $u \in \Lambda_\varepsilon^{\perp}$ is chosen to minimize the variance $\text{Var}\left(\mathcal{J}(\varphi^{\text{F,eff}})\right) + \text{Var}(\mathcal{J}(u)) + 2\text{Cov}\left(\mathcal{J}(u), \mathcal{J}(\varphi^{\text{F,eff}})\right)$. Nevertheless, the RAL estimator associated with $\varphi^{\text{opt},\pi}$ achieves an asymptotic variance no larger than that of any other RAL estimator arising from an augmentation in $IF^{\pi}(\varphi^{F,\text{eff}})$. We refer to Appendix A for a visualization.

4.2. Optimized Observed-Data Estimator $\hat{\xi}_n$ and Optimized Propensity π^{opt}

Efficient RAL estimators $\hat{\beta}_{Mt}$ and $\hat{\beta}_{rM}$ are obtained by solving estimating equations associated with the efficient influence function. Concretely, they correspond to weighted linear equations on the partial datasets defined by the indicators $I(\Delta^{(i)} \geq 2)$ and $I(\Delta^{(i)} = \infty)$, respectively. Knowing the optimal estimator for each model $\mathcal{M}_\pi$ allows solving a restricted version of Problem 5 on the subset of observed-data augmentations of $\varphi^{F,\text{eff}}$, in particular $IF^{\pi}(\varphi^{F,\text{eff}})$. This restriction is purely technical, and Figure 6 in Appendix A illustrates its implied admissible influence function set.

Lemma 9 *The observed-data optimized RAL estimators $\hat{\beta}_{Mt,n}$ and $\hat{\beta}_{rM,n}$ are solutions to a sequence of linear equations on the full and partially observed datasets. Their product estimator $\hat{\xi}_n = \hat{\beta}_{rM,n}\hat{\beta}_{Mt,n}$ is the optimized RAL estimator for the causal effect ξ in Model $\mathcal{M}_\pi$. The exact formulas are provided in Appendix B.3.*

Note that the equations determining the estimators $\hat{\beta}_{rC,n}$ and $\hat{\beta}_{rM,n}$ are conceptually different from a sequence of weighted linear regressions.

Lemma 10 *The asymptotic variance $\text{Var}_\infty(\hat{\xi}_n; \pi)$ of the optimized estimator $\hat{\xi}_n$ for the causal effect ξ in Model $\mathcal{M}_\pi$, $\pi_1 \in \mathcal{C}(\mathbb{R}^{d_C+1}, (0,1])$, $\pi_2 \in \mathcal{C}(\mathbb{R}^{d_C+d_M+1}, (0,1])$, is determined by*

$$\text{Var}\left(\varphi^{opt,\pi}_{\xi}\right) = \text{E}\left[\frac{\varepsilon_t^2\beta_{rM}\text{Var}(\varepsilon_M)\,\beta_{rM}^{\top}}{\text{Var}(\varepsilon_t)^2\,\pi_1(X_{C,t})}\right] + \text{E}\left[\frac{\left(\varepsilon_M^{\top}\text{Var}(\varepsilon_M)^{-1}\,\beta_{Mt}\right)^2\text{Var}(\varepsilon_r \mid \varepsilon_t)}{\pi_1(X_{C,t})\pi_2(X_{C,t,M})}\right].$$

Theorem 11 *Let X follow model $\mathcal{M}_\pi$. Let the base cost c_0, cost functions c_1, c_2 as well as the average budget b_0 be according to the specifications of Problem 5. Under the additional restriction that admissible influence functions φ satisfy $\varphi \in IF^\pi(\varphi^{F,\mathit{eff}})$, Problem 5 is uniquely solvable almost everywhere. Define the leverages $g_1(X_{C,t}) := \beta_{rM}\mathrm{Var}(\varepsilon_M)\,\beta_{rM}^\top \frac{\varepsilon_t^2}{\mathrm{Var}(\varepsilon_t)^2}$, and $g_2(X_{C,t,M}) := \mathrm{Var}(\varepsilon_r \mid \varepsilon_t)\left(\varepsilon_M^\top \mathrm{Var}(\varepsilon_M)^{-1}\,\beta_{Mt}\right)^2$. Then the optimal propensity π^* is given by*

$$(\pi_1^*, \pi_2^*) = \begin{cases} \left(\min\left(1, \max\left(\sqrt{\frac{g_1}{\lambda c_1}}, \sqrt{\frac{g_1 + \mathrm{E}[g_2|\varepsilon_{C,t}]}{\lambda(c_1 + \mathrm{E}[c_2|\varepsilon_{C,t}])}}\right)\right), \min\left(1, \sqrt{\frac{g_2 c_1}{g_1 c_2}}\right)\right), & g_1 < \lambda c_1, \\ \left(1, \min\left(1, \sqrt{\frac{g_2}{\lambda c_2}}\right)\right), & g_1 \geq \lambda c_1, \end{cases}$$

where the constant $\lambda > 0$ is chosen to satisfy the problem's budget constraint

$$\mathrm{E}[c_0 + (\pi_1 c_1)(X_{C,t}) + \pi_1(X_{C,t})(\pi_2 c_2)(X_{C,t,M})] = b_0.$$

Several practical considerations accompany this optimality result. First, the conditional expectation $\mathrm{E}[g_2|\varepsilon_{C,t}]$ in the first-stage propensity π_1^* can be approximated by sampling from the empirical distribution function $\hat{F}_n(\varepsilon_M)$. Second, when the solution π^* lies in the interior of the feasible set, the budget constraint admits a closed form expression for the Lagrange multiplier,

$$\lambda^{\text{interior},*} = \frac{\mathrm{E}\left[\sqrt{g_1 c_1} + \sqrt{g_2 c_2}\right]^2}{(b_0 - c_0)^2},$$

which provides a useful initialization for numerical optimization. In this interior solution, the propensity π_1^* scales linearly with $(b_0 - c_0)$, resulting in the asymptotic variance $\mathrm{Var}_\infty^\pi(\hat{\xi}_n)$ decreasing with $1/(b_0 - c_0)$ as a function of the average invested budget b_0. Third, in regions where the second-stage leverage dominates, so $g_2 c_1 > g_1 c_2$, the optimal propensity satisfies $\pi_2^* = 1$, indicating that $X_{M,r}$ will always be sampled jointly. Fourth, practical design planning in finite samples requires further optimization techniques across different budget levels, with sample sizes adjusted to maintain a fixed total cost (Mareis, 2025). Finally, designs in which the sampling decision may depend only on a restricted subset of variables, as in causal-fairness settings where sensitive attributes must be protected (Kilbertus et al., 2018), also fall within our framework. In such cases, the propensity expressions in Theorem 11 are replaced by versions that integrate out the disallowed components, in the same way the unobserved error ε_M is handled.

4.3. Optimized Back-Door Estimation

Front-door estimation implicitly contains a back-door estimation problem on the subvector $X_{C,t,M}$. By isolating the contribution of the optimized influence function $\varphi^{\text{opt}}_{\beta_{Mt}}$ from Lemma 8, one obtains an analogous problem for the back-door setting. The resulting propensity has the same structure as in Theorem 11, but depends only on the first-stage sampling decision.

Corollary 12 *Let $c_0 > 0$ be the cost of measuring $X_{C,t}$ and let the measurable, non-negative cost function $c \in \mathcal{C}(\mathbb{R}^{d_C+1}, \mathbb{R}_{>0})$ represent the cost of measuring X_M after $X_{C,t}$ is already observed. For an average sample cost of $b_0 \in (c_0, c_0 + \mathrm{E}[c_1(X_{C,t})]]$, consider the optimization problem*

$$\underset{\substack{\pi \in \mathcal{C}(\mathbb{R}^{d_C+1}) \\ \hat{\beta}_{Mt,n} \in \mathcal{E}_{RAL}(\beta_{Mt}, \mathcal{M}_\pi)}}{\text{argmin}} \quad \mathrm{Var}_\infty(\hat{\beta}_{Mt,n}; \pi) \qquad \textit{s.t.} \quad \begin{array}{l} 0 < \pi(X_{C,t}) \leq 1 \textit{ a.e.}, \\ \mathrm{E}[c_0 + (\pi c_1)(X_{C,t})] = b_0 \end{array}$$

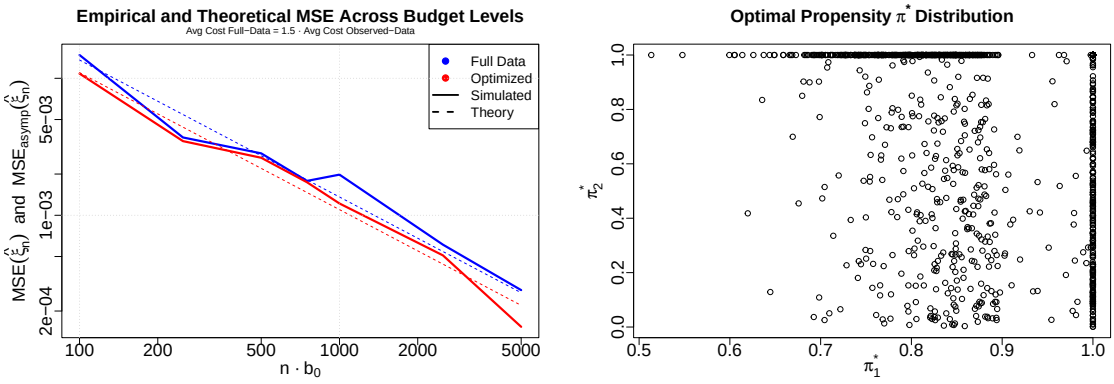

Figure 2: Average empirical and theoretical MSE of the full-data and the optimized observed-data causal effect estimators across varying levels of budget $n \cdot b_0$ are presented in the left panel. In the right panel, the optimized propensity π^* distribution on $1,000$ samples is visualized, contrasting the full-data propensity $\pi \equiv (1,1)^\top$.

with the restriction that $\varphi \in IF^{\pi}(\varphi^{F,eff}_{\beta_{Mt}})$. *If* X_M *is one-dimensional, the optimal propensity is*

$$\pi^* = \min\left(1, \mathrm{E}\left[\frac{\varepsilon_t^2 \mathrm{Var}(\varepsilon_M)}{\mathrm{Var}(\varepsilon_t)^2 \, \lambda c_1}\right]\right),$$

where $\lambda > 0$ *is chosen to satisfy the budget constraint. The corresponding estimator for* $\hat{\beta}_{Mt,n}$ *coincides with the construction in Lemma 9.*

5. Computational Results

5.1. Simulation Study

For our simulations, we study the exemplary front-door model given in Appendix C. The model features non-Gaussian errors on $X_{C,t,r}$ and Gaussian errors on X_M, multivariate confounder and mediator variables with $d_C = 2$, $d_M = 3$, a constant cost function c_2 and a non-constant cost function $c_1(X_{C,t}) = 0.1 \cdot \|X_{C,t}\|_2$. The study examines calibration, computational sensitivity as well as the dependence of the asymptotic variance in Lemma 10 and its optimization in Theorem 11 on model parameters and nuisance components.

The first research question concerns calibration: How accurate does the theoretical asymptotic distribution of the causal effect estimator approximate the finite-sample mean-squared error (MSE), and what factors influence this reliability? Figure 2 compares in the left panel the theoretical asymptotic MSE with the average empirical MSE of both full-data and optimized observed-data estimators. The method is well calibrated even at modest sample sizes. There is a substantial increase in efficiency, or equivalently decrease in asymptotic MSE, when comparing the full-data design to the partial-data design at fixed budget levels, indicated by the x-axis. The optimal propensity function π^* is sensitive towards the input information and attains all four possible configurations

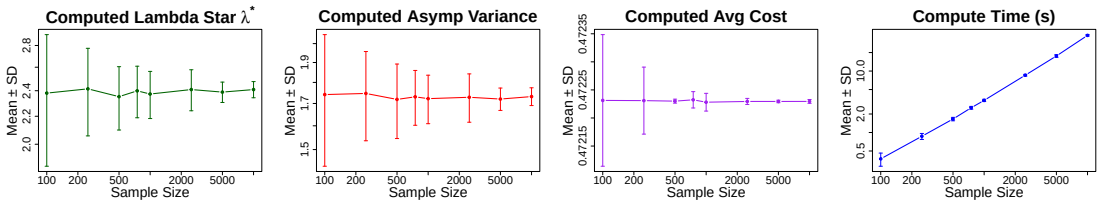

Figure 3: Plots of the mean ± 1SD of the optimized tuning parameter λ^*, the computed asymptotic optimized variance $\mathrm{Var}_\infty(\hat{\xi}_n; \pi^*)$, the computed average sample cost $\mathrm{E}[c_0 + \pi_1^* c_1 + \pi_1^* \pi_2^* c_2]$ and compute time for sample sizes between 100 and $10,000$.

$((0,1),(0,1))$, $(1,(0,1))$, $((0,1),1)$, and $(1,1)$. While Figure 2 reports averages over 50 replications, the uncertainty of a single asymptotic variance computation remains of interest.

Figure 3 displays the variability of the tuning parameter λ^*, the asymptotic optimized variance $\mathrm{Var}_\infty(\hat{\xi}_n; \pi^*)$, the realized average sample cost $\mathrm{E}[c_0 + \pi_1^* c_1 + \pi_1^* \pi_2^* c_2]$, and the computational time. Although we find the average cost to be highly reliable, the variance in optimized asymptotic exhibits non-negligible variability. In critical applications, we recommend quantifying the computational uncertainty, e.g., via bootstrapping, and protecting against it by considering conservative asymptotic variance estimates.

Next, we investigate how the marginal parameters and nuisance components affect the relative efficiency $\mathrm{Var}_\infty(\hat{\xi}_n; \pi^*)/\mathrm{Var}_\infty(\hat{\xi}_n; 1)$ and asymptotic variance, to identify scenarios where partial-measurement designs are particularly advantageous. A relative efficiency of 80% means that, under the same budget, the optimized partial-measurements design achieves a 20% reduction in asymptotic variance compared to full measurements. All experiments are summarized in Figure 4 and we restrict our discussion to the plots with major trends. The general guideline is that partial-measurements designs are favorable when β_{Mt} is difficult to estimate or when β_{rM} is comparatively easy to learn. Such settings allow the design to exploit variance reduction through sample efficient estimation of β_{rM}. This guideline is supported by the following findings.

- Decreasing the magnitude of parameter $\|\beta_{Mt}\|$ (Figure 4 [2nd row, 1st col]), as well as reducing the variance of ε_t (Figure 4 [4, 1]) or increasing the variance of ε_M (Figure 4 [3, 2]) make the estimation of β_{Mt} more challenging and increase the benefit of partial measurements.

- Increasing the parameter $\|\beta_{rM}\|$ (Figure 4 [2, 2]) or decreasing the variance of ε_r (Figure 4 [4, 2]) strengthens the signal along the path $X_M \to X_r$, enabling β_{rM} to be estimated from relatively fewer samples and again enabling efficiency gains through partial measurements.

- Increasing the variance of ε_C (Figure 4 [2, 3]) has a qualitatively different effect. It results in noisier back-door paths, making both β_{rM} and β_{Mt} harder to estimate and thereby degrading the performance of sensitive, leverage-based partial-measurements designs.

- Finally, reducing the confounding between ε_t and ε_r (Figure 4 [4, 3]) transitions the system to an unconfounded setting, where full-data measurements of $X_{t,r}$ suffice and the benefit of partial measurements diminishes.

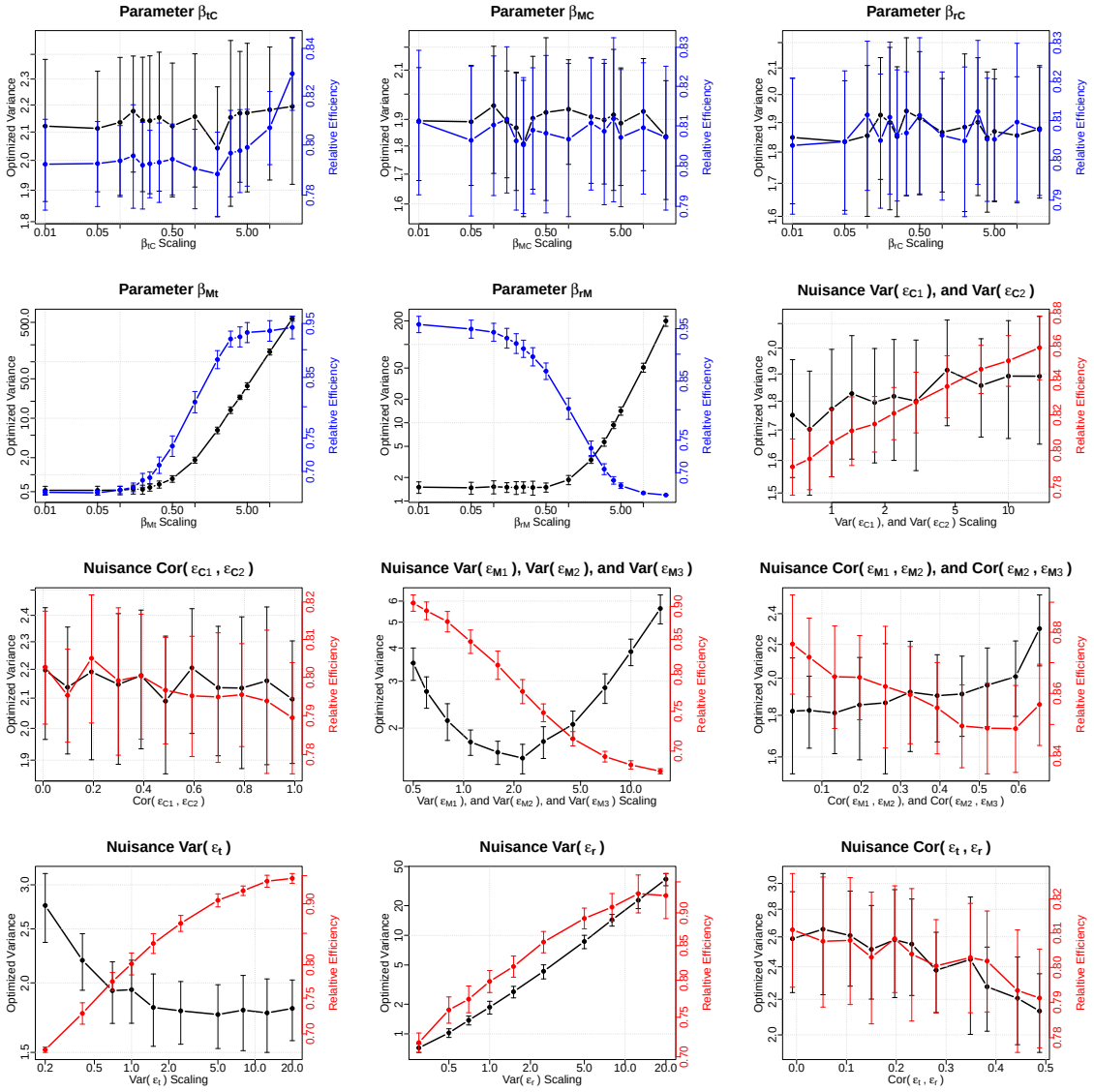

Figure 4: The plots demonstrate the variation (mean $\pm$1SD) in asymptotic optimized variance $\mathrm{Var}_\infty(\hat{\xi}_n^{\mathrm{opt}}; \pi^*)$ (black) and relative efficiency $\mathrm{Var}_\infty(\hat{\xi}_n; \pi^*)/\mathrm{Var}_\infty(\hat{\xi}_n; 1)$ (red / blue), and optimized variance by individual perturbations of the parameters and nuisance. For optimization, we specified the average cost of the observed-data estimator to be $2/3$ of the full-data average cost, resulting in the relative efficiency scaling. All perturbations of parameters and nuisance are listed in Table 2. The x-axis represents either the multiplicative scaling factor or the resulting (average) correlation(s). The y-axes differ in scale.

Dataset	X_C	X_t	X_M	X_r	Oversampling Percentage	Relative Efficiency
Sachs	PKA, PKC	Raf	Erk	Akt	135%	0.7662
Sachs	PKA	Mek	Erk	Akt	154%	0.6802
MIMIC	Age, BMI	Insulin	Δ Glucose		115%	0.8987
causalAssembly	S1mp3	S1mp5	S1mp4		109%	0.9470

Table 1: Results of Theorem 11 andCorollary 12 applied to three datasets. The cost functions are specified to be $c_1 \equiv 1, c_2 \equiv 1$ with base cost $c_0 = 1 + d_C$ for Sachs and causalAssembly. Only for MIMIC we set $c_0 = 1$ as *BMI* and *Age* measurements are negligible.

5.2. Applications

We demonstrate the advantages of partial-measurement designs on two real-world datasets and one semi-synthetic dataset drawn from biology, medicine and industry.

Sachs dataset. The dataset of Sachs et al. (2005) (853 samples; source: 10.5281/zenodo.7679091, Mareis et al. (2025)) records protein and phospholipid expression levels in human cells. The original study includes both observational and interventional data, yielding one of the few ground truth causal directed acyclic graphs (DAG) describing the influence between measurements. Within this graph, we determined two front-door configurations suitable for our analysis.

MIMIC-IV dataset. Medical professionals administer insulin to control elevated blood sugar levels of diabetes patients. In observational data, however, the effect of insulin is confounded by patient characteristics such as age and body mass index (BMI). To study this setting, we processed the electronic health dataset MIMIC-IV (Johnson et al. (2023)) of the Beth Israel Deaconess Medical Center, generating a dataset of size 999 samples on these four variables.

Industrial assembly dataset. The third dataset is generated using the Python package `causal-Assembly` (Göbler et al., 2024). The authors combined real-world assembly line measurements and expert knowledge to construct a graphical model. Their package provides a semi-synthetic sampling method in which the imposed graphical assumptions are satisfied by construction. Consequently, our back-door dataset on $3,000$ samples from `station1` meets the requirements.

Table 1 reports the oversampling percentage and relative efficiency obtained by applying Theorem 11, respectively Corollary 12. Oversampling quantifies the required sample size under a partial measurement to match the total cost of a full-data design. For example, in the causalAssembly experiment, a partial-measurement dataset of $1.09 \cdot n$ yields a 5.3% variance reduction relative to a full-measurement dataset of size n. Overall, the results confirm that exploiting parameter uncertainty through optimized experimental design can substantially reduce the estimation variance compared to the naive full-measurement design.

6. Conclusion

The central contribution of this work is a unified framework that combines semiparametric efficiency theory and causal graphical modeling to address a problem that arises naturally in practice: how to allocate a limited measurement budget across study stages so as to minimize the variance of

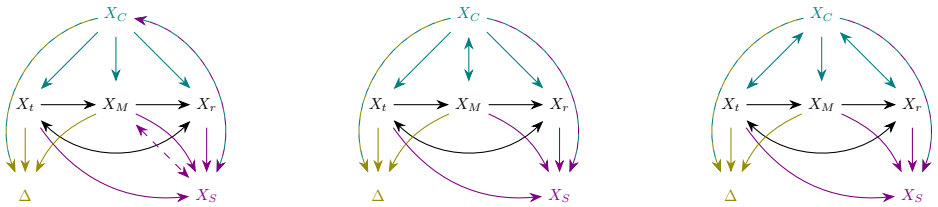

Figure 5: These causal graphs allow for identification (Drton et al., 2011) of linear causal effect ξ. Directed edges result in non-zero elements in the design matrix B, whereas bidirected edges indicate dependencies in the error vector ε.

a causal effect estimator. The RAL estimators and optimized designs we derive show that substantial efficiency gains over full-measurement strategies are achievable without sacrificing inferential validity, a finding with direct relevance to experimental planning in medicine, biology, and industry.

The framework rests on three modeling choices whose relaxation we consider the most consequential directions for future work. First, the linearity assumption, which we regard as the strongest restriction imposed. It was essential for establishing a clean theoretical framework and deriving closed-form optimal designs, while offering a clear interpretation and serving as first-order approximations to more complex systems (Hastie et al., 2009; Pearl, 2013). Appendix D confirms that the optimized design retains its efficiency advantage under mild nonlinearity, though efficiency gains are expected to diminish as departures become more pronounced. Second, the present sampling schemes follow the intrinsic causal order. Designs that depart from this ordering would give rise to conceptually distinct augmented influence functions and may unlock additional efficiency gains, though the theoretical development would be substantially more involved. Here, we also see the possibility of sequential propensity updates offering additional finite-sample gains. And finally, the imposed causal front-door structure. Although the characterization of efficient observed-data influence functions beyond the augmentation class $IF^{\pi}(\varphi^{F,\text{eff}})$ remains open, extensions to identified causal graphs, including those visualized in Figure 5, are feasible and promising.

Acknowledgments

Leopold Mareis received funding from the German Research Foundation (DFG) under the Mathematical Research Data Initiative (project No. 460135501). Mathias Drton received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 883818).

References

Rohit Bhattacharya, Razieh Nabi, Ilya Shpitser, and James M Robins. Identification in Missing Data Models Represented by Directed Acyclic Graphs. In *Uncertainty in Artificial Intelligence*, pages 1149–1158. PMLR, 2020.

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.

Mathias Drton, Rina Foygel, and Seth Sullivant. Global Identifiability of Linear Structural Equation Models. *The Annals of Statistics*, 39(2):865–886, 2011.

Konstantin Göbler, Tobias Windisch, Mathias Drton, Tim Pychynski, Martin Roth, and Steffen Sonntag. `causalAssembly`: Generating Realistic Production Data for Benchmarking Causal Discovery. In Francesco Locatello and Vanessa Didelez, editors, *Proceedings of the Third Conference on Causal Learning and Reasoning*, volume 236 of *Proceedings of Machine Learning Research*, pages 609–642. PMLR, 01–03 Apr 2024.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer, New York, second edition, 2009. Data mining, inference, and prediction.

Christian L. Hines and Oliver J. Hines. Automatic Debiasing of Neural Networks via Moment-Constrained Learning. In Biwei Huang and Mathias Drton, editors, *Proceedings of the Fourth Conference on Causal Learning and Reasoning*, volume 275 of *Proceedings of Machine Learning Research*, pages 390–405. PMLR, 07–09 May 2025.

Oliver Hines, Oliver Dukes, Karla Diaz-Ordaz, and Stijn Vansteelandt. Demystifying Statistical Learning Based on Efficient Influence Functions. *The American Statistician*, 76(3):292–304, 2022.

Kazufumi Ito and Karl Kunisch. *Lagrange Multiplier Approach to Variational Problems and Applications*, volume 15 of *Advances in Design and Control*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2008.

Alistair E W Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, Li-Wei H Lehman, Leo A Celi, and Roger G Mark. MIMIC-IV, a Freely Accessible Electronic Health Record Dataset. *Scientific Data*, 10(1):1, January 2023.

Yonghan Jung, Jin Tian, and Elias Bareinboim. Estimating Identifiable Causal Effects through Double Machine Learning. In *Proceedings of the Thirty-Fifth AAAI Conference on Artificial Intelligence*, pages 12113–12122. AAAI Press, 2021.

Niki Kilbertus, Adria Gascon, Matt Kusner, Michael Veale, Krishna Gummadi, and Adrian Weller. Blind Justice: Fairness with Encrypted Sensitive Attributes. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2630–2639. PMLR, 10–15 Jul 2018.

Marloes Maathuis, Mathias Drton, Steffen Lauritzen, and Martin Wainwright, editors. *Handbook of Graphical Models*. Chapman & Hall/CRC Handbooks of Modern Statistical Methods. CRC Press, Boca Raton, FL, 2019.

Leopold Mareis. Optimizing Experimental Design for Causal Effect Estimation with Partial Measurements. In Xiao-Hua Zhou and Jinzhu Jia, editors, *Causal Inference*, pages 74–85, Singapore, 2025. Springer Nature Singapore.

Leopold Mareis, Stephan Haug, and Mathias Drton. MaRDI's Zenodo Community for Graphical Modeling and Causal Inference. In *Proceedings of the 2nd Conference on Research Data Infrastructure (CoRDI)*, Aachen, Germany, August 2025.

Karthika Mohan and Judea Pearl. Graphical Models for Processing Missing Data. *Journal of the American Statistical Association*, 116(534):1023–1037, 2021.

Whitney K. Newey. Semiparametric Efficiency Bounds. *Journal of Applied Econometrics*, 5(2): 99–135, 1990.

Judea Pearl. Linear Models: A Useful "Microscope" for Causal Analysis. *Journal of Causal Inference*, 1(1):155–170, 2013.

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.

James M. Robins, Andrea Rotnitzky, and Lue Ping Zhao. Estimation of Regression Coefficients When Some Regressors Are Not Always Observed. *Journal of the American Statistical Association*, 89(427):846–866, 1994.

Karen Sachs, Omar Perez, Dana Pe'er, Douglas A Lauffenburger, and Garry P Nolan. Causal Protein-Signaling Networks Derived from Multiparameter Single-Cell Data. *Science*, 308(5721): 523–529, April 2005.

Maximilian Seitzer, Bernhard Schölkopf, and Georg Martius. Causal Influence Detection for Improving Efficiency in Reinforcement Learning. *Advances in Neural Information Processing Systems*, 34:22905–22918, 2021.

Zhiqiang Tan. Bounded, Efficient and Doubly Robust Estimation with Inverse Weighting. *Biometrika*, 97(3):661–682, 2010.

Anastasios A. Tsiatis. *Semiparametric Theory and Missing Data*. Springer Series in Statistics. Springer, New York, 2006.

Aad van der Vaart. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1998.

Aad van der Vaart. Higher Order Tangent Spaces and Influence Functions. *Statistical Science*, 29 (4):679–686, 2014.

Henrik von Kleist, Alireza Zamanian, Ilya Shpitser, and Narges Ahmidi. Evaluation of Active Feature Acquisition Methods for Time-Varying Feature Settings. *Journal of Machine Learning Research*, 26(60):1–84, 2025.

Janine Witte, Leonard Henckel, Marloes H. Maathuis, and Vanessa Didelez. On Efficient Adjustment in Causal Graphs. *Journal of Machine Learning Research*, 21(246):1–45, 2020.

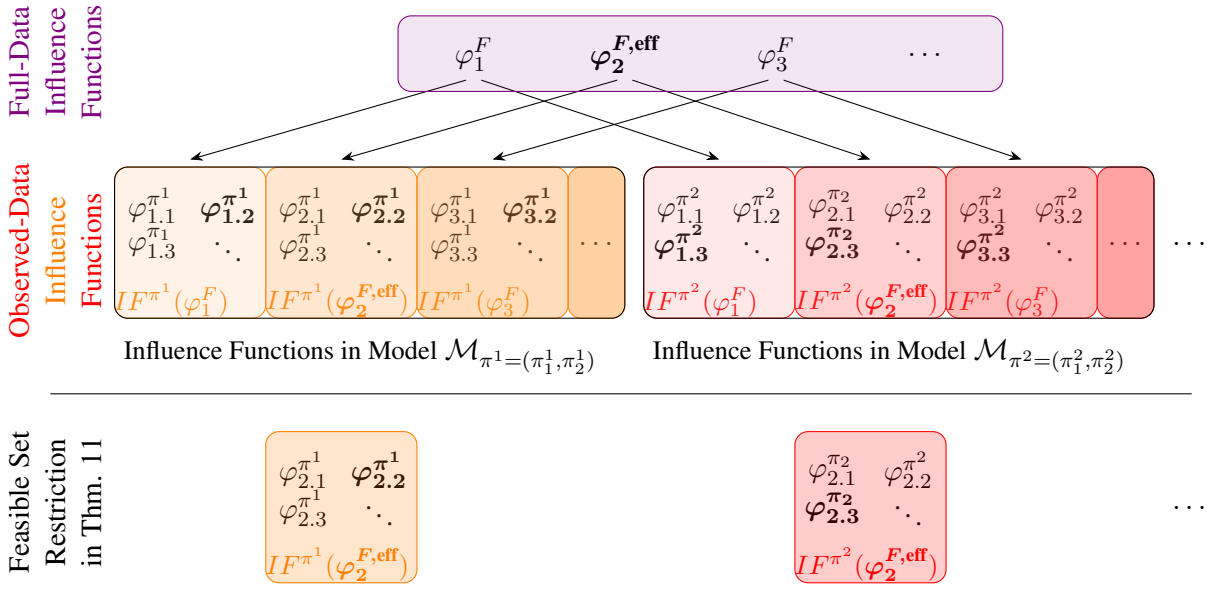

Figure 6: Logical structure of the full-data and observed-data influence functions, augmentation classes, and the feasible set used in the optimization.

Appendix A. Structure of the Results

The following points summarize the logical structure of the results and clarify how the main lemmas and theorems relate to the influence-function relations shown in Figure 6.

- It is known that the affine space $\varphi_0^F + \mathscr{T}^\perp$ characterizes all full-data influence functions, given in Figure 6 by the violet set.
- Theorem 6 identifies the full-data efficient influence function, shown in Figure 6 as $\boldsymbol{\varphi_2^{F,\text{eff}}}$.
- Lemma 7 provides a construction of augmented influence functions for each observed-data model $\mathcal{M}_{(\pi_1,\pi_2)}$. Each full-data influence function φ^F generates a class $IF^\pi(\varphi^F)$ of augmented observed-data influence functions. The disjoint union of all generated classes spans the set of augmented observed-data influence functions in $\mathcal{M}_{(\pi_1,\pi_2)}$. Each arrow in Figure 6 illustrates the generation of such a class.
- Lemma 8 determines the most efficient representative within each augmentation class. These are the bold observed-data influence functions in Figure 6, for example $\boldsymbol{\varphi_{2.2}^{\pi_1}}$ in $IF^{\pi^1}(\varphi_2^{F,\text{eff}})$ or $\boldsymbol{\varphi_{1.3}^{\pi_2}}$ in $IF^{\pi^2}(\varphi_1^F)$.
- The main Problem 5 seeks the minimal-asymptotic-variance RAL estimator $\hat{\xi}_n$ over all models $\mathcal{M}_{\pi_1,\pi_2}$. Since the asymptotic variance equals $\mathrm{E}[\varphi^2]$, the problem is equivalent to minimizing the length of the observed-data influence function over all models. The feasible set

corresponds to the entire second row of Figure 6. If such an observed-data influence function were identified, the experimenter could then sample data from the corresponding optimal model and use the associated estimator for cost-efficient causal effect estimation.

- For technical reasons, however, we must restrict the search for observed-data influence functions to a smaller feasible set, namely to the augmentation classes generated by the efficient full-data influence function. In Figure 6, this is depicted in the third row.

- Lemma 10 computes, for each model $\mathcal{M}_{(\pi_1,\pi_2)}$, the length of the most efficient observed-data influence function within $IF^{\pi}(\varphi^{F,\text{eff}})$. This equals the asymptotic variance of the corresponding estimator.

- Theorem 11 then identifies the optimal influence function φ^{opt} within this restricted feasible set, based on Lemma 10. As discussed in the main text, φ^{opt} might not be globally efficient. This would be the case if $\mathrm{E}[(\boldsymbol{\varphi^{\pi_1}_{1.2}})^2] < \mathrm{E}[(\varphi^{\text{opt}})^2] \leq \mathrm{E}[(\boldsymbol{\varphi^{\pi_1}_{2.2}})^2]$ in Figure 6.

- Although not necessarily observed-data efficient across all models, φ^{opt} remains compelling. It improves (a) on all influence functions in its own augmentation class $IF^{\pi^*}(\varphi^{F,\text{eff}})$ and (b) on all other models' augmentation class of the observed-data efficient influence function. If $\varphi^{\text{opt}} = \boldsymbol{\varphi^{\pi_2}_{2.3}}$ in Figure 6, this implies (a) $\mathrm{E}[(\boldsymbol{\varphi^{\pi_2}_{2.3}})^2] < \mathrm{E}[(\boldsymbol{\varphi^{\pi_2}_{2.1}})^2]$ and (b) $\mathrm{E}[(\boldsymbol{\varphi^{\pi_2}_{2.3}})^2] < \mathrm{E}[(\boldsymbol{\varphi^{\pi_1}_{2.2}})^2]$.

Appendix B. Proofs

B.1. Proof of Theorem 6: Efficient Influence Function in Model $\mathcal{M}_1$

Theorem 6: Denote the residual error $\varepsilon_r - \mathrm{E}[\varepsilon_r|\varepsilon_t]$ by $\varepsilon_r^{\perp}$. In Model $\mathcal{M}_1$, the 0-indexed row and column of $\beta \in \mathcal{B}$ are structurally zero, and the efficient full-data influence function $\varphi^{F,\text{eff}}$ therefore only has non-zero components on the $(t, M, r) \times (C, t, M)$ block. This remaining block is given by

$$\varphi^{F,\text{eff}}_{\beta_{(t,M,r)(C,t,M)}} = \begin{pmatrix} \varepsilon_t \varepsilon_C^{\top} \mathrm{Var}(\varepsilon_C)^{-1} & 0 & 0 \\ \varepsilon_M (\varepsilon_C^{\top} \mathrm{Var}(\varepsilon_C)^{-1} - \varepsilon_t \mathrm{Var}(\varepsilon_t)^{-1} \beta_{tC}) & \varepsilon_M \varepsilon_t \mathrm{Var}(\varepsilon_t)^{-1} & 0 \\ \varepsilon_r^{\perp} (\varepsilon_C^{\top} \mathrm{Var}(\varepsilon_C)^{-1} - \varepsilon_M^{\top} \mathrm{Var}(\varepsilon_M)^{-1} (\beta_{Mt}\beta_{tC} + \beta_{MC})) & 0 & \varepsilon_r^{\perp} \varepsilon_M^{\top} \mathrm{Var}(\varepsilon_M)^{-1} \end{pmatrix}.$$

Consequently, the full-data efficient influence function in $\mathcal{M}_1$ for the causal effect ξ is

$$\varphi_{\xi}^{F,\text{eff}} = \beta_{rM} \varepsilon_M \varepsilon_t \mathrm{Var}(\varepsilon_t)^{-1} + \varepsilon_r^{\perp} \varepsilon_M^{\top} \mathrm{Var}(\varepsilon_M)^{-1} \beta_{Mt}.$$

We need an auxiliary lemma for the proof. It derives the orthogonal complement of the nuisance tangent space $\Lambda_{\varepsilon}^{\perp}$. This characterization is needed to construct the efficient score and, after normalization, the efficient influence function.

Lemma 13 *Let $\varepsilon_r^\perp = \varepsilon_r - \mathrm{E}[\varepsilon_r \,|\, \varepsilon_t]$. The orthogonal complement to the ε nuisance tangent space is given by*

$$\Lambda_\varepsilon^\perp := \left\{ (g_C^\top, g_t, g_M^\top, g_r(\varepsilon_t)) \begin{pmatrix} \varepsilon_C \\ \varepsilon_t \\ \varepsilon_M \\ \varepsilon_r^\perp \end{pmatrix} \;\middle|\; \begin{array}{l} g_C \in \mathbb{R}^{d_C}, \\ g_t \in \mathbb{R}, \\ g_M \in \mathbb{R}^{d_M}, \\ g_r \in L^2(\mathbb{R}) \end{array} \right\}. \tag{2}$$

Proof Using the factorization of the error term in Equation (1), the nuisance tangent space Λ_ε decomposes as $\Lambda_{\varepsilon_C} \oplus \Lambda_{\varepsilon_{t,r}} \oplus \Lambda_{\varepsilon_M}$. Only the component $\Lambda_{\varepsilon_{t,r}}$ needs further refinement and we instead use the representation $\Lambda_{\varepsilon_{t,r}} = \Lambda_{\varepsilon_t} \oplus \Lambda_{\varepsilon_{r|t}}$. Let $\zeta_j : \mathbb{R}^d \to \mathbb{R}^{d_j}, \quad \zeta : x \mapsto x_j$ denote the projection onto the components for all $j \in \{C, t, M, r\}$. Similar to (Tsiatis, 2006), Thm. 4.6 and Thm 4.7, the nuisance tangent spaces for $j \in \{C, t, M\}$ are

$$\begin{aligned} \Lambda_{\varepsilon_j} &= \left\{ h \in L_0^2(\mathbb{R}^d) \,\middle|\, \exists h' \in L_0^2(\mathbb{R}^{d_j}) : h = h' \circ \zeta_j \wedge \mathrm{E}[h\varepsilon_j] = 0 \right\} \quad \text{, and} \\ \Lambda_{\varepsilon_{r|t}} &= \left\{ h \in L^2(\mathbb{R}^d) \,\middle|\, \exists h' \in L^2(\mathbb{R}^2) : h = h' \circ \zeta_{t,r} \wedge \forall \varepsilon_t : \mathrm{E}[h \,|\, \varepsilon_t] = 0, \mathrm{E}[h\varepsilon_r \,|\, \varepsilon_t] = 0 \right\}. \end{aligned}$$

Any concatenation in the restriction, e.g., $h' \circ \zeta_C$, must be understood that the function h is only allowed to depend on the entries in respective component, here C. The tangent spaces $(\Lambda_{\varepsilon_C}, \Lambda_{\varepsilon_t}, \Lambda_{\varepsilon_M})$ are pairwise orthogonal as they act on different components and so is $\Lambda_{\varepsilon_{r|t}}$ to $(\Lambda_{\varepsilon_C}, \Lambda_{\varepsilon_M})$. For $h_1 \in \Lambda_{\varepsilon_t}$ and $h_2 \in \Lambda_{\varepsilon_{r|t}}$ we find orthogonality via $\mathrm{E}\big[h_1^\top h_2\big] = \mathrm{E}\big[\mathrm{E}\big[h_1^\top h_2 \mid \varepsilon_t\big]\big] = \mathrm{E}\big[h_1^\top \mathrm{E}[h_2 \mid \varepsilon_t]\big] = 0$ as h_1 is $\sigma(\varepsilon_t)$-measurable and $\mathrm{E}[h_2 \mid \varepsilon_t] = 0$. Next, we show that the orthogonal space to Λ_ε is given by Equation (2). Since the nuisance tangent space Λ_ε is a direct sum, its orthogonal complement is the intersection of the orthogonal complements of its components. For $j \in \{C, t, M\}$, consider the proposed orthogonal spaces

$$\Lambda_{\varepsilon_j,\text{prop}}^\perp = \left\{ h \in L^2(\mathbb{R}^d) \,\middle|\, \exists g_j \in \mathbb{R}^{d_j} \text{ with } \mathrm{E}[h|\varepsilon_j] = g_j^\top \varepsilon_j \right\}.$$

To see $\Lambda_{\varepsilon_j,\text{prop}}^\perp \subseteq \Lambda_{\varepsilon_j}^\perp$, note that for any $a \in \Lambda_{\varepsilon_j}$ and $h \in \Lambda_{\varepsilon_j,\text{prop}}^\perp$, we find $\mathrm{E}[ah] = \mathrm{E}[a\mathrm{E}[h \,|\, \varepsilon_j]] = g_j^\top \mathrm{E}[a\varepsilon_j] = 0$. To show $\Lambda_{\varepsilon_j,\text{prop}}^\perp \supseteq \Lambda_{\varepsilon_j}^\perp$, let $h \in L_0^2(\mathbb{R}^d)$ be arbitrary. We define the function $h_1 := (h - \mathrm{E}[h \,|\, \varepsilon_j]) + \mathrm{E}[h\varepsilon_j]^\top \mathrm{Var}(\varepsilon_j)^{-1} \varepsilon_j \in \Lambda_{\varepsilon_j,\text{prop}}^\perp$ and argue that $h - h_1 \in \Lambda_{\varepsilon_j}$. The function $h - h_1 = \mathrm{E}[h \,|\, \varepsilon_j] - \mathrm{E}[h\varepsilon_j]^\top \mathrm{Var}(\varepsilon_j)^{-1} \varepsilon_j$ is by construction $\sigma(\varepsilon_j)$-measurable and as also $\mathrm{E}[h - h_1] = 0$ as well as $\mathrm{E}[(h - h_1)\varepsilon_j] = 0$, we conclude $\Lambda_{\varepsilon_j,\text{prop}}^\perp = \Lambda_{\varepsilon_j}^\perp$. Analogously, we define for the conditional component

$$\Lambda_{\varepsilon_{r|t},\text{prop}}^\perp = \left\{ h \in L^2(\mathbb{R}^d) \,\middle|\, \exists g_0, g_r \in L^2(\mathbb{R}) : \mathrm{E}[h|\varepsilon_{t,r}] = g_0(\varepsilon_t) + g_r(\varepsilon_t)^\top \varepsilon_r^\perp \right\}.$$

To see $\Lambda_{\varepsilon_{r|t},\text{prop}}^\perp \subseteq \Lambda_{\varepsilon_{r|t}}^\perp$, note that for any $a \in \Lambda_{\varepsilon_{r|t}}$ and $h \in \Lambda_{\varepsilon_{r|t},\text{prop}}^\perp$, we find

$$\begin{aligned} \mathrm{E}[ah] &= \mathrm{E}[a\mathrm{E}[h \,|\, \varepsilon_{t,r}]] = \mathrm{E}[ag_0(\varepsilon_t)] + \mathrm{E}\Big[ag_r(\varepsilon_t)^\top \varepsilon_r^\perp\Big] \\ &= \mathrm{E}[\mathrm{E}[a|\varepsilon_t]\, g_0(\varepsilon_t)] + \mathrm{E}\Big[ag_r(\varepsilon_t)^\top \varepsilon_r\Big] - \mathrm{E}\Big[ag_r(\varepsilon_t)^\top \mathrm{E}[\varepsilon_r|\varepsilon_t]\Big] \\ &= 0 + \mathrm{E}\Big[g_r(\varepsilon_t)^\top \mathrm{E}[a\varepsilon_r|\varepsilon_t]\Big] - \mathrm{E}\Big[\mathrm{E}[a|\varepsilon_t]\, g_r(\varepsilon_t)^\top \mathrm{E}[\varepsilon_r|\varepsilon_t]\Big] = 0 - 0. \end{aligned}$$

To show $\Lambda^{\perp}_{\varepsilon_{r|t},\text{prop}} \supseteq \Lambda^{\perp}_{\varepsilon_{r|t}}$, let $h \in L^2_0(\mathbb{R}^d)$ be arbitrary. We define the function

$$h_1 := (h - \mathrm{E}[h \,|\, \varepsilon_{t,r}]) + \mathrm{E}[h \,|\, \varepsilon_t] + \mathrm{E}\Big[h\varepsilon_r^{\perp} \,\Big|\, \varepsilon_t\Big]^{\top} \mathrm{Var}(\varepsilon_r \,|\, \varepsilon_t)^{-1} \varepsilon_r^{\perp} \in \Lambda^{\perp}_{\varepsilon_{r|t},\text{prop}}$$

and argue that $h - h_1 \in \Lambda_{\varepsilon_{r|t}}$. The function $h - h_1$ is by construction $\sigma(\varepsilon_{t,r})$-measurable. Using $\mathrm{E}\big[\varepsilon_r(\varepsilon_r^{\perp})^{\top} \,\big|\, \varepsilon_t\big] = \mathrm{Var}(\varepsilon_r \,|\, \varepsilon_t)$ we find

$$\begin{aligned}
\mathrm{E}[h - h_1 \,|\, \varepsilon_t] &= \mathrm{E}\bigg[\mathrm{E}[h \,|\, \varepsilon_{t,r}] - \mathrm{E}[h \,|\, \varepsilon_t] - \mathrm{E}\Big[h\varepsilon_r^{\perp} \,\Big|\, \varepsilon_t\Big]^{\top} \mathrm{Var}(\varepsilon_r \,|\, \varepsilon_t)^{-1} \varepsilon_r^{\perp} \,\bigg|\, \varepsilon_t\bigg] \\
&= 0 - \mathrm{E}\Big[h\varepsilon_r^{\perp} \,\Big|\, \varepsilon_t\Big]^{\top} \mathrm{Var}(\varepsilon_r \,|\, \varepsilon_t)^{-1} \mathrm{E}\Big[\varepsilon_r^{\perp} \,\Big|\, \varepsilon_t\Big] = 0 \quad \text{, and} \\
\mathrm{E}[(h - h_1)\varepsilon_r \,|\, \varepsilon_t] &= \mathrm{E}\bigg[\mathrm{E}[h \,|\, \varepsilon_{t,r}]\, \varepsilon_r - \mathrm{E}[h \,|\, \varepsilon_t]\, \varepsilon_r - \mathrm{E}\Big[h\varepsilon_r^{\perp} \,\Big|\, \varepsilon_t\Big]^{\top} \mathrm{Var}(\varepsilon_r \,|\, \varepsilon_t)^{-1} \varepsilon_r^{\perp}\varepsilon_r \,\bigg|\, \varepsilon_t\bigg] \\
&= \mathrm{E}[h(\varepsilon_r - \mathrm{E}[\varepsilon_r|\varepsilon_t])\varepsilon_r \,|\, \varepsilon_t] - \mathrm{E}\Big[\varepsilon_r(\varepsilon_r^{\perp})^{\top} \,\Big|\, \varepsilon_t\Big] \mathrm{Var}(\varepsilon_r \,|\, \varepsilon_t)^{-1} \mathrm{E}\Big[h\varepsilon_r^{\perp} \,\Big|\, \varepsilon_t\Big] = 0.
\end{aligned}$$

Intersecting the orthogonal complements $\Lambda^{\perp}_{\varepsilon_C}, \Lambda^{\perp}_{\varepsilon_t}, \Lambda^{\perp}_{\varepsilon_M}, \Lambda^{\perp}_{\varepsilon_{t|r}}$ yields the expression for $\Lambda^{\perp}_{\varepsilon}$ in the lemma. The linear ε_t term in $\Lambda^{\perp}_{\varepsilon_t}$ eliminates the functional term $g_0(\varepsilon_t)$ from the conditional component $\Lambda^{\perp}_{\varepsilon_{t|r}}$, completing the characterization. ■

Proof [Theorem 6] To obtain the efficient score for parameter β_{ij}, $i, j \in [d]$, in $\mathcal{M}_1$, we project the full-data score $S^F_{\beta_{ij}} = (\partial/\partial\beta_{ij}) \log(p_\varepsilon)$ onto the nuisance tangent space Λ_ε. The projection takes the form

$$\Pi(S^F_{\beta_{ij}}|\Lambda_\varepsilon) = \begin{pmatrix} \mathrm{E}\big[S_{\beta_{ij}}\varepsilon_C\big] \\ \mathrm{E}\big[S_{\beta_{ij}}\varepsilon_t\big] \\ \mathrm{E}\big[S_{\beta_{ij}}\varepsilon_M\big] \\ \mathrm{E}\big[S_{\beta_{ij}}\varepsilon_r^{\perp}|\varepsilon_t\big] \end{pmatrix}^{\top} \mathrm{E}\left[\begin{pmatrix} \varepsilon_C\varepsilon_C^{\top} & 0 & 0 & 0 \\ 0 & \varepsilon_t^2 & 0 & 0 \\ 0 & 0 & \varepsilon_M\varepsilon_M^{\top} & 0 \\ 0 & 0 & 0 & \varepsilon_r^2 \end{pmatrix}\right]^{-1} \begin{pmatrix} \varepsilon_C \\ \varepsilon_t \\ \varepsilon_M \\ \varepsilon_r^{\perp} \end{pmatrix}$$

The expectations $\mathrm{E}\big[S_{\beta_{ij}}\varepsilon_k\big]$ follow from the centered the error terms $\varepsilon_C, \varepsilon_t, \varepsilon_M, \varepsilon_r^{\perp}$ and the structural equation. Let $i \neq r$ and let $p_{\varepsilon_i,0}$ denote the true density of ε_i. We have

$$\int \left(x_i - \beta_{ij}x_j - \beta_{i(-j)}x_{-j}\right) p_{\varepsilon_i,0} \left(x_i - \beta_{ij}x_j - \beta_{i(-j)}x_{-j}\right) \,\mathrm{d}x_i = 0$$

for all $\beta \in \mathcal{B}$ and $x_{-i} = x_{1,\ldots,i-1,i+1,\ldots,d} \in \mathbb{R}^{d-1}$. Note that $\beta_{i,i} = 0$. Taking the derivative w.r.t. β_{ij} at the truth $\beta_{ij,0}$ and interchanging integration and differentiation yields for $\beta_{ij,0} \neq 0$

$$\int (-x_j) p_{\varepsilon_i,0}(\varepsilon_i) \,\mathrm{d}\varepsilon_i + \int \varepsilon_i S_{\beta_{ij}}(\varepsilon_i) p_{\varepsilon_i,0}(\varepsilon_i) \,\mathrm{d}\varepsilon_i = 0 \Leftrightarrow X_j = \mathrm{E}_{\varepsilon_i}[\varepsilon_i S_{\beta_{ij}}]. \tag{3}$$

If $\beta_{ij,0}$ is however equal to 0, this derivation gives $\mathrm{E}_{\varepsilon_i}[\varepsilon_i S_{\beta_{ij}}] = 0$. Similarly, we get $\mathrm{E}_{\varepsilon_i}[\varepsilon_i S_{\beta_{kj}}] = 0$ for all $k \neq i$ as ε_i expressed in X is not depending on β_{kj}. Under $i = r$, we can rely for all ε_t on

$$\int \left(x_r - \beta_{r,}x - \mathrm{E}[\varepsilon_r|\varepsilon_t]\right) p_{\varepsilon_r|\varepsilon_t,0} \left(x_r - \beta_{r,}x, \varepsilon_t\right) \,\mathrm{d}x_r = 0.$$

Now we differentiate $\partial/\partial\beta_{rj}$ to evaluate at $\beta_{rj,0} \neq 0$ and get $X_j = \mathrm{E}_{\varepsilon_r|\varepsilon_t}[\varepsilon_r^\perp S_{\beta_{rj}}|\varepsilon_t]$ from

$$\int (-x_j) p_{\varepsilon_r|\varepsilon_t,0}(\varepsilon_r \mid \varepsilon_t)\, \mathrm{d}\varepsilon_r + \int \varepsilon_r^\perp S^{r|t}_{\beta_{rj}}(\varepsilon_r \mid \varepsilon_t) p_{\varepsilon_r|\varepsilon_t,0}(\varepsilon_r \mid \varepsilon_t)\, \mathrm{d}\varepsilon_r = 0$$
$$\Leftrightarrow \int (-x_j) p_{\varepsilon_r|\varepsilon_t,0}(\varepsilon_r \mid \varepsilon_t)\, \mathrm{d}\varepsilon_r + \int \varepsilon_r^\perp S_{\beta_{rj}}(\varepsilon) p_{\varepsilon_r|\varepsilon_t,0}(\varepsilon_r \mid \varepsilon_t)\, \mathrm{d}\varepsilon_r = 0$$

as the score decomposed due to independencies into $S(\varepsilon) = S^{r|t}(\varepsilon_r|\varepsilon_t) + \sum_{j\in\{C,t,M\}} S^j(\varepsilon_j)$. The remaining scores $S_{\beta_{kj}}$ all have zero conditional expectation with the residual response $\varepsilon_r^\perp$ given ε_t. The efficient influence function φ_β is obtained by orthogonalizing the projected score $\Pi(S^F_{\beta_{ij}}|\Lambda_\varepsilon)$ with respect to the parameter space β and normalizing by the inverse Fisher information matrix $\mathrm{E}\big[S^{F,\mathrm{eff}}(S^{F,\mathrm{eff}})^\top\big]^{-1}$. This yields for β_{tC} the full-data efficient influence function

$$\begin{aligned}\varphi^{F,\mathrm{prop}}_{\beta_{tC}} &= (\varepsilon_t X_C^\top \mathrm{E}\big[\varepsilon_t^2\big]^{-1}) \mathrm{E}\Big[\mathrm{E}\big[\varepsilon_t^2\big]^{-1} X_C \varepsilon_t^2 X_C^\top \mathrm{E}\big[\varepsilon_t^2\big]^{-1}\Big]^{-1} \\ &= \varepsilon_t X_C^\top \mathrm{Var}(X_C)^{-1} = \varepsilon_t \varepsilon_C^\top \mathrm{Var}(\varepsilon_C)^{-1}\,.\end{aligned}$$

In this step, we have orthogonalized the components within $\varphi^{F,\mathrm{prop}}_{\beta_{tC}}$ such that we find

$$\mathrm{E}\Big[S^\top_{\beta_{tC}} \varphi^{F,\mathrm{prop}}_{\beta_{tC}}\Big] = \mathrm{E}\Big[\mathrm{E}\Big[S^\top_{\beta_{tC}} \varepsilon_t \,\Big|\varepsilon_C\Big]\, \varepsilon_C^\top \mathrm{Var}(\varepsilon_C)^{-1}\Big] = \mathrm{E}\Big[X_C \varepsilon_C^\top \mathrm{Var}(\varepsilon_C)^{-1}\Big] = \mathrm{Id},$$

where $\mathrm{E}\Big[S^\top_{\beta_{tC}} \varepsilon_t \,\Big|\varepsilon_C\Big] = X_C$ follows as a result of $\mathrm{E}\Big[S^\top_{\beta_{tC}} \varepsilon_t\Big] = X_C$ from Equation (3). Using the relation $\mathrm{E}\Big[S^\top_{\beta_{ij}} \varepsilon_t\Big] = 0$ for all $i \neq t$ or $j \notin C$ we obtain $\mathrm{E}\Big[S^\top_{\beta_{ij}} \varphi^{F,\mathrm{prop}}_{\beta_{tC}}\Big] = 0$ for all remaining parameter components. So $\varphi^{F,\mathrm{prop}}_{\beta_{tC}}$ is an influence function by Tsiatis (2006), Thm. 4.2. As $\varphi^{F,\mathrm{prop}}_{\beta_{tC}}$ lies after projection in the tangent space $\mathscr{T}$, it must be the efficient influence function by (Tsiatis, 2006), Thm 4.3, proving the efficiency of $\varphi^{F,\mathrm{prop}}_{\beta_{tC}} = \varphi^{F,\mathrm{eff}}_{\beta_{tC}}$ in Model $\mathcal{M}_1$ for the parameter β_{tC}. Analogously we obtain $\varphi^{F,\mathrm{eff}}_{\beta_{Mt}} = \varepsilon_M \varepsilon_t \mathrm{Var}(\varepsilon_t)^{-1}$. When we however focus on the remaining components of β, say β_{MC}, scores of other β components corresponding to incoming edges of M, so $S_{\beta_{Mt}}$ have non-zero expectation with ε_M. Correcting for these components is the orthogonalization in the parameter space. In particular, the efficient influence function for β_{MC} reads as

$$\varphi^{F,\mathrm{eff}}_{\beta_{MC}} = \varepsilon_M (\varepsilon_C^\top \mathrm{Var}(\varepsilon_C)^{-1} - \varepsilon_t \mathrm{Var}(\varepsilon_t)^{-1}\, \beta_{tC}).$$

For any $i \in [d_M], j \in [d_C]$, we obtain, with the unit vectors e_i, e_j,

$$\begin{aligned}\mathrm{E}\Big[S_{\beta_{M_i C_j}} \varphi^{F,\mathrm{eff}}_{\beta_{MC}}\Big] &= \mathrm{E}\Big[\mathrm{E}\Big[S_{\beta_{M_i C_j}} \varepsilon_M \,\Big|\, \varepsilon_{C,t}\Big]\, (\varepsilon_C^\top \mathrm{Var}(\varepsilon_C)^{-1} - \varepsilon_t \mathrm{Var}(\varepsilon_t)^{-1}\, \beta_{tC})\Big] \\ &= \mathrm{E}\Big[e_i e_j^\top X_C (\varepsilon_C^\top \mathrm{Var}(\varepsilon_C)^{-1} - \varepsilon_t \mathrm{Var}(\varepsilon_t)^{-1}\, \beta_{tC})\Big] = e_i e_j^\top\end{aligned}$$

as $\mathrm{E}[X_C \varepsilon_t] = 0$. Now we can inspect the inner product of $\varphi^{F,\mathrm{eff}}_{\beta_{MC}}$ with the score $S_{\beta_{Mt}}$ to find

$$\begin{aligned}\mathrm{E}\Big[S_{\beta_{M_i t}} \varphi^{F,\mathrm{eff}}_{\beta_{MC}}\Big] &= \mathrm{E}\Big[\mathrm{E}\Big[S_{\beta_{M_i t}} \varepsilon_M \,\Big|\, \varepsilon_{C,t}\Big]\, (\varepsilon_C^\top \mathrm{Var}(\varepsilon_C)^{-1} - \varepsilon_t \mathrm{Var}(\varepsilon_t)^{-1}\, \beta_{tC})\Big] \\ &= \mathrm{E}\Big[e_i X_t (\varepsilon_C^\top \mathrm{Var}(\varepsilon_C)^{-1} - \varepsilon_t \mathrm{Var}(\varepsilon_t)^{-1}\, \beta_{tC})\Big] \\ &= \mathrm{E}\Big[e_i (\beta_{tC}\varepsilon_C + \varepsilon_t)(\varepsilon_C^\top \mathrm{Var}(\varepsilon_C)^{-1} - \varepsilon_t \mathrm{Var}(\varepsilon_t)^{-1}\, \beta_{tC})\Big] = 0.\end{aligned}$$

The crucial point is the fact that X_t decomposes to $\beta_{tC}\varepsilon_C + \varepsilon_t$. Using that the expectation of any other score component multiplied with ε_M is zero, we have thus shown that the proposed function is orthogonal in the parameter nuisance space. As $\varphi^{F,\text{eff}}_{\beta_{MC}}$ lies as a result of linearity in the nuisance tangent space $\mathscr{T}$, it must be efficient for β_{MC} in $\mathcal{M}_1$. Using $X_j = \mathrm{E}_{\varepsilon_r|\varepsilon_t}[\varepsilon_r^\perp S_{\beta_{rj}}|\varepsilon_t]$ we find $\varphi^{F,\text{eff}}_{\beta_{rM}} = \varepsilon_r^\perp \varepsilon_M^\top \mathrm{Var}(\varepsilon_M)^{-1}$, which is also orthogonal to all functions in $\Lambda_\varepsilon^\perp$. Analogously proceeding, we summarize the full-data efficient influence function for $\beta_{(t,M,r)(C,t,M)}$ in $\mathcal{M}_1$ to be $\varphi^{F,\text{eff}}_{\beta_{(t,M,r)(C,t,M)}} =$

$$\begin{pmatrix} \varepsilon_t\varepsilon_C^\top \mathrm{Var}(\varepsilon_C)^{-1} & 0 & 0 \\ \varepsilon_M(\varepsilon_C^\top \mathrm{Var}(\varepsilon_C)^{-1} - \varepsilon_t \mathrm{Var}(\varepsilon_t)^{-1}\beta_{tC}) & \varepsilon_M\varepsilon_t \mathrm{Var}(\varepsilon_t)^{-1} & 0 \\ \varepsilon_r^\perp(\varepsilon_C^\top \mathrm{Var}(\varepsilon_C)^{-1} - \varepsilon_M^\top \mathrm{Var}(\varepsilon_M)^{-1}(\beta_{Mt}\beta_{tC} + \beta_{MC})) & 0 & \varepsilon_r^\perp\varepsilon_M^\top \mathrm{Var}(\varepsilon_M)^{-1} \end{pmatrix}.$$

The component $\beta_{S,}$ is irrelevant in the efficient estimation of ξ and can therefore be considered as nuisance; see also Bhattacharya et al. (2020). Finally, by linearity of the causal effect $\xi = \beta_{rM}\beta_{Mt}$, the efficient influence function for ξ must be

$$\varphi^{F,\text{eff}}_{\xi} = \beta_{rM}\varepsilon_M\varepsilon_t \mathrm{Var}(\varepsilon_t)^{-1} + \varepsilon_r^\perp\varepsilon_M^\top \mathrm{Var}(\varepsilon_M)^{-1}\beta_{Mt},$$

the corresponding linear combination of the efficient influence functions for β_{rM} and β_{Mt}. This function is orthogonal to all nuisance directions and therefore efficient in $\mathcal{M}_1$. ■

B.2. Proof of Lemma 8: Efficient Influence Function in Model $\mathcal{M}_\pi$

Lemma 8: Let $\pi_1 \in \mathcal{C}(\mathbb{R}^{d_C+1}, (0,1])$ and $\pi_2 \in \mathcal{C}(\mathbb{R}^{d_C+d_M+1}, (0,1])$ be given. Under the Model $\mathcal{M}_\pi$, the optimal observed-data influence-function augmentation of the full-data efficient influence function for the $(t,M,r)\times(C,t,M)$ block of the parameter matrix β and for the causal effect ξ are

$$\varphi^{\text{opt},\pi}_{\beta_{(t,M,r)(C,t,M)}}(X) = \begin{pmatrix} \mathrm{diag}(1) & 0 & 0 \\ 0 & \frac{\mathrm{diag}(I(\Delta\geq 2))}{\pi_1(X_{C,t})} & 0 \\ 0 & 0 & \frac{\mathrm{diag}(I(\Delta=\infty))}{\pi_1(X_{C,t})\pi_2(X_{C,t,M})} \end{pmatrix} \varphi^{F,\text{eff}}_{\beta_{(t,M,r)(C,t,M)}}(X), \quad \text{and}$$

$$\varphi^{\text{opt},\pi}_{\xi}(X) = \frac{I(\Delta\geq 2)\beta_{rM}\varepsilon_M\varepsilon_t \mathrm{Var}(\varepsilon_t)^{-1}}{\pi_1(X_{C,t})} + \frac{I(\Delta=\infty)\varepsilon_r^\perp\varepsilon_M^\top \mathrm{Var}(\varepsilon_M)^{-1}\beta_{Mt}}{\pi_1(X_{C,t})\pi_2(X_{C,t,M})}.$$

Proof We define the coarsened model tangent space Λ_π as the tangent space in Model $\mathcal{M}_\pi$ with respect to the propensity function π. The corresponding observed-data augmentation space is

$$\Lambda_{2,\pi} = \left\{ \sum_{\delta\in\{1,2\}} \frac{I(\Delta=\delta) - (1-\pi_\delta)I(\Delta\geq\delta)}{\Pi_{j=1}^{\delta}\pi_j}(h_\delta \circ G_\delta) \,\middle|\, h_1 \in L_0^2(\mathbb{R}^{d_C+1}, \mathbb{R}), h_2 \in L_0^2(\mathbb{R}^{d_C+1+d_M}, \mathbb{R}) \right\}.$$

For any full-data influence function φ^F, the set of all corresponding augmented observed-data influence functions in Model $\mathcal{M}_\pi$ is the functional space

$$IF^\pi(\varphi^F) = \left\{ \left[\frac{I(\Delta=\infty)\varphi^F}{\pi_1\pi_2} + h \right] - \Pi\left([\,\cdot\,]|\Lambda_\pi\right) \,\middle|\, \forall h \in \Lambda_{2,\pi} \right\}.$$

Applying Tsiatis (2006), Thm 10.1 and Thm 10.4, to our setting yields the variance-minimizing, optimal element in $IF^{\pi}(\varphi^F)$. Denoting this optimal augmented full-data influence function by φ^*, its observed-data representation is

$$\mathcal{J}(\varphi^*) = \frac{I(\Delta=\infty)\varphi^*}{\pi_1\pi_2} + \frac{I(\Delta=1)-(1-\pi_1)I(\Delta\geq 1)}{\pi_1}\mathrm{E}[\varphi^* \mid \varepsilon_{C,t}] + \frac{I(\Delta=2)-(1-\pi_2)I(\Delta\geq 2)}{\pi_1\pi_2}\mathrm{E}[\varphi^* \mid \varepsilon_{C,t,M}]\,.$$

Each component in the full-data efficient influence function $\varphi_\beta^{F,\text{eff}}$ satisfies

$$\mathrm{E}\left[\varphi_{\beta_{ij}}^{F,\text{eff}}|\varepsilon_{C,t}\right] \in \{0, \varphi_{\beta_{ij}}^{F,\text{eff}}\}, \quad \mathrm{E}\left[\varphi_{\beta_{ij}}^{F,\text{eff}}|\varepsilon_{C,t,M}\right] \in \{0, \varphi_{\beta_{ij}}^{F,\text{eff}}\} \quad , i,j \in [d].$$

Consequently, the optimal full-data influence function within the augmentation space $IF^{\pi}(\varphi_\beta^{F,\text{eff}})$, and analogously within $IF^{\pi}(\varphi_\xi^{F,\text{eff}})$, simplifies to

$$\varphi^{\text{opt}}_{\beta_{(t,M,r)(C,t,M)}}(X;\pi) = \begin{pmatrix} \text{diag}(1) & 0 & 0 \\ 0 & \frac{\text{diag}(I(\Delta\geq 2))}{\pi_1(X_{C,t})} & 0 \\ 0 & 0 & \frac{\text{diag}(I(\Delta=\infty))}{\pi_1(X_{C,t})\pi_2(X_{C,t,M})} \end{pmatrix} \varphi^{F,\text{eff}}_{\beta_{(t,M,r)(C,t,M)}}(X)$$

and

$$\varphi_\xi^{\text{opt}}(X;\pi) = \frac{I(\Delta\geq 2)\beta_{rM}\varepsilon_M\varepsilon_t\mathrm{Var}(\varepsilon_t)^{-1}}{\pi_1(X_{C,t})} + \frac{I(\Delta=\infty)\varepsilon_r^{\perp}\varepsilon_M^{\top}\mathrm{Var}(\varepsilon_M)^{-1}\beta_{Mt}}{\pi_1(X_{C,t})\pi_2(X_{C,t,M})}.$$

It is important to note that φ_β might not be the efficient influence function in the observed-data Model $\mathcal{M}_\pi$. Every full-data influence function lies in the affine space $\varphi_\beta^{F,\text{eff}} + \Lambda_\varepsilon^{\perp}$ as characterized in Equation (2). Because the operator $\mathcal{J}$ is linear, the efficient observed-data influence function for β must take the form $\varphi_\beta^{F,\text{eff}} + u$, where $u \in \Lambda_\varepsilon^{\perp}$ is chosen to minimize the variance

$$\mathrm{Var}\left(\mathcal{J}(\varphi_\beta^{\text{F,eff}})\right) + \mathrm{Var}(\mathcal{J}(u)) + 2\mathrm{Cov}\left(\mathcal{J}(u), \mathcal{J}(\varphi_\beta^{\text{F,eff}})\right).$$

■

B.3. Proof of Lemma 9: Optimized Estimators are Solutions to Linear Equations

Lemma 9: The observed-data optimized RAL estimators $\hat{\beta}_{Mt,n}$ and $\hat{\beta}_{rM,n}$ are solutions to a sequence of linear equations on the full and partially observed datasets. Their product estimator $\hat{\xi}_n = \hat{\beta}_{rM,n}\hat{\beta}_{Mt,n}$ is the optimized RAL estimator for the causal effect ξ in Model $\mathcal{M}_\pi$. The exact formulas are provided in Appendix B.3.

Proof We prove the statement by deriving the optimized estimator for $\beta_{(t,M,r)(C,t,M)}$ and subsequently applying the delta method to $\xi = \phi(\beta) \equiv \beta_{rM}\beta_{Mt}$. Let $\hat{\beta}_n$ denote the solution to the d^2 estimating equations $0^{d^2} = \sum_i \varphi_\beta^{\text{opt},\pi}(X^{(i)};\theta)$ on the observed-data influence $\varphi_\beta^{\text{opt},\pi}$ function from

Lemma 8. This estimator satisfies $\sqrt{n}(\hat{\beta}_n - \beta) \xrightarrow{P_{\theta;\pi}} \mathcal{N}_{d^2}\left(0, \text{Var}(\varphi_\beta^{\text{opt},\pi})\right)$ (Tsiatis (2006), Chapter 3.3). Using the structural equation model $X = \beta X + \varepsilon$, the relevant estimating equations take the form

$$\begin{pmatrix} 0^{d_C} \\ 0^{d_M} \\ 0^{d_M \times d_C} \\ 0^{1\times d_M} \\ 0^{1\times d_C} \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^n \left(X_t^{(i)} - \beta_{Ct} X_C^{(i)}\right) X_C^{(i)\top} \text{Var}(\varepsilon_C)^{-1} \\ \sum_{i=1}^n \frac{I(\Delta\geq 2)}{\pi_1(X_{C,t}^{(i)})} \left(X_M^{(i)} - \beta_{Mt} X_t^{(i)} - \beta_{MC} X_C^{(i)}\right) \hat{\varepsilon}_t^{(i)} \text{Var}(\varepsilon_t)^{-1} \\ \sum_{i=1}^n \frac{I(\Delta\geq 2)}{\pi_1(X_{C,t,M}^{(i)})} \left(X_M^{(i)} - \beta_{Mt} X_t^{(i)} - \beta_{MC} X_C^{(i)}\right)^\top \left(X_C^{(i)\top} \text{Var}(\varepsilon_C)^{-1} - \hat{\varepsilon}_t^{(i)} \text{Var}(\varepsilon_t)^{-1} \beta_{tC}\right) \\ \sum_{i=1}^n \frac{I(\Delta=\infty)}{\pi_1(X_{C,t}^{(i)})\pi_2(X_{C,t,M}^{(i)})} \left(X_r^{(i)} - \beta_{rM} X_M^{(i)} - \beta_{rC} X_C^{(i)} - \text{E}\left[X_r^{(i)} - \beta_{rM} X_M^{(i)} - \beta_{rC} X_C^{(i)} | \varepsilon_t\right]\right)^\top \hat{\varepsilon}_M^{(i)} \text{Var}(\varepsilon_M)^{-1} \\ \sum_{i=1}^n \frac{I(\Delta=\infty)}{\pi_1(X_{C,t}^{(i)})\pi_2(X_{C,t,M}^{(i)})} \left(X_r^{(i)} - \beta_{rM} X_M^{(i)} - \beta_{rC} X_C^{(i)} - \text{E}\left[X_r^{(i)} - \beta_{rM} X_M^{(i)} - \beta_{rC} X_C^{(i)} | \varepsilon_t\right]\right)^\top \left(X_C^{(i)\top} \text{Var}(\varepsilon_C)^{-1} - \hat{\varepsilon}_M^{(i)} \text{Var}(\varepsilon_M)^{-1} (\beta_{Mt}\beta_{tC} + \beta_{MC})\right) \end{pmatrix}.$$

Each block corresponds to a linear equation system and can be solved iteratively. The first block uniquely determines the closed-form estimator for β_{tC} to be

$$\hat{\beta}_{tC,n} = \left(\sum_{i=1}^n X_t^{(i)} X_C^{(i)\top}\right)\left(\sum_{i=1}^n X_C^{(i)} X_C^{(i)\top}\right)^{-1}.$$

This allows us to replace to specify all treatment error terms as $\hat{\varepsilon}_t^{(i)} = X_t^{(i)} - \hat{\beta}_{Ct,n} X_C^{(i)}$. The next two equation systems have to be solved simultaneously. Let $W_i = I(\Delta \geq 2)/\pi_1(X_{C,t}^{(i)})$. Define the weighted empirical correlations

$$E_{Mt} = \sum_{i=1}^n W_i X_M^{(i)} \hat{\varepsilon}_t^{(i)}, \qquad E_{tt} = \sum_{i=1}^n W_i X_t^{(i)} \hat{\varepsilon}_t^{(i)}, \qquad E_{Ct} = \sum_{i=1}^n W_i X_C^{(i)} \hat{\varepsilon}_t^{(i)},$$
$$E_{MC} = \sum_{i=1}^n W_i X_M^{(i)} \hat{\varepsilon}_C^{(i)\top}, \qquad E_{tC} = \sum_{i=1}^n W_i X_t^{(i)} \hat{\varepsilon}_C^{(i)\top}, \qquad E_{CC} = \sum_{i=1}^n W_i X_C^{(i)} \hat{\varepsilon}_C^{(i)\top}.$$

Subtracting β_{tC} times the second system from the third system yields the joint solution to the linear equation

$$\begin{pmatrix} \hat{\beta}_{Mt,n} \\ \hat{\beta}_{MC,n} \end{pmatrix} = \begin{pmatrix} E_{tt} & E_{Ct} \\ E_{tC} & E_{CC} \end{pmatrix}^{-1} \begin{pmatrix} E_{Mt} \\ E_{MC} \end{pmatrix}.$$

The mediating residuals are $\hat{\varepsilon}_M^{(i)} = X_M^{(i)} - \hat{\beta}_{Mt,n} X_t^{(i)} - \hat{\beta}_{MC,n} X_C^{(i)}$. For the final set of equations, we incorporate the orthogonality restriction between ε_t and $\varepsilon_r^\perp$. Since $\varphi_\beta^{\text{opt}}$ is orthogonal to all elements in $\Lambda_{r|t}$, only ε_t-linear components of $\text{E}\left[\varepsilon_r^{(i)} | \varepsilon_t\right]$ contribute to the estimation. Hence, we can replace this conditional expectation by its best linear predictor $\gamma \hat{\varepsilon}_t^{(i)}$. Let $W_i = I(\Delta = \infty)/\pi_1(X_{C,t}^{(i)})\pi_2(X_{C,t,M}^{(i)})$ and define $Z_1^{(i)} = (X_C^{(i)}, \hat{\varepsilon}_t^{(i)}, \hat{\varepsilon}_M^{(i)})$, $Z_2^{(i)} = (X_C^{(i)}, \hat{\varepsilon}_t^{(i)}, X_M^{(i)})$, and $\theta = (\beta_{rC}^\top, \gamma, \beta_{rM}^\top)^\top$. The last two equation systems for the estimators of β_{rC} and β_{rM} are jointly solved by

$$\begin{pmatrix} \sum_i W_i X_C^{(i)} (X_r^{(i)} - Z_2^{(i)\top}\theta) = 0^{d_C} \\ \sum_i W_i \hat{\varepsilon}_t^{(i)} (X_r^{(i)} - Z_2^{(i)\top}\theta) = 0 \\ \sum_i W_i \hat{\varepsilon}_M^{(i)} (X_r^{(i)} - Z_2^{(i)\top}\theta) = 0^{d_M} \end{pmatrix} \Leftrightarrow \begin{pmatrix} \hat{\beta}_{rC,n} \\ \hat{\gamma}_n \\ \hat{\beta}_{rM,n} \end{pmatrix} = (\sum_i W_i Z_1^{(i)} Z_2^{(i)\top})^{-1} (\sum_i W_i Z_1^{(i)} X_r^{(i)}).$$

Using orthogonality of the efficient influence function components, i.e., $\mathrm{Cov}\left(\varphi^{\mathrm{opt},\pi}_{\beta_{Mt}}, \varphi^{\mathrm{opt},\pi\top}_{\beta_{rM}}\right) = 0$, we obtain the joint asymptotic distribution

$$\sqrt{n}\left(\begin{pmatrix}\hat\beta_{Mt,n}\\ \hat\beta^\top_{rM,n}\end{pmatrix} - \begin{pmatrix}\beta_{Mt}\\ \beta^\top_{rM}\end{pmatrix}\right) \xrightarrow{P_{\theta;\pi}} \mathcal{N}_{2d_M}\left(0, \begin{pmatrix}\mathrm{Var}\left(\varphi^{\mathrm{opt},\pi}_{\beta_{Mt}}\right) & 0\\ 0 & \mathrm{Var}\left(\varphi^{\mathrm{opt},\pi}_{\beta_{rM}}\right)\end{pmatrix}\right).$$

Applying the delta method to $\hat\xi_n = \hat\beta_{rM,n}\hat\beta_{Mt,n}$ with $\xi_n = \beta_{rM}\hat\beta_{Mt}$ yields the asymptotic

$$\sqrt{n}(\hat\xi_n - \xi) \xrightarrow{P_{\theta;\pi}} \mathcal{N}\left(0, \nabla\phi(\beta)^\top \begin{pmatrix}\mathrm{Var}\left(\varphi^{\mathrm{opt},\pi}_{\beta_{Mt}}\right) & 0\\ 0 & \mathrm{Var}\left(\varphi^{\mathrm{opt},\pi}_{\beta_{rM}}\right)\end{pmatrix}\nabla\phi(\beta)\right),$$

where $\nabla\phi(\beta)^\top = (\beta_{rM}, \beta^\top_{Mt})$. Thus the asymptotic variance of $\hat\xi_n = \hat\beta_{rM,n}\hat\beta_{Mt,n}$ coincides with the derived efficiency bound $\mathrm{Var}\left(\varphi^{\mathrm{opt},\pi}_{\xi}\right)$ within $IF^\pi(\varphi^{F,\mathrm{eff}}_\xi)$, as stated in Lemma 8. ■

B.4. Proof of Lemma 10: Variance of Efficient Effect Estimator in Model $\mathcal{M}_\pi$

Lemma 10: The asymptotic variance $\mathrm{Var}_\infty(\hat\xi_n;\pi)$ of the optimized estimator $\hat\xi_n$ for the causal effect ξ in Model $\mathcal{M}_\pi$, $\pi_1 \in \mathcal{C}(\mathbb{R}^{d_C+1},(0,1])$, $\pi_2 \in \mathcal{C}(\mathbb{R}^{d_C+d_M+1},(0,1])$, is determined by

$$\mathrm{Var}\left(\varphi^{\mathrm{opt},\pi}_\xi\right) = \mathrm{E}\left[\frac{\varepsilon_t^2\beta_{rM}\mathrm{Var}(\varepsilon_M)\,\beta^\top_{rM}}{\mathrm{Var}(\varepsilon_t)^2\,\pi_1(X_{C,t})}\right] + \mathrm{E}\left[\frac{\left(\varepsilon_M^\top\mathrm{Var}(\varepsilon_M)^{-1}\beta_{Mt}\right)^2\mathrm{Var}(\varepsilon_r\,|\,\varepsilon_t)}{\pi_1(X_{C,t})\pi_2(X_{C,t,M})}\right].$$

Proof Using the coarsening probabilities $\mathcal{P}(\Delta \geq 2\,|\,\varepsilon_{C,t}) = \pi_1(X_{C,t})$, $\mathcal{P}(\Delta = \infty\,|\,\varepsilon_{C,t,M}) = \pi_1(X_{C,t})\pi_2(X_{C,t,M})$, and the observation that $\mathrm{E}\left[\varepsilon_r^\perp\varepsilon_t\varepsilon_M\varepsilon_M^\top\right] = 0$ is cancelling mixed terms, we get

$$\begin{aligned}
&\mathrm{Var}\left(\varphi^{\mathrm{opt},\pi}_\xi\right) =\\
&\quad= \mathrm{E}\left[\frac{I(\Delta\geq 2)}{\pi_1^2}\left(\beta_{rM}\varepsilon_M\varepsilon_t\mathrm{Var}(\varepsilon_t)^{-1}\right)^2\right] + \mathrm{E}\left[\frac{I(\Delta=\infty)}{\pi_1^2\pi_2^2}\left(\varepsilon_r^\perp\varepsilon_M^\top\mathrm{Var}(\varepsilon_M)^{-1}\beta_{Mt}\right)^2\right]\\
&\quad= \mathrm{E}\left[\frac{\varepsilon_t^2}{\mathrm{Var}(\varepsilon_t)^2\,\pi_1^2}\mathrm{E}[I(\Delta\geq 2)\,|\,\varepsilon_{C,t}]\,\beta_{rM}\mathrm{E}\left[\varepsilon_M\varepsilon_M^\top\,|\,\varepsilon_{C,t}\right]\beta^\top_{rM}\right]\\
&\qquad+ \mathrm{E}\left[\frac{\left(\varepsilon_M^\top\mathrm{Var}(\varepsilon_M)^{-1}\beta_{Mt}\right)^2}{\pi_1^2\pi_2^2}\mathrm{E}[I(\Delta=\infty)\,|\,\varepsilon_{C,t,M}]\,\mathrm{E}\left[\varepsilon_r^{\perp 2}\,|\,\varepsilon_{C,t,M}\right]\right]\\
&\quad= \mathrm{E}\left[\frac{\varepsilon_t^2\beta_{rM}\mathrm{Var}(\varepsilon_M)\,\beta^\top_{rM}}{\mathrm{Var}(\varepsilon_t)^2\,\pi_1(X_{C,t})}\right] + \mathrm{E}\left[\frac{\left(\varepsilon_M^\top\mathrm{Var}(\varepsilon_M)^{-1}\beta_{Mt}\right)^2\mathrm{Var}(\varepsilon_r\,|\,\varepsilon_t)}{\pi_1(X_{C,t})\pi_2(X_{C,t,M})}\right].
\end{aligned}$$

■

B.5. Proof of Theorem 11: Proof of the Optimization Problem

Theorem 11: Let X follow model $\mathcal{M}_\pi$. Let the base cost c_0, cost functions c_1, c_2 as well as the average budget b_0 be according to the specifications of Problem 5. Under the additional restriction that admissible influence functions φ satisfy $\varphi \in IF^\pi(\varphi^{F,\text{eff}})$, Problem 5 is uniquely solvable almost everywhere. Define the leverages $g_1(X_{C,t}) := \beta_{rM}\text{Var}(\varepsilon_M)\,\beta_{rM}^\top \frac{\varepsilon_t^2}{\text{Var}(\varepsilon_t)^2}$, and $g_2(X_{C,t,M}) := \text{Var}(\varepsilon_r \mid \varepsilon_t)\left(\varepsilon_M^\top \text{Var}(\varepsilon_M)^{-1}\beta_{Mt}\right)^2$. Then the optimal propensity π^* is given by

$$(\pi_1^*, \pi_2^*) = \begin{cases} \left(\min\left(1, \max\left(\sqrt{\frac{g_1}{\lambda c_1}}, \sqrt{\frac{g_1 + \text{E}[g_2|\varepsilon_{C,t}]}{\lambda(c_1 + \text{E}[c_2|\varepsilon_{C,t}])}}\right)\right), \min\left(1, \sqrt{\frac{g_2 c_1}{g_1 c_2}}\right)\right), & g_1 < \lambda c_1, \\ \left(1, \min\left(1, \sqrt{\frac{g_2}{\lambda c_2}}\right)\right), & g_1 \geq \lambda c_1, \end{cases}$$

where the constant $\lambda > 0$ is chosen to satisfy the problem's budget constraint

$$\text{E}[c_0 + (\pi_1 c_1)(X_{C,t}) + \pi_1(X_{C,t})(\pi_2 c_2)(X_{C,t,M})] = b_0.$$

Proof The functions g_1 and g_2 are defined so that $\text{Var}\left(\varphi_\xi^{\text{opt}}\right) = \text{E}[g_1/\pi_1 + g_2/(\pi_1\pi_2)]$. Let $\lambda \in \mathbb{R}$ and let $\nu_1, \nu_2, \kappa_1, \kappa_2$ be measurable functions. The Lagrangian $\mathcal{L}(\pi, \lambda, \nu, \kappa)$ associated with this optimization problem is

$$\mathcal{L}(\pi, \lambda, \nu, \kappa) := \int \frac{g_1}{\pi_1} + \frac{g_2}{\pi_1\pi_2} + \lambda(c_1\pi_1 + c_2\pi_1\pi_2 - b_0) - \nu_1\pi_1 + \kappa_1(\pi_1 - 1) \tag{4}$$

$$- \nu_2\pi_2 + \kappa_2(\pi_2 - 1)\,\text{d}\mathcal{P}_\theta, \tag{5}$$

where we suppress the index $\mathcal{P}_{\theta,\pi}$ and write $\mathcal{P}_\theta$, since π affects only the distribution of the sampling stage variable Δ (Ito and Kunisch, 2008). First, we consider a variation in the second-stage propensity. For $h \in \mathcal{L}^2$, define $\pi_{\gamma;2} = (\pi_1, \pi_2 + \gamma h)$ and compute the derivative of the Lagrangian at $\gamma = 0$:

$$\left.\frac{\partial \mathcal{L}(\pi_{\gamma;2}, \lambda, \nu, \kappa)}{\partial \gamma}\right|_{\gamma=0} = \int \left(-\frac{g_2}{\pi_1\pi_2^2} + \lambda c_2\pi_1 - \nu_2 + \kappa_2\right) h\,\text{d}\mathcal{P}_\theta$$

For any optimal π, this expression must vanish for all h, and therefore the bracketed term must vanish almost everywhere. For interior points of the constraint set, complementary slackness yields the pointwise condition

$$-\frac{g_2}{\pi_1\pi_2^2} + \lambda c_2\pi_1 = 0 \quad \Leftrightarrow \quad \pi_2^2 = \frac{g_2}{\lambda c_2\pi_1^2}. \tag{6}$$

Analogously, a variation on π_1 yields the necessary condition

$$-\frac{g_1}{\pi_1^2} - \frac{g_2}{\pi_1^2\pi_2} + \lambda(c_1 + c_2\pi_2) = 0 \quad \Leftrightarrow \quad \pi_1^2 = \frac{g_1 + g_2/\pi_2}{\lambda(c_1 + c_2\pi_2)}. \tag{7}$$

Substituting the expression for π_2 from Equation (6) gives

$$g_2/\pi_2 = \sqrt{g_2\lambda c_2}\pi_1, c_2\pi_2 = \sqrt{g_2 c_2}/(\pi_1\sqrt{\lambda})$$

and therefore Equation (7) simplifies to $\pi_1^2 = g_1/(\lambda c_1) =: \tilde{\pi}_1^2 > 0$. Using this representation in Equation (6) yields $\pi_2^2 = (g_2 c_1)/(g_1 c_2) =: \tilde{\pi}_2^2 > 0$. Whenever both $\tilde{\pi}_1^2$ and $\tilde{\pi}_2^2$ lie in the interior of the feasible set, the pair $\tilde{\pi}_{1,2}$ is optimal. If either exceeds 1, the corresponding component must be clipped at the boundary and the KKT conditions adjust accordingly. When $\pi_2^* = 1$ or $\pi_1^* = 1$ the Equations (7) and (6) reduce to $\pi_1^2 = \frac{g_1+g_2}{\lambda(c_1+c_2)}$ and $\pi_2^2 = \frac{g_2}{\lambda c_2}$, yielding mixed solutions. Finally, if both $\tilde{\pi}_1^2$ and $\tilde{\pi}_1^2$ exceed 1, no mixed solution is feasible, since the right-hand sides of Equations (6) and (7) are non-decreasing with decreasing as π_1 and π_2, decrease. Thus the solution is $\pi^* = (1, 1)$. Collecting all cases, every solution of the Lagrangian (4) must satisfy the following pointwise conditions:

$$(\pi_1, \pi_2) = \begin{cases} \left(\sqrt{\frac{g_1}{\lambda c_1}}, \sqrt{\frac{g_2 c_1}{g_1 c_2}}\right) & \text{, where } g_1 < \lambda c_1 \text{ , and } g_2 c_1 < g_1 c_2 \\ \left(\min\left(1, \sqrt{\frac{g_1+g_2}{\lambda(c_1+c_2)}}\right), 1\right) & \text{, where } g_1 < \lambda c_1 \text{ , and } g_2 c_1 \geq g_1 c_2 \\ \left(1, \min\left(1, \sqrt{\frac{g_2}{\lambda c_2}}\right)\right) & \text{, where } g_1 \geq \lambda c_1 \end{cases}$$

Rearranging terms and using continuity at $g_2 c_1 = g_1 c_2$, this can be equivalently expressed as

$$(\pi_1, \pi_2) = \begin{cases} \left(\min\left(1, \max\left(\sqrt{\frac{g_1}{\lambda c_1}}, \sqrt{\frac{g_1+g_2}{\lambda(c_1+c_2)}}\right)\right), \min\left(1, \sqrt{\frac{g_2 c_1}{g_1 c_2}}\right)\right) & \text{, where } g_1 < \lambda c_1 \\ \left(1, \min\left(1, \sqrt{\frac{g_2}{\lambda c_2}}\right)\right) & \text{, where } g_1 \geq \lambda c_1 \end{cases}.$$

However, since c_2 and g_2 depend on ε_M, they are not available when computing π_1. Thus the above representation it is not admissible. Its conditional version,

$$(\pi_1^*, \pi_2^*) = \begin{cases} \left(\min\left(1, \max\left(\sqrt{\frac{g_1}{\lambda c_1}}, \sqrt{\frac{g_1+\mathrm{E}[g_2|\varepsilon_{C,t}]}{\lambda(c_1+\mathrm{E}[c_2|\varepsilon_{C,t}])}}\right)\right), \min\left(1, \sqrt{\frac{g_2 c_1}{g_1 c_2}}\right)\right) & , g_1 < \lambda c_1 \\ \left(1, \min\left(1, \sqrt{\frac{g_2}{\lambda c_2}}\right)\right) & , g_1 \geq \lambda c_1 \end{cases}$$

is admissible and optimal. Note that (π_1^*, π_2^*) is continuous and π_2^* remains unchanged. Optimality follows by applying the law of iterated expectation to the Lagrangian in Equation (4). Finally, the optimal multiplier λ^* is determined by enforcing the budget constraint $\mathrm{E}[c_0 + (\pi_1 c_1)(X_{C,t}) + \pi_1(X_{C,t})(\pi_2 c_2)(X_{C,t,M})] = b_0$ which concludes the solution to a restricted version of Problem 5. ■

Appendix C. Simulation Study. Setup and Experiments

This section describes the front-door model and the experiments analysed in Section 5.1. Our ground truth parameter matrix, with dimension specifications $d_C = 2$, $d_M = 3$, is given by

$$\beta = \left(\begin{array}{ccc|ccc|c|c} 0 & 0 & 0 & 0 & 0 & 0 & 0 & \cdot \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \cdot \\ \hline 0.5 & -0.2 & 0 & 0 & 0 & 0 & 0 & \cdot \\ \hline 0.3 & 0.1 & 0.7 & 0 & 0 & 0 & 0 & \cdot \\ 0.5 & 0.2 & 0.2 & 0 & 0 & 0 & 0 & \cdot \\ -0.1 & 0.3 & 0.1 & 0 & 0 & 0 & 0 & \cdot \\ \hline 0.2 & -0.1 & 0 & 0.5 & 0.4 & -0.3 & 0 & \cdot \\ \hline \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \end{array}\right)$$

This specifications determines the causal effect ξ to be $\beta_{rM}\beta_{Mt} = 0.4$. The data is now generated by the structural equations $X = (I - \beta)^{-1}\varepsilon$ where the error distributions have the form

$$\varepsilon_C \sim t_2\left(\begin{pmatrix}0\\0\end{pmatrix}, \sigma_C, df = 5\right), \varepsilon_{tr} \sim t_2\left(\begin{pmatrix}0\\0\end{pmatrix}, \sigma_{tr}, df = 5\right) \text{ and } \varepsilon_M \sim \mathcal{N}_3\left(\begin{pmatrix}0\\0\end{pmatrix}, \sigma_M\right) \text{ with}$$

$$\sigma_C = \begin{pmatrix}1 & 0.7\\ 0.7 & 1.5\end{pmatrix}, \sigma_{tr} = \begin{pmatrix}1 & -0.5\\ -0.5 & 1.5\end{pmatrix}, \text{ and } \sigma_M = \begin{pmatrix}1 & 0.3 & 0\\ 0.3 & 1.5 & -0.5\\ 0 & -0.5 & 1\end{pmatrix}.$$

We fixed the cost functions to scale with the norm and to be constant, i.e.,

$$c_1(x_{Ct}) = 0.1 \cdot \|x_{Ct}\|_2, \quad c_2(x_{CtM}) = 0.5.$$

Experiment to Figure 2: We created 50 replications of datasets with sample sizes

$$N = (100, 250, 500, 750, 1000, 2500, 5000, 7500)$$

under $\pi_1 \equiv 1, \pi_2 \equiv 1$. For each dataset, the estimate $\hat{\xi}_{\text{naive}}$, average sample cost c_{naive} and asymptotic variance v_{naive} under $\pi_1 \equiv 1, \pi_2 \equiv 1$ were determined. We computed the optimal propensities π^* at a budget of $b_0 = c_{\text{naive}}/1.5$ and determined the estimate $\hat{\xi}_{\text{opt}}$ and optimized asymptotic variance v_{opt} on dataset created by the optimal propensities π^* of sizes $1.5 \cdot N$ to counteract the smaller budget. In Figure 2, the left plot portrays the empirical mean-squared errors $MSE(\hat{\xi}_{\text{naive}}), MSE(\hat{\xi}_{\text{opt}})$ and their theoretical counterparts $MSE_{\text{naive}} = v_{\text{naive}}, MSE_{\text{opt}} = v_{\text{opt}}$ at different budget levels $n \cdot b_0$. The right plot visualizes the distribution of the optimized propensity π^* on $1,000$ samples.
Experiment to Figure 3: As in the experiment to Figure 2, we generated 50 replications of datasets with sample sizes N under $\pi_1 \equiv 1, \pi_2 \equiv 1$. For each dataset, we optimized for λ^*, computed the asymptotic variance v_{opt}, its cost c_{naive} and stored the computational time τ. The four plots in Figure 3 show the means at the various sample sizes along with the uncertainty measure of $\pm$ one standard deviation. Note that all computations were run on an Apple M2 chip with 16GB of RAM.
Experiment to Figure 4: In this experiment, we varied one parameter or nuisance of our data generation process at a time and studied its effects. For reference, we listed the exact alterations in Table 2. For each alteration, we generated 50 datasets of size 500 and computed the asymptotic variance v_{naive} and average sample cost c_{naive}. The budget for the optimization problem was set to be $b_0 = c_{\text{naive}}/1.5$ giving the asymptotic optimized variance v_{opt}. Here, we did not generate a second larger dataset, but instead scaled the ratio of the asymptotic variances accordingly to get the relative efficiency $v_{\text{opt}}/(1.5 \cdot v_{\text{naive}})$. A relative efficiency of, say, 80% indicates that under the same budget, the asymptotic optimized partial-measurements variance is 20% smaller than the always-measuring asymptotic variance. The 12 plots showcase the means with the uncertainty of $\pm$ one standard deviation against either the scaling factor α or the resulting (average) correlation(s). Note that the scaling ranges were carefully chosen so as to generate positive definite matrices.
Computational Note: To approximate the conditional expectations in Theorem 11 at an evaluation point $x_{C,t}$, we simulated realizations of $X_{C,t,M}$ using a subsample of ε_M, applied the functional, and then computed the weighted average of the subsample's propensity.

Appendix D. Model Misspecification. Non-Linear Data

This appendix extends the analysis underlying Figure 4 by examining efficiency gains under deliberate model misspecification. The data-generating mechanism is identical to the linear setting

Scaled Parameter / Nuisance	Multiplicative scaling range α
$\beta_{tC} = \alpha \cdot (0.5, -0.2)$	$(\frac{1}{100}, \frac{5}{100}, \frac{10}{100}, \frac{15}{100}, \frac{1}{5}, \frac{1}{4}, \frac{1}{3}, \frac{1}{2}, 1, 2, 3, 4, 5, 10, 20)$
$\beta_{MC} = \alpha \cdot \begin{pmatrix} 0.3 & 0.1 \\ 0.5 & 0.2 \\ -0.1 & 0.3 \end{pmatrix}$	$(\frac{1}{100}, \frac{5}{100}, \frac{10}{100}, \frac{15}{100}, \frac{1}{5}, \frac{1}{4}, \frac{1}{3}, \frac{1}{2}, 1, 2, 3, 4, 5, 10, 20)$
$\beta_{rC} = \alpha \cdot (0.2, -0.1)$	$(\frac{1}{100}, \frac{5}{100}, \frac{10}{100}, \frac{15}{100}, \frac{1}{5}, \frac{1}{4}, \frac{1}{3}, \frac{1}{2}, 1, 2, 3, 4, 5, 10, 20)$
$\beta_{Mt} = \alpha \cdot \begin{pmatrix} 0.7 \\ 0.2 \\ 0.1 \end{pmatrix}$	$(\frac{1}{100}, \frac{5}{100}, \frac{10}{100}, \frac{15}{100}, \frac{1}{5}, \frac{1}{4}, \frac{1}{3}, \frac{1}{2}, 1, 2, 3, 4, 5, 10, 20)$
$\beta_{rM} = \alpha \cdot (0.5, 0.4, -0.3)$	$(\frac{1}{100}, \frac{5}{100}, \frac{10}{100}, \frac{15}{100}, \frac{1}{5}, \frac{1}{4}, \frac{1}{3}, \frac{1}{2}, 1, 2, 3, 4, 5, 10, 20)$
$\sigma_C = \begin{pmatrix} 1\alpha & 0.7 \\ 0.7 & 1.5\alpha \end{pmatrix}$	$(0.6, 0.75, 1, 1.3, 1.75, 2.25, 3, 4.5, 7, 10, 15)$
$\sigma_C = \begin{pmatrix} 1 & 0.7\alpha \\ 0.7\alpha & 1.5 \end{pmatrix}$	`seq(0, sqrt(3), length.out = 11)`
$\sigma_M = \begin{pmatrix} 1\alpha & 0.3 & 0 \\ 0.3 & 1.5\alpha & -0.5 \\ 0 & -0.5 & 1\alpha \end{pmatrix}$	$(0.5, 0.6, 0.8, 1.1, 1.6, 2.25, 3, 4.5, 7, 10, 15)$
$\sigma_M = \begin{pmatrix} 1 & 0.3\alpha & 0 \\ 0.3\alpha & 1.5 & -0.5\alpha \\ 0 & -0.5\alpha & 1 \end{pmatrix}$	`seq(0, 2, by = 0.2)`
$\sigma_{tr} = \begin{pmatrix} 1\alpha & -0.5 \\ -0.5 & 1.5 \end{pmatrix}$	$(0.2, 0.4, 0.7, 1, 1.5, 2.5, 5, 8, 12.5, 20)$
$\sigma_{tr} = \begin{pmatrix} 1 & -0.5 \\ -0.5 & 1.5\alpha \end{pmatrix}$	$(0.3, 0.5, 0.7, 1, 1.5, 2.5, 5, 8, 12.5, 20)$
$\sigma_{tr} = \begin{pmatrix} 1 & -0.5\alpha \\ -0.5\alpha & 1.5 \end{pmatrix}$	`-seq(0, 1.2, length.out = 11)`

Table 2: Alterations in the experiments to Figure 4.

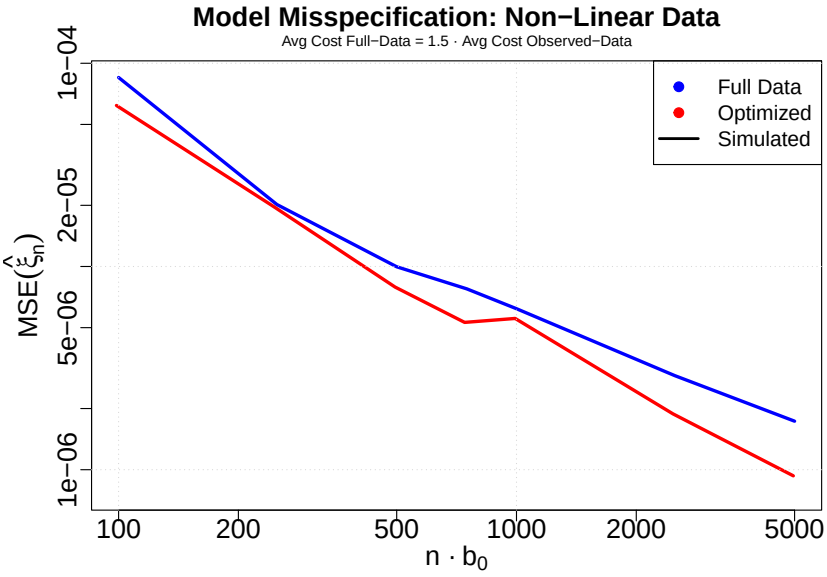

Figure 7: Mean squared error of the estimated causal effect under quadratic data-generating mechanisms, evaluated and averaged at ten treatment values in $[-0.1, 0.1]$ over 50 replications.

(details in Appendix C) except for the structural equation governing the mediator X_M, which is replaced by the quadratic expression

$$\begin{pmatrix} X_{M_1} \\ X_{M_2} \\ X_{M_3} \end{pmatrix} = \begin{pmatrix} 0.7 & -0.1 \\ 0.2 & -0.2 \\ 0.1 & -0.4 \end{pmatrix} \begin{pmatrix} X_t \\ X_t^2 \end{pmatrix} + \begin{pmatrix} \varepsilon_{M_1} \\ \varepsilon_{M_2} \\ \varepsilon_{M_3} \end{pmatrix}.$$

Under this specification, the causal effect of X_t on X_r becomes a quadratic function of the treatment. To accommodate this, we estimate the causal effect using quadratic regression models while keeping the design optimization step unchanged. That is, the optimized partial-data design is still constructed under the (incorrect) assumption of linearity. This experiment therefore analyzes the robustness of the optimized design when the estimation problem becomes non-linear.

We evaluate the causal effect at ten equally spaced treatment values in the interval $[-0.1, 0.1]$. Figure 7 reports the mean squared error (MSE), averaged over 50 replications and all evaluation points, across a range of budget levels. The display illustrates that the optimized partial-data design continues to outperform the full-data design, yielding smaller MSEs despite the slight misspecification. As the non-linear component of the data-generating process however becomes more pronounced, we expect the misalignment between the design stage and the estimation stage to increase, potentially reducing the efficiency gains.