Orthogonal Statistical Learning and Inference for Function-Valued Parameters with Heterogeneous Data Sources

Jaewon Lim

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2026

Reading Committee:
Alex Luedtke, Chair
Marco Carone
Ting Ye

Program Authorized to Offer Degree:
Biostatistics

University of Washington

Abstract

Orthogonal Statistical Learning and Inference for
Function-Valued Parameters with Heterogeneous Data Sources

Jaewon Lim

Chair of the Supervisory Committee:
Alex Luedtke
Department of Statistics

This dissertation develops methods for estimation of and inference on function-valued parameters when the available data are heterogeneous, drawn from partially aligned sources or missing at random by study design. Neyman-orthogonal losses and one-step estimation allow nuisance functions to be estimated with flexible machine learning methods without compromising the first-order behavior of the resulting estimators and statistical tests.

Chapter 1 studies estimation of causal dose-response functions by combining partially aligned data sources. I propose a data fusion framework that estimates the population risk of the dose-response functions using a Neyman-orthogonal loss. Then I show that data fusion necessarily improves worst-case performance in a minimax sense. Chapter 2 develops nonparametric tests of whether a function-valued parameter belongs to a closed linear subspace, such as the space of linear functions or of additive functions. Each null hypothesis is membership in a subspace, and I construct asymptotically linear estimators of the kernel embedding of the projected parameter. Chapter 3 provides a general orthogonal statistical learning framework for function-valued parameters under two-phase sampling. The argument of the function is observed only on a gold-standard subsample. I propose a corrected loss that preserves Neyman orthogonality, and the price of the incomplete design is shown to be second order.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I would like to express my gratitude to all the people I met during my PhD journey. It took a full five years, but it was fun and every moment was meaningful.

I am especially grateful to my advisor, Alex Luedtke, for guiding me in becoming an independent researcher. Alex always took my naive ideas and intuitions seriously and worked with me to turn them into concrete research questions. He never pushed me in a particular direction, but instead guided me as I learned to find my own way. From him, I also learned the importance of being thorough and clear in research, lessons that I hope to carry with me throughout my career.

I am also grateful to my committee members, Marco Carone and Ting Ye, whose thoughtful feedback helped me learn how to develop and strengthen my research. I was fortunate to have opportunities to work as a research assistant and apply statistics to real-world problems, and I thank my mentors Tom Fleming, Robyn McClelland, and Ali Shojaie for their guidance and support. I also thank all my friends and colleagues in the department for creating such a supportive environment throughout these years.

Most importantly, I thank my parents for their unconditional love and support throughout my life.

DEDICATION

to my parents

Chapter 1

ESTIMATION OF CAUSAL DOSE-RESPONSE FUNCTIONS UNDER DATA FUSION

1.1 Introduction

The causal dose-response function (CDRF) maps a level of continuous or multi-valued exposure to a corresponding expected counterfactual outcome. The exposure can be a potentially beneficial treatment, such as the duration of training (Kluve et al., 2012) or the frequency of therapy sessions (Robinson et al., 2020), or a harmful contaminant, such as the concentration of air pollution (Dominici et al., 2002). CDRFs are also referred to as causal dose-response curves (Díaz & van der Laan, 2013), dose-response functions (Bonvini & Kennedy, 2022), average dose-response functions (Galvao & Wang, 2015), or average structural functions (Blundell & Powell, 2004). Estimating a CDRF is challenging because, typically, few, if any, individuals in a dataset have received most dose levels (Bahadori et al., 2022).

Extensive work has been devoted to estimating CDRFs. Many existing approaches rely on parametric assumptions. An early formal approach, introduced by Robins et al. (2000), was based on a marginal structural model, which fits a parametric model using inverse-probability weighting. This approach requires the parametric model to contain the true CDRF. Later, the assumption of correct model specification was relaxed by projecting CDRFs onto the function space implied by a particular parametric model for more robust estimation (Neugebauer & van der Laan, 2007). Ai et al. (2021) introduced a weighted optimization scheme for estimating CDRFs, employing weights suggested by Robins et al. (2000). Yiu and Su (2018) constructed balancing weights that render pretreatment variables unassociated with treatment assignment after weighting. Hirano and Imbens (2004) and Imai and van Dyk (2004) used parametric generalized propensity scores to estimate CDRFs, and Fong et al.

(2018) proposed a covariate-balancing generalized propensity score methodology.

More recently, a growing body of literature has focused on estimating CDRFs using nonparametric methods and orthogonal statistical learning (Bonvini & Kennedy, 2022; Foster & Syrgkanis, 2023). Flores (2007) considered a plug-in, kernel-based estimator of CDRFs. Singh et al. (2023) explored kernel ridge regression to develop simple, closed-form estimators of CDRFs. Galvao and Wang (2015) proposed a semiparametric, two-step estimator that first estimates a ratio of conditional densities and then estimates the CDRF. Bonvini and Kennedy (2022) and Kennedy et al. (2016) introduced a two-step estimation procedure in which nuisance functions are estimated and then a pseudo-outcome is regressed on the exposure variable. Westling et al. (2020) proposed a nonparametric estimator for monotone CDRFs. Given multiple candidate CDRF estimators, early work demonstrated how to use cross-validation to select among them, with corresponding oracle guarantees (Díaz & van der Laan, 2013; van der Laan & Dudoit, 2003).

Recent approaches include the estimation of CDRFs in nonstandard settings. Huang and Zhang (2023) studied settings in which the continuous treatment variable is measured with error. They proposed a method based on weighting and generalized empirical likelihood techniques. Zeng et al. (2024) examined scenarios in which the primary outcome is missing for some proportion of the data. They developed a two-step estimation procedure similar to that of Bonvini and Kennedy (2022), but their method constructs pseudo-outcomes by leveraging both labeled and unlabeled data through a surrogate outcome.

Previous approaches have considered only samples originating from a single set of trial or observational data. This limitation makes the estimation of CDRFs practically difficult because data from a single source may not be sufficient for precise estimation across all exposure levels. To address this issue, we consider data fusion, which enables the combination of information from distinct sources that are partially aligned with the target distribution (Bareinboim & Pearl, 2016). Data fusion has been applied to the transportability of causal effects (Bareinboim & Pearl, 2014; Dahabreh & Hernán, 2019; Dahabreh et al., 2022; Hernán & VanderWeele, 2011; Pearl & Bareinboim, 2011; Rudolph & van der Laan, 2016; Stuart

et al., 2015). Qiu et al. (2024) combined information from an auxiliary population to address dataset shift and to estimate the target population risk. Sun and Miao (2022) combined information from two data sources—one containing treatment information and the other outcome information—to estimate average treatment effects via an instrumental variable. Graham et al. (2026), Li and Luedtke (2023), and Li et al. (2025) developed general frameworks for the efficient estimation of smooth, finite-dimensional parameters under data fusion.

We apply a general form of data fusion, leveraging one or more sources. Our first contribution is as follows:

1. We derive a Neyman-orthogonal loss for estimating the CDRF in data fusion settings.

Evaluating this loss can be computationally expensive. To overcome this challenge:

2. We propose a stochastic approximation of the loss that retains Neyman orthogonality. When used with kernel ridge regression, it yields a closed-form estimator.

3. We establish finite-sample regret upper bounds for both kernel ridge regression and empirical risk minimizers. These bounds become tighter as additional sources are incorporated into the analysis.

Since upper bounds can be loose, the above results do not guarantee that incorporating additional data sources will necessarily improve performance. Our final contribution is to investigate this question:

4. We characterize settings in which using data fusion necessarily reduces the risk of estimating CDRFs with high probability. In doing so, we establish a high-probability worst-case lower bound for the risk of estimating CDRFs without data fusion.

1.2 Data fusion framework and estimation algorithm

1.2.1 Statistical setting and parameters of interest

We begin by defining the parameter of interest and its associated risk. Let $(X, A, Y) \sim Q^0$, where $X \in \mathcal{X} \subset \mathbb{R}^r$ denotes the covariates, $A \in \mathcal{A} = [0,1]^d$ the exposure, $Y \in \mathcal{Y} \subset \mathbb{R}$ the outcome of an experiment, and Q^0 the target distribution. Our parameter of interest, the CDRF, maps a realized exposure a to the expected value of the potential outcome Y^a. Under the assumptions discussed in Westling et al. (2020), the CDRF can be identified as

$$\theta_{Q^0}(a) = \mathrm{E}[Y^a] = \int \mathrm{E}[Y \mid X = x, A = a]\; Q^0_X(dx).$$

The target distribution is assumed to belong to a model $\mathcal{Q}$ of distributions for the random vector (X, A, Y)—some of our later results will impose moment and smoothness restrictions on this model, but for now we leave it unspecified. For any $Q \in \mathcal{Q}$, we denote by Q_X the marginal distribution of X, by $Q_{A|X}$ the conditional distribution of A given X, and by $Q_{Y|X,A}$ the conditional distribution of Y given (X, A).

We evaluate performance by integrating a squared error loss with respect to μ, a pre-specified measure on $\mathcal{A}$ that dominates $Q^0_{A|X}$ with Q^0_X-probability one. Hence, the risk at $\theta : \mathcal{A} \to \mathbb{R}$ is

$$\mathcal{L}_{Q^0}(\theta) := \int \big[\theta(a) - \theta_{Q^0}(a)\big]^2 \,\mu(da) = \int \mathrm{E}_{Q^0_X}\Big[\mathrm{E}_{Q^0_{Y|X,A}}\big[\theta(A) - Y \,\big|\, X, A = a\big]\Big]^2 \mu(da).$$

Our data do not consist of draws directly from the target distribution Q^0. Instead, they contain samples from k data sources, formalized as observing n independent copies of $(X, A, Y, S) \sim P^0$. Here, S is a categorical random variable with support $[k]$ indicating the data source. For any distribution P, we define $P_X(\cdot \mid s)$ as the conditional distribution of $X \mid S = s$, $P_{A|X}(\cdot \mid x, s)$ as that of $A \mid X = x, S = s$, and $P_{Y|X,A}(\cdot \mid x, a, s)$ as that of $Y \mid X = x, A = a, S = s$. We assume that all such conditional distributions are well defined on common measurable spaces.

Under the following condition, we establish a connection between the conditional distributions of P^0 and Q^0. Let $\mathcal{S}_X$ and $\mathcal{S}_Y$ denote known collections of data sources whose

distributions of X and $Y \mid X, A$ align with Q^0_X and $Q^0_{Y|X,A}$, respectively, in the sense defined below.

Condition 1.1 (Data Fusion Conditions). All of the following hold:

(a) (Positive correctly aligned sources) There exist constants $M_\xi, M_\eta \geq 1$ such that

$$P^0(S \in \mathcal{S}_X) \;\geq\; M_\xi^{-1}, \quad P^0(S \in \mathcal{S}_Y) \;\geq\; M_\eta^{-1}.$$

(b) (Sufficient overlap) $Q^0_{X,A}(\cdot)$ is absolutely continuous with respect to $P^0_{X,A}(\cdot \mid S = s)$ for all $s \in \mathcal{S}_Y$. Moreover, there exists $c_1 \geq 1$ such that

$$c_1^{-1} \;\leq\; \frac{dQ^0_{X,A}}{dP^0_{X,A}(\cdot \mid S \in \mathcal{S}_Y)}(x, a) \;\leq\; c_1, \quad Q^0\text{-a.e. } (x, a).$$

(c) (Exchangeability of conditionals) For all $s \in \mathcal{S}_X$, $P^0_X(\cdot \mid S = s) = Q^0_X(\cdot)$. For all $s \in \mathcal{S}_Y$,

$$P^0_{Y|X,A}(\cdot \mid X = x, A = a, S = s) \;=\; Q^0_{Y|X,A}(\cdot \mid X = x, A = a), \quad Q^0_{X,A}\text{-a.e.}$$

The statistical model $\mathcal{P}$ is defined as the set of all distributions P^0 on some measurable space such that (P^0, Q^0) satisfy Condition 1.1 for some $Q^0 \in \mathcal{Q}$. Condition 1.1 can be relaxed in the sense that it suffices for there to exist at least one $\underline{Q} \in \mathcal{Q}$ satisfying the condition, even if the true Q^0 fails to meet Condition 1.1(b), as discussed in Li and Luedtke (2023). This relaxation is justified because the risk does not depend on the distribution of $A \mid X$, which is not identifiable. Condition 1.1(a) guarantees that the probabilities of correctly and partially aligned sources are uniformly bounded away from zero across $\mathcal{P}$. Conditions 1.1(b) and 1.1(c) are used to identify the risk through the observed distribution P^0. Specifically, Condition 1.1(b) ensures that the conditional distribution $P^0_{Y|X,A}(\cdot \mid x, a, S = s)$ is uniquely defined up to Q^0-null sets, while Condition 1.1(c) enables us to link the risk defined by the conditional distributions of Q^0 to those of P^0. Moreover, we require Condition 1.1(b) to define a Neyman–orthogonal loss for the risk in the next section.

By Theorem 1 in Li and Luedtke (2023), Conditions 1.1(b) and 1.1(c) allow us to identify θ_{Q^0} and $\mathcal{L}_{Q^0}(\theta)$ with the following functions indexed by P^0:

$$\mathcal{L}_{Q^0}(\theta) = L_{P^0}(\theta) := \int \mathrm{E}_{P^0}\Big[\mathrm{E}_{P^0}\big[\theta(A) - Y \,\big|\, X, A = a, S \in \mathcal{S}_Y\big]\Big| \, S \in \mathcal{S}_X\Big]^2 \mu(da).$$

In view of this identification result, $\mathcal{L}_{Q^0}(\theta)$ can be estimated using data drawn from P^0. The same applies to the global risk minimizer θ_{Q^0}, which is identified with

$$\theta_{Q^0}(a) = \theta_0(a) := \mathrm{E}_{P^0}\Big[\mathrm{E}_{P^0}\big[Y \,\big|\, X = x, A = a, S \in \mathcal{S}_Y\big]\Big| \, S \in \mathcal{S}_X\Big].$$

1.2.2 Derivation of Neyman-orthogonal loss

Our goal is to find a minimizer of the population risk, denoted by $L_{P^0}(\theta)$, within a function class $\Theta \subseteq L^2(\mu)$. Because the population risk involves nuisance parameters, we aim to construct a Neyman–orthogonal loss function such that the estimation error of the nuisance parameters has a fourth-order impact on the oracle population risk (Curth et al., 2021; Foster & Syrgkanis, 2023). This property is particularly beneficial because estimators of some nuisance parameters on which the problem relies typically do not converge at the $n^{-1/2}$ rate when estimated using machine-learning methods (Sugiyama et al., 2012).

A candidate for a Neyman-orthogonal loss is identified as the loss function whose empirical mean corresponds to a one-step estimator of $L_{P^0}(\theta)$, $L_{OS,\widehat{P}_n}(\theta) := L_{\widehat{P}_n}(\theta) + \mathbb{P}_n D^*_{\widehat{P}_n}(\theta)$, based on an initial estimate $\widehat{P}_n$ of P^0, where $D^*_P(\theta)$ denotes a gradient of $P \mapsto L_P(\theta)$ at P (Bickel, 1982). This estimator adjusts for the bias of the plug-in estimator $L_{\widehat{P}_n}(\theta)$ (Pfanzagl, 1982). By Corollary 1 in Li and Luedtke (2023), $D^*_P(\theta)$ can be obtained as the projection of $D^*_Q(\theta)$ onto the tangent space of $\mathcal{P}$ at P.

The resulting one-step estimator of $L_{P^0}(\theta)$ is given by the empirical mean

$$\mathbb{P}_n\left[\ell^{\mathrm{OS}}_{P^0}(\theta, g_0; \cdot)\right],$$

where $\ell^{\mathrm{OS}}_{P^0}(\theta, g_0; (x, a, y, s)) := \mathrm{E}_{B\sim\mu}[\ell_{P^0}(\theta, g_0; (x, a, y, s, B))]$ and

$$\ell_{P^0}(\theta, g_0; (x, a, y, s, b)) := \big(\theta(b) - \theta_0(b)\big)^2 + 2\xi_0 \, \mathbf{1}(s \in \mathcal{S}_X)\big(\theta(b) - \theta_0(b)\big)\big(\theta_0(b) - m_0(x, b)\big)$$

$$+ 2\eta_0\, w_0(x,a)\, \mathbf{1}(s \in \mathcal{S}_Y)\big(\theta(a) - \theta_0(a)\big)\big(m_0(x,a) - y\big), \tag{1.1}$$

with nuisance parameters defined as

$$g_0 : (a,x) \mapsto (\xi_0, \eta_0, w_0(a,x), m_0(a,x), \theta_0(a)),$$
$$\xi_0 := \frac{1}{P^0(S \in \mathcal{S}_X)}, \quad \eta_0 := \frac{1}{P^0(S \in \mathcal{S}_Y)}, \quad w_0(x,a) := \frac{p^0(x \mid S \in \mathcal{S}_X)}{p^0(x \mid S \in \mathcal{S}_Y)} \cdot \frac{1}{p^0(a \mid x, S \in \mathcal{S}_Y)}$$
$$m_0(x,a) := \mathrm{E}_{P^0}[Y \mid x, a, S \in \mathcal{S}_Y].$$

We define $p^0(x \mid S \in \mathcal{S}_X)$ and $p^0(x \mid S \in \mathcal{S}_Y)$ as densities with respect to a common dominating measure on $\mathcal{X}$. Similarly, $p(a \mid x, S \in \mathcal{S}_Y)$ denotes the conditional density of A given X and $S \in \mathcal{S}_Y$ with respect to the known measure μ on $\mathcal{A}$.

We employ a stochastic approximation of the one-step estimator because it involves an expectation with respect to the known measure μ, which makes direct optimization computationally challenging in practice. To approximate this expectation, we draw n independent samples $B_1, \ldots, B_n \sim \mu$ and denote a realization of the full data vector by $z := (x, a, y, s, b)$. For notational convenience, we denote the (oracle) loss function by $\ell_{P^0}(\theta, g_0; z)$, as defined in (1.1), for the remainder of the chapter.

Distributions in $\mathcal{Q}$ are required to satisfy the following moment conditions:

Condition 1.2 (Constraints on Model for Target Distribution). For each $Q \in \mathcal{Q}$:

(a) (Bounded conditional mean) $|m_0(x,a)| \leq 1$ $Q_X \times \mu$-almost everywhere (a.e.). In particular, this implies that $\big|\theta_0(a)\big| = \big|\mathrm{E}_{Q_X}\big[m_0(x,a)\big]\big| \leq 1$ μ-a.e.

(b) (Sub-exponential tails) There exist constants $\sigma, L > 0$ such that, for all integers $m \geq 2$,

$$\int |y - \theta_Q(a)|^m \, Q(dy \mid x, a) \leq \tfrac{1}{2} m!\, \sigma^2 L^{m-2}.$$

(c) (Bounded density ratio between μ and $Q_{A|X}$) There exists a constant $c_\mu \geq 1$ such that for Q_X-almost every x,

$$\mu \ll Q_{A|X}(\cdot \mid x) \quad \text{and} \quad c_\mu^{-1} \leq \frac{d\mu}{dQ_{A|X}(\cdot \mid x)}(a) \leq c_\mu, \quad Q_{A|X}(\cdot \mid x)\text{-a.e. } a.$$

The bounded conditional mean assumption in Condition 1.2(a) can be enforced, without loss of generality, by rescaling the outcome Y whenever the conditional mean is uniformly bounded by some finite constant. The sub-exponential tail condition in Condition 1.2(b) allows the outcome to be unbounded while ensuring that its tail probabilities are appropriately controlled. More precisely, for each (X, A) it implies that

$$P_Q\big(|Y - \theta_Q(A)| \geq t \,\big|\, X = x, A = a\big) \leq 2\exp(-t/L), \quad t \gg 0,$$

by Proposition 2.7.1 in Vershynin (2018). Condition 1.2(c) states that the Radon–Nikodym derivative of μ with respect to $Q_{A|X}(\cdot \mid x)$ is uniformly bounded away from zero and infinity over $\mathcal{Q}$. Combining Conditions 1.1(b) and 1.2(c), we also obtain that, for any $P^0 \in \mathcal{P}$, $M_w^{-1} \leq w_0(x, a) \leq M_w$ for $M_w := c_1 c_\mu \geq 1$. Hence, w_0 is also uniformly bounded away from zero and infinity.

We define the nuisance parameter space $\mathcal{G}$ (with $g_0 \in \mathcal{G}$) associated with $\mathcal{P}$ as the set of tuples $g = (\xi, \eta, w, m, \tau)$, where $\xi, \eta \in (0, 1]$, $w, m : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$, and $\tau : \mathcal{A} \to \mathbb{R}$, satisfying the following condition.

Condition 1.3 (Uniform Boundedness of Nuisances). For each $g \in \mathcal{G}$, the constants M_ξ, M_η, M_w serve as universal bounds for any $P^0 \in \mathcal{P}$ such that:

(a) (Positive probability for relevant sources) $\xi \leq M_\xi$ and $\eta \leq M_\eta$.

(b) (Bounded density ratio and conditional mean)

$$M_w^{-1} \leq w(x, a) \leq M_w, \qquad |m(x, a)| \leq 1 \quad Q_X^0 \times \mu\text{-a.e.}, \qquad |\tau(a)| \leq 1 \quad \mu\text{-a.e.}.$$

Condition 1.3(a) requires that the source probability estimators are uniformly bounded away from zero over $\mathcal{P}$. Condition 1.3(b) requires that the density ratio estimators of w_0 are uniformly bounded away from zero and infinity, and that the estimators of the conditional mean m_0 and the CDRF θ_0 are uniformly bounded by one. We also define a constant $M_\lambda := M_w^{-1}(M_\eta + M_w + 2)$. The loss function defined in (1.1) can be extended to any $g \in \mathcal{G}$

by replacing g_0 with g. Building on this extension, we define the expected loss as a function of both the parameter of interest and the nuisance parameters:

$$L_{P^0}(\theta, g) := \mathrm{E}_{\widetilde{P}^0}\big[\ell_{P^0}(\theta, g; Z)\big],$$

where $\widetilde{P}^0 := P^0 \times \mu$ denotes the product measure on the extended space $Z := (X, A, Y, S, B)$. In particular, when evaluated at the true nuisance parameters, the loss reduces to the target functional:

$$L_{P^0}(\theta, g_0) = L_{P^0}(\theta).$$

1.2.3 Estimation algorithms

Using the loss function defined above, we propose two estimation algorithms—a generic two-stage estimation strategy and one based on kernel ridge regression—both of which employ sample splitting (Chernozhukov et al., 2018; Foster & Syrgkanis, 2023). We divide the sample into two non-overlapping subsets, estimate nuisance parameters using the first subset, and then estimate the target parameter based on the estimated nuisance parameters using the remaining subset.

Algorithm 1.1 implements empirical risk minimization over a function class $\Theta \subseteq L^2(\mu)$ subject to the constraint $\sup_{\theta\in\Theta} \|\theta\|_{L^\infty(\mu)} \leq 1$. The sampling weights ξ_0 and η_0 are estimated as the inverses of the empirical probabilities of $S \in \mathcal{S}_X$ and $S \in \mathcal{S}_Y$, respectively. The density ratio $w_0(x, a)$ is estimated using direct density-ratio methods such as Unconstrained Least-Squares Importance Fitting (uLSIF) or the Kullback–Leibler Importance Estimation Procedure (KLIEP) (Sugiyama et al., 2012). The outcome regression $m_0(x, a)$ is fit using a flexible learning algorithm, for example, Super Learner (via the `SuperLearner` package).

Algorithm 1.2 uses kernel ridge regression. Let $\mathcal{H}$ denote a reproducing kernel Hilbert space (RKHS) with norm $\|\cdot\|_{\mathcal{H}}$ and kernel $\mathcal{K}$. We consider the constrained RKHS ball

$$\mathcal{H}_c = \big\{\theta \in \mathcal{H} \,\big|\, \|\theta\|_{\mathcal{H}} \leq c,\ \|\theta\|_{L^\infty(\mu)} \leq 1\big\},$$

where c is a hyperparameter selected by cross-validation. Given z, the kernel ridge regression problem becomes an empirical loss minimization over the constrained class $\mathcal{H}_c$.

Algorithm 1.1 Estimation of CDRF over a general function class Θ

1: **Input:** Partition the sample z^n into two disjoint subsamples z_1^n and z_2^n.

2: **Nuisance estimation:** Estimate nuisance parameters $\widehat{g}$ using only z_1^n:

$$\widehat{\xi} := \frac{1}{\widehat{P}_n(S \in \mathcal{S}_X)}, \qquad \widehat{\eta} := \frac{1}{\widehat{P}_n(S \in \mathcal{S}_Y)}, \qquad \widehat{w}(x,a) := \text{direct estimator of } w_0(x,a),$$

$$\widehat{m}(x,a) := \mathrm{E}_{\widehat{P}_n}[Y \mid X = x, A = a, S \in \mathcal{S}_Y], \qquad \widehat{\tau}_0(a) := \mathrm{E}_{\widehat{P}_n}[\widehat{m}(X,a) \mid S \in \mathcal{S}_X].$$

3: **Target estimation:** Find the empirical-risk minimizer $\widehat{\theta}$ using only z_2^n:

$$\widehat{\theta} = \arg\min_{\theta \in \Theta} \mathbb{P}_n^{(2)}\big[\ell_{P^0}(\theta, \widehat{g}; \cdot)\big],$$

where $\mathbb{P}_n^{(2)}$ denotes the empirical distribution based on z_2^n.

1.3 Excess risk guarantees and worst-case efficiency gains from using data fusion

1.3.1 Overview

This section summarizes and formalizes our theoretical findings regarding the impact of data fusion on estimator performance. We first derive oracle excess risk bounds for empirical risk minimizers, both in general settings and in kernel-ridge regression. These results demonstrate that, when nuisance functions are estimated with sufficient accuracy, the excess risk is dominated by the target estimation error. We then show that incorporating data fusion reduces the Lipschitz constant of the loss function, which in turn tightens the upper bound on the excess risk. Lastly, we establish that the high-probability worst-case upper bound on the risk under data fusion can fall strictly below the high-probability worst-case lower bound attainable without data fusion. Together, these results provide both predictive and worst-case efficiency guarantees for leveraging partially aligned data sources. Figure 1.1 summarizes these relationships between excess risk bounds and worst-case guarantees under data fusion and no fusion.

Algorithm 1.2 Estimation of the CDRF using kernel ridge regression

1: **Input:** Two disjoint subsamples z_1^n and z_2^n obtained by splitting the full sample z^n.

2: **Nuisance estimation:** Estimate nuisance parameters using only z_1^n as described in Algorithm 1.1, yielding $\widehat{g}$.

3: **Target estimation:**

1. For each $i = 1, \ldots, |z_2^n|$, define

$$u_i = \mathbf{1}(s_i \in \mathcal{S}_Y)\,\widehat{\eta}\,\widehat{w}(x_i, a_i)\big(\widehat{m}(x_i, a_i) - y_i\big),$$
$$v_i = \mathbf{1}(s_i \in \mathcal{S}_X)\,\widehat{\xi}\,\big(\widehat{\tau}(b_i) - \widehat{m}(x_i, b_i)\big) - \widehat{\tau}(b_i).$$

2. Define the Gram matrix $\mathbf{K} = (K_{ij}) \in \mathbb{R}^{2|z_2^n| \times 2|z_2^n|}$, where $K_{ij} = \mathcal{K}(a_i', a_j')/|z_2^n|$ and $a' = (a_i) \cup (b_i)$ for $i = 1, \ldots, |z_2^n|$.

3. Use cross-validation (Algorithm A.1) to select λ_n and compute the kernel weights:

$$\widehat{\beta} = -\frac{\mathbf{u}}{\lambda_n \sqrt{|z_2^n|}}, \qquad \widehat{\gamma} = -(\mathbf{K}_{22} + \lambda_n \mathbf{I})^{-1}\left(\frac{\mathbf{v}}{\sqrt{|z_2^n|}} + \mathbf{K}_{21}\widehat{\beta}\right).$$

4. Construct the closed-form kernel-ridge estimator $\widehat{\theta}$ as

$$\widehat{\theta}(\cdot) = \frac{1}{\sqrt{|z_2^n|}} \sum_{j=1}^{|z_2^n|} \big[\widehat{\beta}_j \mathcal{K}(\cdot, a_j) + \widehat{\gamma}_j \mathcal{K}(\cdot, b_j)\big].$$

1.3.2 Excess risk bound

This subsection establishes oracle excess risk bounds for the estimators obtained from Algorithms 1.1 and 1.2. We show that, for a general function class $\Theta \subseteq L^2(\mu)$, the nuisance parameter estimation error affects the excess risk only at the fourth order. As a result, when the nuisance estimation error is sufficiently small, the oracle estimation error is dominated by the target estimation error.

Let $\theta^* \in \arg\min_{\theta \in \Theta} L_{P^0}(\theta, g_0)$ denote a minimizer of the oracle risk. Throughout, we

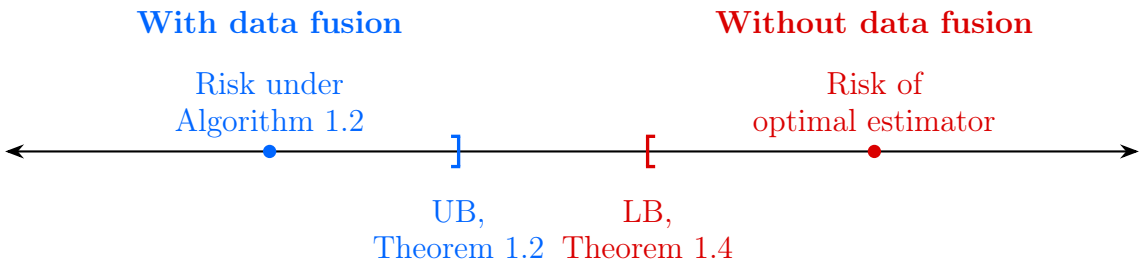


Figure 1.1: Number line demonstrating approach for establishing data fusion necessarily improves worst-case performance. The key idea is to show that a **high-probability worst-case upper bound (UB)** of the risk for kernel ridge regression with data fusion lies below a **high-probability worst-case lower bound (LB)** for any estimator without it.

equip Θ with the $L^2(\mu)$ norm and, depending on the context, $\mathcal{G}$ with either the $L^2(Q_X^0 \times \mu; \ell^2)$ or $L^4(Q_X^0 \times \mu; \ell^2)$ norm. For $g \in \mathcal{G}$ and $1 \leq p < \infty$, we define

$$\|g\|_{L^p(Q_X^0 \times \mu; \ell^2)} := \left(\mathrm{E}_{(X,A) \sim Q_X^0 \times \mu} \left[\|g(X, A)\|_2^p \right] \right)^{1/p}.$$

The nuisance estimation error of $\widehat{g}$ is then quantified as $\|\widehat{g} - g_0\|_{L^2(Q_X^0 \times \mu; \ell^2)}$ or $\|\widehat{g} - g_0\|_{L^4(Q_X^0 \times \mu; \ell^2)}$.

We next show that the oracle excess risk, $L_{P^0}(\widehat{\theta}, g_0) - L_{P^0}(\theta^*, g_0)$, is bounded by the sum of the excess risk evaluated at $\widehat{g}$, $L_{P^0}(\widehat{\theta}, \widehat{g}) - L_{P^0}(\theta^*, \widehat{g})$, and a fourth-order term quantifying the nuisance estimation error, where $\widehat{\theta}$ and $\widehat{g}$ are the estimators produced by Algorithm 1.1.

Lemma 1.1 (Oracle Excess Risk Bound in Terms of Target and Nuisance Errors)**.** *Assume Conditions 1.1 to 1.3 hold. If $\widehat{\theta}$ and $\widehat{g}$ are the target estimator and nuisance estimators from Algorithm 1.1, then*

$$L_{P^0}(\widehat{\theta}, g_0) - L_{P^0}(\theta^*, g_0) \leq 2 \left(L_{P^0}(\widehat{\theta}, \widehat{g}) - L_{P^0}(\theta^*, \widehat{g}) \right) + 2M_\lambda^2 \|\widehat{g} - g_0\|_{L^4(Q_X^0 \times \mu; \ell^2)}^4.$$

The proof is given in the appendix. It relies on the Neyman orthogonality of the loss, as well as second-order smoothness, strong convexity, and higher-order smoothness conditions, as stated in Lemma A.3. This lemma shows that the oracle excess risk is bounded by the sum of two components: the target estimation error and the fourth power of the nuisance

estimation error. Thus, when the nuisance estimation error is smaller than the one-fourth power of the target estimation error, the excess risk bound is dominated by the target estimation error alone.

We now analyze the oracle excess risk for the estimators produced by Algorithms 1.1 and 1.2. The difficulty of estimating the CDRF depends on the complexity of the function class Θ, which we quantify via the local Rademacher complexity. For a measure $\mathcal{D}$, define

$$\mathcal{R}_{n,\mathcal{D}}(\Theta, \delta) := \mathrm{E}_{\epsilon_{1:n}, a_{1:n} \sim \mathcal{D}} \left[\sup_{\theta \in \Theta : \|\theta\|_{L^2(\mu)} \leq \delta} \left| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i \theta(a_i) \right| \right],$$

where $\epsilon_1, \ldots, \epsilon_n$ are independent Rademacher random variables.

For any $\theta' \in \Theta$, define the star hull of Θ around θ' as

$$\mathrm{star}(\Theta, \theta') := \{ t\theta + (1-t)\theta' \mid \theta \in \Theta,\, t \in [0,1] \}.$$

We now present the oracle excess risk bound for the estimator obtained from Algorithm 1.1. To control deviations of the empirical loss from its expectation, we apply Talagrand's concentration inequality, leveraging the fact that the loss is Lipschitz in its first argument, with a Lipschitz constant that depends on the true nuisance parameters with high probability. We define

$$B_{P^0}(\delta) := 4(1+\delta)^2 \big(1 + \xi_0 + C_{\sigma,\delta}\, \eta_0 \|w_0\|_\infty\big),$$

where

$$C_{\sigma,\delta} := 1 + \max\Big\{ \sigma\sqrt{2\log(8/\delta)},\, 2L\log(8/\delta) \Big\},$$

and $\|\cdot\|_\infty$ denotes the essential supremum with respect to $Q_X^0 \times \mu$. Throughout, we write $A \lesssim B$ to mean that $A \leq CB$ for some universal constant $C > 0$.

Theorem 1.1 (Excess Risk Bound for Algorithm 1.1)**.** *Suppose Conditions 1.1 to 1.3 hold. Fix $\delta \in (0,1)$, and let $(\widehat{\theta}, \widehat{g})$ be as in Algorithm 1.1, with $\|\widehat{g} - g_0\|_{L^4(Q_X^0 \times \mu; \ell^2)} = o_p(1)$ and*

$\|\widehat{w} - w_0\|_{L^\infty(Q_X^0 \times \mu; \ell^2)} = o_p(1)$. *Suppose* $\delta_n^2 = \Omega(n^{-1} \log\log n)$, *and let* δ_n *be any solution to the inequality*

$$\max\left\{\mathcal{R}_{n,Q^0}\big(\mathrm{star}(\Theta - \theta^*, 0), r\big),\ \mathcal{R}_{n,\mu}\big(\mathrm{star}(\Theta - \theta^*, 0), r\big)\right\} \le r^2.$$

Then there exists $N(\delta) \in \mathbb{N}$ *such that, for all* $n \ge N(\delta)$, *the following holds with probability at least* $1 - \delta/2$: *for any realization* $z := (x, a, b, y, s)$ *of* $Z \sim \widetilde{P}^0$ *and any* $\theta_1, \theta_2 \in \Theta$,

$$\left|\ell_{P^0}(\theta_1, \widehat{g}; z) - \ell_{P^0}(\theta_2, \widehat{g}; z)\right| \le B_{P^0}(\delta) \left\|\big(\theta_1(b) - \theta_2(b),\ \theta_1(a) - \theta_2(a)\big)\right\|_2. \tag{1.2}$$

Furthermore, with probability at least $1 - \delta$, *the oracle excess risk satisfies*

$$L_{P^0}(\widehat{\theta}, g_0) - L_{P^0}(\theta^*, g_0) \ \lesssim\ B_{P^0}(\delta)^2 \left(\delta_n^2 + \frac{\log(1/\delta)}{n}\right) \ +\ 2M_\lambda^2 \big\|\widehat{g} - g_0\big\|^4_{L^4(Q_X^0 \times \mu; \ell^2)}.$$

Next, we consider the excess risk bound for kernel ridge regression. We follow the setup introduced in Nie and Wager (2021) and Zhang et al. (2023). Assume a separable RKHS $\mathcal{H}$ on the compact set $\mathcal{A}$ with a strictly positive-definite, continuous kernel $\mathcal{K}: \mathcal{A} \times \mathcal{A} \to \mathbb{R}$. Define the integral operator $T_\mathcal{K}: L^2(\mu) \to L^2(\mu)$ by

$$(T_\mathcal{K}\theta)(a) := \int_\mathcal{A} \mathcal{K}(a, t)\, \theta(t)\, d\mu(t).$$

This operator is self-adjoint, positive, and compact (Steinwart & Scovel, 2012). By the spectral theorem for self-adjoint compact operators and Mercer's decomposition (Cucker & Smale, 2001),

$$k(a, t) = \sum_{j=1}^\infty \sigma_j e_j(a) e_j(t), \qquad T_\mathcal{K}(\cdot) = \sum_{j=1}^\infty \sigma_j \langle \cdot, e_j \rangle_{L^2(\mu)} e_j,$$

where $\{\sigma_j\}_{j=1}^\infty$ is a non-increasing summable sequence of eigenvalues and $\{e_j\}_{j=1}^\infty$ is a corresponding $L^2(\mu)$-orthonormal system of eigenfunctions. For $s \ge 0$, we denote the fractional integral operator $T_\mathcal{K}^s : L^2(\mu) \to L^2(\mu)$ and its image, called an s-power space $\mathcal{H}^s$ of $\mathcal{H}$, as

$$T_\mathcal{K}^s(\cdot) := \sum_{j=1}^\infty \sigma_j^s \langle \cdot, e_j \rangle_{L^2(\mu)} e_j, \qquad \mathcal{H}^s := \left\{ \sum_{j=1}^\infty \sigma_j^{s/2} a_j e_j \ :\ \sum_{j=1}^\infty a_j^2 < \infty \right\}.$$

It holds that $\mathcal{H}^1 \subseteq \{[h] : h \in \mathcal{H}\}$, with $[h]$ a μ-equivalence class of functions. Moreover, for $0 < s_1 < s_2$, $\mathcal{H}^{s_2} \subseteq \mathcal{H}^{s_1} \subseteq \mathcal{H}^0 \subseteq L^2(\mu)$. Functions in $\mathcal{H}^s$ with larger s are naturally interpreted as having higher smoothness with respect to the RKHS $\mathcal{H}$.

Moreover, we impose the following assumptions on $\mathcal{H}$ and its associated kernel $\mathcal{K}$.

Condition 1.4 (RKHS Conditions). The following conditions hold:

(a) (Kernel function) The reproducing kernel $\mathcal{K} : \mathcal{A} \times \mathcal{A} \to \mathbb{R}$ is continuous and strictly positive definite, so that the Gram matrix is invertible for distinct points.

(b) (Eigenvalue decay at least polynomial) There exist constants $p \in (0, 1)$ and $G < \infty$ such that the eigenvalues $\{\sigma_j\}_{j=1}^{\infty}$ of the kernel integral operator satisfy $\sigma_j \leq G j^{-1/p}$ for all $j \geq 1$.

(c) (Bounded eigenfunctions) The eigenfunctions e_j corresponding to σ_j are uniformly bounded by a constant $M_e < \infty$ not depending on j.

Condition 1.4(a) ensures the strict positive-definiteness of the kernel, and hence the invertibility of the Gram matrix for distinct inputs, guaranteeing the closed-form estimator in Algorithm 1.2. By the spectral theorem (e.g., Theorem 2 of Cucker and Smale, 2001), the eigenvalues of a compact, self-adjoint, positive operator $T_{\mathcal{K}}$ converge to zero. Condition 1.4(b) further requires that this decay is at least polynomial, $\sigma_j \lesssim j^{-1/p}$ with $p \in (0, 1)$. The eigenvalue-decay rate reflects the smoothness of the kernel and of the associated RKHS (see, e.g., Rasmussen and Williams, 2005). Finally, Condition 1.4(c) assumes that the eigenfunctions are uniformly bounded to control the local Rademacher complexity (cf. Lemma 5 in Nie and Wager (2021)).

The following condition imposes both that $\mathcal{Q}$ is not too large—satisfying both the moment conditions from Condition 1.2 and an additional smoothness condition—and that it is also large enough so that estimating θ_0 is challenging.

Condition 1.5 (Characterization of $\mathcal{Q}$). $\mathcal{Q}$ consists of all distributions satisfying Condition 1.2 and the following source condition: there exist constants $\alpha \in (0, 1/2)$ and $R < \infty$ such that, for any $Q^0 \in \mathcal{Q}$, the CDRF θ_0 satisfies

$$\|T_{\mathcal{K}}^{\alpha}\theta_0\|_{\mathcal{H}} \leq R.$$

Condition 1.5 is referred to as a *source condition* and requires that the true CDRF be sufficiently smooth, belonging to a $(1-2\alpha)$-power space where $0 < 1-2\alpha < 1$ (Fischer & Steinwart, 2020).

Given a nuisance estimator $\widehat{g}$, the use of kernel ridge regression in Algorithm 1.2 produces an estimator $\widehat{\theta}$. Since the penalty parameter λ is chosen by cross-validation, it is random, and consequently the corresponding ball radius c is also random. Nonetheless, as shown by Mitchell and van de Geer (2009) and van der Laan and Dudoit (2003), the cross-validation selector is asymptotically equivalent to the oracle selector. At the population level, Algorithm 1.2 is equivalent to empirical risk minimization over the constrained RKHS ball $\mathcal{H}_c$ centered at the origin with radius $c \geq 1$, which depends only on λ. Combining these facts, the oracle constrained RKHS ball minimizer is asymptotically equivalent to the estimator from Algorithm 1.2. For clarity, we focus on this constrained formulation throughout the remainder of the chapter.

Theorem 1.2 (Excess Risk Bound for Algorithm 1.2)**.** *Assume that Conditions 1.1 to 1.5 hold. Fix $\delta \in (0,1)$, and let $(\widehat{\theta}, \widehat{g})$ be as in Algorithm 1.2, such that $\|\widehat{g} - g_0\|_{L^2(Q_X^0 \times \mu; \ell^2)} = o_p(1)$ and $\|\widehat{w} - w_0\|_{L^\infty(Q_X^0 \times \mu; \ell^2)} = o_p(1)$. Then there exists $N(\delta) \in \mathbb{N}$ such that, for all $n \geq N(\delta)$, with probability at least $1-\delta$,*

$$\begin{aligned} &L_{P^0}(\widehat{\theta}, g_0) - L_{P^0}(\theta^*, g_0) \\ &\quad \lesssim B_{P^0}(\delta)^2 \left(c^{\frac{2p}{1+p}} (n \log(n)^{-2})^{-\frac{1}{1+p}} + \frac{\log(1/\delta)}{n} \right) + M_\lambda^2 c^{\frac{2p}{1+p}} \|\widehat{g} - g_0\|_{L^2(Q_X^0 \times \mu; \ell^2)}^{\frac{4}{1+p}}. \end{aligned}$$

Moreover, suppose $\widehat{g}$ satisfies $\|\widehat{g} - g_0\|_{L^2(Q_X^0 \times \mu; \ell^2)} = o_p\big((n \log(n)^{-2})^{-1/4}\big)$, and the regularization parameter is chosen as

$$c \propto \left(B_{P^0}(\delta)^{-2(1+p)} n \log(n)^{-2} \right)^{\frac{\alpha}{p+(1-2\alpha)}}.$$

Then there exists $N'(\delta) \in \mathbb{N}$ *such that, for all* $n \geq N'(\delta)$*, with probability at least* $1-\delta$*,*

$$L_{P^0}(\widehat{\theta}, g_0) = L_{P^0}(\widehat{\theta}, g_0) - L_{P^0}(\theta_0, g_0) \lesssim \left(B_{P^0}(\delta)^{-2(1+p)} n \log(n)^{-2}\right)^{-\frac{1-2\alpha}{p+(1-2\alpha)}}.$$

Theorem 1.2 is a special case of Theorem 1.1 with Θ an RKHS ball $\mathcal{H}_c$. The proof derives a valid critical radius via Lemma 5 of Nie and Wager (2021), which bounds the local Rademacher complexity in terms of c, δ, and n. It also requires a weaker condition on the nuisance rates (L^2 rather than L^4 in Theorem 1.1). For the second part, the approximation error is absorbed into the excess risk term, yielding an estimation error bound relative to the global minimizer θ_0. Moreover, Theorem 1.2 provides not only a high-probability upper bound on the risk but—under a stronger uniform rate for the nuisance estimators as in Theorem 1.4—also yields a bound uniform over $\mathcal{P}$, obtained by taking the supremum over $P \in \mathcal{P}$.

1.3.3 Efficiency gain and prediction guarantee

We now show that performing data fusion reduces the excess risk bound. Specifically, we compare the excess risk bounds under two settings: (i) using data fusion to leverage the full sample, and (ii) not using data fusion and instead only using target samples—that is, those from sources in $\mathcal{S}_X \cap \mathcal{S}_Y$. When not using data fusion, the (oracle) Neyman-orthogonal loss reduces to:

$$\begin{aligned}\underline{\ell}_{P^0}(\theta, g_0; z) := (\theta(b) - \underline{\tau}_0(b))^2 + 2\underline{\xi}_0 \mathbf{1}(s \in \mathcal{S}_X \cap \mathcal{S}_Y)\,(\theta(b) - \underline{\tau}_0(b))\,(\underline{\tau}_0(b) - \underline{m}_0(x,b)) \\ + 2\underline{\eta}_0 \underline{w}_0(x,a)\mathbf{1}(s \in \mathcal{S}_X \cap \mathcal{S}_Y)\,(\theta(a) - \underline{\tau}_0(a))\,(\underline{m}_0(x,a) - y)\,,\end{aligned}$$

where $z := (x, a, b, y, s)$ and the nuisance components are defined as

$$\underline{\xi}_0 := \frac{1}{P^0(S \in \mathcal{S}_X \cap \mathcal{S}_Y)}, \quad \underline{\eta}_0 := \frac{1}{P^0(S \in \mathcal{S}_X \cap \mathcal{S}_Y)}, \quad \underline{w}_0(x,a) := \frac{1}{p^0(a \mid x, S \in \mathcal{S}_X \cap \mathcal{S}_Y)},$$

$$\underline{m}_0(x,a) := \mathrm{E}_{P^0}\left[Y \mid X = x, A = a, S \in \mathcal{S}_X \cap \mathcal{S}_Y\right], \quad \underline{\tau}_0(a) := \mathrm{E}_{P^0}\left[\underline{m}_0(X,a) \mid S \in \mathcal{S}_X \cap \mathcal{S}_Y\right],$$

where $p^0(a \mid x, S \in \mathcal{S}_X \cap \mathcal{S}_Y)$ denotes the conditional density of A given X and $S \in \mathcal{S}_X \cap \mathcal{S}_Y$ with respect to the known measure μ on $\mathcal{A}$.

In this restricted setting, we analogously define the nuisance estimators $\widehat{\underline{\xi}}, \widehat{\underline{\eta}}, \widehat{\underline{w}}, \widehat{\underline{m}}$, and $\widehat{\underline{\tau}}$. Let $(\widehat{\theta}, \widehat{g})$ denote the estimators obtained from Algorithm 1.2 with data fusion, and $(\underline{\theta}, \underline{g})$ the corresponding estimators without data fusion, where $\underline{g} : (a, x) \mapsto (\widehat{\underline{\xi}}, \widehat{\underline{\eta}}, \widehat{\underline{w}}(a, x), \widehat{\underline{m}}(a, x), \widehat{\underline{\tau}}(a))$.

For Theorem 1.3 below, we impose the same nuisance estimation and cross-validation conditions as in the second statement of Theorem 1.2. Specifically, suppose the nuisance estimators satisfy

$$\|\widehat{g} - g_0\|_{L^2(Q_X^0 \times \mu; \ell^2)} = o_p\big((n \log(n)^{-2})^{-1/4}\big), \qquad \|\widehat{w} - w_0\|_{L^\infty(Q_X^0 \times \mu; \ell^2)} = o_p(1),$$
$$\|\underline{g} - g_0\|_{L^2(Q_X^0 \times \mu; \ell^2)} = o_p\big((n \log(n)^{-2})^{-1/4}\big), \qquad \|\widehat{\underline{w}} - w_0\|_{L^\infty(Q_X^0 \times \mu; \ell^2)} = o_p(1),$$

and the regularization parameter is chosen as

$$c \propto \left(B_{P^0}(\delta)^{-2(1+p)} n \log(n)^{-2}\right)^{\frac{\alpha}{p+(1-2\alpha)}}.$$

Theorem 1.3 (Data Fusion Improves the Excess Risk Upper Bound)**.** *Suppose the nuisance estimators and the regularization parameter satisfy the conditions listed directly above the theorem. Further assume that Conditions 1.1 to 1.5 hold, and that Conditions 1.1 and 1.3 also hold with fusion sets* $\underline{\mathcal{S}}_X = \underline{\mathcal{S}}_Y = \mathcal{S}_X \cap \mathcal{S}_Y$.

Fix $\delta \in (0, 1)$. *Then there exists* $N(\delta)$ *such that, for all* $n \geq N(\delta)$, *the* $(1 - \delta)$*-probability excess risk bounds under data fusion and without data fusion are both of order* $(n \log(n)^{-2})^{-\frac{1-2\alpha}{p+(1-2\alpha)}}$. *Moreover, they satisfy*

$$\frac{\text{Excess-risk bound under data fusion}}{\text{Excess-risk bound without data fusion}} = \left(\frac{1 + \xi_0 + C_{\sigma,\delta}\, \eta_0 \|w_0\|_\infty}{1 + \underline{\xi}_0 + C_{\sigma,\delta}\, \underline{\eta}_0 \|\underline{w}_0\|_\infty}\right)^{\frac{2(1+p)(1-2\alpha)}{p+(1-2\alpha)}} \leq 1,$$

where $C_{\sigma,\delta} := 1 + \max\Big\{\sigma\sqrt{2\log(8/\delta)},\, 2L\log(8/\delta)\Big\}$. *The inequality is strict if*

$$P^0\big(S \in \mathcal{S}_X \setminus \mathcal{S}_Y\big) + \operatorname*{ess\,inf}_{(X,A,Y)\sim P^0} P^0\big(S \in \mathcal{S}_Y \setminus \mathcal{S}_X \mid A, X, S \in \mathcal{S}_Y\big) > 0.$$

The condition for the strict inequality is slightly stronger than merely requiring the existence of some partially aligned sources; that is, $P^0\big(S \in \mathcal{S}_X \triangle \mathcal{S}_Y\big) = P^0\big(S \in \mathcal{S}_X \setminus \mathcal{S}_Y\big) + P^0\big(S \in \mathcal{S}_Y \setminus \mathcal{S}_X\big) > 0$. The key to establishing the above result is showing that using data fusion

necessarily reduces the loss's Lipschitz constant. We have shown that the excess-risk bound decreases under data fusion by applying Theorem 1.2.

Beyond excess risk upper bounds, we show that using data fusion necessarily improves performance in a minimax sense, under appropriate conditions. Concretely, we show a high-probability worst-case upper bound of the risk under data fusion is smaller than a worst-case lower bound of the risk without data fusion, provided there are sufficiently more data that are only partially aligned with the target distribution than are perfectly aligned. To formalize this result, we impose an additional assumption on the eigenvalue decay of the kernel integral operator, requiring it to be bounded both above and below by a polynomial rate.

Condition 1.6 (Eigenvalue Decay at Exactly Polynomial Rate). There exist $p \in (0,1)$ and constants $G_1, G_2 > 0$ such that, for all $j \geq 1$,

$$G_1 j^{-1/p} \leq \sigma_j \leq G_2 j^{-1/p}.$$

In the following lemma, we provide minimax lower bounds for an estimator $\underline{\theta}$ that does not use data fusion. Formally, any such $\underline{\theta}$ takes as input a sample $\{(X_i, A_i, Y_i, S_i)\}_{i=1}^n$ and is measurable with respect to the σ-algebra generated by $\{(X_i, A_i, Y_i)\}_{i:S_i \in \mathcal{S}_X \cap \mathcal{S}_Y}$.

Lemma 1.2 (Minimax Lower Bound without Data Fusion)**.** *Assume that Conditions 1.1 to 1.6 hold. Suppose* $1 - 2\alpha \geq p$. *Fix an estimator* $\underline{\theta}$ *that does not use data fusion and* $\delta \in (0,1)$. *Let* $N_\delta < \infty$ *be as defined in* (A.6) *in the appendix. There exists a constant* $c_{\mathcal{Q}} > 0$ *depending on* $\mathcal{Q}$ *only such that, for all* (n, n_{XY}) *such that* $n \geq n_{XY} \geq N_\delta$, *there exists* $P \in \mathcal{P}$ *such that* $P(S \in \mathcal{S}_X \cap \mathcal{S}_Y) = n_{XY}/n$ *and, with probability at least* $1 - 2\delta$,

$$L_P(\underline{\theta}, g_0) \geq c_{\mathcal{Q}} \delta \left(1 + \sqrt{3\log(1/\delta)/n_{XY}}\right)^{-\frac{1-2\alpha}{p+(1-2\alpha)}} n_{XY}^{-\frac{1-2\alpha}{p+(1-2\alpha)}}. \tag{1.3}$$

In the above lemma, n_{XY} denotes the expected number of perfectly aligned samples among n i.i.d. observations drawn from P. The proof proceeds by lower bounding a minimax risk over $\mathcal{P}$ by a worst-case Bayes risk over $\mathcal{Q}$ via a maximin argument. We then construct a finite family $\{Q_0, \ldots, Q_M\} \subset \mathcal{Q}$ whose CDRFs are well separated, while the corresponding

product measures are close in Kullback–Leibler divergence; applying Fano's inequality to this packing yields the desired bound.

We combine the minimax lower bound for the no–fusion estimator from Lemma 1.2 with the uniform excess risk upper bound for the data-fusion estimator from Theorem 1.2. The next theorem formalizes the cases when data fusion strictly improves the worst-case risk.

Let $(\widehat{\theta}, \widehat{g})$ be the estimators in Algorithm 1.2 and $\underline{\theta}$ be any estimator that does not use data fusion. In the following result, we assume the nuisance estimator $\widehat{g}$ satisfies the following uniform condition (stronger than Theorem 1.2): for any $\varepsilon > 0$, there exists $M_g < \infty$ such that

$$\limsup_{n\to\infty} \sup_{P\in\mathcal{P}} P^n\Big(\|\widehat{g} - g_0\|_{L^2(Q_X^0\times\mu;\,\ell^2)} > M_g \left(n \log(n)^{-2}\right)^{-1/4} \Big) \le \varepsilon. \tag{1.4}$$

We further suppose the regularization parameter is chosen as

$$c \propto \left(n \log(n)^{-2}\right)^{\frac{\alpha}{p+(1-2\alpha)}}. \tag{1.5}$$

Theorem 1.4 (Sufficient Conditions for Data Fusion to Improve Worst-Case Performance)**.** *Suppose $(\widehat{\theta}, \widehat{g})$ satisfy the conditions above, c satisfies (1.5), Conditions 1.1 to 1.6 hold, and $1 - 2\alpha \ge p$. Fix $\delta \in (0, 1/2]$ and $h > 0$. Suppose the* ***non-data fusion sample size*** *is sufficiently large and the* ***data fusion sample size*** *is sufficiently larger, in that, for N_δ as defined in (A.10),*

$$N_\delta \le \boldsymbol{n_{XY}} \le \tfrac{n}{(\log n)^{2+h}} \le \boldsymbol{n}.$$

Then, for $\mathcal{P}_{XY} := \{P \in \mathcal{P} : P(S \in \mathcal{S}_X \cap \mathcal{S}_Y) = n_{XY}/n\}$ and $r(\delta, n) := c_{\mathcal{Q}} \delta n^{-\frac{1-2\alpha}{p+(1-2\alpha)}}/4$,

$$\sup_{P\in\mathcal{P}_{XY}} P^n\Big\{ L_P(\underline{\theta}, g_0) \ge r(\delta, n) \Big\} - \sup_{P\in\mathcal{P}_{XY}} P^n\Big\{ L_P(\widehat{\theta}, g_0) > r(\delta, n)/\log(n)^{\frac{h(1-2\alpha)}{2[p+(1-2\alpha)]}} \Big\} \;\ge\; 1 - 2\delta. \tag{1.6}$$

To our knowledge, this is the first work to formally demonstrate that data fusion improves performance in a statistical learning problem. Notably, this improvement holds regardless of the estimation algorithm used in the non-fusion setting. To establish this result, we use that the first term on the left-hand side is approximately one by Lemma 1.2, and that, by taking

the supremum of the excess-risk bound in Theorem 1.2, the second term is approximately zero.

The suprema in the theorem may be attained at different distributions, which may appear to complicate interpretation. However, this also makes it possible to exhibit a particular distribution under which using data fusion improves performance. Concretely, letting $P_\star$ denote a (near) maximizer of the first supremum in (1.6), it follows that

$$P_\star^n \left\{ \frac{L_{P_\star}(\widehat{\theta}, g_0)}{L_{P_\star}(\underline{\theta}, g_0)} \leq \log(n)^{-\frac{h(1-2\alpha)}{2[p+(1-2\alpha)]}} \right\} \geq 1 - 3\delta.$$

In other words, for sufficiently large n, there exists a distribution under which using data fusion substantially outperforms not using data fusion.

1.4 Numerical experiments

We evaluate the empirical performance of our proposed data fusion kernel ridge CDRF estimator. Our goal is to assess whether incorporating additional partially aligned datasets via data fusion improves predictive performance, even when the reference measure is different from the data-generating measure or the true CDRF does not belong to the chosen RKHS. Throughout we focus on the RKHS with a Laplace kernel and bandwidth chosen according to the median heuristic (Garreau et al., 2018).

We conducted Monte Carlo simulations with 500 runs under various configurations. The data are generated as follows: $S_X \sim \text{Bernoulli}(0.5)$, $S_Y \sim \text{Bernoulli}(0.5)$, $S_X \perp\!\!\!\perp S_Y$, $X \mid S_X = 1 \sim \mathcal{N}(\mu_0 = \frac{1}{3}(1,1,1)^\top,\ \Sigma_0 = 0.09 \cdot I_3)$,

$$X \mid S_X = 0 \sim \mathcal{N}\left(\mu_1 = \tfrac{1}{6}(1,1,1)^\top,\ \Sigma_1 = \begin{bmatrix} 0.25 & 0.1 & 0.1 \\ 0.1 & 0.25 & 0.1 \\ 0.1 & 0.1 & 0.25 \end{bmatrix} \right),$$

$A \mid X \sim \text{Beta}\left(1 + 1/[1+\exp(\langle X, 1\rangle)],\ 1 + 1/[1+\exp(\langle X, 1\rangle)]\right)$, $Y \mid X, A, S_Y = 1 \sim \mathcal{N}(\theta_0(A) \cdot \langle X, 1\rangle,\ 0.1^2)$, and $Y \mid X, A, S_Y = 0 \sim \mathcal{N}(\widetilde{\theta}_0(A) \cdot \langle X, 1\rangle,\ 0.1^2)$, where $\langle X, 1\rangle$ denotes the row sum of covariates X, and $\theta_0(a)$ is the true CDRF.

We consider three distinct functional forms for $\theta_0(a)$ across simulation sets, along with corresponding misspecified counterparts $\tilde{\theta}_0(a)$ used to generate conditional outcomes. The first setting has a Gaussian-shaped CDRF, with $\theta_0(a) = \phi(a; 0.5, 0.25^2) - 1$ and $\widetilde{\theta}_0(a) = 0.5\left(\phi(a; 1, 0.25^2) - 1\right)$ for $\phi(a; \mu, \sigma^2)$ the Gaussian density with mean μ and variance σ^2. The second has CDRFs defined by trigonometric functions, with $\theta_0(a) = 5\sin(3a) + 3\cos(10a)$ and $\widetilde{\theta}_0(a) = 0.5[\cos(3a) + \sin(10a)]$. The third has discontinuous CDRFs defined as

$$\theta_0(a) = \begin{cases} \left(a^{0.5} + 0.1\right), & \text{if } a < 0.5, \\ 0.5\left(a^4 + 1\right), & \text{if } a \geq 0.5, \end{cases} \qquad \widetilde{\theta}_0(a) = \begin{cases} \left(\log(1+a)\right)^{0.5}, & \text{if } a < 0.3, \\ 0.1\left(\cos(3a) + 1\right), & \text{if } a \geq 0.3. \end{cases}$$

The discontinuous CDRFs are of interest since—unlike the other two CDRFs considered (Wendland, 2004, Corollary 10.48)—they are not contained in the RKHS induced by the Laplacian kernel.

For all three true and misspecified CDRFs, we evaluated performance under three different reference measures for risk assessment. The first is the uniform distribution over $[0, 1]$, which assigns equal importance to all parts of the dose-response curve. The second is Beta$(5, 5)$, a bell-shaped distribution emphasizing the center of the range, thereby making estimation accuracy near $a = 0.5$ more important. The third is Beta$(0.5, 0.5)$, a U-shaped distribution that places more weight on boundary regions, which may be useful when treatment effects at the boundaries are of particular interest.

For each scenario described above and a range of sample sizes n, we computed the median empirical risk of the estimated CDRFs with and without data fusion. Following the procedure in Algorithm 1.2, we randomly split the dataset into two halves. The first half was used to estimate nuisance parameters, while the second half was used for target estimation and risk evaluation.

Specifically, we began by estimating the source probabilities $P^0(S \in \mathcal{S}_X)$ and $P^0(S \in \mathcal{S}_Y)$ using empirical proportions. We then estimated the density ratio $(x, a) \mapsto w_0(x, a)$ via the `densratio` package in R, implementing the unconstrained Least-Squares Importance Fitting (uLSIF) method (Kanamori et al., 2009). For the conditional outcome regression

$(x, a) \mapsto m_0(x, a)$, we used the SuperLearner package (van der Laan et al., 2007), combining mean regression, random forests, XGBoost, and elastic net. To choose the regularization parameter λ for kernel ridge regression, we performed cross-validation over a grid $\{0.0001, 0.0051, \ldots, 0.0301\}$, selecting the value minimizing estimated risk, as detailed in Algorithm A.1.

Given the estimated nuisance components and the selected λ, we computed the closed-form estimator of the CDRF as in Algorithm 1.2. We further evaluated performance by computing the empirical risk, defined as the mean squared error between the estimated and true CDRFs, where the evaluation points a were sampled from the designated reference measure in each setting. We also report the percentage reduction in median risk achieved through data fusion.

The results are summarized in Table 1.1. Across all settings, incorporating data fusion improves performance. This includes the discontinuous CDRF case in Table 1.1c, where the true CDRF does not lie in the RKHS of the estimator. Nonetheless, the proposed method still achieved good performance even at moderate sample sizes, demonstrating robustness to model misspecification. The tables also reveal robustness to the choice of reference measure. Importantly, our method benefits from data fusion even in cases where the reference measure differs from the data-generating distribution.

1.5 Discussion

Our data fusion framework can also be applied to the estimation of function-valued parameters other than CDRFs. As an extension, it can be used to estimate any smooth summary functional of such parameters—even when the associated risk is not of the mean integrated squared error type. This broadens the applicability of data fusion beyond traditional risk measures.

By employing a stochastic approximation, we successfully construct a Neyman-orthogonal loss function for such risks that avoids taking expectations over the parameter's indexing variable. The resulting orthogonal loss is almost the same as the loss derived directly from the

nonparametric efficient influence function, up to an $O_p(1)$ term. While similar constructions may be possible for other parameters, a general theoretical discussion remains to be developed.

Although CDRF estimation under data fusion shows both strong theoretical guarantees and favorable empirical performance, several caveats remain. First, the estimation algorithms assume correct knowledge of which part of each data source aligns with the target distribution. Second, the methods are tailored for prediction, not for statistical inference. Finally, while the regularization parameter λ is data-dependent, the radius c of the RKHS ball is fixed. In practice, kernel ridge regression behaves similarly to RKHS-ball-constrained optimization when the sample size is sufficiently large.

There are several promising directions for future work. One is to address federated learning settings where individual-level data from each clinical trial are private and cannot be shared, but summary statistics can be exchanged (Han et al., 2025). Another is to extend statistical inference to non-pathwise differentiable parameters under data fusion (Luedtke & Chung, 2024). Lastly, it would be interesting to explore the use of neural networks or other function classes for estimating CDRFs, as in the orthogonal statistical learning framework (Foster & Syrgkanis, 2023).

Table 1.1: Median risks and percent reduction from using data fusion across sample sizes.

(a) Gaussian CDRF, with risks scaled by $\times 10^{-3}$

Sample Size	Uniform(0, 1)			Beta(5, 5)			Beta(0.5, 0.5)		
	Fusion	No Fusion	(%)	Fusion	No Fusion	(%)	Fusion	No Fusion	(%)
100	90.2	145.3	38	259.8	265.3	2	169.4	251.7	33
200	48.9	82.8	41	128.9	189.4	32	98.2	158.0	38
400	16.9	37.6	55	23.9	81.3	71	30.1	82.4	64
800	6.7	10.3	35	7.4	15.4	52	10.1	17.9	44
1600	3.1	4.4	30	3.6	5.3	31	4.7	6.6	30
3200	1.8	2.3	19	2.4	3.2	25	2.7	3.4	21

(b) Trigonometric CDRF, with risks scaled by $\times 10^{-2}$

Sample Size	Uniform(0, 1)			Beta(5, 5)			Beta(0.5, 0.5)		
	Fusion	No Fusion	(%)	Fusion	No Fusion	(%)	Fusion	No Fusion	(%)
100	367.3	549.4	33	804.1	907.6	11	519.6	855.2	39
200	213.6	306.2	30	439.1	689.0	36	334.7	499.2	33
400	78.8	149.2	47	164.5	350.5	53	124.7	294.4	58
800	36.4	51.2	29	51.1	61.7	17	39.5	70.7	44
1600	15.8	23.1	32	21.5	29.0	26	15.3	26.6	43
3200	8.3	12.6	34	11.4	17.4	34	7.5	11.6	35

Table 1.1: Median risks and percent reduction from using data fusion across sample sizes (continued).

(c) Discontinuous CDRF, with risks scaled by $\times 10^{-3}$

Sample Size	Uniform(0, 1)			Beta(5, 5)			Beta(0.5, 0.5)		
	Fusion	No Fusion	(%)	Fusion	No Fusion	(%)	Fusion	No Fusion	(%)
100	143.7	198.8	28	190.7	214.4	11	208.4	275.2	24
200	63.4	111.9	43	80.3	136.2	41	95.5	169.2	44
400	31.6	45.8	31	32.5	56.8	43	49.6	67.8	27
800	19.5	26.1	26	19.7	25.1	21	29.8	32.4	8
1600	10.8	16.1	33	13.3	15.8	16	17.5	19.2	8
3200	6.7	9.2	27	8.5	10.0	15	9.9	12.6	21

Chapter 2

TESTING SUBSPACE HYPOTHESES ABOUT FUNCTION-VALUED PARAMETERS VIA KERNEL EMBEDDINGS

2.1 Introduction

Testing hypotheses about function-valued parameters is a recurring problem in statistics, and especially in causal inference. Examples include whether a conditional average treatment effect is zero (Luedtke et al., 2019), whether a dose-response curve is constant (Hudson et al., 2024; Westling, 2022), and whether a treatment alters the counterfactual outcome distribution beyond its mean (Jain & Luedtke, 2026; Kennedy, 2023). The obstacle is that such parameters are typically not smooth enough to be estimated at the parametric rate and standard estimators on which a test could be based are unavailable (Luedtke & Chung, 2024).

A first line of work performs inference on the function-valued parameter directly. Luedtke et al. (2019) translate equality of two unknown functions into equality of the distributions they induce on the data, and measure the latter by the maximum mean discrepancy (Gretton et al., 2006). Hudson et al. (2026) propose a restricted score test that evaluates the risk derivative over a class of smooth directions, and Hudson et al. (2024) apply it to the causal dose-response function. The flat null of the dose-response curve is also tested by Westling (2022) through its integrated form. Others project the parameter onto a finite-dimensional model under a generalized distance (Kennedy, 2023) or onto a series space whose dimension grows with the sample size (Wang et al., 2026).

A second line of work estimates or tests after embedding the target into a reproducing kernel Hilbert space (RKHS). Jain and Luedtke (2026) embed conditional outcome distributions

to test conditional distributional treatment effects. Morzywolek et al. (2025) embed local variable importance measures for treatment effect heterogeneity into an RKHS to obtain the efficient influence function, and test the null of no importance. Luedtke and Chung (2024) provide a general framework of one-step estimation and statistical inference for pathwise differentiable Hilbert-valued parameters.

A common theme runs through these problems. Each asks whether the function-valued parameter belongs to a closed linear subspace, such as the set of constant functions, additive functions, or functions that do not depend on a given subset of covariates. We develop a unified framework for testing such subspace hypotheses. Our contributions are as follows:

1. We propose a nonparametric test of whether a pathwise differentiable function-valued parameter belongs to a pre-specified closed linear subspace, in the regime where the projected parameter need not admit an efficient influence function.

2. We smooth the projected parameter by kernel embedding. We characterize exactly when the embedded parameter admits an efficient influence function in an RKHS codomain. When it fails to, we embed it further into L^2 space. A universal kernel ensures that each embedded null remains equivalent to the original one. Then we construct one-step estimators of the embedded parameters.

3. Based on the weak convergence of these estimators, we construct Wald-type and norm-based tests with bootstrap critical values, together with confidence sets that remain valid whether or not the null holds.

2.2 Statistical test of function-valued parameter

Let $Z \sim P^0$, and let f be a known coarsening map. Define $X := f(Z)$, and let $P_X^0 := f_\sharp P^0$ denote the induced distribution of X. We seek to test whether the distribution-dependent function $\nu(P^0)$ satisfies

$$H_0 : \nu(P^0) \in \mathcal{S}, \tag{2.1}$$

where $\mathcal{S}$ is a closed linear subspace of $L^2(P_X^0)$ that does not depend on P_X^0. Let $\mathcal{S}_0^\perp$ denote the orthogonal complement of $\mathcal{S}$ in $L^2(P_X^0)$, and let $\Pi_0^\perp$ denote the corresponding orthogonal projection onto $\mathcal{S}_0^\perp$, so that H_0 is equivalent to $\Pi_0^\perp\big(\nu(P^0)\big) = 0$. Our tests are based on a kernel embedding of this projected parameter. Let $\mathcal{K}$ be a bounded universal kernel on $\mathcal{X}$ with $\kappa^2 := \sup_{x\in\mathcal{X}} \mathcal{K}(x,x) < \infty$ and reproducing kernel Hilbert space $\mathcal{H}$. Define

$$\gamma(P^0) := \Big[x' \mapsto \int \mathcal{K}(x',x)\, \Pi_0^\perp\big(\nu(P^0)\big)(x)\, dP_X^0(x)\Big] \in \mathcal{H}, \qquad \Gamma(P^0) := [\gamma(P^0)]_{P_X^0} \in L^2(P_X^0). \tag{2.2}$$

Since $\mathcal{K}$ is universal, the associated kernel integral operator is injective (Sriperumbudur et al., 2011), and hence

$$H_0 \iff \gamma(P^0) = 0 \text{ in } \mathcal{H} \iff \Gamma(P^0) = 0 \text{ in } L^2(P_X^0).$$

The two characterizations differ only in the codomain in which the embedded parameter is viewed, and this choice has consequences for inference: as we show in Section 2.2.2, the efficient influence function may fail to exist for the $\mathcal{H}$-valued parameter γ while existing for its $L^2(P_X^0)$-valued counterpart Γ. Examples of $\nu(P^0)$ include the conditional average treatment effect (CATE) (Nie & Wager, 2021), the causal dose-response function (CDRF) (Hirano & Imbens, 2004; Imai & van Dyk, 2004), and the counterfactual density function (Kennedy, 2023).

2.2.1 Testing membership in a closed linear subspace

Let $\mathcal{P}$ be a collection of distributions on a Polish space $(\mathcal{Z}, \mathcal{B})$, where $\mathcal{Z} \subset \mathbb{R}^d$, and let $\nu : \mathcal{P} \to L^2(P_X)$ denote a function-valued parameter. For each $P \in \mathcal{P}$, let $P_X := f_\sharp P$ denote the distribution of $X = f(Z)$ under $Z \sim P$; the case $P = P^0$ recovers P_X^0 above. We assume throughout that the distributions in $\mathcal{P}$ are mutually equivalent and that the induced marginal density ratios are uniformly bounded: for all $P, P' \in \mathcal{P}$, $\operatorname{ess\,sup} dP_X/dP'_X < \infty$. In particular, writing $\rho_P := dP_X^0/dP_X$, both ρ_P and ρ_P^{-1} are essentially bounded, so the norms $\|\cdot\|_{L^2(P_X)}$ and $\|\cdot\|_{L^2(P_X^0)}$ are equivalent for every $P \in \mathcal{P}$. For each $P \in \mathcal{P}$, the space $L^2(P_X)$ is a real separable Hilbert space with inner product $\langle g, h\rangle_{L^2(P_X)} := \int g(x)h(x)\, dP_X(x)$.

Let $\mathcal{T} \subset \mathbb{R}^{\mathcal{X}}$ be such that each $g \in \mathcal{T}$ belongs to $L^2(P_X)$ for every $P \in \mathcal{P}$. For each $P \in \mathcal{P}$, write $[\mathcal{T}]_{P_X} := \{[g]_{P_X} : g \in \mathcal{T}\}$, where $[g]_{P_X}$ denotes the P_X-equivalence class of g, and let $\mathrm{cl}_{P_X}[\mathcal{T}]$ denote the closure of $[\mathcal{T}]_{P_X}$ in $L^2(P_X)$. Since the induced distributions are mutually equivalent, the notion of equivalence does not depend on the particular choice of P_X, and under the bounded density ratio assumption above the closure $\mathrm{cl}_{P_X}[\mathcal{T}]$ is the same for all $P \in \mathcal{P}$. We denote this common closed linear subspace by $\mathcal{S}$.

For each $P \in \mathcal{P}$, let $\mathcal{S}_P^{\perp}$ denote the orthogonal complement of $\mathcal{S}$ in $L^2(P_X)$, that is,

$$\mathcal{S}_P^{\perp} := \left\{ h \in L^2(P_X) : \langle h, s \rangle_{L^2(P_X)} = 0 \text{ for all } s \in \mathcal{S} \right\}.$$

Since $\mathcal{S}$ is a closed linear subspace of $L^2(P_X)$, we have $L^2(P_X) = \mathcal{S} \oplus \mathcal{S}_P^{\perp}$. Let $\Pi_P^{\perp} : L^2(P_X) \to \mathcal{S}_P^{\perp}$ denote the corresponding orthogonal projection onto $\mathcal{S}_P^{\perp}$, and $\Pi_P := \mathrm{Id} - \Pi_P^{\perp}$ the projection onto $\mathcal{S}$. The specialization to $P = P^0$ recovers $\mathcal{S}_0^{\perp} = \mathcal{S}_{P^0}^{\perp}$ and $\Pi_0^{\perp} = \Pi_{P^0}^{\perp}$ defined above.

Example 2.1 (Choices of the subspace $\mathcal{S}$)**.** The following examples illustrate several interesting choices of the subspace $\mathcal{S}$ and the corresponding projection $\Pi_0^{\perp}$ onto its orthogonal complement $\mathcal{S}_0^{\perp}$:

(i) The space of constant functions,

$$\mathcal{S} = \{x \mapsto c : c \in \mathbb{R}\},$$

for which

$$\Pi_0^{\perp}(h)(x) = h(x) - \mathrm{E}_{P_X^0}[h(X)].$$

This corresponds to testing whether $\nu(P^0)$ is constant, that is, whether it does not vary with the input level (Hudson et al., 2024; Kennedy, 2023; Westling, 2022).

(ii) The space of linear functions (with intercept),

$$\mathcal{S} = \{x \mapsto a + b^{\top} x : a \in \mathbb{R},\ b \in \mathbb{R}^d\},$$

for which

$$\Pi_0^{\perp}(h)(x) = h(x) - \mathrm{E}_{P_X^0}[h(X)] - (x - \mathrm{E}_{P_X^0}[X])^{\top}\{\mathrm{Cov}_{P_X^0}(X)\}^{-1}\mathrm{E}_{P_X^0}\{(X - \mathrm{E}_{P_X^0}[X])h(X)\}.$$

This corresponds to testing whether $\nu(P^0)$ is linear, and nests the constant case (i).

(iii) The generalized additive function space,

$$\mathcal{S} = \mathrm{cl}_{L^2(P_X^0)}\left\{ x \mapsto c + \sum_{i=1}^{d} g_i(x_i) : c \in \mathbb{R},\ g_i \in L^2(P_{X_i}^0),\ \mathrm{E}_{P_X^0}\{g_i(X_i)\} = 0,\ x = (x_1, \ldots, x_d) \right\}.$$

$\mathcal{S}$ contains the constants and the linear functions, so it nests the constant case (i) and the linear case (ii). Testing $\nu(P^0) \in \mathcal{S}$ amounts to testing for the absence of interactions among the coordinates. When $X_1, \ldots, X_d$ are independent under P_X^0, the centered marginal subspaces $\mathcal{S}_i^0 := \{g_i(X_i) : \mathrm{E}_{P_X^0}\{g_i(X_i)\} = 0\}$ are mutually orthogonal and the projection admits the closed form

$$\Pi_0^{\perp}(h)(x) = h(x) - \sum_{i=1}^{d} \mathrm{E}_{P_X^0}\{h(X) \mid X_i = x_i\} + (d-1)\,\mathrm{E}_{P_X^0}\{h(X)\}.$$

Without independence the marginal subspaces are no longer orthogonal and the projection has no closed form; with $c = \mathrm{E}_{P_X^0}\{h\}$, the centered components are characterized by the coupled normal equations

$$g_j = \mathrm{E}_{P_X^0}\Big\{ h - \textstyle\sum_{i \neq j} g_i \;\Big|\; X_j \Big\} - \mathrm{E}_{P_X^0}\{h\}, \qquad j = 1, \ldots, d,$$

which are solved by backfitting (Buja et al., 1989; Hastie & Tibshirani, 1986).

(iv) The space of functions that depend only on a subset of coordinates $\mathcal{V} \subseteq [d]$, namely

$$\mathcal{S} = \big\{ g : \mathbb{R}^d \to \mathbb{R} : g(x) = g(x') \text{ whenever } x_{\mathcal{V}} = x'_{\mathcal{V}} \big\}.$$

In this case,

$$\Pi_0^{\perp}(h)(x) = h(x) - \mathrm{E}\{h(X) \mid X_{\mathcal{V}} = x_{\mathcal{V}}\}.$$

This corresponds to testing whether $\nu(P^0)$ depends only on the coordinates in $\mathcal{V}$ (Morzywolek et al., 2025).

The reformulation above replaces H_0 with the hypothesis that a projected parameter vanishes: testing $\nu(P^0) \in \mathcal{S}$ is equivalent to testing $\Pi_0^{\perp}\big(\nu(P^0)\big) = 0$ in $L^2(P_X^0)$. However, an efficient influence function for this projected parameter is generally unavailable. Viewed as an $L^2(P_X^0)$-valued parameter, $P \mapsto \Pi_P^{\perp}\{\nu(P)\}$ typically does not admit one since its efficient influence operator acts on a test direction w by pointwise multiplication, so for fixed z the map $w \mapsto \dot{\nu}_P^*(\Pi_P^{\perp} w)(z)$ fails to be a bounded linear functional on $L^2(P_X^0)$ whenever it involves a point evaluation functional (Luedtke & Chung, 2024). Without an efficient influence function, no regular asymptotically linear estimator of the projected parameter exists (Luedtke & Chung, 2024, Appendix H). This motivates smoothing the projected parameter through the kernel embedding (2.2). In the next subsection, we show that the kernel embedded parameters can admit efficient influence functions.

2.2.2 Smoothness condition on parameters

We now generalize the embedded parameter of (2.2) to a generic distribution in the model. For a generic $\widetilde{P} \in \mathcal{P}$, let $\widetilde{P}_X := f_\sharp \widetilde{P}$, and define $\gamma : \mathcal{P} \to \mathcal{H}$ by

$$\gamma(\widetilde{P}) := \Big[x' \mapsto \int \mathcal{K}(x', x)\, \Pi_{\widetilde{P}}^{\perp}\big(\nu(\widetilde{P})\big)(x)\, d\widetilde{P}_X(x)\Big], \tag{2.3}$$

so that $\gamma(P^0)$ recovers (2.2). To study pathwise differentiability of γ at P^0, we work in the fixed ambient space $L^2(P_X^0)$. Since the induced distributions $\{P_X : P \in \mathcal{P}\}$ are mutually equivalent, we identify P_X-almost sure and P_X^0-almost sure equivalence classes whenever needed. In particular, let

$$\mathcal{B}(P_X^0) := \{u \in L^\infty(P_X^0) : \|u\|_{L^\infty(P_X^0)} \le m\} \subset L^2(P_X^0),$$

for some $m \in (0, \infty)$. We assume that $\nu(P) \in \mathcal{B}(P_X^0)$ for all $P \in \mathcal{P}$. This uniform bound serves two purposes: it provides the regularity needed to differentiate the projection primitive below, and it ensures the square integrability of the influence functions in Theorems 2.1 and 2.2.

Then, for any $P \in \mathcal{P}$, γ can be expressed as the composition

$$\gamma(P) = \theta_3(P, \theta_2(P, \theta_1(P, 0))), \tag{2.4}$$

where the primitive maps are defined as follows:

$$\begin{aligned}
&\theta_1 : \mathcal{P} \times \{0\} \to L^2(P_X^0), && \theta_1(P, 0) = \nu(P), \\
&\theta_2 : \mathcal{P} \times \mathcal{B}(P_X^0) \to L^2(P_X^0), && \theta_2(P, u) = \Pi_P^{\perp}(u), \\
&\theta_3 : \mathcal{P} \times L^2(P_X^0) \to \mathcal{H}, && \theta_3(P, u) = \Big[x' \mapsto \int \mathcal{K}(x', x) u(x)\, dP_X(x)\Big].
\end{aligned}$$

Here, θ_1 is the primitive corresponding to the parameter ν. Its degenerate second argument encodes that ν takes no function-valued input. The map θ_2 applies the $L^2(P_X)$-orthogonal projection onto $\mathcal{S}_P^{\perp}$. Although $\Pi_P^{\perp}(u)$ belongs to $\mathcal{S}_P^{\perp}$ as an element of $L^2(P_X)$, we view it as an element of the fixed ambient space $L^2(P_X^0)$ through the common equivalence-class representation. The map θ_3 is the kernel embedding primitive.

To prove the pathwise differentiability of γ, we first recall the notion of total pathwise differentiability as discussed in (Luedtke, 2026; Luedtke & Chung, 2024). Let $\mathcal{U} \subseteq \mathcal{T}'$ and $\mathcal{V} \subseteq \mathcal{W}$ be subsets of ambient Hilbert spaces $\mathcal{T}'$ and $\mathcal{W}$, respectively, and consider a map $\theta : \mathcal{P} \times \mathcal{U} \to \mathcal{V}$ and a parameter $\vartheta : \mathcal{P} \to \mathcal{V}$. For $u \in \mathcal{U}$ and $t \in \mathcal{T}'$, let $\mathscr{P}(u, \mathcal{U}, t)$ denote the collection of paths $\{u_\epsilon : \epsilon \in [0, 1]\} \subseteq \mathcal{U}$ satisfying $\|u_\epsilon - u - \epsilon t\|_{\mathcal{T}'} = o(\epsilon)$. Define the tangent set of $\mathcal{U}$ at u by $\check{\mathcal{U}}_u := \{t \in \mathcal{T}' : \mathscr{P}(u, \mathcal{U}, t) \neq \emptyset\}$, and let $\dot{\mathcal{U}}_u$ denote the $\mathcal{T}'$-closure of the linear span of $\check{\mathcal{U}}_u$. Similarly, let $\check{\mathcal{P}}_P := \{s \in L^2(P) : \mathscr{P}(P, \mathcal{P}, s) \neq \emptyset\}$, where $\mathscr{P}(P, \mathcal{P}, s)$ denotes the collection of quadratic-mean differentiable one-dimensional submodels $\{P_\epsilon : \epsilon \in [0, 1]\} \subseteq \mathcal{P}$ with score s at $\epsilon = 0$. Let $\dot{\mathcal{P}}_P$ denote the $L^2(P)$-closure of the linear span of $\check{\mathcal{P}}_P$.

We say that θ is totally pathwise differentiable at (P, u) if there exists a continuous linear operator $\dot{\theta}_{P,u} : \dot{\mathcal{P}}_P \oplus \dot{\mathcal{U}}_u \to \mathcal{W}$ such that, for every $s \in \check{\mathcal{P}}_P$, $t \in \check{\mathcal{U}}_u$, every path $\{P_\epsilon\} \in \mathscr{P}(P, \mathcal{P}, s)$, and every path $\{u_\epsilon\} \in \mathscr{P}(u, \mathcal{U}, t)$,

$$\Big\| \theta(P_\epsilon, u_\epsilon) - \theta(P, u) - \epsilon\, \dot{\theta}_{P,u}(s, t) \Big\|_{\mathcal{W}} = o(\epsilon).$$

This differential operator is unique; we denote its Hermitian adjoint by $\dot{\theta}_{P,u}^* : \mathcal{W} \to \dot{\mathcal{P}}_P \oplus \dot{\mathcal{U}}_u$. We say that ϑ is pathwise differentiable at P if there exists a continuous linear operator

$\dot{\vartheta}_P : \dot{\mathcal{P}}_P \to \mathcal{W}$ such that, for every $s \in \check{\mathcal{P}}_P$ and every path $\{P_\epsilon\} \in \mathscr{P}(P, \mathcal{P}, s)$,

$$\left\| \vartheta(P_\epsilon) - \vartheta(P) - \epsilon\, \dot{\vartheta}_P(s) \right\|_{\mathcal{W}} = o(\epsilon).$$

The operator $\dot{\vartheta}_P$ is called the local parameter of ϑ at P, and its Hermitian adjoint $\dot{\vartheta}_P^* : \mathcal{W} \to \dot{\mathcal{P}}_P$ the efficient influence operator.

In our framework, we assume that, for each $P \in \mathcal{P}$, the parameter ν is pathwise differentiable at P as a map into $L^2(P_X)$, with local parameter $\dot{\nu}_P : \dot{\mathcal{P}}_P \to L^2(P_X)$ and efficient influence operator $\dot{\nu}_P^* : L^2(P_X) \to \dot{\mathcal{P}}_P$, the adjoint taken with respect to the $L^2(P_X)$ inner product. The parameter ν admits an efficient influence function at P if and only if, for P-almost every z, the map $w \mapsto \dot{\nu}_P^*(w)(z)$ is a bounded linear functional on $L^2(P_X)$ (Luedtke & Chung, 2024). Throughout the paper, we do not assume that such an efficient influence function exists.

Lemma 2.1 (Pathwise differentiability and efficient influence operator of γ)**.** *Fix $P \in \mathcal{P}$, and suppose that the conditions stated above hold. Then the parameter γ is pathwise differentiable at P as an $\mathcal{H}$-valued parameter. Moreover, for every $w \in \mathcal{H}$, the Hermitian adjoint of the local parameter of γ is given by*

$$\dot{\gamma}_P^*(w)(z) = \Pi_P^\perp\{\nu(P)\}(x)\, \Pi_P^\perp(w)(x) - E_{P_X}\left[\Pi_P^\perp\{\nu(P)\}(X)\, \Pi_P^\perp(w)(X)\right] + \dot{\nu}_P^*\{\Pi_P^\perp(w)\}(z).$$

The preceding lemma establishes pathwise differentiability of the parameter γ and identifies its efficient influence operator. For one-step estimation, it is useful to know whether this operator can be represented by an $\mathcal{H}$-valued efficient influence function. The reproducing property of the RKHS provides a natural candidate for this representer, and the following result shows that any efficient influence function must coincide with it.

Theorem 2.1 (Efficient influence function for the $\mathcal{H}$-embedded parameter)**.** *Suppose that the conditions of Lemma 2.1 hold. For $x' \in \mathcal{X}$, let $\mathcal{K}_{x'} := \mathcal{K}(\cdot, x')$, and for $z \in \mathcal{Z}$ with*

$x = f(z)$ *define*

$$\begin{aligned}\phi_P(z)(x') &:= \dot{\gamma}_P^*(\mathcal{K}_{x'})(z) \\ &= \Pi_P^{\perp}\{\nu(P)\}(x)\,\Pi_P^{\perp}\{[\mathcal{K}_{x'}]_{P_X}\}(x) - E_{P_X}\left[\Pi_P^{\perp}\{\nu(P)\}(X)\,\Pi_P^{\perp}\{[\mathcal{K}_{x'}]_{P_X}\}(X)\right] \\ &\quad + \dot{\nu}_P^*\left(\Pi_P^{\perp}\{[\mathcal{K}_{x'}]_{P_X}\}\right)(z).\end{aligned} \tag{2.5}$$

Both of the following hold:

(i) *If γ admits an efficient influence function at P, then it equals ϕ_P P-almost surely.*

(ii) *If $z \mapsto \phi_P(z)$ defines an element of $\mathcal{H}$ for P-almost every z and $E_P\{\|\phi_P(Z)\|_{\mathcal{H}}^2\} < \infty$, then ϕ_P is the efficient influence function of γ at P: for every $w \in \mathcal{H}$, $\dot{\gamma}_P^*(w)(z) = \langle w, \phi_P(z)\rangle_{\mathcal{H}}$ for P-almost every z.*

Theorem 2.1 evaluates the efficient influence operator at the kernel sections to produce the natural $\mathcal{H}$-valued candidate, and part (i) shows that this candidate is the only possibility: if it fails to define an element of $\mathcal{H}$, no efficient influence function exists. We use a separable version of the efficient influence process throughout, as in Luedtke and Chung (2024). Whether the candidate lies in $\mathcal{H}$ depends on the interplay between the projection and the RKHS geometry: even though $\mathcal{K}_{x'} \in \mathcal{H}$, the projected feature $\Pi_P^{\perp}\{[\mathcal{K}_{x'}]_{P_X}\}$ is only guaranteed to lie in $L^2(P_X)$.

For finite-dimensional structural subspaces, no difficulty arises. If $\mathcal{S} = \operatorname{span}(q_1, \ldots, q_m)$ with each q_j admitting a representative in $\mathcal{H}$, then $\Pi_P^{\perp}\{[\mathcal{K}_{x'}]_{P_X}\}$ is the difference between $\mathcal{K}_{x'}$ and a finite linear combination of the q_j, and hence lies in $\mathcal{H}$. This covers the constant and linear examples. The situation is more delicate for infinite-dimensional structural subspaces. In the additive example with dependent covariates, the projection admits no closed-form expression: it is characterized only implicitly by the coupled normal equations, and is computed in practice by backfitting. No representation of $\Pi_P^{\perp}\{[\mathcal{K}_{x'}]_{P_X}\}$ as an element of $\mathcal{H}$ is then available, and we do not know of general conditions under which one exists. Rather than resolving this question for each structural subspace separately, we next embed the parameter into $L^2(P_X^0)$, where the required representer exists under a single spectral condition.

This motivates viewing the embedded parameter in the fixed ambient space $L^2(P_X^0)$. Define the inclusion primitive

$$\theta_4 : \mathcal{P} \times \mathcal{H} \to L^2(P_X^0), \qquad \theta_4(P,t) = [t]_{P_X^0},$$

which maps an RKHS element to its P_X^0-equivalence class, and define the $L^2(P_X^0)$-embedded parameter by $\Gamma := \theta_4 \circ (\mathrm{id}, \gamma)$, that is, $\Gamma(\widetilde{P}) := [\gamma(\widetilde{P})]_{P_X^0}$ for $\widetilde{P} \in \mathcal{P}$. This recovers $\Gamma(P^0)$ from (2.2). For every $t \in \mathcal{H}$ and $w \in L^2(P_X^0)$, the reproducing property yields $\langle [t]_{P_X^0}, w \rangle_{L^2(P_X^0)} = \langle t, T_{\mathcal{K},P^0} w \rangle_{\mathcal{H}}$, so $\theta_4(P,\cdot)$ is a bounded linear map with Hermitian adjoint $T_{\mathcal{K},P^0}$. In particular, θ_4 does not depend on its first argument, and, being bounded and linear in its second, it is totally pathwise differentiable with $\dot{\theta}_{4,P,t}(s,t') = [t']_{P_X^0}$. The chain rule of Luedtke (2026) then yields the following result.

Lemma 2.2 (Pathwise differentiability and efficient influence operator of Γ)**.** *Fix $P \in \mathcal{P}$ and suppose that the conditions of Lemma 2.1 hold. Then Γ is pathwise differentiable at P as an $L^2(P_X^0)$-valued parameter, with efficient influence operator*

$$\dot{\Gamma}_P^* = \dot{\gamma}_P^* \circ T_{\mathcal{K},P^0}.$$

The composition with $T_{\mathcal{K},P^0}$ smooths the test directions. Whereas $\dot{\gamma}_P^*$ is evaluated at arbitrary elements of $\mathcal{H}$, the operator $\dot{\Gamma}_P^*$ is evaluated only at directions of the form $T_{\mathcal{K},P^0} w$, whose coordinates in the Mercer basis are damped by the eigenvalues of $T_{\mathcal{K},P^0}$. The following result shows that this damping is precisely what the existence of an efficient influence function requires.

Theorem 2.2 (Efficient influence function for the $L^2(P_X^0)$-embedded parameter)**.** *Suppose that the conditions of Lemma 2.1 hold at P. Let $(e_k)_{k\geq 1}$ be an orthonormal basis of $L^2(P_X^0)$, and define $\psi_{P,k} := \Pi_P^{\perp}\big([T_{\mathcal{K},P^0} e_k]_{P_X}\big)$ and, for $z \in \mathcal{Z}$ with $x = f(z)$,*

$$\Phi_P(z) := \Phi_{P,12}(z) + \Phi_{P,3}(z), \qquad \Phi_{P,3}(z) := \sum_{k\geq 1} \dot{\nu}_P^*(\psi_{P,k})(z)\, e_k, \tag{2.6}$$

where

$$\Phi_{P,12}(z) := \Big[x' \mapsto \Pi_P^{\perp}\{\nu(P)\}(x)\, \Pi_P^{\perp}\{[\mathcal{K}_{x'}]_{P_X}\}(x) - E_{P_X}\big[\Pi_P^{\perp}\{\nu(P)\}(X)\, \Pi_P^{\perp}\{[\mathcal{K}_{x'}]_{P_X}\}(X)\big]\Big]_{P_X^0}.$$

Both of the following hold:

(i) *If* Γ *admits an efficient influence function at* P*, then it equals* Φ_P P*-almost surely, and*

$$E_P\Big[\sum_{k\geq 1}\big\{\dot{\nu}_P^*(\psi_{P,k})(Z)\big\}^2\Big] < \infty \tag{2.7}$$

holds.

(ii) *If* (2.7) *holds, then the series defining* $\Phi_{P,3}(z)$ *converges in* $L^2(P_X^0)$ *for* P*-almost every* z*, and* Φ_P *is the efficient influence function of* Γ *at* P*.*

In particular, Γ *admits an efficient influence function at* P *if and only if* (2.7) *holds, and the left side of* (2.7) *does not depend on the choice of orthonormal basis.*

Condition (2.7) quantifies the benefit of the additional embedding. Evaluated directly at basis directions, the efficient influence operator of the projected parameter involves point evaluations, whose squared coordinates fail to be summable when $\mathcal{S}_P^{\perp}$ is infinite-dimensional. After the embedding, each direction is first passed through $T_{\mathcal{K},P^0}$, and the spectral decay of its eigenvalues tempers that divergence. This term plays the role of the β-regularized efficient influence operator of Luedtke and Chung (2024), with the smoothing supplied intrinsically by the embedding rather than by an externally chosen truncation, so no adaptive truncation of the Mercer expansion is required. Moreover, whereas the regularized parameter Γ_β of Luedtke and Chung (2024) alters the estimand itself, the embedding leaves the null hypothesis intact: by the injectivity of $T_{\mathcal{K},P^0}$, $\Gamma(P^0) = 0$ remains exactly equivalent to H_0.

2.2.3 Weak convergence of one step estimators

Throughout this subsection we assume that the conditions of Theorem 2.1 and Theorem 2.2 hold at every $P \in \mathcal{P}$, so that γ and Γ admit efficient influence functions. We write $\phi_n := \phi_{\widehat{P}_n}$, $\phi_0 := \phi_{P^0}$, and analogously Φ_n, Φ_0. When only Γ admits an efficient influence function, all statements below concerning $\widehat{\Gamma}_n$ remain valid, and those concerning $\widehat{\gamma}_n$ are vacuous. We present the estimators and conditions under sample splitting, in which $\widehat{P}_n$ is constructed on an

auxiliary sample independent of $Z_1, \ldots, Z_n$. The cross-fitted versions used in Algorithm 2.1 and our simulations follow in the standard way (Luedtke & Chung, 2024).

We propose one-step estimators of γ and Γ, defined by

$$\widehat{\gamma}_n := \gamma(\widehat{P}_n) + P_n\phi_n, \qquad \widehat{\Gamma}_n := \Gamma(\widehat{P}_n) + P_n\Phi_n,$$

where $P_n\phi_n := \frac{1}{n}\sum_{i=1}^n \phi_n(Z_i)$ denotes the empirical average over the evaluation sample, and analogously for $P_n\Phi_n$. Asymptotic linearity of these estimators reduces to the negligibility of two error terms: writing $\mathcal{R}_n^\gamma := \gamma(\widehat{P}_n) + P^0\phi_n - \gamma(P^0)$ for the remainder term and $\mathcal{D}_n^\gamma := (P_n - P^0)(\phi_n - \phi_0)$ for the drift term, the triangle inequality gives

$$\left\|\widehat{\gamma}_n - \gamma(P^0) - P_n\phi_0\right\|_{\mathcal{H}} \le \|\mathcal{R}_n^\gamma\|_{\mathcal{H}} + \|\mathcal{D}_n^\gamma\|_{\mathcal{H}},$$

so $\widehat{\gamma}_n$ is asymptotically linear with influence function ϕ_0 whenever the right side is $o_p(n^{-1/2})$. The same decomposition holds for $\widehat{\Gamma}_n$ with $(\Gamma, \Phi, \|\cdot\|_{L^2(P_X^0)})$ in place of $(\gamma, \phi, \|\cdot\|_{\mathcal{H}})$. The following conditions make the remainder and drift terms negligible.

Condition 2.1 (Sufficient conditions for asymptotic linearity of $\widehat{\gamma}_n$). Let $\widehat{\nu} := \nu(\widehat{P}_n)$, $\nu_0 := \nu(P^0)$, $\widehat{\Pi}^\perp := \Pi^\perp_{\widehat{P}_n}$, and define

$$\omega_n := \sup_{\|w\|_{\mathcal{H}} \le 1} \left\|(\widehat{\Pi}^\perp - \Pi_0^\perp)w\right\|_{L^2(P_X^0)}.$$

The following hold:

(a) *(One-step remainder of ν.)*

$$\sup_{\|w\|_{\mathcal{H}} \le 1} \left| P_X^0\{(\widehat{\nu} - \nu_0)\,\Pi_0^\perp w\} + P^0\dot{\nu}^*_{\widehat{P}_n}(\Pi_0^\perp w) \right| = o_p(n^{-1/2}).$$

(b) *(Projection-nuisance product rate)* $\omega_n\big(\omega_n + \|\widehat{\nu} - \nu_0\|_{L^2(P_X^0)}\big) = o_p(n^{-1/2})$;

(c) *(Drift.)* $\|\phi_n - \phi_0\|_{L^2(P^0;\mathcal{H})} = o_p(1)$.

Condition 2.1(a) is a second-order remainder requirement for ν, imposed only along directions in $\mathcal{S}_0^\perp$. It is implied by the corresponding requirement for one-step estimation

of ν itself. Condition 2.1(b) controls the estimation of the projection. Because $\mathcal{S}$ is a fixed subspace, the first-order contribution of $\Delta := \widehat{\Pi}^{\perp} - \Pi_0^{\perp}$ to the remainder vanishes for any P^0, whether or not H_0 holds, and what remains is of product order in Δ and $\delta := \widehat{\nu} - \nu_0$. Moreover, since the remainder is evaluated against test directions in the unit ball of $\mathcal{H}$ and Δ is a difference of orthogonal projections, every remaining term can be bounded with Δ acting on a smooth direction, so its error need be controlled only in the restricted norm ω_n rather than in operator norm, which would be impossible for infinite-dimensional $\mathcal{S}$. Condition 2.1(b) holds, for example, whenever ω_n and $\|\delta\|_{L^2(P_X^0)}$ are each $o_p(n^{-1/4})$, and it accommodates trade-offs between the two rates. Condition 2.1(c) ensures the drift term is negligible, and consistency of the nuisance estimators suffices (Luedtke & Chung, 2024, Lemma 3).

Condition 2.2 (Sufficient conditions for asymptotic linearity of $\widehat{\Gamma}_n$). With the notation of Condition 2.1, the following hold:

(a) *(One-step remainder of ν.)*

$$\sup_{\|u\|_{L^2(P_X^0)} \le 1} \left| P_X^0\{(\widehat{\nu} - \nu_0)\, \Pi_0^{\perp}(T_{\mathcal{K},P^0} u)\} + P^0 \dot{\nu}^*_{\widehat{P}_n}\left(\Pi_0^{\perp}(T_{\mathcal{K},P^0} u)\right) \right| = o_p(n^{-1/2});$$

(b) *(Projection–nuisance product rate.)* $\omega_n\big(\omega_n + \|\widehat{\nu} - \nu_0\|_{L^2(P_X^0)}\big) = o_p(n^{-1/2})$;

(c) *(Drift.)* $\|\Phi_n - \Phi_0\|_{L^2(P^0; L^2(P_X^0))} = o_p(1)$.

Condition 2.2(b) coincides with Condition 2.1(b). The projection error enters both paths through the same restricted norm ω_n, since the test directions of the Γ-path lie in a κ-multiple of the unit ball of $\mathcal{H}$. For the same reason, since $\|T_{\mathcal{K},P^0} u\|_{\mathcal{H}} \le \kappa \|u\|_{L^2(P_X^0)}$, Condition 2.2(a) is implied by Condition 2.1(a). Whenever ϕ_0 exists, Condition 2.2(c) is implied by Condition 2.1(c), since $\Phi_P = [\phi_P]_{P_X^0}$ and $\|[h]_{P_X^0}\|_{L^2(P_X^0)} \le \kappa \|h\|_{\mathcal{H}}$. In this sense, the embedding only smooths the test directions, and estimating Γ is no harder than estimating γ. When only Γ admits an efficient influence function, Condition 2.2 stands on its own.

Conditions 2.1 and 2.2 render the remainder and drift terms negligible, so each estimator is asymptotically linear with its efficient influence function as influence function. Since these influence functions are Bochner square integrable by construction, a Hilbert-valued central limit theorem (van der Vaart & Wellner, 1996, Example 1.8.5) applies to the leading term and yields a tight Gaussian limit; regularity follows by the argument of Luedtke and Chung (2024, Theorem 2). We state the two results in parallel.

Theorem 2.3 (Weak convergence of $\widehat{\gamma}_n$)**.** *Fix $P^0 \in \mathcal{P}$, suppose the conditions of Theorem 2.1(ii) hold at P^0, so that γ admits the efficient influence function $\phi_0 \in L_0^2(P^0; \mathcal{H})$, and let Condition 2.1 hold. Then $\widehat{\gamma}_n$ is asymptotically linear,*

$$\widehat{\gamma}_n - \gamma(P^0) = \frac{1}{n}\sum_{i=1}^{n} \phi_0(Z_i) + o_p(n^{-1/2}) \qquad \text{in } \mathcal{H},$$

and is regular. Consequently,

$$\sqrt{n}\{\widehat{\gamma}_n - \gamma(P^0)\} \rightsquigarrow \mathbb{G}_\gamma \qquad \text{in } \mathcal{H},$$

where $\mathbb{G}_\gamma$ is a tight $\mathcal{H}$-valued Gaussian random element with, for each $g \in \mathcal{H}$,

$$\langle \mathbb{G}_\gamma, g\rangle_{\mathcal{H}} \sim N\big(0,\ E_{P^0}[\langle \phi_0(Z), g\rangle_{\mathcal{H}}^2]\big).$$

The same argument applies to $\widehat{\Gamma}_n$, with values in the fixed ambient space $L^2(P_X^0)$.

Theorem 2.4 (Weak convergence of $\widehat{\Gamma}_n$)**.** *Fix $P^0 \in \mathcal{P}$, suppose the conditions of Theorem 2.2(ii) hold at P^0, so that Γ admits the efficient influence function $\Phi_0 \in L_0^2(P^0; L^2(P_X^0))$, and let Condition 2.2 hold. Then $\widehat{\Gamma}_n$ is asymptotically linear,*

$$\widehat{\Gamma}_n - \Gamma(P^0) = \frac{1}{n}\sum_{i=1}^{n} \Phi_0(Z_i) + o_p(n^{-1/2}) \qquad \text{in } L^2(P_X^0),$$

and is regular. Consequently,

$$\sqrt{n}\{\widehat{\Gamma}_n - \Gamma(P^0)\} \rightsquigarrow \mathbb{G}_\Gamma \qquad \text{in } L^2(P_X^0),$$

where $\mathbb{G}_\Gamma$ is a tight $L^2(P_X^0)$-valued Gaussian random element with, for each $g \in L^2(P_X^0)$,

$$\langle \mathbb{G}_\Gamma, g\rangle_{L^2(P_X^0)} \sim N\big(0,\ E_{P^0}[\langle \Phi_0(Z), g\rangle_{L^2(P_X^0)}^2]\big).$$

The limits are determined by their marginals. Writing $\Sigma_0 : g \mapsto E_{P^0}[\langle \Phi_0(Z), g\rangle_{L^2(P_X^0)} \Phi_0(Z)]$ for the covariance operator of Φ_0, the Gaussian limit $\mathbb{G}_\Gamma$ has marginal variance $\langle \Sigma_0 g, g\rangle_{L^2(P_X^0)}$ in each direction g, and Σ_0 is trace class since $\mathrm{tr}(\Sigma_0) = E_{P^0}\|\Phi_0(Z)\|^2 < \infty$; the analogous statements hold for $\mathbb{G}_\gamma$. The test statistics of the next subsection are quadratic forms in $\widehat{\Gamma}_n$ weighted by estimates of this operator.

2.2.4 Confidence sets and hypothesis tests

The construction of the test and confidence set is identical for the two embedded parameters, so we treat them in a unified notation: let $(\eta, \widehat{\eta}_n, \Psi, \mathcal{W}, \mathbb{G}_\bullet)$ denote either $(\gamma, \widehat{\gamma}_n, \phi, \mathcal{H}, \mathbb{G}_\gamma)$ or $(\Gamma, \widehat{\Gamma}_n, \Phi, L^2(P_X^0), \mathbb{G}_\Gamma)$, with $\mathbb{G}_\bullet$ the Gaussian limit of Theorem 2.3 or Theorem 2.4. Recall that H_0 is equivalent to $\eta(P^0) = 0$ in $\mathcal{W}$.

We base inference on the quadratic functional $W(h; \Omega) := \langle \Omega h, h\rangle_{\mathcal{W}}$, where Ω is a continuous, self-adjoint, positive-definite operator on $\mathcal{W}$. Two choices are of interest:

- *Norm-squared (spherical):* $\Omega = I$, giving $W(h; I) = \|h\|_{\mathcal{W}}^2$. This is the kernel analogue of a maximum mean discrepancy and requires no covariance estimation.

- *Wald-type:* $\Omega = \Omega_0$, a regularized inverse of the covariance operator Σ_0, e.g. $\Omega_0 = [(1-\lambda)\Sigma_0 + \lambda I]^{-1}$ for fixed $\lambda \in (0, 1)$. This adapts to the local geometry of Σ_0, yielding higher power in low-variance directions.

The test statistic is $T_n := n\, W(\widehat{\eta}_n; \Omega_n)$, where Ω_n is an estimator of Ω_0. Under H_0 we have $\eta(P^0) = 0$, so Theorem 2.3 (resp. 2.4) and the continuous mapping theorem give $T_n \rightsquigarrow W(\mathbb{G}_\bullet; \Omega_0)$, as will be verified in the proof of Theorem 2.5 below. The limit is a weighted chi-square $\sum_k \tau_k N_k^2$, with weights (τ_k) the eigenvalues of $\Sigma_0^{1/2}\Omega_0\Sigma_0^{1/2}$ and hence unknown, so we estimate its $(1-\alpha)$-quantile by the bootstrap.

By asymptotic linearity, $\sqrt{n}\{\widehat{\eta}_n - \eta(P^0)\} = n^{-1/2}\sum_{i=1}^n \Psi_0(Z_i) + o_p(1)$, so the limit $\mathbb{G}_\bullet$ is the Gaussian limit of the leading influence-function term, with covariance determined by the law of $\Psi_0(Z)$ under P^0. This motivates approximating its law by re-weighting the fixed

evaluations $\Psi_n^{j(i)}(Z_i)$. The nuisances are never refit, and these evaluations are computed once and reused across replications. Formally, let ζ_n denote the $(1-\alpha)$-quantile of the conditional distribution of $W(\mathbb{G}_n^{(1)}; \Omega_n)$ given $Z_1, \ldots, Z_n$, where $\mathbb{G}_n^{(1)}$ is the multiplier process defined in Algorithm 2.1. This threshold is a deterministic function of the data, and Algorithm 2.1 returns a Monte Carlo approximation.

Algorithm 2.1 Bootstrap threshold: Monte Carlo approximation of ζ_n

Require: cross-fitted influence function estimates Ψ_n^j on folds $j \in \{1, 2\}$ with index sets I_1, I_2 of sizes n_1, n_2, with $j(i)$ the fold whose nuisance estimate was constructed without Z_i; operator estimate Ω_n; replications B

for $b = 1, \ldots, B$ **do**
  for all $j \in \{1, 2\}$ **do**
    draw $\big(\xi_i^{(b)}\big)_{i \in I_j} \sim \text{Multinomial}\big(n_j;\, 1/n_j, \ldots, 1/n_j\big) - 1$ independently
  end for
  $\mathbb{G}_n^{(b)} \leftarrow \dfrac{1}{\sqrt{n}} \sum_{i=1}^n \xi_i^{(b)}\, \Psi_n^{j(i)}(Z_i)$
  $T_n^{(b)} \leftarrow W\big(\mathbb{G}_n^{(b)}; \Omega_n\big)$
end for
return $\widehat{\zeta}_{n,\alpha}$, the empirical $(1-\alpha)$-quantile of $\{T_n^{(b)}\}_{b=1}^B$.

Because the multipliers are zero-centered multinomial counts drawn within each fold, this scheme reproduces the empirical bootstrap of Luedtke and Chung (2024, Section 4.2). $\xi_i^{(b)} + 1$ is the number of times observation i appears in a resample of size n_j from the fold-j empirical distribution. In practice the test is carried out with $\widehat{\zeta}_{n,\alpha}$ in place of ζ_n. Its Monte Carlo error is negligible for large B, so the results below are stated for the exact threshold ζ_n and apply equally to $\widehat{\zeta}_{n,\alpha}$. Inverting this threshold yields a confidence set for the embedded parameter, and evaluating it at $h = 0$ yields the test of H_0. The following results follow directly from Theorems 4 and 3(i) of Luedtke and Chung (2024).

Theorem 2.5 (Confidence set validity)**.** *Suppose the conditions of Theorem 2.3 (resp. 2.4)*

hold, $\|\Psi_0\|_{L_0^2(P^0;\mathcal{W})} > 0$*, the operators* Ω_n, Ω_0 *are continuous, self-adjoint, and positive-definite with* $\|\Omega_n - \Omega_0\|_{\mathrm{op}} = o_p(1)$*, and* $\max_j \|\Psi_n^j - \Psi_0\|_{L_0^2(P^0;\mathcal{W})} = o_p(1)$*. Then* $\zeta_n \xrightarrow{p} \zeta_{1-\alpha}$*, the* $(1-\alpha)$*-quantile of* $W(\mathbb{G}_\bullet; \Omega_0)$*, and*

$$C_n := \left\{ h \in \mathcal{W} : \; n\, W(\widehat{\eta}_n - h; \Omega_n) \le \zeta_n \right\}$$

satisfies $\lim_{n\to\infty} P_{P^0}^n\{\eta(P^0) \in C_n\} = 1 - \alpha$ *for every* $P^0 \in \mathcal{P}$*, whether or not* H_0 *holds.*

Theorem 2.6 (Validity of the test)**.** *Under the conditions of Theorem 2.5, the test that rejects* H_0 *when* $T_n > \zeta_n$ *satisfies:*

(i) (Size) *for every* P^0 *with* $\nu(P^0) \in \mathcal{S}$*,* $\lim_{n\to\infty} P_{P^0}^n(T_n > \zeta_n) = \alpha$*;*

(ii) (Consistency) *for every fixed* P^0 *with* $\nu(P^0) \notin \mathcal{S}$*,* $\lim_{n\to\infty} P_{P^0}^n(T_n > \zeta_n) = 1$*.*

By construction, the level-α test rejects H_0 if and only if $0 \notin C_n$, so Theorem 2.6(i) is the $h = 0$ instance of the coverage guarantee of Theorem 2.5. We emphasize that the coverage in Theorem 2.5 holds at every $P^0 \in \mathcal{P}$: the confidence set remains informative under the alternative, where procedures that exploit the null structure provide no distributional guarantees.

2.3 Examples of function-valued parameters ν_P

We illustrate the framework with three examples of function-valued parameters. For each, we give the parameter and its efficient influence operator.

1. **Conditional variance.** For $Z = (X, Y)$, the conditional variance function $\nu(P) : x \mapsto \mathrm{Var}_P(Y \mid X = x)$, viewed as a map into $L^2(P_X^0)$. Its efficient influence operator is

$$\dot{\nu}_P^*(w)(z) = \frac{dP_X^0}{dP_X}(x)\, w(x) \left[\big(y - E_P(Y \mid X = x)\big)^2 - \mathrm{Var}_P(Y \mid X = x) \right], \qquad w \in L^2(P_X^0).$$

2. **Conditional causal effect functions.** For $Z = (X, A, Y) \sim P \in \mathcal{P}$, define the outcome regressions $\mu_{P,a}(x) := \mathrm{E}_P[Y \mid A = a, X = x], \quad a \in \{0, 1\}$, and let $m :$

$\mathbb{R}^2 \to \mathbb{R}$, $(b, c) \mapsto m(b, c)$, be a known, continuously differentiable contrast function with partial derivatives $\partial_b m$ and $\partial_c m$. We consider the function-valued parameter $\nu : \mathcal{P} \to L^2(P_X^0)$ defined by

$$\nu(P)(x) := m(\mu_{P,1}(x),\, \mu_{P,0}(x))\,.$$

Notable choices include $m(b, c) = b - c$, which yields the conditional average treatment effect (CATE), and $m(b, c) = \log(b/c)$, the conditional causal log risk ratio. By the chain rule of Luedtke (2026), the efficient influence operator $\dot{\nu}_P^* : L^2(P_X^0) \to L^2(P)$ is, for any $w \in L^2(P_X^0)$,

$$\dot{\nu}_P^*(w)(z) = w(x)\,\frac{dP_X^0}{dP_X}(x)\Bigg[\partial_b m\big(\mu_{P,1}(x), \mu_{P,0}(x)\big)\,\frac{\mathbb{1}\{a=1\}}{g_P(1 \mid x)}\big\{y - \mu_{P,1}(x)\big\} + \partial_c m\big(\mu_{P,1}(x), \mu_{P,0}(x)\big)\,\frac{\mathbb{1}\{a=0\}}{g_P(0 \mid x)}\big\{y - \mu_{P,0}(x)\big\}\Bigg],$$

where $g_P(a \mid x) := P(A = a \mid X = x)$.

Example 2.2 (Longitudinal modified treatment policies)**.** Our framework also covers longitudinal parameters whose efficient influence operators are built from sequential regressions, such as the LMTP curve of Díaz et al. (2023). The definition and the verification of our conditions require substantial additional notation and are deferred to Appendix B.1.3.

2.4 Simulation

2.4.1 Simulation setup

We observe n i.i.d. copies of $Z = (X, A, Y)$ with $A \mid X \sim \mathrm{Bern}\{g_0(X)\}$ and $Y = \mu_0(X) + A\,\tau_0(X) + \varepsilon$, $\varepsilon \sim N(0, 1)$, where $X \in [-1, 1]^3$ is drawn from a Gaussian copula with equicorrelation $\rho = 0.5$: each coordinate is exactly $\mathrm{Unif}[-1, 1]$ but the coordinates are dependent, so that the additive and subset projections admit no closed form. We test $H_0 \colon \tau_0 \in \mathcal{S}$ for the three subspaces of Items (ii) to (iv), with $\tau_0 = \tau_{\mathrm{null}} + \delta\, c_\perp$: under each null a fixed $\tau_{\mathrm{null}} \in \mathcal{S}$

Table 2.1: Simulation nulls: CATEs under H_0, alternative directions, and the EIF guarantee of each path at $\rho = 0.5$.

				EIF guarantee	
$\mathcal{S}$	$\tau_{\text{null}}(x)$	smooth c	rough c	γ (RKHS)	Γ (L^2)
linear	$1 + x_1 - x_2 + 0.5x_3$	x_1x_2	$\cos(3\pi x_1)$	Thm 2.1	Thm 2.2
additive	$\sin(\pi x_1) + x_2^2 + 0.5x_3$	x_1x_2	$\cos(3\pi x_1)\, x_2$	—	Thm 2.2
subset, $V{=}\{1,2\}$	$\sin(\pi x_1) + x_2^2 + x_1x_2$	x_2x_3	$\cos(3\pi x_3)$	—	Thm 2.2

that violates the other two nulls, and alternatives along a smooth and a rough direction $c_\perp \in \mathcal{S}^\perp$, unit-normalized in $L^2(P_X^0)$ so that δ is the L^2 distance from τ_0 to $\mathcal{S}$. Exact forms of g_0, μ_0, the directions, and all implementation details are in Appendix B.10.

All cells share the augmented ANOVA kernel $\mathcal{K}^*(x, x') = \prod_{i=1}^3 \{1 + x_i x_i' + \exp(-(x_i - x_i')^2/2\sigma^2)\}$, $\sigma = 2 - \sqrt{2}$, whose RKHS contains the constant and coordinate functions. Estimation follows Section 2.2.3: two-fold cross-fitting, the spherical weighting and $\Omega_n = [(1-\lambda)\Sigma_n + \lambda I]^{-1}$ with $\lambda = 0.5$ (Section 4.2 and Appendix D of Luedtke & Chung, 2024), and thresholds from $B = 999$ multiplier bootstrap replicates at level 0.05. For the linear null we report both codomains; for the additive and subset nulls, where Theorem 2.1 provides no $\mathcal{H}$-valued efficient influence function, we report the Γ-path only, as our theory covers no other cells. The main grid crosses $n \in \{500, 1000, 2000, 4000\}$ and $\delta \in \{0, 0.25, 0.5, 0.75, 1\}$, with 1000 replications for size cells ($\delta = 0$) and 500 for power cells; rough-direction alternatives are extended to $\delta \in \{1.5, 2, 3, 5\}$ and $n = 8000$, with full results in Appendix B.10. A sensitivity analysis over $\lambda \in \{0.1, 0.9\}$ is reported there as well.

Table 2.2: Empirical size at $\delta = 0$ with estimated nuisances (cross-fitted propensity from SuperLearner, kernel ridge outcomes; 1000 replications, binomial standard errors in parentheses)

		γ (RKHS)		Γ (L^2)	
$\mathcal{S}$	n	sph.	Wald	sph.	Wald
linear	500	0.055 (0.007)	0.053 (0.007)	0.063 (0.008)	0.059 (0.007)
	1000	0.063 (0.008)	0.058 (0.007)	0.055 (0.007)	0.060 (0.008)
	2000	0.065 (0.008)	0.060 (0.008)	0.052 (0.007)	0.061 (0.008)
	4000	0.049 (0.007)	0.044 (0.006)	0.052 (0.007)	0.049 (0.007)
additive	500	—	—	0.051 (0.007)	0.055 (0.007)
	1000	—	—	0.058 (0.007)	0.059 (0.007)
	2000	—	—	0.052 (0.007)	0.058 (0.007)
	4000	—	—	0.063 (0.008)	0.055 (0.007)
subset	500	—	—	0.058 (0.007)	0.051 (0.007)
	1000	—	—	0.047 (0.007)	0.053 (0.007)
	2000	—	—	0.059 (0.007)	0.053 (0.007)
	4000	—	—	0.062 (0.008)	0.061 (0.008)

2.4.2 Simulation results

Table 2.2 reports empirical size with estimated nuisances. All cells lie within two binomial standard errors of the nominal level 0.05 across both codomains, both weightings, and all sample sizes, for each of the three nulls. Table 2.3 reports power, against the smooth alternative directions in the left panel and the rough directions in the right panel. The full grids are in Appendix B.10. Against smooth alternatives, power increases in both n and δ throughout, and every cell reaches power 1.00 by $\delta = 0.5$ at $n = 1000$. Two patterns are

Table 2.3: Power against the smooth and rough alternative directions with estimated nuisances (500 replications; Monte Carlo standard errors at most 0.02). Smooth alternatives are shown at $n \in \{500, 1000\}$, rough alternatives require larger signals and samples and are shown at $n \in \{4000, 8000\}$.

		Smooth				Rough			
		$\delta = 0.25$		$\delta = 0.5$		$\delta = 1$		$\delta = 5$	
Cell	Ω_n	n=500	1000	n=500	1000	n=4000	8000	n=4000	8000
linear, γ	sph.	0.51	0.88	0.98	1.00	0.76	1.00	1.00	1.00
	Wald	0.42	0.79	0.96	1.00	0.97	1.00	1.00	1.00
linear, Γ	sph.	0.50	0.87	0.98	1.00	0.57	0.89	0.97	1.00
	Wald	0.49	0.85	0.98	1.00	0.57	0.90	1.00	1.00
additive, Γ	sph.	0.35	0.73	0.87	1.00	0.06	0.04	0.07	0.05
	Wald	0.41	0.78	0.92	1.00	0.05	0.05	0.21	0.35
subset, Γ	sph.	0.14	0.38	0.61	0.98	0.10	0.09	0.13	0.27
	Wald	0.31	0.73	0.91	1.00	0.18	0.31	1.00	1.00

worth noting. First, the two codomains perform nearly identically for the linear null, consistent with the embedding interpretation of Section 2.2.2: the Γ-path only smooths the test directions, and against a smooth alternative little is lost. Second, the comparison between weightings depends on the null. For the linear null, the spherical statistic is slightly more powerful at the smallest signal (0.51 versus 0.42 at $\delta = 0.25$, $n = 500$, γ-path), whereas for the additive and, especially, the subset null the Wald-type weighting yields sizable gains (0.31 versus 0.14 in the subset cell). The alternative direction has low variance under Σ_0, which the regularized inverse upweights.

Against the rough directions, larger signals and samples are required, and both patterns are more pronounced. The codomains now separate: for the linear null, the γ-path retains substantially more power than the Γ-path (0.97 versus 0.57 for the Wald-type statistic at $\delta = 1$, $n = 4000$), reflecting that the L^2-embedding applies one additional layer of spectral damping to the test directions. The gap between weightings also widens: the Wald-type test attains power 1.00 by $\delta = 5$ in the subset cell while the spherical test does not exceed 0.27 even at $n = 8000$, and in the additive cell only the Wald-type statistic recovers as n grows. We return to the interpretation of this behavior, which reflects the spectral damping inherent to the kernel embedding, in Section 2.5.

2.5 Discussion

We proposed a nonparametric test of whether a pathwise differentiable function-valued parameter belongs to a closed linear subspace of $L^2(P_X^0)$, covering structural hypotheses such as additivity or dependence on a subset of covariates. The kernel embedding converts the projected parameter, which in general admits no efficient influence function, into one that does: the spectral decay of the kernel integral operator damps the test directions enough for the efficient influence operator to admit a representer. Any bounded universal kernel supplies this smoothing, and the validity of the test does not depend on which one is used, since universality guarantees the exact equivalence with H_0 (Sriperumbudur et al., 2011). Its power against a given alternative, however, does depend on the kernel, whose spectral profile determines how strongly rough alternatives are damped. The choice of kernel is therefore a choice of weighting over alternatives rather than a nuisance.

Chapter 3

ORTHOGONAL STATISTICAL LEARNING FROM TWO-PHASE SAMPLING

3.1 Introduction

Two-phase sampling arises when a measurement is too costly to collect on every unit. The design dates back to survey sampling, where an inexpensive auxiliary variable is recorded for all units and the survey variable for a subsample chosen by the investigator (Neyman, 1938). Two-stage case-control studies in epidemiology follow the same design, with an expensive exposure measurement made on an outcome-dependent subsample (Breslow & Cain, 1988; White, 1982). Recent work revives this design for machine-learning workflows, where the expensive measurement is a label and the proxy a model prediction: prediction-powered inference corrects inference for finite-dimensional parameters defined by convex minimization, using predictions on a large unlabeled sample (Angelopoulos et al., 2023, 2024), and subsequent work collects the labels adaptively (Kluger & Bates, 2026; Zrnic & Candès, 2024). These works concern finite-dimensional parameters. More recently, Sundberg et al. (2026) study conditional average treatment effects when the treatment is missing at random.

Separately, nonparametric estimation of function-valued parameters has grown rapidly. Such parameters typically do not admit efficient influence functions, so one-step estimation and targeted learning (Newey, 1994; Pfanzagl, 1982; van der Laan & Rubin, 2006) do not apply. One remedy is to recast estimation as risk minimization whose minimizer is robust to nuisance estimation error. Examples include the dose-response function (Bonvini & Kennedy, 2022; Díaz & van der Laan, 2013), conditional average treatment effects (Nie & Wager, 2021), and counterfactual densities (Kennedy, 2023). Foster and Syrgkanis (2023) give the general theory of orthogonal statistical learning, which makes the target estimation

insensitive to first-order nuisance error, so that flexible machine-learning nuisance estimators can be plugged in. Although two-phase designs have been combined with causal targets for scalar functionals, no existing work estimates a function-valued parameter whose argument is the expensive measurement.

In this paper, we make the following contributions.

1. We construct a corrected loss, computable from the full two-phase sample, that augments a given Neyman-orthogonal base loss in the style of augmented inverse probability weighting.

2. We show that the corrected loss inherits the regularity conditions of the base loss with explicit constants, and that the corrected risk is doubly robust in the two added nuisances.

3. We bound the excess risk of the empirical risk minimizer and show that nuisance estimation error can be absorbed into the target estimation error.

4. We illustrate the theory on MIMIC-IV data, estimating a dose-response function with ejection fraction as the expensive phase-2 variable and a chest radiograph embedding as the phase-1 variable.

3.2 Problem formulation

Following the two-phase sampling literature (Breslow & Wellner, 2007), variables recorded for every unit belong to phase 1 and the expensive measurement to phase 2. Let $W \in \mathcal{W}$ be the phase-1 variable, recorded for every unit; let $X \in \mathbb{R}^d$ collect structured covariates; and let $Y \in \mathbb{R}$ be an outcome, both also observed for everyone. The variable of interest, $V \in \mathcal{V} \subseteq \mathbb{R}$, is the argument at which the target function is evaluated. Obtaining it requires a costly gold-standard measurement, so it is ascertained for only a small fraction of units. The observed data are therefore $Z := (W, X, Y, S, SV)$, where the product SV encodes

that V is available only when $S = 1$. We observe $Z_1, \ldots, Z_n$ drawn i.i.d. from a distribution $P^0 \in \mathcal{P}$, where $\mathcal{P}$ is a statistical model specified in Section 3.2.1.

For a distribution $P \in \mathcal{P}$ and a measurable function h we write $Ph := \int h\, dP$, and $\mathbb{P}_n$ denotes the empirical distribution of $Z_1, \ldots, Z_n$, so that $\mathbb{P}_n h = n^{-1} \sum_{i=1}^n h(Z_i)$. Subscripts on P denote marginal and conditional laws. Let $L^2(P)$ be the space of square-integrable functions with inner product $\langle h_1, h_2 \rangle_P = P(h_1 h_2)$ and norm $\|h\|_P = (Ph^2)^{1/2}$, and let $L^2_0(P)$ be its subspace of mean-zero functions. We write $\|h\|_{L^\infty(P)} = \operatorname{ess\,sup} |h|$ for the supremum norm. Finally, $a \lesssim b$ means $a \le Cb$ for a constant C not depending on n, and $a \asymp b$ means $a \lesssim b \lesssim a$.

In our sampling design, given (W, X, Y) the indicator $S \in \{0, 1\}$ is drawn by exogenous randomization with probability

$$\rho_0(W, X, Y) \;:=\; P^0\big(S = 1 \,\big|\, W, X, Y\big),$$

without reference to V. Two properties then hold by design. First, the gold-standard measurement is missing at random:

$$S \perp\!\!\!\perp V \,\big|\, (W, X, Y). \tag{3.1}$$

Second, every unit retains a nonvanishing chance of ascertainment, $\rho_0 > 0$; since the investigator controls the sampling mechanism, this positivity condition can be enforced at the design stage.

3.2.1 Objective

We consider a function-valued parameter $\theta : \mathcal{P} \to L^2(P_V)$ and aim to estimate its value $\theta_0 := \theta(P^0)$ at the true distribution, searching over a convex class $\Theta \subset L^2(P_V)$ with $\sup_{\theta \in \Theta} \|\theta\|_{L^\infty(P_V)} \le 1$. The parameter θ_0 is not assumed to admit an efficient influence function, so neither a one-step correction (Newey, 1994; Pfanzagl, 1982) nor targeted learning (van der Laan & Rubin, 2006) is available. We instead characterize θ_0 as a risk minimizer.

Since the risk generally involves nuisance parameters, we consider a Neyman-orthogonal loss (Foster & Syrgkanis, 2023), a loss whose risk has vanishing cross-derivative at the class

minimizer and the true nuisance. Let $\ell(\theta, g; (v, x, y))$ be such a loss, with nuisance parameter $g = (g_1, \ldots, g_d) \in \mathcal{G}$, where each component g_j is a real-valued function of a subvector of (v, x); by trivial extension we regard g as a map $\mathcal{V} \times \mathcal{X} \to \mathbb{R}^d$, and equip $\mathcal{G}$ with a pre-norm $\|\cdot\|_{\mathcal{G}}$. Write $L_P(\theta, g) := \mathrm{E}_P[\ell(\theta, g; (V, X, Y))]$ for the associated risk. We require that the loss identify the target at the truth,

$$\theta_0 \;=\; \operatorname*{arg\,min}_{\theta \in L^2(P_V)} L_{P^0}(\theta, g_0),$$

where $g_0 \in \mathcal{G}$ denotes the true nuisance value, and we denote by $\theta^* \in \arg\min_{\theta \in \Theta} L_{P^0}(\theta, g_0)$ a minimizer over the class. Such losses are readily available in examples: the doubly robust pseudo-outcome construction of Bonvini and Kennedy (2022) yields one for the dose-response function, and the R-learner loss does so for the conditional average treatment effect (Nie & Wager, 2021).

Under the two-phase sampling design, $\ell(\theta, g; (v, x, y))$ is not directly usable: evaluating it requires the phase-2 variable V and is therefore possible only on $\{S = 1\}$. Our correction, constructed below, recovers the risk from the full sample; it requires the conditional law of V given the observed data and the sampling probability to be well behaved, so we first delimit the model over which the guarantees are stated. Fix a finite measure μ on $\mathcal{V}$ and define

$$\begin{aligned} \mathcal{P} \;:=\; \Big\{ P : P_{V,X,Y} \in \mathcal{P}_{\text{base}};\;\; S \perp\!\!\!\perp V \,\Big|\, (W, X, Y) \text{ under } P; \\ P_{V|W,X,Y} \ll \mu \text{ with density } c_0 \le q_P \le C;\;\; P(S{=}1 \mid W, X, Y) \ge \rho_{\min} \Big\}, \end{aligned} \tag{3.2}$$

where $\mathcal{P}_{\text{base}}$ collects the laws of (V, X, Y) under which the base loss is well-defined, and $c_0 > 0, C < \infty$ and $\rho_{\min} > 0$ are fixed constants. We assume $P^0 \in \mathcal{P}$ throughout and write $q_0 := q_{P^0}$; the true sampling probability ρ_0 was defined in Section 3.2. The conditions on q_P and on the sampling probability restrict P^0 and are the two-phase analogues of the overlap conditions familiar from treatment effect estimation.

We now construct the corrected loss. On units with $S = 0$ the loss is not computable, but a surrogate for its conditional mean given the observed data is: with a working conditional density q drawn from the class $\mathcal{Q} := \{q : c_0 \le q \le C\}$ of densities with respect to μ, and a

working sampling probability ρ drawn from $\mathcal{R} := \{\rho : \rho_{\min} \le \rho \le 1\}$, define

$$\begin{aligned}\tilde{\ell}(\theta, g, \rho, q;\, z) := \frac{s}{\rho(w,x,y)}\,\ell(\theta, g;\,(v,x,y)) \\ + \left(1 - \frac{s}{\rho(w,x,y)}\right)\int \ell(\theta, g;\,(v',x,y))\; q(v' \mid w,x,y)\,\mu(dv'),\end{aligned} \tag{3.3}$$

where the first term vanishes when $s = 0$ by the encoding of Z, and the integral in the second term is computable for every unit. The restrictions on $\mathcal{Q}$ and $\mathcal{R}$ entail no loss of generality and render the density ratio and the inverse weight uniformly bounded, $q/q_0 \le C/c_0$ and $S/\rho \le \rho_{\min}^{-1}$. By disintegration, at $q = q_0$ the integral equals $\mathrm{E}_{P^0}[\ell(\theta, g;\,(V,X,Y)) \mid W,X,Y]$, the natural choice of augmentation, and $\mathbb{P}_n\tilde{\ell}(\theta, g, \rho_0, q_0)$ is precisely the design-based AIPW estimator of the risk $L_{P^0}(\theta, g)$ under nonparametric $\mathcal{P}$ (Robins et al., 1994; Tsiatis, 2006). The first term reweights the gold subsample so that it represents the target population; the second is an augmentation term, leaving the risk unchanged while reducing its variance.

We may view $\tilde{\ell}$ as a new loss with the enlarged nuisance parameter $\eta := (g, \rho, q)$, with risk $\tilde{L}_P(\theta, \eta) := \mathrm{E}_{Z\sim P}[\tilde{\ell}(\theta, \eta;\, Z)]$. The two auxiliary nuisances play asymmetric roles. At the true sampling probability the risk coincides with the base risk: for every $\theta \in \Theta$, $g \in \mathcal{G}$, and $q \in \mathcal{Q}$,

$$\tilde{L}_{P^0}\big(\theta, g, \rho_0, q\big) \;=\; L_{P^0}(\theta, g),$$

so the correction shifts neither the target minimizer nor the identification of θ_0, and misspecifying q alone only costs efficiency.

3.2.2 Estimation algorithms

We propose a two-stage estimation procedure based on sample splitting: the first subsample is used to estimate the nuisance parameters, and the second to estimate the target. Sample splitting of this kind is standard in orthogonal statistical learning: it renders the estimated nuisance independent of the target-stage sample, so that $\widehat{\eta}$ may be treated as fixed in the analysis of the second stage (Foster & Syrgkanis, 2023; Nie & Wager, 2021). Under the corrected loss the natural target-stage learner is empirical risk minimization over Θ; we state the procedure below, leaving the nuisance estimation methods unspecified.

Algorithm 3.1 ERM over a function class Θ under two-phase sampling

1: **Input:** data $z^n = (z_1, \ldots, z_n)$; partition z^n into two disjoint subsamples z_1^n and z_2^n.

2: **Nuisance estimation:** using only z_1^n, construct a nuisance estimator $\widehat{\eta} = (\widehat{g}, \widehat{\rho}, \widehat{q})$.

3: **Target estimation:** find the empirical risk minimizer over Θ using only z_2^n:

$$\widehat{\theta} \in \arg\min_{\theta \in \Theta} \mathbb{P}_n^{(2)} \big[\tilde{\ell}(\theta, \widehat{\eta}; \cdot)\big],$$

where $\mathbb{P}_n^{(2)}$ denotes the empirical distribution of z_2^n.

In practice, cross-fitting allows every observation to enter the target stage: split $[n]$ into K folds, construct $\widehat{\eta}^{(-k)}$ on the complement of fold I_k, and minimize the aggregated objective $\theta \mapsto n^{-1} \sum_{k=1}^{K} \sum_{i \in I_k} \tilde{\ell}(\theta, \widehat{\eta}^{(-k)}; z_i)$. This avoids discarding z_1^n at the target stage, while preserving the independence between each $\widehat{\eta}^{(-k)}$ and the fold on which it is evaluated. Our guarantees are stated for a single split to keep the analysis simple; standard arguments extend them to the cross-fitted version.

Algorithm 3.1 leaves nuisance estimation flexible: the guarantees in Section 3.3 require only that $\widehat{\eta}$ be constructed on samples disjoint from those of the target stage. For components of g taking v as an argument, the user has a choice. The simplest option is to fit them on the gold subsample alone unless this incurs a complete case bias, which together with the reduced sample size enters the bound of Theorem 3.1 through $\|\widehat{g} - g_0\|_{\mathcal{G}}$ and may be acceptable when the design renders it small. The alternative applies the correction (3.3) to the loss identifying each such component, just as at the target stage. In that case the sample is divided into three parts: the first estimates $\widehat{\rho}$ and $\widehat{q}$, neither of which requires a correction, since ρ_0 is a regression of the binary S on fully observed data and q_0 is identified from the gold subsample by (3.1); the second estimates the components involving v from the corrected losses; their union estimates the remaining components; and the third is reserved for the target stage.

3.3 Theoretical guarantees

3.3.1 Neyman orthogonality

We collect here the conditions under which the results of this section are stated. The first is imposed on the base loss and is inherited from the orthogonal statistical learning framework of Foster and Syrgkanis (2023); the second is specific to the corrected loss. Throughout, derivatives are directional (Gâteaux) derivatives along line segments, and $\mathrm{star}(\Theta, \theta^*) := \{t\theta + (1-t)\theta^* : \theta \in \Theta,\, t \in [0,1]\}$.

Condition 3.1 (Base loss regularity). There exist constants $\lambda, \kappa, \beta_1, \beta_2 \geq 0$ and $r \in [0,1)$ such that, for all $\theta, \bar{\theta} \in \Theta$, $g \in \mathcal{G}$, and $\bar{g} \in \mathrm{star}(\mathcal{G}, g_0)$:

(a) (Neyman orthogonality) $D_g D_\theta L_{P^0}(\theta^*, g_0)[\theta - \theta^*,\, g - g_0] = 0$;

(b) (First-order optimality) $D_\theta L_{P^0}(\theta^*, g_0)[\theta - \theta^*] \geq 0$;

(c) (Smoothness in the target) $D_\theta^2 L_{P^0}(\bar{\theta}, g_0)[\theta - \theta^*, \theta - \theta^*] \leq \beta_1 \left\|\theta - \theta^*\right\|_\Theta^2$;

(d) (Higher-order smoothness)

$$\left|D_g^2 D_\theta L_{P^0}(\theta^*, \bar{g})[\theta - \theta^*,\, g - g_0,\, g - g_0]\right| \leq \beta_2 \left\|\theta - \theta^*\right\|_\Theta^{1-r} \left\|g - g_0\right\|_\mathcal{G}^2;$$

(e) (Strong convexity) $D_\theta^2 L_{P^0}(\bar{\theta}, g)[\theta - \theta^*, \theta - \theta^*] \geq \lambda \left\|\theta - \theta^*\right\|_\Theta^2 - \kappa \left\|g - g_0\right\|_\mathcal{G}^{4/(1+r)}$.

Condition 3.1 comprises the regularity conditions under which Theorem 1 of Foster and Syrgkanis (2023) bounds the excess risk by the sum of a target estimation error and a nuisance estimation error. Conditions 3.1(a) to 3.1(c) involve derivatives in θ alone and transfer to the corrected risk with unchanged constants through the identity $\tilde{L}_{P^0}(\cdot, g, \rho_0, q) = L_{P^0}(\cdot, g)$ for every g and q. Conditions 3.1(d) and 3.1(e) concern nuisance directions and must be re-established for the enlarged nuisance, which is where the following moment conditions enter. Throughout, c_0, C, and $\rho_{\min}$ denote the constants fixed by the model (3.2) and the classes $\mathcal{Q}$ and $\mathcal{R}$. The requirement $q_P \geq c_0$ constrains the choice of μ: it holds, for instance,

when $\mathcal{V}$ is compact, μ is Lebesgue measure, and the conditional density of V is bounded away from zero.

Condition 3.2 (Moment bounds on loss derivatives). There exist constants $\beta_3, \beta_4 \geq 0$ such that, for all $\theta \in \Theta$, $g \in \mathcal{G}$, and $\bar{\theta} \in \mathrm{star}(\Theta, \theta^*)$:

(a) (Hessian moment) $\mathrm{E}_{P^0}\big[\big(D_\theta^2 \ell(\bar{\theta}, g;\, Z)[\theta - \theta^*, \theta - \theta^*]\big)^2\big]^{1/2} \leq \beta_3 \, \|\theta - \theta^*\|_\Theta$;

(b) (Score moment) $\mathrm{E}_{P^0}\big[\big(D_\theta \ell(\theta^*, g;\, Z)[\theta - \theta^*]\big)^2\big]^{1/2} \leq \beta_4 \, \|\theta - \theta^*\|_\Theta$.

Condition 3.2 is the price of the correction. The corrected loss couples the θ-derivatives of the base loss with the multiplicative random weight S/ρ, so risk-level bounds such as Condition 3.1(c) no longer suffice: separating the weight from the derivative by the Cauchy–Schwarz inequality requires second moments of the pointwise derivatives themselves. The condition is mild for squared-error losses, including both examples of Section 3.2.1, where the class bound $\sup_{\theta \in \Theta} \|\theta\|_{L^\infty(P_V)} \leq 1$, bounded outcomes, and overlap for the base problem yield explicit constants.

Condition 3.1 and Condition 3.2 together imply that the corrected loss satisfies the corresponding conditions with respect to the enlarged nuisance (Lemma 3.1 below). The excess risk upper bound is further quantified by two error measures: for $\rho \in \mathcal{R}$ and $q \in \mathcal{Q}$, define

$$e_\rho(\rho) \;:=\; \mathrm{E}_{P^0}\Big[\Big(\frac{\rho - \rho_0}{\rho}\Big)^4\Big]^{1/4}, \qquad e_q(q) \;:=\; \mathrm{E}_{P^0}\big[\chi_q(W, X, Y)^4\big]^{1/4},$$

where $\chi_q(w, x, y)^2 := \int \big(q - q_0\big)^2(v' \mid w, x, y)\, q_0(v' \mid w, x, y)^{-1}\, \mu(dv')$ is the conditional χ^2-divergence between q and q_0. Both measures are bounded since $e_\rho(\rho) \leq \rho_{\min}^{-1}$ and $e_q(q)^2 \leq \mu(\mathcal{V})\, C^2/c_0$ by the model and class bounds and both vanish at the truth.

Lemma 3.1 (Regularity conditions of corrected loss)**.** *Suppose Condition 3.1 and Condition 3.2 hold. Then for all $\theta \in \Theta$, $\eta = (g, \rho, q) \in \mathcal{G} \times \mathcal{R} \times \mathcal{Q}$, and $\bar{\theta} \in \mathrm{star}(\Theta, \theta^*)$:*

(a) (Neyman orthogonality) $D_\eta D_\theta \tilde{L}_{P^0}(\theta^*, \eta_0)[\theta - \theta^*,\, \eta - \eta_0] = 0$;

(b) (First-order optimality) $D_\theta \tilde{L}_{P^0}(\theta^*, \eta_0)[\theta - \theta^*] \geq 0$;

(c) (Smoothness in the target) $D_\theta^2 \tilde{L}_{P^0}(\bar{\theta}, \eta_0)[\theta - \theta^*, \theta - \theta^*] \le \beta_1 \|\theta - \theta^*\|_\Theta^2$;

(d) (First-order bias) $\left| D_\theta \tilde{L}_{P^0}(\theta^*, g, \rho, q)[\theta - \theta^*] - D_\theta L_{P^0}(\theta^*, g)[\theta - \theta^*] \right| \le \beta_4 \, e_\rho(\rho) \, e_q(q) \, \|\theta - \theta^*\|_\Theta$;

(e) (Strong convexity) $D_\theta^2 \tilde{L}_{P^0}(\bar{\theta}, \eta)[\theta - \theta^*, \theta - \theta^*] \ge \frac{3\lambda}{4} \|\theta - \theta^*\|_\Theta^2 - \kappa \, \|g - g_0\|_{\mathcal{G}}^{4/(1+r)} - \frac{\beta_3^2}{\lambda} \, e_\rho(\rho)^2 \, e_q(q)^2$.

Lemmas 3.1(a) to 3.1(c) hold with the constants of Condition 3.1 unchanged: the identity $\tilde{L}_{P^0}(\cdot, g, \rho_0, q) = L_{P^0}(\cdot, g)$ holds as an equality of functions on Θ for every g and q, so derivatives along directions preserved by the identity coincide for the two risks. Lemma 3.1(d) plays, for the auxiliary directions, the role that Condition 3.1(d) plays for g: in the proof of the main theorem, the term through which nuisance error enters is $D_\theta \tilde{L}_{P^0}(\theta^*, \widehat{\eta})[\widehat{\theta} - \theta^*]$, and its deviation from the base counterpart is exactly bilinear in $(\rho - \rho_0, \, q - q_0)$, so no Taylor expansion in the auxiliary nuisances is needed and the price is the product $e_\rho \, e_q$ rather than a squared joint error; the g-direction is then handled by Conditions 3.1(a) and 3.1(d) at the base level. In Lemma 3.1(e), the curvature constant degrades from λ to $3\lambda/4$; the lost quarter absorbs, via Young's inequality, the Hessian perturbation induced by the random weight. The coefficient κ on the g-error is unchanged, and the auxiliary nuisances enter only through the additive doubly robust product $e_\rho^2 \, e_q^2$.

Lemma 3.2 (Oracle excess risk bound in terms of target and nuisance errors)**.** *Suppose Condition 3.1 and Condition 3.2 hold. If $\widehat{\theta}$ and $\widehat{\eta} = (\widehat{g}, \widehat{\rho}, \widehat{q})$ are the target and nuisance estimators from Algorithm 3.1, then*

$$\left\|\widehat{\theta} - \theta^*\right\|_\Theta^2 \le \frac{4}{\lambda}\Big(\tilde{L}_{P^0}(\widehat{\theta}, \widehat{\eta}) - \tilde{L}_{P^0}(\theta^*, \widehat{\eta})\Big) + C_g \, \|\widehat{g} - g_0\|_{\mathcal{G}}^{4/(1+r)} + C_{\rho q} \, e_\rho(\widehat{\rho})^2 \, e_q(\widehat{q})^2, \quad (3.4)$$

where C_g and $C_{\rho q}$ depend only on $(\lambda, \kappa, \beta_2, \beta_3, \beta_4, r)$; for $r = 0$ one may take $C_g = \frac{8\beta_2^2}{\lambda^2} + \frac{2\kappa}{\lambda}$ and $C_{\rho q} = \frac{8\beta_3^2 + 16\beta_4^2}{\lambda^2}$. Consequently,

$$L_{P^0}(\widehat{\theta}, g_0) - L_{P^0}(\theta^*, g_0) \le \frac{\beta_1}{2}\left[\frac{4}{\lambda}\Big(\tilde{L}_{P^0}(\widehat{\theta}, \widehat{\eta}) - \tilde{L}_{P^0}(\theta^*, \widehat{\eta})\Big) + C_g \, \|\widehat{g} - g_0\|_{\mathcal{G}}^{4/(1+r)} + C_{\rho q} \, e_\rho(\widehat{\rho})^2 \, e_q(\widehat{q})^2\right]. \quad (3.5)$$

Lemma 3.2 decomposes the oracle excess risk into three parts: the excess of the corrected risk at the estimated nuisances, which is the quantity the target stage controls, base nuisance error $\|\widehat{g} - g_0\|_{\mathcal{G}}^{4/(1+r)}$; and the doubly robust product $e_\rho^2\, e_q^2$. The first two terms reproduce the decomposition of Theorem 1 of Foster and Syrgkanis (2023) for the base problem: the price of two-phase sampling is confined to the third term. Its product form means the price vanishes when either auxiliary nuisance is estimated consistently, and permits a trade-off between the rates of $\widehat{\rho}$ and $\widehat{q}$. The bound is deterministic and requires no consistency of any nuisance estimator. The terms simply quantify the cost of whatever errors remain.

3.3.2 Estimation performance

Lemma 3.2 holds for an arbitrary pair $(\widehat{\theta}, \widehat{\eta})$, and the only stochastic quantity on its right-hand side is the corrected excess risk at the estimated nuisance, $\tilde{L}_{P^0}(\widehat{\theta}, \widehat{\eta}) - \tilde{L}_{P^0}(\theta^*, \widehat{\eta})$. This subsection bounds that quantity for the empirical risk minimizer of Algorithm 3.1, through the localized analysis of Foster and Syrgkanis (2023) carried out on the corrected loss class. The analysis requires that the variance of a loss difference be bounded by a multiple of $\|\theta - \theta^*\|_\Theta^2$. For the base loss this is immediate when the loss is Lipschitz in $\theta(v)$. A corrected loss difference carries the weight S/ρ and an integral of the base difference over the phase-2 variable, so the bound does not follow from the Lipschitz property alone; it is proved in the appendix from the following condition together with the bounds defining the model (3.2).

Condition 3.3 (Lipschitz base loss). There exists $B < \infty$ such that, for all $\theta, \theta' \in \Theta$, $g \in \mathcal{G}$, and P^0-almost every (v, x, y),

$$\big|\ell(\theta, g;\, (v, x, y)) - \ell(\theta', g;\, (v, x, y))\big| \;\le\; B\,\big|\theta(v) - \theta'(v)\big|.$$

The condition converts a difference of losses into a pointwise difference of the two functions. In the proof of Theorem 3.1, this bounds the supremum norm and the variance of the corrected loss differences, the two quantities through which Talagrand's concentration inequality controls the deviation of the empirical risk.

Now fix $\eta \in \mathcal{G} \times \mathcal{R} \times \mathcal{Q}$ and write $f_\theta := \tilde{\ell}(\theta, \eta; \cdot) - \tilde{\ell}(\theta^*, \eta; \cdot)$. The corrected loss-difference class and its localization at radius $\delta > 0$ are

$$\tilde{\mathcal{F}}_\eta := \{f_\theta : \theta \in \Theta\}, \qquad \tilde{\mathcal{F}}_\eta(\delta) := \{f_\theta : \|\theta - \theta^*\|_\Theta \le \delta\},$$

and $\mathfrak{R}_n(\mathcal{F}) := \mathrm{E}[\sup_{f \in \mathcal{F}} |n^{-1} \sum_{i=1}^n \epsilon_i f(Z_i)|]$ denotes the Rademacher complexity of a class $\mathcal{F}$, with $\epsilon_1, \ldots, \epsilon_n$ independent Rademacher variables. A *sub-root* function is a nondecreasing $\psi : (0, \infty) \to [0, \infty)$ with $\delta \mapsto \psi(\delta)/\delta$ nonincreasing, and its critical radius is $r_\psi := \inf\{\delta > 0 : \psi(\delta) \le \delta^2/\tilde{b}\}$, where $\tilde{b} := 4B/\rho_{\min}$.

Theorem 3.1 (Excess risk of the corrected empirical risk minimizer)**.** *Suppose Conditions 3.1 to 3.3 hold. Fix $\delta \in (0, 1/2]$. Let $\widehat{\eta}, \widehat{\theta}$ be the estimators from Algorithm 3.1. Let ψ be any sub-root function with*

$$\psi(u) \;\ge\; \mathfrak{R}_n\big(\mathrm{star}(\tilde{\mathcal{F}}_{\widehat{\eta}}(u), 0)\big) \qquad \textit{for all } u > 0,$$

and let r_ψ be its critical radius. Then there is a constant c depending only on $(B, \lambda, \rho_{\min}, C/c_0)$ such that, with probability at least $1 - \delta$,

$$\big\|\widehat{\theta} - \theta^*\big\|_\Theta^2 \;\le\; c\Big(r_\psi^2 + \frac{\log(1/\delta)}{n}\Big) \;+\; 2C_g \,\|\widehat{g} - g_0\|_{\mathcal{G}}^{4/(1+r)} \;+\; 2C_{\rho q}\, e_\rho(\widehat{\rho})^2\, e_q(\widehat{q})^2,$$

with $C_g, C_{\rho q}$ as in Lemma 3.2, and the oracle excess risk $L_{P^0}(\widehat{\theta}, g_0) - L_{P^0}(\theta^*, g_0)$ is bounded by $\beta_1/2$ times the same right-hand side.

The critical radius is where the two-phase design enters the bound. Write $\Delta_\theta := \ell(\theta, \widehat{g}; \cdot) - \ell(\theta^*, \widehat{g}; \cdot)$ for a base loss difference; every element of $\tilde{\mathcal{F}}_{\widehat{\eta}}$ decomposes as

$$f_\theta \;=\; \Delta_\theta \;+\; \Big(\frac{S}{\widehat{\rho}} - 1\Big)\Big(\Delta_\theta - \int \Delta_\theta(v', X, Y)\, \widehat{q}(v' \mid W, X, Y)\, \mu(dv')\Big),$$

so the localized complexity of $\tilde{\mathcal{F}}_{\widehat{\eta}}$ is at most that of the base class $\{\Delta_\theta\}$, the quantity governing Theorem 3 of Foster and Syrgkanis (2023), plus that of the weighted residual class, and r_ψ is bounded by the sum of the two critical radii. The second radius carries the whole dependence on the design. At $\rho_0 \equiv 1$ the weight vanishes identically, the residual class

is $\{0\}$, and Theorem 3.1 reduces to the guarantee for the base problem. As the sampling budget $\bar{\rho} := \mathrm{E}_{P^0}[\rho_0]$ shrinks, the second radius takes over, and its scale is set by an identity: at the design sampling probability, and with the conditional mean $\mathrm{E}_{P^0}[\Delta_\theta \mid W, X, Y]$ as the augmentation,

$$\mathrm{E}_{P^0}\big[f_\theta(Z)^2\big] \;=\; \mathrm{E}_{P^0}\big[\Delta_\theta(V,X,Y)^2\big] \;+\; \mathrm{E}_{P^0}\Big[\frac{1-\rho_0}{\rho_0}\,\mathrm{Var}\big(\Delta_\theta \,\big|\, W,X,Y\big)\Big],$$

so the price of missingness attaches not to the loss but to the share of its variance that the observed data do not explain, inflated by the inverse budget. When the localized complexity of a class is a function of its $L^2(P^0)$ metric alone, inflating the variance of a loss class by a factor is equivalent to shrinking the sample by that factor: for a class spanned by d functions the squared critical radius scales as d/n, and for a Hölder ball of smoothness $s > 1/2$ as $n^{-2s/(2s+1)}$. The nuisance terms of Theorem 3.1 enter only through the squared product $e_\rho(\widehat{\rho})^2\, e_q(\widehat{q})^2$ and the power $\|\widehat{g} - g_0\|_{\mathcal{G}}^{4/(1+r)}$, so whenever the nuisance errors decay faster than $n^{-1/4}$ these terms are of lower order and the target term r_ψ^2 governs the bound. The critical radius of the corrected loss then replaces n by an effective sample size. Since $f_{\theta^*} = 0$, the localization uses only the norms $\|f_\theta\|_{P^0}$, which the identity above computes exactly; writing γ_ρ for the supremum over $\theta \neq \theta^*$ of its second term relative to its first,

$$\gamma_\rho \;:=\; \sup_{\theta\in\Theta,\,\theta\neq\theta^*} \frac{\mathrm{E}_{P^0}\big[\{(1-\rho_0)/\rho_0\}\,\mathrm{Var}\big(\Delta_\theta \,\big|\, W,X,Y\big)\big]}{\mathrm{E}_{P^0}\big[\Delta_\theta(V,X,Y)^2\big]},$$

the corrected class carries the noise of the base class inflated by $1+\gamma_\rho$, and the bound behaves as that of the base problem at the effective sample size $n_{\mathrm{eff}} = n/(1+\gamma_\rho)$. The design enters γ_ρ through the average of $(1-\rho_0)/\rho_0$ weighted by where the unexplained variance sits, not through its floor. Two special cases factor the design out. Bounding $(1-\rho_0)/\rho_0$ by its worst case gives $\gamma_\rho \le \gamma\,(1-\rho_{\min})/\rho_{\min}$, where $\gamma := \sup_{\theta\neq\theta^*} \mathrm{E}[\mathrm{Var}(\Delta_\theta \,|\, W,X,Y)]/\mathrm{E}[\Delta_\theta^2] \in [0,1]$ is the largest share of the variance of a base loss difference that the observed data do not explain. For a constant design $\rho_0 \equiv \bar{\rho}$ the factorization is exact and

$$n_{\mathrm{eff}} \;=\; \frac{n\,\bar{\rho}}{\bar{\rho} + \gamma\,(1-\bar{\rho})},$$

equal to $n\bar{\rho}$ for a phase-1 variable carrying no information about the phase-2 variable and to n for one that determines it. Beyond such target classes one can sharpen the excess risk bound by using localization in the weighted metric of the correction term.

3.4 Data illustration: an LVEF-mortality curve from MIMIC-IV chest radiographs

We illustrate the estimator on clinical data in which the phase-2 variable is expensive to measure and an image phase-1 variable is available for every unit: echocardiographic left ventricular ejection fraction (LVEF) paired with chest radiographs in MIMIC-IV. The goal of this section is to examine what the estimator recovers under a two-phase design in which the missingness of LVEF is deliberately controlled. Since LVEF is continuous, the target is the dose-response function, for which orthogonal losses are well developed (Bonvini & Kennedy, 2022; Díaz & van der Laan, 2013). We take the loss of Bonvini and Kennedy (2022) as our base loss and extend it to two-phase sampling through the corrected loss (3.3).

3.4.1 Data

We construct a cohort from MIMIC-IV and MIMIC-CXR (Johnson et al., 2019, 2023). Echocardiographic left ventricular ejection fraction (LVEF) measurements are paired with the nearest chest radiograph of the same subject within ±3 days; keeping each subject's earliest paired study yields $n = 13{,}537$ subjects, one row each. The phase-2 variable is $V = \text{LVEF}$, rescaled to $[0, 1]$; the phase-1 variable W is the 1,376-dimensional embedding of the paired radiograph from a pretrained chest X-ray foundation model (Sellergren et al., 2023), reduced to its leading 200 principal components; the covariates X collect age, sex, nine comorbidity indicators derived from pre-radiograph hospitalizations (Charlson categories mapped from ICD codes; Quan et al., 2005), insurance, and prior-admission count ($d_x = 14$); and the outcome Y is one-year all-cause mortality (prevalence 23.7%). The phase-1 variable is modestly informative: regularized regression of V on (W, X) attains an out-of-fold R^2 of 0.16–0.19. Our theory does not require the phase-1 variable to be informative, but the practical value of

the correction depends on how much of the variance of the loss the phase-1 variable explains. This illustration examines whether a phase-1 variable of this modest strength already yields gains over weighting alone.

3.4.2 Design

A gold-standard measurement of V is available for every subject. We evaluate two-phase procedures against a full-data benchmark by artificial masking, in the spirit of semi-synthetic evaluations. Let $\hat{\theta}_{\text{full}}$ denote the estimator from full data. This estimator serves as the fixed reference curve against which all two-phase estimators are compared. We then generate sampling indicators $S_i \sim \text{Bernoulli}\{\rho_0(W_i, Y_i)\}$ with $\operatorname{logit} \rho_0 = \alpha_0 + 0.8\,(u^\top W) + 0.5\,Y$, where u is the leading singular direction of the centered embedding matrix. The intercept is solved so that the cohort mean of ρ_0 equals a target sampling budget $\bar{\rho}$, clipped below at 0.02. The design is missing at random by construction, $S \perp\!\!\!\perp V \mid (W, X, Y)$, and is outcome-dependent, so unweighted complete-case analysis is inconsistent. The true ρ_0 is withheld from all estimators except a marked known-design benchmark.

For each budget $\bar{\rho} \in \{0.05, 0.10, 0.15, 0.30, 0.50\}$ we draw 500 sampling replications. Each replication re-splits the sample for cross-fitting and re-estimates every nuisance. The proposed estimator is Algorithm 3.1 applied verbatim. The sampling probability $\hat{\rho}$ is fit by ℓ_1-penalized logistic regression of S on (W, X, Y) with penalty chosen by cross-validation. The auxiliary density $\hat{q}$ is a homoscedastic Gaussian working model, $\mathcal{N}\{\hat{\mu}_V(W, X, Y),\, \hat{\sigma}^2\}$, where the location $\hat{\mu}_V$ is a gradient-boosted regression of V on (W, X, Y) fit on the gold subsample and $\hat{\sigma}^2$ is the pooled residual variance; the resulting density is truncated to $[c_0, C]$ and renormalized so that $\hat{q} \in \mathcal{Q}$. The outcome regression $\hat{\mu}(x, v)$ is fit by the corrected weighted least squares on the basis $[\, B(v) \mid x \mid v\, x\,]$, where B is a cubic B-spline basis with interior knots at $\{0.2, 0.4, 0.6, 0.8\}$; the marginalized components $\hat{\pi}(v \mid x)$, $\hat{\varpi}(v)$, and $\hat{m}(v)$ are obtained by averaging $\hat{q}$ and $\hat{\mu}$ over the sample, with population averages replaced by empirical averages over the nuisance-stage data. The target class is the fixed sieve $\{\theta = B\beta\}$ on the same spline basis, identical across all estimators. Nuisances are estimated without

touching the target fold, which enters only through the corrected empirical risk.

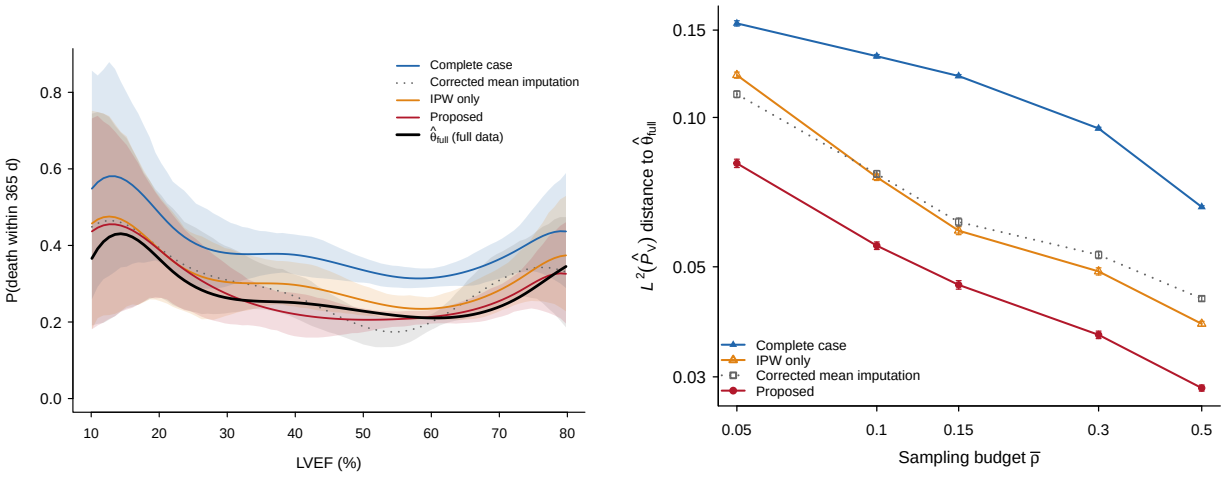
Every comparator shares the same target class and differs in the loss construction. The *complete-case* minimizes the unweighted empirical risk over the phase-2 units. *IPW* minimizes the corrected empirical risk with the augmentation term removed. *Mean imputation* replaces the phase-2 variable of every unit by $\widehat{E}[V \mid W, X, Y]$ and fits on the full sample. *Corrected mean imputation* retains the full corrected loss but collapses $\hat{q}$ to a point mass at its conditional mean, so it keeps the weighting and degrades only the centering distribution.

3.4.3 Results

Figure 3.1a shows the fitted curves at $\bar{\rho} = 0.15$. Figure 3.1b and Table 3.1 report the same summary, the mean $L^2(\widehat{P}_V)$ distance to $\hat{\theta}_{\text{full}}$ over the 500 replications across the sampling budget. First, the reference curve itself is clinically credible: it is U-shaped in LVEF, reproducing the association documented in large echocardiography cohorts, where mortality risk rises both for reduced and for supranormal ejection fraction with a nadir near 60–70% (Chang et al., 2025; Wehner et al., 2020). Second, the proposed estimator attains the smallest distance at every budget, and every method improves as $\bar{\rho}$ grows. Relative to IPW, the augmentation term in our corrected loss reduces the error by 22% at $\bar{\rho} = 0.15$ ($5.91 \to 4.60$) and by 34% at $\bar{\rho} = 0.05$ ($12.17 \to 8.08$). Corrected mean imputation, the same loss with $\hat{q}$ collapsed to a point mass at its mean, beats IPW only at the smallest budget and loses from $\bar{\rho} = 0.10$ onward. The value of the augmentation lies in the conditional distribution, not the conditional mean.

3.5 Discussion

We developed a corrected Neyman-orthogonal loss for the estimation of function-valued parameters under two-phase sampling. The framework concerns estimation rather than statistical inference. One route to inference is to estimate the target directly through a regularized one-step estimator (Luedtke & Chung, 2024) based on a regularized efficient influence operator. Extending that construction would require the operator and its regularization to be



(a) Fitted curves, $\bar{\rho} = 0.15$.

(b) $L^2(\widehat{P}_V)$ distance by sampling budget.

Figure 3.1: (a) Fitted mortality–LVEF curves at $\bar{\rho} = 0.15$ against the full-data reference $\hat{\theta}_{\text{full}}$; shaded bands are pointwise 10th–90th percentiles across the 500 sampling replications. (b) $L^2(\widehat{P}_V)$ distance to $\hat{\theta}_{\text{full}}$ by sampling budget $\bar{\rho}$ (log scale); points are means over 500 sampling replications.

Table 3.1: Mean $L^2(\widehat{P}_V)$ distance to the full-data reference curve over 500 sampling replications (Monte Carlo standard errors in parentheses).

	Sampling budget $\bar{\rho}$				
Estimator	0.05	0.10	0.15	0.30	0.50
Complete case	15.47 (0.19)	13.29 (0.11)	12.12 (0.08)	9.50 (0.05)	6.60 (0.04)
IPW only	12.17 (0.17)	7.58 (0.11)	5.91 (0.11)	4.89 (0.08)	3.84 (0.04)
Corrected mean imputation	11.14 (0.17)	7.70 (0.12)	6.16 (0.11)	5.28 (0.09)	4.31 (0.05)
Proposed	8.08 (0.16)	5.51 (0.10)	4.60 (0.09)	3.65 (0.06)	2.85 (0.04)

derived under two-phase sampling, which we leave to future work.

We have taken the argument to be univariate, but this can be extended to multivariate arguments at the cost of a larger critical radius. More broadly, both the argument and the phase-1 variable may be unstructured, handled through embeddings from a pre-trained model. With a fixed encoder the embedding is a deterministic transformation and the theory applies unchanged. When the encoder is itself estimated, it must be fit on a separate fold and frozen.

Our simulations indicate that a phase-1 variable of modest strength already improves estimation. This suggests enriching the phase-1 variable, but the enrichment is not free. As the dimension of phase-1 variables grows, the positivity conditions on the sampling probability and on the conditional density become difficult to satisfy. One can apply dimension reduction to the phase-1 variable to retain its information about the phase-2 variable.

BIBLIOGRAPHY

Ai, C., Linton, O., Motegi, K., & Zhang, Z. (2021). A unified framework for efficient estimation of general treatment models. *Quantitative Economics*, *12*(3), 779–816.

Angelopoulos, A. N., Bates, S., Fannjiang, C., Jordan, M. I., & Zrnic, T. (2023). Prediction-powered inference. *Science*, *382*(6671), 669–674.

Angelopoulos, A. N., Duchi, J. C., & Zrnic, T. (2024). PPI++: Efficient prediction-powered inference. *arXiv preprint arXiv:2311.01453*.

Bahadori, T., Tchetgen Tchetgen, E. J., & Heckerman, D. (2022). End-to-end balancing for causal continuous treatment–effect estimation. *Proceedings of the 39th International Conference on Machine Learning*, *162*, 1313–1326.

Bareinboim, E., & Pearl, J. (2014). Transportability from multiple environments with limited experiments: Completeness results. In *Advances in neural information processing systems* (Vol. 27).

Bareinboim, E., & Pearl, J. (2016). Causal inference and the data-fusion problem. *Proceedings of the National Academy of Sciences*, *113*(27), 7345–7352.

Bickel, P. J. (1982). On adaptive estimation. *The Annals of Statistics*, *10*(3), 647–671.

Blundell, R. W., & Powell, J. L. (2004). Endogeneity in semiparametric binary response models. *The Review of Economic Studies*, *71*(3), 655–679.

Bonvini, M., & Kennedy, E. H. (2022). Fast convergence rates for dose–response estimation. *arXiv preprint arXiv:2207.11825*.

Bousquet, O. (2003). Concentration Inequalities for Sub-Additive Functions Using the Entropy Method. In *Stochastic Inequalities and Applications* (pp. 213–247). Birkhäuser Basel.

Breslow, N. E., & Cain, K. C. (1988). Logistic regression for two-stage case-control data. *Biometrika*, *75*(1), 11–20.

Breslow, N. E., & Wellner, J. A. (2007). Weighted likelihood for semiparametric models and two-phase stratified samples, with application to Cox regression. *Scandinavian Journal of Statistics*, *34*(1), 86–102.

Buja, A., Hastie, T., & Tibshirani, R. (1989). Linear Smoothers and Additive Models. *The Annals of Statistics*, *17*(2), 453–510.

Chang, H.-C., Huang, W.-M., Lin, L.-Y., Lee, C.-W., Tseng, C.-H., Yu, W.-C., Cheng, H.-M., Chiang, C.-E., Chen, C.-H., & Sung, S.-H. (2025). A U-Shaped Relationship Between Left Ventricular Ejection Fraction and Risk of Worsening Heart Failure. *European Journal of Heart Failure*, *27*(12), 2938–2947.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., & Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, *21*(1), C1–C68.

Cucker, F., & Smale, S. (2001). On the mathematical foundations of learning. *Bulletin of the American Mathematical Society*, *39*, 1–49.

Curth, A., Alaa, A. M., & van der Schaar, M. (2021). Estimating structural target functions using machine learning and influence functions. *arXiv preprint arXiv:2008.06461*.

Dahabreh, I. J., & Hernán, M. A. (2019). Extending inferences from a randomized trial to a target population. *European Journal of Epidemiology*, *34*(8), 719–722.

Dahabreh, I. J., Robertson, S. E., Petito, L. C., Hernán, M. A., & Steingrimsson, J. A. (2022). Efficient and robust methods for causally interpretable meta-analysis: Transporting inferences from multiple randomized trials to a target population. *Biometrics*, *79*(2), 1057–1072.

Díaz, I., & van der Laan, M. J. (2013). Targeted data adaptive estimation of the causal dose–response curve. *Journal of Causal Inference*, *1*(2), 171–192.

Díaz, I., Williams, N., Hoffman, K. L., & Schenck, E. J. (2023). Nonparametric Causal Effects Based on Longitudinal Modified Treatment Policies. *Journal of the American Statistical Association*, *118*(542), 846–857.

Dominici, F., Daniels, M., Zeger, S. L., & Samet, J. M. (2002). Air pollution and mortality: Estimating regional and national dose–response relationships. *Journal of the American Statistical Association*, *97*(457), 100–111.

Fischer, S., & Steinwart, I. (2020). Sobolev norm learning rates for regularized least-squares algorithms. *Journal of Machine Learning Research*, *21*(205), 1–38.

Flores, C. A. (2007). *Estimation of dose–response functions and optimal doses with a continuous treatment* (tech. rep. No. 0707). University of Miami, Department of Economics.

Fong, C., Hazlett, C., & Imai, K. (2018). Covariate balancing propensity score for a continuous treatment: Application to the efficacy of political advertisements. *The Annals of Applied Statistics*, *12*(1), 156–177.

Foster, D. J., & Syrgkanis, V. (2023). Orthogonal statistical learning. *The Annals of Statistics*, *51*(3), 879–908.

Galvao, A. F., & Wang, L. (2015). Uniformly semiparametric efficient estimation of treatment effects with a continuous treatment. *Journal of the American Statistical Association*, *110*(512), 1528–1542.

Garreau, D., Jitkrittum, W., & Kanagawa, M. (2018). Large sample analysis of the median heuristic. *arXiv preprint arXiv:1707.07269*.

Graham, E., Carone, M., & Rotnitzky, A. (2026). Towards a unified theory for semiparametric data fusion with individual-level data. *The Annals of Statistics*, *54*(3), 1398–1424.

Gretton, A., Borgwardt, K., Rasch, M., Schölkopf, B., & Smola, A. (2006). A Kernel Method for the Two-Sample-Problem. In *Advances in Neural Information Processing Systems* (Vol. 19). MIT Press.

Han, L., Hou, J., Cho, K., Duan, R., & Cai, T. (2025). Federated adaptive causal estimation (FACE) of target treatment effects. *Journal of the American Statistical Association*, *120*(551), 1503–1516.

Hastie, T., & Tibshirani, R. (1986). Generalized Additive Models. *Statistical Science*, *1*(3), 297–310.

Hernán, M. A., & VanderWeele, T. J. (2011). Compound treatments and transportability of causal inference. *Epidemiology*, *22*(3), 368–377.

Hirano, K., & Imbens, G. W. (2004). The propensity score with continuous treatments. In *Applied Bayesian modeling and causal inference from incomplete-data perspectives* (pp. 73–84). John Wiley & Sons, Ltd.

Huang, W., & Zhang, Z. (2023). Nonparametric estimation of the continuous treatment effect with measurement error. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *85*(2), 474–496.

Hudson, A., Geng, E. H., Odeny, T. A., Bukusi, E. A., Petersen, M. L., & van der Laan, M. J. (2024). An approach to nonparametric inference on the causal dose–response function. *Journal of Causal Inference*, *12*(1), 20240001.

Hudson, A., Carone, M., & Shojaie, A. (2026). Inference on function-valued parameters using a restricted score test. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, qkag043.

Imai, K., & van Dyk, D. A. (2004). Causal inference with general treatment regimes. *Journal of the American Statistical Association*, *99*(467), 854–866.

Jain, S., & Luedtke, A. (2026). Conditional distributional treatment effects: Doubly robust estimation and testing. *arXiv preprint arXiv:2603.16829*.

Johnson, A. E. W., Pollard, T. J., Berkowitz, S. J., Greenbaum, N. R., Lungren, M. P., Deng, C.-y., Mark, R. G., & Horng, S. (2019). MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*, *6*(1), 317.

Johnson, A. E. W., Bulgarelli, L., Shen, L., Gayles, A., Shammout, A., Horng, S., Pollard, T. J., Hao, S., Moody, B., Gow, B., Lehman, L.-w. H., Celi, L. A., & Mark, R. G. (2023). MIMIC-IV, a freely accessible electronic health record dataset. *Scientific Data*, *10*(1), 1.

Kanamori, T., Hido, S., & Sugiyama, M. (2009). A least-squares approach to direct importance estimation. *Journal of Machine Learning Research*, *10*(48), 1391–1445.

Kennedy, E. H. (2023). Semiparametric doubly robust targeted double machine learning: A review. *arXiv preprint arXiv:2203.06469*.

Kennedy, E. H., Ma, Z., McHugh, M. D., & Small, D. S. (2016). Non-parametric methods for doubly robust estimation of continuous treatment effects. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *79*(4), 1229–1245.

Kluger, D. M., & Bates, S. (2026). M-estimation under two-phase multiwave sampling with applications to prediction-powered inference. *arXiv preprint arXiv:2602.16933*.

Kluve, J., Schneider, H., Uhlendorff, A., & Zhao, Z. (2012). Evaluating continuous training programmes by using the generalized propensity score. *Journal of the Royal Statistical Society Series A: Statistics in Society*, *175*(2), 587–617.

Lang, S. (1996). *Differential and Riemannian Manifolds* (3rd ed). Springer New York.

Li, S., & Luedtke, A. (2023). Efficient estimation under data fusion. *Biometrika*, *110*(4), 1041–1054.

Li, S., Gilbert, P. B., Duan, R., & Luedtke, A. (2025). Data fusion using weakly aligned sources. *Journal of the American Statistical Association*, *120*(552), 2569–2579.

Luedtke, A. (2026). Simplifying debiased inference via automatic differentiation and probabilistic programming. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *88*(1), 313–329.

Luedtke, A., & Chung, I. (2024). One-step estimation of differentiable Hilbert-valued parameters. *The Annals of Statistics*, *52*(4), 1534–1563.

Luedtke, A., Carone, M., & van der Laan, M. J. (2019). An Omnibus Non-Parametric Test of Equality in Distribution for Unknown Functions. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *81*(1), 75–99.

Mendelson, S., & Neeman, J. (2010). Regularization in kernel learning. *The Annals of Statistics*, *38*(1), 526–565.

Mitchell, C., & van de Geer, S. (2009). General oracle inequalities for model selection. *Electronic Journal of Statistics*, *3*, 176–204.

Morzywolek, P., Gilbert, P. B., & Luedtke, A. (2025). Inference on local variable importance measures for heterogeneous treatment effects. *arXiv preprint arXiv:2510.18843*.

Neugebauer, R., & van der Laan, M. J. (2007). Nonparametric causal effects based on marginal structural models. *Journal of Statistical Planning and Inference*, *137*(2), 419–434.

Newey, W. K. (1994). The asymptotic variance of semiparametric estimators. *Econometrica*, *62*(6), 1349–1382.

Neyman, J. (1938). Contribution to the Theory of Sampling Human Populations. *Journal of the American Statistical Association*, *33*(201), 101–116.

Nie, X., & Wager, S. (2021). Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika*, *108*(2), 299–319.

Pearl, J., & Bareinboim, E. (2011). Transportability of causal and statistical relations: A formal approach. *2011 IEEE 11th International Conference on Data Mining Workshops*, 540–547.

Pfanzagl, J. (1982). *Contributions to a general asymptotic statistical theory* (Vol. 13). Springer.

Qiu, H., Tchetgen Tchetgen, E. J., & Dobriban, E. (2024). Efficient and multiply robust risk estimation under general forms of dataset shift. *The Annals of Statistics*, *52*(4), 1796–1824.

Quan, H., Sundararajan, V., Halfon, P., Fong, A., Burnand, B., Luthi, J.-C., Saunders, L. D., Beck, C. A., Feasby, T. E., & Ghali, W. A. (2005). Coding algorithms for defining

comorbidities in ICD-9-CM and ICD-10 administrative data. *Medical Care*, *43*(11), 1130–1139.

Rasmussen, C. E., & Williams, C. K. I. (2005). *Gaussian processes for machine learning*. The MIT Press.

Robins, J. M., Rotnitzky, A., & Zhao, L. P. (1994). Estimation of Regression Coefficients When Some Regressors Are Not Always Observed. *Journal of the American Statistical Association*, *89*(427), 846–866.

Robins, J. M., Hernán, M. Á., & Brumback, B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology*, *11*(5), 550–560.

Robinson, L., Delgadillo, J., & Kellett, S. (2020). The dose-response effect in routinely delivered psychological therapies: A systematic review. *Psychotherapy Research*, *30*(1), 79–96.

Rudolph, K. E., & van der Laan, M. J. (2016). Robust estimation of encouragement design intervention effects transported across sites. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *79*(5), 1509–1525.

Sellergren, A., Kiraly, A., Pollard, T., Weng, W.-H., Liu, Y., Uddin, A., & Chen, C. (2023). Generalized Image Embeddings for the MIMIC Chest X-Ray dataset. *PhysioNet*.

Singh, R., Xu, L., & Gretton, A. (2023). Kernel methods for causal functions: Dose, heterogeneous and incremental response curves. *Biometrika*, *111*(2), 497–516.

Smale, S., & Zhou, D.-X. (2003). Estimating the approximation error in learning theory. *Analysis and Applications*, *1*(1), 17–41.

Sriperumbudur, B. K., Fukumizu, K., & Lanckriet, G. R. G. (2011). Universality, characteristic kernels and RKHS embedding of measures. *Journal of Machine Learning Research*, *12*, 2389–2410.

Steinwart, I., & Scovel, C. (2012). Mercer's theorem on general domains: On the interaction between measures, kernels, and RKHSs. *Constructive Approximation*, *35*(3), 363–417.

Stuart, E. A., Bradshaw, C. P., & Leaf, P. J. (2015). Assessing the generalizability of randomized trial results to target populations. *Prevention Science*, *16*(3), 475–485.

Sugiyama, M., Suzuki, T., & Kanamori, T. (2012). *Density ratio estimation in machine learning*. Cambridge University Press.

Sun, B., & Miao, W. (2022). On semiparametric instrumental variable estimation of average treatment effects through data fusion. *Statistica Sinica*, *32*, 569–590.

Sundberg, J., Ghani, R., Ben-Michael, E., & Kennedy, E. (2026). Learning who to treat when treatment is missing. *arXiv preprint arXiv:2607.14346*.

Tsiatis, A. (2006). *Semiparametric Theory and Missing Data*. Springer New York.

Tsybakov, A. B. (2009). *Introduction to nonparametric estimation* (1st ed.). Springer.

van der Laan, M. J., & Dudoit, S. (2003). *Unified cross-validation methodology for selection among estimators and a general cross-validated adaptive epsilon-net estimator: Finite sample oracle inequalities and examples* (tech. rep. No. 130). U.C. Berkeley Division of Biostatistics Working Paper Series.

van der Laan, M. J., & Rubin, D. (2006). Targeted Maximum Likelihood Learning. *The International Journal of Biostatistics*, *2*(1).

van der Laan, M. J., Polley, E. C., & Hubbard, A. E. (2007). Super learner. *Statistical Applications in Genetics and Molecular Biology*, *6*(1).

van der Vaart, A. W., & Wellner, J. A. (1996). *Weak Convergence and Empirical Processes*. Springer New York.

Vershynin, R. (2018). *High-dimensional probability: An introduction with applications in data science*. Cambridge University Press.

Wald, A. (1945). Sequential tests of statistical hypotheses. *The Annals of Mathematical Statistics*, *16*(2), 117–186.

Wang, R., Osom, A., & Zhang, B. (2026). Generalized projection tests for function-valued parameters with applications to testing structural causal assumptions. *arXiv preprint arXiv:2603.13681*.

Wehner, G. J., Jing, L., Haggerty, C. M., Suever, J. D., Leader, J. B., Hartzel, D. N., Kirchner, H. L., Manus, J. N. A., James, N., Ayar, Z., Gladding, P., Good, C. W., Cleland, J. G. F., & Fornwalt, B. K. (2020). Routinely reported ejection fraction

and mortality in clinical practice: Where does the nadir of risk lie? *European Heart Journal*, *41*(12), 1249–1257.

Wendland, H. (2004). *Scattered data approximation*. Cambridge University Press.

Westling, T. (2022). Nonparametric Tests of the Causal Null With Nondiscrete Exposures. *Journal of the American Statistical Association*, *117*(539), 1551–1562.

Westling, T., Gilbert, P., & Carone, M. (2020). Causal isotonic regression. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *82*(3), 719–747.

White, J. E. (1982). A two stage design for the study of the relationship between a rare exposure and a rare disease. *American Journal of Epidemiology*, *115*(1), 119–128.

Yiu, S., & Su, L. (2018). Covariate association eliminating weights: A unified weighting framework for causal effect estimation. *Biometrika*, *105*(3), 709–722.

Zeng, Z., Arbour, D., Feller, A., Addanki, R., Rossi, R., Sinha, R., & Kennedy, E. H. (2024). Continuous treatment effects with surrogate outcomes. *Proceedings of the 41st International Conference on Machine Learning*.

Zhang, H., Li, Y., Lu, W., & Lin, Q. (2023). On the optimality of misspecified kernel ridge regression. *Proceedings of the 40th International Conference on Machine Learning*, *202*, 41331–41353.

Zrnic, T., & Candès, E. J. (2024). Active statistical inference. *Proceedings of the 41st International Conference on Machine Learning*, *235*, 62993–63010.

Appendix A

SUPPLEMENTARY MATERIALS FOR CHAPTER 1

Algorithm A.1 Cross validation in kernel ridge regression

1: **Input**: The sample z_n, number of folds K, values of λ.

2: **Cross validation**:

1. Partition the sample z^n into training set z_k^n and test set z_{-k}^n for $k = 1, \ldots, K$.
2. For each value of λ, do the following steps.
3. For each k, construct nuisance estimators only using z_k^n as in Algorithm 1.1.
4. With the same k, estimate the following parameters by plugging in test set (z_{-k}^n) into the nuisance parameters in the previous step.

$$
\begin{aligned}
u_i &= \mathbf{1}(s_i \in \mathcal{S}_Y)\, \widehat{\eta}\, \widehat{w}(a_i, x_i)\, (\widehat{m}(a_i, x_i) - y_i)\,, \\
v_i &= \mathbf{1}(s_i \in \mathcal{S}_X)\, \widehat{\xi}\, (\mathrm{E}_X\left[\widehat{m}(b_i, x_i)\right] - \widehat{m}(b_i, x_i)) - \mathrm{E}_X\left[\widehat{m}(b_i, x_i)\right] \\
\widehat{\beta} &= -\frac{\mathbf{u}}{\lambda_n \sqrt{|z_k^n|}},\ \widehat{\gamma} = -\left(\mathbf{K}_{22} + \lambda_n \mathbf{I}_{|z_k^n|}\right)^{-1} \left(\frac{\mathbf{v}}{\sqrt{|z_k^n|}} + \mathbf{K}_{21}\widehat{\beta}\right) \\
\widehat{\theta}(\cdot) &= \frac{1}{\sqrt{|z_k^n|}} \sum_{j=1}^{|z_k^n|} \left[\widehat{\beta}_j \mathcal{K}(\cdot, a_j) + \widehat{\gamma}_j \mathcal{K}(\cdot, b_j)\right].
\end{aligned}
$$

5. Calculate the risk for the specific k as follows.

$$
\widehat{R}_k(\widehat{\theta}) = \frac{1}{|z_{-k}^n|} \sum_{i=1}^{|z_{-k}^n|} \left[\widehat{\theta}(b_i)^2 + 2v_i\widehat{\theta}(b_i) + 2u_i\widehat{\theta}(a_i)\right] + \lambda_n \|\widehat{\theta}\|_{\mathcal{H}}.
$$

6. Calculate the cross-validated risk by taking the average of $\widehat{R}_k$ over $k = 1, \ldots, K$.
7. Going back to step 2, choose the λ_n showing the smallest cross-validated risk.

A.1 Derivation of Neyman-orthogonal loss and a stochastic approximation thereof (Lemma A.1)

We now derive a Neyman-orthogonal loss for CDRFs. We proceed in several steps. First, we provide the form of the nonparametric canonical gradient of the risk at Q^0 with respect to the tangent space of the target distributions. Next, we project it onto the corresponding tangent space of P^0. Finally, we construct a one-step estimator of the risk, along with an approximate version that facilitates its minimization. We provide the nonparametric canonical gradients and construction of one-step estimators under two scenarios: (i) when data fusion is fully utilized, and (ii) when data fusion is not used and estimation is based solely on samples in the intersection $S \in \mathcal{S}_X \cap \mathcal{S}_Y$.

If $\mathcal{Q}$ is locally nonparametric, standard calculations or a chain rule (Kennedy, 2023; Luedtke, 2026) show $Q \mapsto \mathcal{L}_Q(\theta)$ at Q^0 is pathwise differentiable with canonical gradient

$$D^*_{Q^0}(\theta) : (a, x, y) \mapsto 2 \int (\theta(a) - \theta_0(a)) (\theta_0(a) - m_0(x, a)) \, d\mu(a) + 2 \frac{d\mu_A}{dQ^0_{A|X}}(a \mid x) (\theta(a) - \theta_0(a)) \big(m_0(x, a) - y\big).$$

By Corollary 1 in Li and Luedtke (2023), the canonical gradient of $P \mapsto L_P(\theta)$ at P^0 under the data fusion setting defined by fusion sets $(\mathcal{S}_X, \mathcal{S}_Y)$ is

$$D^*_{P^0}(\theta) : (a, x, y, s) \mapsto \mathbf{1}(s \in \mathcal{S}_X) \cdot 2\xi_0 \int (\theta(a) - \theta_0(a)) (\theta_0(a) - m_0(x, a)) \, d\mu(a) + \mathbf{1}(s \in \mathcal{S}_Y) \cdot 2\eta_0 w_0(x, a) (\theta(a) - \theta_0(a)) (m_0(x, a) - y),$$

where

$$\xi_0 := \frac{1}{P^0(S \in \mathcal{S}_X)}, \quad \eta_0 := \frac{1}{P^0(S \in \mathcal{S}_Y)}, \quad w_0(x, a) := \frac{p^0(x \mid S \in \mathcal{S}_X)}{p^0(x \mid S \in \mathcal{S}_Y)} \frac{1}{p^0(a \mid x, S \in \mathcal{S}_Y)}$$
$$m_0(x, a) := \mathrm{E}_{P^0}[Y \mid x, a, S \in \mathcal{S}_Y]$$

are the true nuisance parameters. Similarly, the canonical gradient of $\mathcal{L}_P(\theta)$ at P^0 without data fusion (i.e., using only the intersection $\mathcal{S}_X \cap \mathcal{S}_Y$) is given by

$$\underline{D}^*_{P^0}(\theta) : (a, x, y, s) \mapsto \mathbf{1}(s \in \mathcal{S}_X \cap \mathcal{S}_Y) \cdot 2\underline{\xi}_0 \int (\theta(a) - \underline{\tau}_0(a)) (\underline{\tau}_0(a) - \underline{m}_0(x, a)) \, d\mu(a)$$

$$+ \mathbf{1}(s \in \mathcal{S}_X \cap \mathcal{S}_Y) \cdot 2\underline{\eta}_0 \underline{w}_0(x,a) \left(\theta(a) - \underline{\tau}_0(a)\right) \left(\underline{m}_0(x,a) - y\right),$$

where

$$\underline{\xi}_0 = \underline{\eta}_0 := \frac{1}{P^0(S \in \mathcal{S}_X \cap \mathcal{S}_Y)}, \quad \underline{w}_0(x,a) := \frac{1}{p^0(a \mid x, S \in \mathcal{S}_X \cap \mathcal{S}_Y)}$$
$$\underline{m}_0(x,a) := \mathrm{E}_{P^0}\left[Y \mid x, a, S \in \mathcal{S}_X \cap \mathcal{S}_Y\right].$$

If $\mathcal{Q}$ is not locally nonparametric, then the above quantities are still gradients that can be used to construct one-step estimators, but they may not be canonical gradients.

Lemma A.1 (Stochastically approximated loss)**.** *The (oracle) pointwise loss function associated with the (nonparametric) one-step estimator of $L_{P^0}(\theta)$ under data fusion is defined as*

$$\ell^{OS}_{P^0}(\theta, g_0; x, a, y, s) := \mathrm{E}_{B\sim\mu}\big[\ell_{P^0}(\theta, g_0; x, a, y, s, B)\big],$$

where the inner loss function $\ell_{P^0}(\theta, g_0; x, a, y, s, b)$ is given by

$$\begin{aligned}\ell_{P^0}(\theta, g_0; x, a, y, s, b) := {} & \left(\theta(b) - \theta_0(b)\right)^2 \\ & + 2\xi_0 \mathbf{1}(s \in \mathcal{S}_X) \left(\theta(b) - \theta_0(b)\right) \left(\theta_0(b) - m_0(x,b)\right) \\ & + 2\eta_0 w_0(x,a) \mathbf{1}(s \in \mathcal{S}_Y) \left(\theta(a) - \theta_0(a)\right) \left(m_0(x,a) - y\right). \end{aligned} \tag{A.1}$$

The empirical mean of this loss, $\mathbb{P}_n\big[\ell_{P^0}(\theta, g_0; \cdot)\big]$, is a stochastic approximation of the oracle one-step estimator of $L_{P^0}(\theta)$. Similarly, the (oracle) loss function without data fusion is defined as

$$\underline{\ell}^{OS}_{P^0}(\theta, g_0; x, a, y, s) := \mathrm{E}_{B\sim\mu}\big[\underline{\ell}_{P^0}(\theta, g_0; x, a, y, s, B)\big],$$

where

$$\begin{aligned}\underline{\ell}_{P^0}(\theta, g_0; x, a, y, s, b) := {} & \left(\theta(b) - \theta_0(b)\right)^2 \\ & + 2\underline{\xi}_0 \mathbf{1}(s \in \mathcal{S}_X \cap \mathcal{S}_Y) \left(\theta(b) - \underline{\tau}_0(b)\right) \left(\underline{\tau}_0(b) - \underline{m}_0(x,b)\right) \\ & + 2\underline{\eta}_0 \underline{w}_0(x,a) \mathbf{1}(s \in \mathcal{S}_X \cap \mathcal{S}_Y) \left(\theta(a) - \underline{\tau}_0(a)\right) \left(\underline{m}_0(x,a) - y\right). \end{aligned} \tag{A.2}$$

Proof of Lemma A.1. We construct a one-step estimator of $L_{P^0}(\theta)$ using $D^*_{P^0}(\theta)$. Given an estimator P of P^0, the one-step expansion of L_{P^0} under data fusion is

$$
\begin{aligned}
&L_{P^0} + \mathbb{P}_n D^*_{P^0}(\theta) \\
&= \int (\theta(a) - \theta_0(a))^2 \, d\mu(a) + \frac{2\xi_0}{n} \sum_{i=1}^{n} \mathbf{1}(S_i \in \mathcal{S}_X) \int (\theta(a) - \theta_0(a)) \, (\theta_0(a) - m_0(X_i, a)) \, d\mu(a) \\
&\quad + \frac{2\eta_0}{n} \sum_{i=1}^{n} \mathbf{1}(S_i \in \mathcal{S}_Y) w_0(X_i, A_i) \, (\theta(A_i) - \theta_0(A_i)) \, (m_0(X_i, A_i) - Y_i) \, .
\end{aligned}
$$

Since the first two terms involve an expectation with respect to the known measure μ, we can approximate them using an artificial sample $B_i \sim \mu$ and the corresponding empirical average:

$$
\begin{aligned}
&\approx \frac{1}{n} \sum_{i=1}^{n} (\theta(B_i) - \theta_0(B_i))^2 + \frac{2\xi_0}{n} \sum_{i=1}^{n} \mathbf{1}(S_i \in \mathcal{S}_X) \, (\theta(B_i) - \theta_0(B_i)) \, (\theta_0(B_i) - m_0(X_i, B_i)) \\
&\quad + \frac{2\eta_0}{n} \sum_{i=1}^{n} \mathbf{1}(S_i \in \mathcal{S}_Y) w_0(A_i, X_i) \, (\theta(A_i) - \theta_0(A_i)) \, (m_0(X_i, A_i) - Y_i) \\
&= \mathbb{P}_n \big[\ell_{P^0}(\theta, g_0; \cdot) \big],
\end{aligned}
$$

where $Z_i = (X_i, A_i, Y_i, S_i, B_i)$. Hence, the approximated loss under data fusion is defined as

$$
\begin{aligned}
\ell_{P^0}(\theta, g_0; (x, a, y, s, b)) := &\, (\theta(b) - \theta_0(b))^2 \\
&+ 2\xi_0 \mathbf{1}(s \in \mathcal{S}_X) \, (\theta(b) - \theta_0(b)) \, (\theta_0(b) - m_0(x, b)) \\
&+ 2\eta_0 w_0(x, a) \mathbf{1}(s \in \mathcal{S}_Y) \, (\theta(a) - \theta_0(a)) \, (m_0(x, a) - y) \, ,
\end{aligned}
$$

with nuisance parameters defined as

$$
\xi_0 := \frac{1}{P^0(S \in \mathcal{S}_X)}, \quad \eta_0 := \frac{1}{P^0(S \in \mathcal{S}_Y)}, \quad w_0(x, a) := \frac{p^0(x \mid S \in \mathcal{S}_X)}{p^0(x \mid S \in \mathcal{S}_Y)} \cdot \frac{1}{p^0(a \mid x, S \in \mathcal{S}_Y)}
$$

$$
m_0(x, a) := \mathrm{E}_{P^0}[Y \mid x, a, S \in \mathcal{S}_Y].
$$

A similar procedure can be applied to define the one-step loss function without data fusion. □

A.2 Closed-form solution of kernel ridge regression (Lemma A.2)

Lemma A.2 (Closed-form solution of kernel ridge regression)**.** *The closed-form estimator $\widehat{\theta}$ obtained in Algorithm 1.2 corresponds to the solution of the kernel ridge regression problem in the RKHS $\mathcal{H}$ induced by the kernel $\mathcal{K}$, with the objective function given by*

$$\mathbb{P}_n\big[\ell_{P^0}(\theta, \widehat{g}; \cdot)\big] + \lambda_n \|\theta\|_{\mathcal{H}}^2,$$

where $\{u_i, v_i\}_{i=1}^n$ are defined as in Algorithm 1.2. The resulting estimator admits the following closed-form expression:

$$\widehat{\theta}(\cdot) = \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \left[\widehat{\beta}_j \mathcal{K}(\cdot, a_j) + \widehat{\gamma}_j \mathcal{K}(\cdot, b_j)\right],$$

where the weights $\widehat{\beta}$ and $\widehat{\gamma}$ are computed as described in Algorithm 1.2.

Proof of Lemma A.2. The empirical objective function for kernel ridge regression is given by

$$\begin{aligned}
\widehat{R}(\theta) &:= \mathbb{P}_n\left[\ell_{P^0}(\theta, \widehat{g}; \cdot)\right] + \lambda_n \|\theta\|_{\mathcal{H}}^2 \\
&= \frac{1}{n} \sum_{i=1}^{n} \left(\theta(b_i) - \widehat{\tau}(b_i)\right)^2 + \frac{2}{n} \sum_{i=1}^{n} \widehat{\xi} \cdot \mathbf{1}(s_i \in \mathcal{S}_X) \left(\theta(b_i) - \widehat{\tau}(b_i)\right) \left(\widehat{\tau}(b_i) - \widehat{m}(x_i, b_i)\right) \\
&\quad + \frac{2}{n} \sum_{i=1}^{n} \widehat{\eta} \cdot \widehat{w}(x_i, a_i) \cdot \mathbf{1}(s_i \in \mathcal{S}_Y) \left(\theta(a_i) - \widehat{\tau}(a_i)\right) \left(\widehat{m}(x_i, a_i) - y_i\right) + \lambda_n \|\theta\|_{\mathcal{H}}^2.
\end{aligned}$$

We reorganize the expression by expanding the quadratic term and regrouping linear and constant components:

$$\widehat{R}(\theta) = \frac{1}{n} \sum_{i=1}^{n} \theta(b_i)^2 + \frac{2}{n} \sum_{i=1}^{n} v_i \cdot \theta(b_i) + \frac{2}{n} \sum_{i=1}^{n} u_i \cdot \theta(a_i) + \frac{1}{n} \sum_{i=1}^{n} C_i + \lambda_n \|\theta\|_{\mathcal{H}}^2,$$

where

$$\begin{aligned}
u_i &:= \widehat{\eta} \cdot \widehat{w}(a_i, x_i) \cdot \mathbf{1}(s_i \in \mathcal{S}_Y) \cdot (\widehat{m}(x_i, a_i) - y_i), \\
v_i &:= \widehat{\xi} \cdot \mathbf{1}(s_i \in \mathcal{S}_X) \cdot (\widehat{\tau}(b_i) - \widehat{m}(x_i, b_i)) - \widehat{\tau}(b_i), \\
C_i &:= \widehat{\tau}(b_i)^2 - 2\widehat{\xi} \cdot \mathbf{1}(s_i \in \mathcal{S}_X) \cdot (\widehat{\tau}(b_i) - \widehat{m}(x_i, b_i)) \cdot \widehat{\tau}(b_i)
\end{aligned}$$

$$- 2\widehat{\eta} \cdot \widehat{w}(x_i, a_i) \cdot \mathbf{1}(s_i \in \mathcal{S}_Y) \cdot (\widehat{m}(x_i, a_i) - y_i) \cdot \widehat{\tau}(a_i).$$

Denote by $\theta \in \mathcal{H}$ of the form

$$\theta(\cdot) = \frac{1}{n} \sum_{j=1}^{n} [\beta_j \mathcal{K}(\cdot, a_j) + \gamma_j \mathcal{K}(\cdot, b_j)].$$

Then the empirical objective becomes

$$\widehat{R}(\theta) = \frac{1}{n} \sum_{i=1}^{n} \left(\sum_{j=1}^{n} \frac{1}{n} \beta_j \mathcal{K}(b_i, a_j) + \gamma_j \mathcal{K}(b_i, b_j) \right)^2 + \frac{2}{n} \sum_{i=1}^{n} v_i \left(\sum_{j=1}^{n} \frac{1}{n} \beta_j \mathcal{K}(b_i, a_j) + \gamma_j \mathcal{K}(b_i, b_j) \right)$$
$$+ \frac{2}{n} \sum_{i=1}^{n} u_i \left(\sum_{j=1}^{n} \left(\frac{1}{n} \beta_j \mathcal{K}(a_i, a_j) + \gamma_j \mathcal{K}(a_i, b_j) \right) \right) + \lambda_n \langle \theta, \theta \rangle_{\mathcal{H}} + \sum_{i=1}^{n} C_i.$$

Let $u := (u_1, \ldots, u_n)^\top$, $v := (v_1, \ldots, v_n)^\top$, and define the block Gram matrix

$$\mathbf{K} = \begin{pmatrix} \mathbf{K}_{11} & \mathbf{K}_{12} \\ \mathbf{K}_{21} & \mathbf{K}_{22} \end{pmatrix} := \frac{1}{n} \begin{pmatrix} \mathcal{K}(a_i, a_j) & \mathcal{K}(a_i, b_j) \\ \mathcal{K}(b_i, a_j) & \mathcal{K}(b_i, b_j) \end{pmatrix}.$$

Then the objective simplifies to:

$$\widehat{R}(\theta) = \frac{1}{n} \left(\beta^\top \mathbf{K}_{12} \mathbf{K}_{21} \beta + \beta^\top \mathbf{K}_{12} \mathbf{K}_{22} \gamma + \gamma^\top \mathbf{K}_{22} \mathbf{K}_{21} \beta + \gamma^\top \mathbf{K}_{22} \mathbf{K}_{22} \gamma \right)$$
$$+ \frac{2}{n} \left(\beta^\top \mathbf{K}_{12} v + \gamma^\top \mathbf{K}_{22} v + \beta^\top \mathbf{K}_{11} u + \gamma^\top \mathbf{K}_{21} u \right)$$
$$+ \frac{\lambda_n}{n} \left(\beta^\top \mathbf{K}_{11} \beta + 2\beta^\top \mathbf{K}_{12} \gamma + \gamma^\top \mathbf{K}_{22} \gamma \right) + \sum_{i=1}^{n} C_i.$$

We compute the gradients of the empirical risk with respect to β and γ as follows:

$$\begin{bmatrix} \frac{\partial}{\partial \beta} \widehat{R}(\theta) \\ \frac{\partial}{\partial \gamma} \widehat{R}(\theta) \end{bmatrix} = \frac{2}{n} \begin{pmatrix} \mathbf{K}_{11} & \mathbf{K}_{12} \\ \mathbf{K}_{21} & \mathbf{K}_{22} \end{pmatrix} \begin{pmatrix} u + \lambda_n \beta \\ \mathbf{K}_{21} \beta + \mathbf{K}_{22} \gamma + v + \lambda_n \gamma \end{pmatrix}.$$

Setting the gradient to zero and using the fact that, by Condition 1.4(a), the block matrix $\mathbf{K}$ is invertible, we obtain that the minimizer $(\widehat{\beta}, \widehat{\gamma})$ satisfies

$$\left(\widehat{\beta}, \widehat{\gamma} \right) = \left(-\frac{u}{\lambda_n}, \ -(\mathbf{K}_{22} + \lambda_n I_n)^{-1} \left(\mathbf{K}_{21} \widehat{\beta} + v \right) \right),$$

where I_n denotes the $n \times n$ identity matrix. $\square$

Notation for functional derivatives. Let $L(\theta, g)$ be a functional defined on $\Theta \times \mathcal{G}$, where $\Theta \subset L^2(\mu)$ and $\mathcal{G} \subset L^2(\mu \times Q_X^0; \ell^2)$.

- The *(Gâteaux) derivative* of L with respect to θ at (θ^*, g_0) in the direction $h \in L^2(\mu)$ is
$$\mathcal{D}_\theta L(\theta^*, g_0)[h] := \frac{d}{dt} L(\theta^* + th, g_0)\Big|_{t=0}.$$
- The derivative of L with respect to g at (θ^*, g_0) in the direction $k \in L^2(\mu \times Q_X^0; \ell^2)$ is
$$\mathcal{D}_g L(\theta^*, g_0)[k] := \frac{d}{dt} L(\theta^*, g_0 + tk)\Big|_{t=0}.$$
- The *second derivative* with respect to θ in directions $h_1, h_2 \in L^2(\mu)$ is
$$\mathcal{D}_\theta^2 L(\theta^*, g_0)[h_1, h_2] := \frac{\partial^2}{\partial s\, \partial t} L(\theta^* + th_1 + sh_2, g_0)\Big|_{t=0,\, s=0}.$$
- The *mixed derivative* with respect to θ and g in directions $h \in L^2(\mu)$ and $k \in L^2(\mu \times Q_X^0; \ell^2)$ is
$$\mathcal{D}_\theta \mathcal{D}_g L(\theta^*, g_0)[h, k] := \frac{\partial^2}{\partial t\, \partial s} L(\theta^* + th,\, g_0 + sk)\Big|_{t=0,\, s=0}.$$

A.3 Properties of the population loss (Lemma A.3)

Lemma A.3 (Properties of the population loss)**.** *The loss function in (1.1) satisfies the following properties. Assume that Θ is convex, and recall that θ^* denotes the minimizer of the oracle risk over Θ.*

(a) ***(First-order optimality)*** *The minimizer θ^* satisfies the first-order optimality condition:*
$$\mathcal{D}_\theta L_{P^0}(\theta^*, g_0)[\theta - \theta^*] = 2\,\mathrm{E}_\mu\left[\big(\theta(A) - \theta^*(A)\big)\big(\theta^*(A) - \theta_0(A)\big)\right] \geq 0,$$
for all $\theta \in L^2(\mu)$.

(b) ***(Neyman orthogonality)*** *The population risk L_{P^0} is Neyman orthogonal:*

$$\mathcal{D}_g \mathcal{D}_\theta L_{P^0}(\theta^*, g_0)[\theta - \theta^*, g - g_0] = 0,$$

for all $\theta \in L^2(\mu)$ and $g \in \mathcal{G}$.

(c) ***(Second-order smoothness and strong convexity)*** *The population risk is second-order smooth and strongly convex in θ:*

$$\mathcal{D}_\theta^2 L_P(\overline{\theta}, g)[\theta - \theta^*, \theta - \theta^*] = 2\,\mathrm{E}_\mu\left[\left(\theta(A) - \theta^*(A)\right)^2\right] = 2\|\theta - \theta^*\|_{L^2(\mu)}^2,$$

for all $\theta \in \Theta$, $\overline{\theta} \in L^2(\mu)$, and $g \in \mathcal{G}$.

(d) ***(Higher-order smoothness)*** *The population risk is higher-order smooth in the nuisance parameter:*

$$\left|\mathcal{D}_g^2 \mathcal{D}_\theta L_P(\theta^*, \overline{g})[\theta - \theta^*, g - g_0, g - g_0]\right| \leq 2M_\lambda \|\theta - \theta^*\|_{L^2(\mu)} \|g - g_0\|_{L^4(Q_X^0 \times \mu; \ell^2)}^2,$$

for all $\theta \in L^2(\mu)$, $g \in \mathcal{G}$, and $\overline{g} \in \mathrm{star}(\mathcal{G}, g_0)$.

Proof of Lemma A.3. We will use the fact that Θ is convex to prove first-order optimality. Since Θ is convex, $t\theta + (1-t)\theta^* \in \Theta$ whenever $\theta, \theta^* \in \Theta$ and $t \in [0,1]$. By the definition of θ^*,

$$\|\theta_0 - \theta^*\|_{L^2(\mu)} = \inf_{\theta \in \Theta} \|\theta_0 - \theta\|_{L^2(\mu)} \leq \|\theta_0 - (t\theta + (1-t)\theta^*)\|_{L^2(\mu)}$$

implies

$$\begin{aligned}
\|\theta_0 - \theta^*\|_{L^2(\mu)}^2 &\leq \|\theta_0 - (t\theta + (1-t)\theta^*)\|_{L^2(\mu)}^2 \\
&= \|\theta_0 - \theta^* - t(\theta - \theta^*)\|_{L^2(\mu)}^2 \\
&= \|\theta_0 - \theta^*\|_{L^2(\mu)}^2 - 2t\langle \theta_0 - \theta^*, \theta - \theta^* \rangle_{L^2(\mu)} + t^2 \|\theta - \theta^*\|_{L^2(\mu)}^2
\end{aligned}$$

Canceling out the first term and for any $t \in (0,1]$, we can rearrange terms and divide both sides by t to find

$$2\langle \theta_0 - \theta^*, \theta - \theta^* \rangle_{L^2(\mu)} \leq t \|\theta - \theta^*\|_{L^2(\mu)}^2.$$

Since above inequality holds for any $t \in [0, 1)$, we can take right limits as $t \to 0+$. Then the right handside goes to 0. This implies that

$$2\langle \theta_0 - \theta^*, \theta - \theta^* \rangle_{L^2(\mu_A)} \leq \lim_{t \to 0+} t \|\theta - \theta^*\|^2_{L^2(\mu_A)} = 0.$$

Then for any $\theta \in \Theta$, the first order derivative of the population risk with respect to θ,

$$\begin{aligned}
&\mathcal{D}_\theta L_{\mathcal{P}^0}(\theta^*, g_0)\,[\theta - \theta^*] \\
&= \frac{d}{dt}\,\mathrm{E}_\mu\left[(\theta^*(A) + t(\theta(A) - \theta^*(A)) - \theta_0(A))^2\right]\Big|_{t=0} \\
&\quad + \frac{d}{dt}\,\mathrm{E}_{P^0}\left[\mathbf{1}(S \in \mathcal{S}_X) \cdot 2\xi_0 \cdot \mathrm{E}_\mu\left[(\theta^*(A) + t(\theta(A) - \theta^*(A)))\,(\theta_0(A) - m_0(X, A))\right]\right]\Big|_{t=0} \\
&\quad + \frac{d}{dt}\,\mathrm{E}_{P^0}\left[\mathbf{1}(S \in \mathcal{S}_Y) \cdot 2\eta_0 w_0(X, A) \cdot (\theta^*(A) + t(\theta(A) - \theta^*(A)))\,(m_0(X, A) - Y)\right]\Big|_{t=0} \\
&= 2\mathrm{E}_\mu\left[(\theta(A) - \theta^*(A))\,(\theta^*(A) - \theta_0(A))\right] \\
&\quad + 2\mathrm{E}_{Q_X^0}\left[\mathrm{E}_\mu\left[(\theta(A) - \theta^*(A))\,(\theta_0(A) - m_0(X, A))\right]\right] \\
&\quad + 2\mathrm{E}_{Q_X^0}\left[\mathrm{E}_\mu\left[\mathrm{E}_{Q^0_{Y|X,A}}\left[(\theta(A) - \theta^*(A))\,(m_0(X, A) - Y)\right]\right]\right] \\
&= 2\mathrm{E}_\mu\left[(\theta(A) - \theta^*(A))\,(\theta^*(A) - \theta_0(A))\right] \geq 0,
\end{aligned}$$

where last two terms in the derivative disappear using the change of order of integration and the last inequality comes from the argument above.

Next, we prove Neyman orthogonality with respect to the nuisance parameters. For any $\theta \in L^2(\mu)$ and nuisance parameter $g \in \mathcal{G}$,

$$\begin{aligned}
&\mathcal{D}_g \mathcal{D}_\theta L_{P^0}(\theta^*, g_0)[\theta - \theta^*, g - g_0] \\
&= \mathrm{E}_\mu\left[\frac{d^2}{dt_1 dt_2}\,(\theta^* + t_2(\theta - \theta^*) - \theta_0 - t_1(\tau - \theta_0))^2\Big|_{(t_1,t_2)=(0,0)}\right] \\
&\quad + 2\,\mathrm{E}_{Q_X^0 \times \mu}\left[\frac{d^2}{dt_1 dt_2}\left(\frac{\xi_0 + t_1(\xi - \xi_0)}{\xi_0}\right)(\theta^* + t_2(\theta - \theta^*) - \theta_0 - t_1(\tau - \theta_0))\right. \\
&\quad \left.\times\,(\theta_0 + t_1(\tau - \theta_0) - (m + t_1(m - m_0)))|_{(t_1,t_2)=(0,0)}\right] \\
&\quad + 2\,\mathrm{E}_{Q_X^0 \times \mu \times Q^0_{Y|X,A}}\left[\frac{d^2}{dt_1 dt_2}\left(\frac{\eta_0 + t_1(\eta - \eta_0)}{\eta_0}\right)\left(\frac{w_0 + t_1(w - w_0)}{w_0}\right)\right.
\end{aligned}$$

$$
\begin{aligned}
&\times \left(\theta^* + t_2(\theta - \theta^*) - \theta_0 - t_1(\tau - \theta_0)\right)\left(m_0 + t_1(m - m_0) - Y\right)\big|_{(t_1,t_2)=(0,0)}\Big] \\
&= -2\,\mathrm{E}_\mu\left[(\theta - \theta^*)(\tau - \theta_0)\right] \\
&\quad + 2\,\mathrm{E}_{Q_X^0 \times \mu}\left[(\theta - \theta^*)\left(\frac{\xi - \xi_0}{\xi_0}(\theta_0 - m_0) + (\tau - \theta_0) - (m - m_0)\right)\right] \\
&\quad + 2\,\mathrm{E}_{Q_X^0 \times \mu \times Q_{Y|X,A}^0}\left[(\theta - \theta^*)\left(\frac{\eta - \eta_0}{\eta_0}(m_0 - Y) + \frac{w - w_0}{w_0}(m_0 - Y) + (m - m_0)\right)\right] \\
&= 0.
\end{aligned}
$$

Next, we derive the second-order derivative with respect to the target parameter and prove second-order smoothness and strong convexity.

$$
\begin{aligned}
\mathcal{D}_\theta^2 L_{P^0}(\overline{\theta}, g_0)[\theta - \theta^*, \theta - \theta^*] &= \mathrm{E}_\mu\left[\frac{d^2}{dt_1 dt_2}\left(\overline{\theta} + t_1(\theta - \theta^*) + t_2(\theta - \theta^*) - \theta_0\right)^2\Big|_{(t_1,t_2)=(0,0)}\right] \\
&= \mathrm{E}_\mu\left[\frac{d}{dt_1}\left(2(\theta - \theta^*)\left(\overline{\theta} + t_1(\theta - \theta^*) - \theta_0\right)\right)\Big|_{t_1=0}\right] \\
&= 2\,\mathrm{E}_\mu\left[\left(\theta(A) - \theta^*(A)\right)^2\right].
\end{aligned}
$$

Lastly, we prove the higher-order smoothness of the risk:

$$
\begin{aligned}
&\mathcal{D}_g^2 \mathcal{D}_\theta L_{P^0}(\theta^*, \overline{g})[\theta - \theta^*, g - g_0, g - g_0] \\
&= \mathrm{E}_\mu\Big[\frac{d^3}{dt_1 dt_2 dt_3}\left(\theta^* + t_3(\theta - \theta^*) - \overline{\tau} - t_1(\tau - \theta_0) - t_2(\tau - \theta_0)\right)^2\Big]_{(0,0,0)} \\
&\quad + 2\mathrm{E}_{Q_X^0 \times \mu}\Big[\frac{d^3}{dt_1 dt_2 dt_3}\left(\frac{\overline{\xi} + t_1(\xi - \xi_0) + t_2(\xi - \xi_0)}{\xi_0}\right) \\
&\qquad \times \left(\theta^* + t_3(\theta - \theta^*) - \overline{\tau} - t_1(\tau - \theta_0) - t_2(\tau - \theta_0)\right) \\
&\qquad \times \left(\overline{\tau} + t_1(\tau - \theta_0) + t_2(\tau - \theta_0) - (\overline{m} + t_1(m - m_0) + t_2(m - m_0))\right)\Big]_{(0,0,0)} \\
&\quad + 2\mathrm{E}_{Q_X^0 \times \mu \times Q_{|A,X}^0}\Big[\frac{d^3}{dt_1 dt_2 dt_3}\left(\frac{\overline{\eta} + t_1(\eta - \eta_0) + t_2(\eta - \eta_0)}{\eta_0}\right)\left(\frac{\overline{w} + t_1(w - w_0) + t_2(w - w_0)}{w_0}\right) \\
&\qquad \times \left(\theta^* + t_3(\theta - \theta^*) - \overline{\tau} - t_1(\tau - \theta_0) - t_2(\tau - \theta_0)\right)\left(\overline{m} + t_1(m - m_0) + t_2(m - m_0) - Y\right)\Big]_{(0,0,0)} \\
&= 4\mathrm{E}_{Q_X^0 \times \mu}\Big[(\theta - \theta^*)\left(\frac{\xi - \xi_0}{\xi_0}\right)\left((\tau - \theta_0) - (m - m_0)\right)\Big]
\end{aligned}
$$

$$+ 4\mathrm{E}_{Q_X^0\times\mu}\left[(\theta-\theta^*)\left(\frac{\eta-\eta_0}{\eta_0}\frac{w-w_0}{w_0}(\overline{m}-m_0)+\frac{\overline{\eta}}{\eta_0}\frac{w-w_0}{w_0}(m-m_0)+\frac{\eta-\eta_0}{\eta_0}\frac{\overline{w}}{w_0}(m-m_0)\right)\right]$$

$$= 2\mathrm{E}_{Q_X^0\times\mu}\left[(\theta-\theta^*)\begin{bmatrix}\xi-\xi_0\\ \eta-\eta_0\\ w-w_0\\ m-m_0\\ \tau-\theta_0\end{bmatrix}^T \underbrace{\begin{bmatrix}0 & 0 & 0 & 0 & \frac{1}{\xi_0}\\ 0 & 0 & \frac{\overline{m}-m_0}{\eta_0 w_0} & \frac{\overline{w}}{\eta_0 w_0} & 0\\ 0 & \frac{\overline{m}-m_0}{\eta_0 w_0} & 0 & \frac{\overline{\eta}}{\eta_0 w_0} & 0\\ 0 & \frac{\overline{w}}{\eta_0 w_0} & \frac{\overline{\eta}}{\eta_0 w_0} & 0 & 0\\ \frac{1}{\xi_0} & 0 & 0 & 0 & 0\end{bmatrix}}_{:=\mathrm{A}}\begin{bmatrix}\xi-\xi_0\\ \eta-\eta_0\\ w-w_0\\ m-m_0\\ \tau-\theta_0\end{bmatrix}\right].$$

Here, we analyze the characteristic polynomial of A to identify its maximum eigenvalue:

$$\begin{aligned}\det(\lambda I - A) &= \left(\lambda^2-\frac{1}{\xi_0^2}\right)\cdot\begin{vmatrix}\lambda & -\frac{\overline{m}-m_0}{\eta_0 w_0} & -\frac{\overline{w}}{\eta_0 w_0}\\ -\frac{\overline{m}-m_0}{\eta_0 w_0} & \lambda & -\frac{\overline{\eta}}{\eta_0 w_0}\\ -\frac{\overline{w}}{\eta_0 w_0} & -\frac{\overline{\eta}}{\eta_0 w_0} & \lambda\end{vmatrix}\\ &= \lambda\left(\lambda^2-\frac{1}{\xi_0^2}\right)\left(\lambda^2-\frac{1}{\eta_0^2 w_0^2}\left\{\overline{\eta}^2+\overline{w}^2+(\overline{m}-m_0)^2\right\}\right)=0.\end{aligned}$$

This implies that the eigenvalues of A are given by

$$\left\{0, \pm\frac{1}{\xi_0}, \pm\frac{\sqrt{\overline{\eta}^2+\overline{w}^2+(\overline{m}-m_0)^2}}{\eta_0 w_0}\right\}.$$

We observe that the maximum eigenvalue is bounded above, since

$$\frac{1}{\xi_0}\le 1, \quad \frac{\sqrt{\overline{\eta}^2+\overline{w}^2+(\overline{m}-m_0)^2}}{\eta_0 w_0}\le \overline{\eta}+\overline{w}+|\overline{m}-m_0|\le M_\eta+M_w+2 := M_\lambda.$$

Since the operator norm $\|A\|$ is equal to the maximum eigenvalue in magnitude, it follows that $\|A\|\le M_\lambda$. Therefore, we conclude that

$$\begin{aligned}\left|\mathcal{D}_g^2\mathcal{D}_\theta L_P(\theta^*,\overline{g})\,[\theta-\theta^*, g-g_0, g-g_0]\right| &\le 2\,\mathrm{E}_\mu\left[(\theta-\theta^*)^2\right]^{1/2}\,\mathrm{E}_{Q_X^0\times\mu}\left[\left((g-g_0)^\top A(g-g_0)\right)^2\right]^{1/2}\\ &\le 2\,\mathrm{E}_\mu\left[(\theta-\theta^*)^2\right]^{1/2}\,\mathrm{E}_{Q_X^0\times\mu}\left[\|A\|^2\,\|g-g_0\|_2^4\right]^{1/2}\\ &\le 2M_\lambda\,\|\theta-\theta^*\|_{L^2(\mu)}\|g-g_0\|^2_{L^4(Q_X^0\times\mu;\ell^2)}.\end{aligned}$$

□

Based on the properties established in the previous lemma, we now derive a risk difference inequality at the true nuisance, showing that it is controlled by the target estimation error at $\widehat{g}$ together with the nuisance estimation error.

A.4 Oracle excess risk bound in terms of target and nuisance errors (proof of Lemma 1.1)

Proof. Our argument adapts the proof of Theorem 1 in Foster and Syrgkanis (2023). We first directly compute the difference in the risk function with respect to θ around θ^*, holding the nuisance function fixed at $\hat{g}$:

$$L_{P^0}(\hat{\theta}, \hat{g}) - L_{P^0}(\theta^*, \hat{g}) = \mathcal{D}_\theta L_{P^0}(\theta^*, \hat{g})[\hat{\theta} - \theta^*] + \|\hat{\theta} - \theta^*\|^2_{L^2(\mu)}. \tag{A.3}$$

We next expand the first-order term $\mathcal{D}_\theta L_{P^0}(\theta^*, \hat{g})[\hat{\theta} - \theta^*]$ around g_0 via Taylor expansion in g:

$$\begin{aligned}\mathcal{D}_\theta L_{P^0}(\theta^*, \hat{g})[\hat{\theta} - \theta^*] = \mathcal{D}_\theta L_{P^0}(\theta^*, g_0)[\hat{\theta} - \theta^*] + \mathcal{D}_g \mathcal{D}_\theta L_{P^0}(\theta^*, g_0)[\hat{\theta} - \theta^*, \hat{g} - g_0] \\ + \tfrac{1}{2}\mathcal{D}_g^2 \mathcal{D}_\theta L_{P^0}(\theta^*, \tilde{g})[\hat{\theta} - \theta^*, \hat{g} - g_0, \hat{g} - g_0]\end{aligned}$$

for some $\tilde{g}$ on the line segment between g_0 and $\hat{g}$. Substituting this into (A.3) yields:

$$\begin{aligned}L_{P^0}(\hat{\theta}, \hat{g}) - L_{P^0}(\theta^*, \hat{g}) = \mathcal{D}_\theta L_{P^0}(\theta^*, g_0)[\hat{\theta} - \theta^*] + \mathcal{D}_g \mathcal{D}_\theta L_{P^0}(\theta^*, g_0)[\hat{\theta} - \theta^*, \hat{g} - g_0] \\ + \frac{1}{2}\mathcal{D}_g^2 \mathcal{D}_\theta L_{P^0}(\theta^*, \tilde{g})[\hat{\theta} - \theta^*, \hat{g} - g_0, \hat{g} - g_0] + \|\hat{\theta} - \theta^*\|^2_{L^2(\mu)}.\end{aligned}$$

By the Neyman orthogonality property, the second term above vanishes. Then applying the smoothness bound on the third derivative and using Hölder and Young's inequality, we obtain:

$$\begin{aligned}&L_{P^0}(\hat{\theta}, \hat{g}) - L_{P^0}(\theta^*, \hat{g}) \\ &\quad \geq \mathcal{D}_\theta L_{P^0}(\theta^*, g_0)[\hat{\theta} - \theta^*] - M_\lambda \|\hat{\theta} - \theta^*\|_{L^2(\mu)} \cdot \|\hat{g} - g_0\|^2_{L^4(Q_X^0 \times \mu; \ell^2)} + \|\hat{\theta} - \theta^*\|^2_{L^2(\mu)} \\ &\quad \geq \mathcal{D}_\theta L_{P^0}(\theta^*, g_0)[\hat{\theta} - \theta^*] - \frac{1}{2}\|\hat{\theta} - \theta^*\|^2_{L^2(\mu)} - \frac{M_\lambda^2}{2}\|\hat{g} - g_0\|^4_{L^4(Q_X^0 \times \mu; \ell^2)} + \|\hat{\theta} - \theta^*\|^2_{L^2(\mu)},\end{aligned}$$

which implies

$$\|\hat{\theta} - \theta^*\|^2_{L^2(\mu)} \leq 2\left(L_{P^0}(\hat{\theta}, \hat{g}) - L_{P^0}(\theta^*, \hat{g})\right) - 2\mathcal{D}_\theta L_{P^0}(\theta^*, g_0)[\hat{\theta} - \theta^*] + M_\lambda^2 \|\hat{g} - g_0\|^4_{L^4(Q_X^0 \times \mu; \ell^2)}. \tag{A.4}$$

Finally, consider the risk difference with respect to θ around θ^* under g_0:

$$L_{P^0}(\hat{\theta}, g_0) - L_{P^0}(\theta^*, g_0) = \mathcal{D}_\theta L_{P^0}(\theta^*, g_0)[\hat{\theta} - \theta^*] + \|\hat{\theta} - \theta^*\|^2_{L^2(\mu)}.$$

Plugging (A.4) into the above yields

$$\begin{aligned}
&L_{P^0}(\hat{\theta}, g_0) - L_{P^0}(\theta^*, g_0) \\
&\quad \leq 2\left(L_{P^0}(\hat{\theta}, \hat{g}) - L_{P^0}(\theta^*, \hat{g})\right) - \mathcal{D}_\theta L_{P^0}(\theta^*, g_0)[\hat{\theta} - \theta^*] + M_\lambda^2 \|\hat{g} - g_0\|^4_{L^4(Q_X^0 \times \mu; \ell^2)} \\
&\quad \leq 2\left(L_{P^0}(\hat{\theta}, \hat{g}) - L_{P^0}(\theta^*, \hat{g})\right) + M_\lambda^2 \|\hat{g} - g_0\|^4_{L^4(Q_X^0 \times \mu; \ell^2)},
\end{aligned}$$

where the last inequality follows from the first-order optimality of θ^*. □

We have established the excess risk bound for the estimators in Algorithm 1.1. We now turn to Algorithm 1.2, for which the bound can be derived in a more direct manner by combining a local Rademacher complexity bound with the approximation error of kernel ridge regression. Finally, by appropriately selecting the cross-validation parameter c, we obtain a simplified bound.

A.5 Excess risk bound for Algorithm 1.2 (proof of Theorem 1.2)

Proof. First, we know that the approximation error is bounded as

$$L_{P^0}(\theta^*, g_0) - L_{P^0}(\theta_0, g_0) = \inf_{\theta \in \Theta} \|\theta_0 - \theta\|^2_{L^2(\mu)} \leq c^{2-\frac{1}{\alpha}} \|T_{\mathcal{K}}^\alpha \theta_0\|^{1/\alpha}_{\mathcal{H}},$$

by Theorem 1.1 of Smale and Zhou (2003). Next, we establish an excess risk bound for our estimator relative to θ^*. Before applying Theorem 1.3, we revisit the higher-order smoothness condition (d) in the proof of Theorem 1.1 to adjust it for the kernel ridge regression setting. By Lemma 5.1 of Mendelson and Neeman (2010), for all $\theta \in \Theta$,

$$\|\theta - \theta^*\|_{L^\infty(\mu)} \leq C_p \|\theta - \theta^*\|^p_{\mathcal{H}} \|\theta - \theta^*\|^{1-p}_{L^2(\mu)}, \tag{A.5}$$

where C_p is a constant depending on A, p, G in Condition 1.4(b). By Hölder's inequality and (A.5),

$$\left|\mathcal{D}_g^2 \mathcal{D}_{\boldsymbol{\theta}} L_P(\boldsymbol{\theta}^*, \overline{g})\left[\boldsymbol{\theta} - \boldsymbol{\theta}^*, g - g_0, g - g_0\right]\right| \leq 2\mathrm{E}_{Q_X^0 \times \mu}\left[\left|(\theta - \theta^*)(g - g_0)^T A (g - g_0)\right|\right]$$

$$
\begin{aligned}
&\leq 2\|(\theta-\theta^*)\|_{L^\infty(\mu)}\mathrm{E}_{Q_X^0\times\mu}\Big[\big|(g-g_0)^T A(g-g_0)\big|\Big] \\
&\leq 2M_\lambda\|(\theta-\theta^*)\|_{L^\infty(\mu)}\|g-g_0\|^2_{L^2(Q_X^0\times\mu,\ell^2)} \\
&\leq 2M_\lambda C_p\|\theta-\theta^*\|^p_{\mathcal{H}}\|\theta-\theta^*\|^{1-p}_{L^2(\mu)}\|g-g_0\|^2_{L^2(Q_X^0\times\mu,\ell^2)} \\
&\leq 2M_\lambda C_p c^p\|\theta-\theta^*\|^{1-p}_{L^2(\mu)}\|g-g_0\|^2_{L^2(Q_X^0\times\mu,\ell^2)} \\
&\leq 2M_\lambda C_p c^p\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|^{1-p}_{L^2(\mu)}\|g-g_0\|^2_{L^2(Q_X^0\times\mu,\ell^2)}.
\end{aligned}
$$

Hence, we apply Theorem 3 of Foster and Syrgkanis (2023) with the parameters

$$
K_2=2,\quad R=\sqrt{2},\quad \beta_1=2,\quad r=p,\quad \lambda=2,\quad \kappa=0,\quad B_1=B_{P^0}(\delta)^2,\quad B_2=(M_\lambda C_p c^p)^{\frac{2}{1+p}},
$$

resulting in the following bound of sample error:

$$
\begin{aligned}
L_{P^0}(\widehat{\theta},g_0)-L_{P^0}(\theta^*,g_0) &\leq C\left(B_{P^0}(\delta)^2\left(\delta_n^2+\frac{\log(\delta^{-1})}{n}\right)+(M_\lambda C_p c^p)^{\frac{2}{1+p}}\|\widehat{g}-g_0\|^{\frac{4}{1+p}}_{L^2(\ell^2,P^0)}\right) \\
&\lesssim B_{P^0}(\delta)^2\left(\delta_n^2+\frac{\log(\delta^{-1})}{n}\right)+c^{\frac{2p}{1+p}}\|\widehat{g}-g_0\|^{\frac{4}{1+p}}_{L^2(\ell^2,P^0)},
\end{aligned}
$$

with probability at least $1-\delta$, where C is a universal constant and δ_n is the critical radius. We will propose a particular solution of the equation defining the critical radius using Lemma 5 of Nie and Wager (2021). Given that $\mathcal{Z}_i=\epsilon_i, w=1, h=\theta$, the local Rademacher complexity is bounded as

$$
R_{n,\mathcal{Z}}(\Theta,\delta)\leq K\log(n)\frac{c^p\delta^{1-p}}{\sqrt{n}},
$$

where K is a constant. Hence, $\delta_n=Kc^{\frac{p}{1+p}}\log(n)^{\frac{1}{1+p}}n^{-\frac{1}{2(1+p)}}$ is a valid critical radius and if we plug this critical radius into the sample bound above,

$$
\begin{aligned}
L_{P^0}(\widehat{\theta},g_0)-L_{P^0}(\theta^*,g_0) &\lesssim B_{P^0}(\delta)^2\left(\delta_n^2+\frac{\log(\delta^{-1})}{n}\right)+c^{\frac{2p}{1+p}}\|\widehat{g}-g_0\|^{\frac{4}{1+p}}_{L^2(Q_X^0\times\mu;\ell^2)} \\
&\lesssim B_{P^0}(\delta)^2\left(c^{\frac{2p}{1+p}}(n\log(n)^{-2})^{-\frac{1}{1+p}}+\frac{\log(\delta^{-1})}{n}\right)+c^{\frac{2p}{1+p}}\|\widehat{g}-g_0\|^{\frac{4}{1+p}}_{L^2(Q_X^0\times\mu;\ell^2)},
\end{aligned}
$$

which is the first result of Theorem. If $\|\widehat{g}-g_0\|_{L^2(\ell^2,P^0)}=O_p(B_{P^0}(\delta)^{2/(1+p)}(n\log(n)^{-2})^{-1/4})$, the third term is absorbed into the first and the second is absorbed into the first for large n,

giving us

$$L_{P^0}(\widehat{\theta}, g_0) - L_{P^0}(\theta^*, g_0) \lesssim B_{P^0}(\delta)^2 c^{\frac{2p}{1+p}} (n \log(n)^{-2})^{-\frac{1}{1+p}}.$$

Then if we look at the excess risk which is the sum of sample error and approximation error, it becomes

$$\begin{aligned} L_{P^0}(\widehat{\theta}, g_0) - L_{P^0}(\theta_0, g_0) &= L_{P^0}(\widehat{\theta}, g_0) - L_{P^0}(\theta^*, g_0) + L_{P^0}(\theta^*, g_0) - L_{P^0}(\theta_0, g_0) \\ &\lesssim B_{P^0}(\delta)^2 c^{\frac{2p}{1+p}} (n \log(n)^{-2})^{-\frac{1}{1+p}} + c^{2-\frac{1}{\alpha}}. \end{aligned}$$

If we choose $c = (B_{P^0}(\delta)^{-2(1+p)} n \log(n)^{-2})^{\frac{\alpha}{p+(1-2\alpha)}}$, then the error bound becomes

$$L_{P^0}(\widehat{\theta}, g_0) - L_{P^0}(\theta_0, g_0) \lesssim (B_{P^0}(\delta)^{-2(1+p)} n \log(n)^{-2})^{-\frac{1-2\alpha}{p+(1-2\alpha)}},$$

which we desired. □

We are now ready to compare the upper bounds on the excess risk with and without data fusion. Since the difference between the two upper bounds is only up to a multiplicative factor of the Lipschitz constant, it suffices to compare the corresponding Lipschitz constants. We denote the Lipschitz constant without data fusion, in analogy to that with data fusion, as

$$\underline{B}_{P^0}(\delta) := 4(1+\delta)^2 \Big(1 + \underline{\xi}_0 + C_{\sigma,\delta}\, \eta_0 \, \|\underline{w}_0\|_\infty\Big).$$

A.6 Lipschitz constant decreases in data fusion (Lemma A.4)

Lemma A.4 (Lipschitz constant decreases in data fusion)**.** *Suppose the conditions of Theorem 1.3 hold. Then there exists $N(\delta) \in \mathbb{N}$ such that, for all $n \geq N(\delta)$, with probability at least $1 - \delta/2$, the following inequalities hold for every realization $z := (x, a, b, y, s)$ of $Z \sim \widetilde{P}^0$ and all $\theta_1, \theta_2 \in \Theta$:*

$$\begin{aligned} \big|\ell_{P^0}(\theta_1, \widehat{g}; z) - \ell_{P^0}(\theta_2, \widehat{g}; z)\big| \;&\leq\; B_{P^0}(\delta) \cdot \big\|\big(\theta_1(a) - \theta_2(a),\, \theta_1(b) - \theta_2(b)\big)\big\|_2\,, \\ \big|\underline{\ell}_{P^0}(\theta_1, \widehat{g}; z) - \underline{\ell}_{P^0}(\theta_2, \widehat{g}; z)\big| \;&\leq\; \underline{B}_{P^0}(\delta) \cdot \big\|\big(\theta_1(a) - \theta_2(a),\, \theta_1(b) - \theta_2(b)\big)\big\|_2\,. \end{aligned}$$

Moreover, the constants satisfy $B_{P^0}(\delta) \leq \underline{B}_{P^0}(\delta)$, with strict inequality whenever

$$P^0\big(S \in \mathcal{S}_X \setminus \mathcal{S}_Y\big) \;+\; \operatorname*{ess\,inf}_{(X,A)\sim P^0} P^0\big(S \in \mathcal{S}_Y \setminus \mathcal{S}_X \,\big|\, A, X, S \in \mathcal{S}_Y\big) \;>\; 0.$$

Proof of Lemma A.4. In the proof of Theorem 1.1, we showed that there exists $N(\delta)$ such that, for all $n \geq N(\delta)$, with probability at least $1 - \delta/2$,

$$\left|\ell_{P^0}(\theta_1, \widehat{g}; z) - \ell_{P^0}(\theta_2, \widehat{g}; z)\right| \;\leq\; B_{P^0}(\delta) \cdot \left\| \big(\theta_1(a) - \theta_2(a),\, \theta_1(b) - \theta_2(b)\big) \right\|_2 .$$

By a similar argument as in the proof of Theorem 1.1, for all $n \geq N(\delta)$,

$$\left|\underline{\ell}_{P^0}(\theta_1, \widehat{g}; z) - \underline{\ell}_{P^0}(\theta_2, \widehat{g}; z)\right| \;\leq\; \underline{B}_{P^0}(\delta) \cdot \left\| \big(\theta_1(a) - \theta_2(a),\, \theta_1(b) - \theta_2(b)\big) \right\|_2 ,$$

with probability at least $1 - \delta/2$. It remains to show $B_{P^0}(\delta) \leq \underline{B}_{P^0}(\delta)$ and to characterize when the inequality is strict.

Step 1. Comparison of ξ_0 terms. Since $S_X \cap S_Y \subset S_X$,

$$\xi_0 = \frac{1}{P^0(S \in S_X)} \;\leq\; \frac{1}{P^0(S \in S_X \cap S_Y)} = \underline{\xi}_0,$$

with strict inequality if $P^0(S \in S_X \setminus S_Y) > 0$.

Step 2. Comparison of $\eta_0 \|w_0\|_\infty$ terms. Let $T_1 := \mathcal{S}_X \cap \mathcal{S}_Y$ and $T_2 := \mathcal{S}_Y \setminus \mathcal{S}_X$. Fix a common dominating measure $\nu \times \mu$ for (X, A) and write all densities with respect to it. For any measurable $E \times F \subset \mathcal{A} \times \mathcal{X}$,

$$\begin{aligned} P^0(A \in E, X \in F, S \in \mathcal{S}_Y) &= P^0(A \in E, X \in F, S \in T_1) + P^0(A \in E, X \in F, S \in T_2) \\ &= \frac{1}{\underline{\eta}_0} \int_{E \times F} p^0(x \mid S \in T_1)\, p^0(a \mid x, S \in T_1)\, d(\nu \times \mu) \\ &\quad + \frac{1}{\eta_0} \int_{E \times F} p^0(a, x \mid S \in \mathcal{S}_Y)\, P^0\big(S \in T_2 \mid a, x, S \in \mathcal{S}_Y\big)\, d(\nu \times \mu), \end{aligned}$$

where we used

$$P^0(A \in E, X \in F, S \in T_2) = \frac{1}{\eta_0} \int_{E \times F} p^0(a, x \mid S \in \mathcal{S}_Y)\, P^0\big(S \in T_2 \mid a, x, S \in \mathcal{S}_Y\big)\, d(\nu \times \mu).$$

Define

$$\varepsilon_{P^0} := \operatorname*{ess\,inf}_{(A,X) \sim P^0} P^0\big(S \in T_2 \;\big|\; A, X, S \in \mathcal{S}_Y\big) \in [0, 1].$$

Then for $\nu \times \mu$-a.e. (a, x) with $S \in \mathcal{S}_Y$,

$$P^0\big(S \in T_2 \mid a, x, S \in \mathcal{S}_Y\big) \;\geq\; \varepsilon_{P^0}.$$

Hence, for all measurable $E \times F$,

$$\begin{aligned} P^0(A \in E, X \in F, S \in \mathcal{S}_Y) \geq \frac{1}{\underline{\eta}_0} \int_{E\times F} p^0(x \mid S \in T_1)\, p^0(a \mid x, S \in T_1)\, d(\nu \times \mu) \\ + \frac{\varepsilon_{P^0}}{\eta_0} \int_{E\times F} p^0(a, x \mid S \in \mathcal{S}_Y)\, d(\nu \times \mu). \end{aligned}$$

But also

$$P^0(A \in E, X \in F, S \in \mathcal{S}_Y) = \frac{1}{\eta_0} \int_{E\times F} p^0(x \mid S \in \mathcal{S}_Y)\, p^0(a \mid x, S \in \mathcal{S}_Y)\, d(\nu \times \mu).$$

Combining and rearranging,

$$\begin{aligned} &(1 - \varepsilon_{P^0})\, \frac{1}{\eta_0} \int_{E\times F} p^0(x \mid S \in \mathcal{S}_Y)\, p^0(a \mid x, S \in \mathcal{S}_Y)\, d(\nu \times \mu) \\ &\geq\; \frac{1}{\underline{\eta}_0} \int_{E\times F} p^0(x \mid S \in T_1)\, p^0(a \mid x, S \in T_1)\, d(\nu \times \mu). \qquad (*) \end{aligned}$$

Since $(*)$ holds for all measurable $E \times F$, we obtain the *pointwise* a.e. bound

$$(1-\varepsilon_{P^0})\, \frac{p^0(x \mid S \in \mathcal{S}_Y)\, p^0(a \mid x, S \in \mathcal{S}_Y)}{\eta_0} \;\geq\; \frac{p^0(x \mid S \in T_1)\, p^0(a \mid x, S \in T_1)}{\underline{\eta}_0} \quad \nu\times\mu\text{-a.e. } (x, a). \qquad (**)$$

Note that $p^0(x \mid S \in T_1) = p^0(x \mid S \in \mathcal{S}_X)$ by Condition 1.1(c).

Taking inverse of $(**)$ and multiply by $(1 - \varepsilon_{P^0})p^0(x \mid S \in T_1)$ yields

$$(1 - \varepsilon_{P^0})\, \underline{\eta}_0\, \underline{w}_0(x, a) \;\geq\; \eta_0\, w_0(x, a) \qquad Q_X^0 \times \mu\text{-a.e. } (x, a),$$

and taking essential suprema over (x, a) gives

$$(1 - \varepsilon_{P^0})\, \underline{\eta}_0\, \|\underline{w}_0\|_\infty \;\geq\; \eta_0\, \|w_0\|_\infty.$$

Step 3. Conclusion. Combining the comparisons for ξ_0 and for $\eta_0\|w_0\|_\infty$ shows

$$B_{P^0}(\delta) \;\le\; \underline{B}_{P^0}(\delta),$$

with strict inequality if $P^0(S \in S_X \setminus S_Y) > 0$ or $\varepsilon_{P^0} > 0$. This completes the proof. □

Given that the Lipschitz constant decreases under data fusion and that the remainder terms in the risk bounds coincide with and without data fusion, we can show that the risk is smaller under data fusion when applying the same estimation algorithm as Algorithm 1.2.

A.7 Data fusion improves the excess risk upper bound (proof of Theorem 1.3)

Proof. Under the assumptions of Theorem 1.2, there exists a universal constant $C > 0$ such that

$$L_{P^0}(\widehat{\theta}, g_0) \;\le\; C\left(B_{P^0}(\delta)^{-2(1+p)}\, n\, \log(n)^{-2}\right)^{-\frac{1-2\alpha}{p+(1-2\alpha)}}.$$

Similarly, since we apply the same estimation algorithm and the remainder terms coincide, the bound without data fusion satisfies

$$L_{P^0}(\underline{\theta}, g_0) \;\le\; C\left(\underline{B}_{P^0}(\delta)^{-2(1+p)}\, n\, \log(n)^{-2}\right)^{-\frac{1-2\alpha}{p+(1-2\alpha)}}.$$

Taking the ratio of the upper bounds and cancelling common $n, \log n$ factors yields

$$\left(\frac{1+\xi_0+C_{\sigma,\delta}\,\eta_0\|w_0\|_\infty}{1+\underline{\xi}_0+C_{\sigma,\delta}\,\underline{\eta}_0\|\underline{w}_0\|_\infty}\right)^{\frac{2(1+p)(1-2\alpha)}{p+(1-2\alpha)}} \;\le\; 1,$$

with strict inequality when

$$P^0\big(S \in \mathcal{S}_X \setminus \mathcal{S}_Y\big) \;+\; \operatorname*{ess\,inf}_{(X,A)\sim P^0} P^0\big(S \in \mathcal{S}_Y \setminus \mathcal{S}_X \mid A, X, S \in \mathcal{S}_Y\big) \;>\; 0.$$

This completes the proof. □

A.8 Minimax lower bound without data fusion (proof of Lemma 1.2)

Define

$$N_\delta := \frac{1}{4}\left(\sqrt{3\log\frac{2}{\delta} \;+\; 4\left(\frac{\delta}{32}\right)^{-1-\frac{(1-2\alpha)}{p}}} \;-\; \sqrt{3\log\frac{2}{\delta}}\right)^2. \tag{A.6}$$

As $\delta \to 0$, $N_\delta = [1 + o(1)] \, (\delta/32)^{-1-(1-2\alpha)/p}$.

Proof of Lemma 1.2. Let $\tau = (\tau_k)_{k=0}^{\infty}$ denote an estimator sequence, where, for each k, τ_k takes as input k i.i.d draws from Q and outputs an estimate of θ_Q; if $k = 0$, then θ_0 takes as input an empty set and returns an arbitrary value, such as the zero function. Define $\underline{\theta}_\tau$ to be the estimator based on n i.i.d draws from P that first selects the $K := \sum_{i=1}^n \mathbf{1}(S_i \in \mathcal{S}_{XY})$ observations from sources in $\mathcal{S}_{XY}$, and then applies τ_K to this subsample of observations; expressed in notation, $\underline{\theta}_\tau(\{(X_i, A_i, Y_i, S_i)\}_{i=1}^n) = \tau_K(\{(X_i, A_i, Y_i)\}_{i:S_i \in \mathcal{S}_{XY}})$. For a distribution Q on $\mathcal{X} \times \mathcal{A} \times \mathcal{Y}$, we denote by $Q^{n_{XY}}$ the product measure on $(\mathcal{X} \times \mathcal{A} \times \mathcal{Y})^{n_{XY}}$ corresponding to n_{XY} i.i.d. samples drawn from Q. Similarly, for $P \in \mathcal{P}$ we write P^n for the product measure on $(\mathcal{X} \times \mathcal{A} \times \mathcal{Y} \times \mathcal{S})^n$ of n i.i.d. copies from P. For any Q in the model $\mathcal{Q}$ defined in Condition 1.5, further define

$$\mathcal{P}(Q) := \left\{ P \in \mathcal{P} : P(X|S \in \mathcal{S}_X) = Q(X), P(Y|X, A, S \in \mathcal{S}_Y) = Q(Y|X, A), P(S \in \mathcal{S}_{XY}) = \tfrac{n_{XY}}{n} \right\},$$

where we have suppressed the dependence on n_{XY}/n in the notation. We also let $\|\cdot\|$ denote the $L^2(\mu)$ norm.

Part 1: Lower bounding the no-data-fusion minimax risk under sampling from P by the worst-case Bayes risk under sampling from Q. Fix $Q \in \mathcal{Q}$, $P \in \mathcal{P}(Q)$ such that $P(A|X, S = s) = Q(A|X)$ for all s, and an estimator sequence τ. For any k and $t > 0$,

$$P^n\{\|\underline{\theta}_\tau - \theta_Q\|^2 \geq t \mid K = k\} = Q^k\{\|\tau_k - \theta_Q\|^2 \geq t\},$$

where we use that $\{(X_i, A_i, Y_i)\}_{i:S_i \in \mathcal{S}_{XY}} | K = k$ is equal in distribution to an iid sample from Q^k. Above we use the convention that $Q^0\{\|\theta_0 - \theta_Q\|^2 \geq t\} = 1\{\|\theta_0 - \theta_Q\|^2 \geq t\}$. Hence, for $r \geq 0$,

$$\begin{aligned} P^n\{\|\underline{\theta}_\tau - \theta_Q\|^2 \geq t\} &\geq P^n\{\|\underline{\theta}_\tau - \theta_Q\|^2 \geq t, K \leq r\} \\ &= \sum_{k=0}^{r} Q^k\{\|\tau_k - \theta_Q\|^2 \geq t\} P^n\{K = k\} \end{aligned}$$

$$\geq \min_{k:0\leq k\leq r} Q^k\{\|\tau_k - \theta_Q\|^2 \geq t\}\ P^n\{K \leq r\}.$$

Taking a supremum over $P \in \mathcal{P}(Q)$, followed by a supremum over $Q \in \mathcal{Q}$ and an infimum over estimator sequences τ yields

$$\begin{aligned}&\inf_{\tau} \sup_{Q\in\mathcal{Q}} \sup_{P\in\mathcal{P}(Q):P(A|X,S)=Q(A|X)} P^n\{\|\underline{\theta}_\tau - \theta_Q\|^2 \geq t\} \\ &\quad\geq \inf_{\tau} \sup_{Q\in\mathcal{Q}} \min_{k:0\leq k\leq r} Q^k\{\|\tau_k - \theta_Q\|^2 \geq t\} \sup_{P\in\mathcal{P}(Q)} P^n\{K \leq r\}.\end{aligned}$$

Upper bounding the left-hand side shows that

$$\inf_{\tau} \sup_{Q\in\mathcal{Q}} \sup_{P\in\mathcal{P}(Q)} P^n\{\|\underline{\theta}_\tau - \theta_Q\|^2 \geq t\} \geq \inf_{\tau} \sup_{Q\in\mathcal{Q}} \min_{k:0\leq k\leq r} Q^k\{\|\tau_k - \theta_Q\|^2 \geq t\} \sup_{P\in\mathcal{P}(Q)} P^n\{K \leq r\}.$$

We will show that the supremum over P lower bounds by some function f applied to n and n_{XY}. Let $r := C_{\delta,v} n_{XY}$, where $C_{\delta,v}$ is the smallest constant makes r as a positive integer and $C_{\delta,v} \geq 1 + \sqrt{3\log(2/\delta)/n_{XY}}$. By choosing r in this way, a multiplicative Chernoff bound yields $P^n\{K \leq r\} \geq 1 - \delta/2$ for each $P \in \mathcal{P}(Q)$, and so the above shows that

$$\begin{aligned}&\inf_{\tau} \sup_{Q\in\mathcal{Q}} \sup_{P\in\mathcal{P}(Q)} P^n\{\|\underline{\theta}_\tau - \theta_Q\|^2 \geq t\} \\ &\geq (1-\delta/2) \inf_{\tau} \sup_{Q\in\mathcal{Q}} \min_{k:0\leq k\leq r} Q^k\{\|\tau_k - \theta_Q\|^2 \geq t\} \\ &\geq (1-\delta/2) \sup_{\Pi} \inf_{\tau} \min_{k:0\leq k\leq r} \int Q^k\{\|\tau_k - \theta_Q\|^2 \geq t\}\Pi(dQ), \qquad \text{(A.7)}\end{aligned}$$

where the final supremum is over priors Π on $\mathcal{Q}$; this inequality—which reflects that the minimax risk is lower bounded by the worst-case Bayes risk—is essentially a consequence of the maximin inequality (Wald, 1945). The right-hand side above can be simplified by using that, for any Π,

$$\inf_{\tau} \min_{k:0\leq k\leq r} \int Q^k\{\|\tau_k - \theta_Q\|^2 \geq t\}\Pi(dQ) \geq \inf_{\tau} \int Q^r\{\|\tau_r - \theta_Q\|^2 \geq t\}\Pi(dQ).$$

The inequality also trivially holds the other way, but that fact will not be used in this proof. To see why the above holds, fix an estimator sequence τ, and let $\overline{\tau}$ be the same sequence except with $\overline{\tau}_r$ an estimator that takes as input r observations and applies $\tau_{j^\star}$ to the first

$j^\star$ of them, where $j^\star := \arg\min_{j:0\le j\le r} \int Q^j\{\|\tau_j - \theta_Q\|^2 \ge t\}\Pi(dQ)$. This construction then implies that

$$\min_{k:0\le k\le r} \int Q^k\{\|\tau_k - \theta_Q\|^2 \ge t\}\Pi(dQ) = \int Q^r\{\|\overline{\tau}_r - \theta_Q\|^2 \ge t\}\Pi(dQ),$$

and taking an infimum over all possible estimators of r observation on the right, followed by over all estimator sequences on the left, implies the claim. Returning to (A.7), this shows that

$$\inf_{\tau} \sup_{Q\in\mathcal{Q}} \sup_{P\in\mathcal{P}(Q)} P^n\{\|\underline{\theta}_\tau - \theta_Q\|^2 \ge t\} \ge (1-\delta/2) \sup_{\Pi} \inf_{\tau} \int Q^r\{\|\tau_r - \theta_Q\|^2 \ge t\}\Pi(dQ). \quad \text{(A.8)}$$

Hence, we have shown that any lower bound on the Bayes risk under a least favorable prior in the problem with r observations drawn from Q also implies a lower bound on the minimax risk under sampling a random number, K, of observations from the target distribution.

Part 2: Using Fano's method to lower bound the worst-case Bayes risk. In Part 3, we will follow arguments from Zhang et al. (2023) to construct $\{Q_0, Q_1, \ldots, Q_M\} \subset \mathcal{Q}$ with $M \ge 2^{r^{p/[p+(1-2\alpha)]}/8}$ such that, for $\theta_j := \theta_{Q_j}$ and to-be-specified constants $c_\delta > 0$ and $\kappa := \delta/4$, the

(i) *dose-response functions are far apart:* $\|\theta_j - \theta_k\| \ge c_\delta\kappa\, r^{-\frac{1-2\alpha}{p+(1-2\alpha)}}$ for all $0 \le j < k \le M$.

(ii) *distributions are close together:* $\frac{1}{M+1}\sum_{i=1}^M D_{\mathrm{KL}}(Q_i^r, Q_0^r) \le \kappa \log M$.

Before giving this construction, we motivate why it will be useful. This motivation begins with Eq. 2.8 from Tsybakov (2009), which shows that, if we take $t = c_\delta\kappa\, r^{-\frac{1-2\alpha}{p+(1-2\alpha)}}/2$, then

$$Q_j^r\{\|\tau_r - \theta_j\|^2 \ge t\} \ge Q_j^r\{\psi_\tau^* \ne j\}, \quad j = 0, 1, \ldots, M,$$

where $\psi_\tau^* := \arg\min_{j\in\{0,1,\ldots,M\}} \|\tau^r - \theta_{Q_j}\|$ is a test based on τ that returns the index of a distribution most likely to have generated the data. Combining the fact that Π_M is a prior

with the above bound and the fact that ψ_τ^* is a test yields that the worst-case Bayes risk from (A.8) lower bounds as

$$\sup_{\Pi} \inf_{\tau} \int Q^r\{\|\tau_r - \theta_Q\|^2 \geq t\}\, \Pi(dQ)$$
$$\geq \inf_{\tau} \frac{1}{M+1} \sum_{j=0}^{M} Q_j^r\{\|\tau_r - \theta_j\|^2 \geq t\} \geq \inf_{\tau} \frac{1}{M+1} \sum_{j=0}^{M} Q_j^r\{\psi_\tau^* \neq j\} \geq \inf_{\psi} \frac{1}{M+1} \sum_{j=0}^{M} Q_j^r\{\psi \neq j\}.$$

The final infimum above is over all tests, which take as input r observations and return an element of $\{0, 1, \ldots, M\}$ indicating a guess of which product measure in $\{Q_0^r, Q_1^r, \ldots, Q_M^r\}$ they were generated by. Tsybakov (2009) refers to the right-hand side as the average probability of error. It can be lower bounded via Fano's method (Corollary 2.6 from that work), which, when combined with the above, yields that

$$\sup_{\Pi} \inf_{\tau} \int Q^r\{\|\tau_r - \theta_Q\|^2 \geq t\}\Pi(dQ) \geq \frac{\log(M+1) - \log 2}{\log M} - \kappa \geq 1 - \frac{\log 2}{\log M} - \kappa.$$

Plugging in $\kappa = \delta/4$ and $M \geq 2^{r^{p/[p+(1-2\alpha)]}/8}$ yields

$$\sup_{\Pi} \inf_{\tau} \int Q^r\{\|\tau_r - \theta_Q\|^2 \geq t\}\, \Pi(dQ) \geq 1 - \frac{\log 2}{\log 2^{r^{p/[p+(1-2\alpha)]}/8}} - \frac{\delta}{4} \geq 1 - 8r^{-p/[p+(1-2\alpha)]} - \frac{\delta}{4}.$$

Since $n_{XY} \geq N_\delta$ as defined in (A.6), $r := C_{\delta,v} n_{XY} \geq (\delta/32)^{-[p+(1-2\alpha)]/p}$, and so the right-hand side is lower bounded by $1 - \delta/2$. Plugging in our chosen t and recalling (A.8) yields that

$$\inf_{\tau} \sup_{Q \in \mathcal{Q}} \sup_{P \in \mathcal{P}(Q)} P^n \left\{ \|\underline{\theta}_\tau - \theta_Q\|^2 \geq \frac{c_\delta \delta}{8} r^{-\frac{1-2\alpha}{p+(1-2\alpha)}} \right\} \geq (1 - \delta/2)(1 - \delta/2) \geq 1 - \delta.$$

Plugging in the fact that $r := C_{\delta,v} n_{XY} \geq (1 + \sqrt{3\log(2/\delta)/n_{XY}}) n_{XY}$ yields that

$$\inf_{\tau} \sup_{Q \in \mathcal{Q}} \sup_{P \in \mathcal{P}(Q)} P^n \left\{ \|\underline{\theta}_\tau - \theta_Q\|^2 \geq \frac{c_\delta \delta}{8} (1 + \sqrt{3\log(2/\delta)/n_{XY}})^{-\frac{1-2\alpha}{p+(1-2\alpha)}} n_{XY}^{-\frac{1-2\alpha}{p+(1-2\alpha)}} \right\} \geq 1 - \delta.$$

Finally, plugging in the inequality $c_\delta/8 \geq c_{\mathcal{Q}}$ from the forthcoming (A.9) for the constant $c_{\mathcal{Q}}$ defined therein, we find that

$$\inf_{\tau} \sup_{Q \in \mathcal{Q}} \sup_{P \in \mathcal{P}(Q)} P^n \left\{ \|\underline{\theta}_\tau - \theta_Q\|^2 \geq c_{\mathcal{Q}} \delta (1 + \sqrt{3\log(2/\delta)/n_{XY}})^{-\frac{1-2\alpha}{p+(1-2\alpha)}} n_{XY}^{-\frac{1-2\alpha}{p+(1-2\alpha)}} \right\} \geq 1 - \delta.$$

The desired conclusion is at hand. Let $\underline{\theta}$ be the estimator from the lemma statement. Since it does not use data fusion, there exists $\underline{\tau}$ such that $\underline{\theta} = \underline{\theta}_{\underline{\tau}}$. Hence, by the above, for any $\epsilon > 0$ there exist $Q \in \mathcal{Q}$ and $P \in \mathcal{P}(Q)$ such that (1.3) holds with probability at least $1-\delta-\epsilon$ under sampling from P^n; in particular, this holds when $\epsilon = \delta$.

Part 3: Constructing $Q_0, Q_1, \ldots, Q_M$. We follow arguments from Zhang et al. (2023). Let $m := \lceil r^{p/[p+(1-2\alpha)]} \rceil$. By the Varshamov-Gilbert lemma (Lemma 2.9 Tsybakov, 2009), there exist binary vectors $w^{(0)}, \ldots, w^{(M)} \in \{0,1\}^m$ for some $M \geq 2^{m/8}$ such that $\sum_{l=1}^m (w_l^{(i)} - w_l^{(j)})^2 \geq \frac{m}{8}$ for all $i \neq j$. Let $\xi := C_\kappa m^{-[p+(1-2\alpha)]/p}$ for

$$C_\kappa := \min\left\{ 2^{-(1-2\alpha)/p} \frac{R}{G_1}, \ \frac{\kappa \overline{\sigma}^2 \log 2}{4}, \ \frac{1}{\left(\sup_l \|e_l\|_\infty\right)^2} \right\},$$

and define

$$\theta_0 \equiv 0, \quad \theta_i := \xi^{1/2} \sum_{l=1}^m w_l^{(i)} e_{m+l}, \quad i = 1, \ldots, M,$$

where $\{e_l\}$ are the orthonormal eigenfunctions of the integral operator $T_\mathcal{K}$, indexed so the corresponding eigenvalues are nonincreasing. Let Q_X be the marginal distribution of X under an arbitrarily chosen distribution in $\mathcal{Q}$—the particular choice made will play no role in this proof. Let Q_i be the distribution of (X, A, Y) that can be sampled from via $X \sim Q_X$, $A \mid X \sim \mu$, and $Y \mid X = x, A = a \sim \mathcal{N}(\theta_i(a), \overline{\sigma}^2)$ where $\overline{\sigma} := \min(\sigma, L)$, where σ and L are defined in Condition 1.2(b). Note that the dose response function of Q_i, θ_{Q_i}, is equal to θ_i. These distributions belong to the model $\mathcal{Q}$ since (i) their Gaussian errors easily satisfy the sub-exponential condition in Condition 1.2(b), (ii) $\mathrm{E}_{Q_i}[Y|A, X]$ is $Q_X^0 \times \mu$-a.s. bounded by 1 since $p \leq 1 - 2\alpha$ ensures $\xi^{1/2} \leq C_\kappa^{1/2} m^{-1} \leq 1/(m \sup_l \|e_l\|_\infty) = 1/(mM_e)$, and (iii) the source condition Condition 1.5 is also satisfied since

$$\|T_\mathcal{K}^\alpha \theta_i\|_\mathcal{H} = \xi \sum_{k=1}^m \sigma_{m+k}^{-(1-2\alpha)} \left(w_k^{(i)}\right)^2 \leq \xi \sum_{k=1}^m \sigma_{2m}^{-(1-2\alpha)} \leq \xi \sum_{k=1}^m G_1 (2m)^{\frac{1-2\alpha}{p}} = C_\kappa G_1 2^{\frac{1-2\alpha}{p}} \leq R.$$

Having constructed $Q_0, Q_1, \ldots, Q_M \in \mathcal{Q}$, it remains to show they satisfy (i) and (ii) from Part 2 of this proof. For (i), we use the orthonormality of $\{e_l\}$, the choice of Varshamov-Gilbert weights, and the fact that $\lceil r^{p/[p+(1-2\alpha)]} \rceil \leq 2r^{p/[p+(1-2\alpha)]}$ since $r \geq 1$ to show that, for

all $i \neq j$,

$$\|\theta_i - \theta_j\|^2_{L^2(\mu)} = \xi \sum_{l=1}^{m} \left(w_l^{(i)} - w_l^{(j)}\right)^2 \geq \xi \, \frac{m}{8} = \frac{(C_\kappa/\kappa)\kappa \; m^{-(1-2\alpha)/p}}{8}$$
$$\geq \frac{(C_\kappa/\kappa)\kappa \; r^{-\frac{1-2\alpha}{p+(1-2\alpha)}}}{8} 2^{-(1-2\alpha)/p} = c_\delta \kappa \; r^{-\frac{1-2\alpha}{p+(1-2\alpha)}},$$

where $c_\delta := \frac{(C_\kappa/\kappa)}{8} 2^{-(1-2\alpha)/p}$. The δ subscript indicates the dependence of c_δ on $\kappa = \delta/4$. When divided by 8 as at the end of Part 2 of this proof, this quantity lower bounds as

$$c_\delta/8 = \min\left\{ 2^{-2(1-2\alpha)/p} R/(16\delta G_1), \; (\overline{\sigma}^2 \log 2) 2^{-(1-2\alpha)/p}/256, \frac{2^{-(1-2\alpha)/p}}{16\delta \left(\sup_l \|e_l\|_\infty\right)^2} \right\}$$
$$\geq \min\left\{ 2^{-2(1-2\alpha)/p} R/(16 G_1), \; (\overline{\sigma}^2 \log 2) 2^{-(1-2\alpha)/p}/256, \frac{2^{-(1-2\alpha)/p}}{16 \left(\sup_l \|e_l\|_\infty\right)^2} \right\} =: c_{\mathcal{Q}}. \tag{A.9}$$

It remains to show (ii). For this, we use that the errors are Gaussian, which yields that, for any $i \neq 0$,

$$D_{\mathrm{KL}}(Q_i^r, Q_0^r) = \frac{r}{2\overline{\sigma}^2} \|\theta_i\|^2_{L^2(\mu)} = \frac{r\xi}{2\overline{\sigma}^2} \sum_{k=1}^{m} \left(w_k^{(i)}\right)^2 \leq \frac{r\xi m}{2\overline{\sigma}^2}$$
$$= \frac{r C_\kappa m^{-\frac{1-2\alpha}{p}}}{2\overline{\sigma}^2} \leq \frac{C_\kappa m}{2\overline{\sigma}^2} \leq \frac{4 C_\kappa \log M}{\overline{\sigma}^2 \log 2} \leq \kappa \log M.$$

As $i \neq 0$ was arbitrary, (ii) holds.

□

A.9 Sufficient conditions for data fusion to improve worst-case performance (proof of Theorem 1.4)

Proof. We will show that, for all n, n_{XY} large enough, the first term in (1.6) is at least $1-\delta$ and the second term is at most δ. For brevity, let $\rho := (1-2\alpha)/[p + (1-2\alpha)] \in (0,1)$.

Step 1. Show $\sup_{P \in \mathcal{P}_{XY}} P^n\big(L_P(\underline{\theta}, g_0) \geq \frac{c_{\mathcal{Q}}}{4} \delta \, n_{XY}^{-\rho}\big) \geq 1-\delta$. By Lemma 1.2, for all (n, n_{XY}) with $n \geq n_{XY} \geq N_1(\delta)$, there exists $P \in \mathcal{P}_{XY}$ with $P(S \in \mathcal{S}_X \cap \mathcal{S}_Y) = n_{XY}/n$ such that

$$P^n\left(L_P(\underline{\theta}, g_0) \;\geq\; \frac{c_{\mathcal{Q}}}{2} \delta \left(1 + \sqrt{3\log(2/\delta)/n_{XY}}\right)^{-\rho} n_{XY}^{-\rho} \right) \;\geq\; 1-\delta,$$

where $c_{\mathcal{Q}} > 0$ is a constant depending on $\mathcal{Q}$. Using that $(1+x)^{-\rho} \geq 1 - \rho x$ for $x \geq 0$, we have

$$\left(1 + \sqrt{3\log(2/\delta)/n_{XY}}\right)^{-\rho} \;\geq\; 1 - \rho\sqrt{3\log(2/\delta)/n_{XY}}.$$

Hence, if

$$n_{XY} \;\geq\; N_2(\delta) \;:=\; \left\lceil \frac{3\rho^2\,\log(2/\delta)}{\delta^2} \right\rceil,$$

then $\left(1 + \sqrt{3\log(2/\delta)/n_{XY}}\right)^{-\rho} \geq 1 - \delta$. Therefore, for all $n \geq n_{XY} \geq \max\{N_1(\delta), N_2(\delta)\}$,

$$P^n\Big(L_P(\underline{\theta}, g_0) \;\geq\; \frac{c}{2}\,\delta(1-\delta)\,n_{XY}^{-\rho}\Big) \;\geq\; 1 - \delta.$$

This shows that, for all $n \geq n_{XY} \geq \max\{N_1(\delta), N_2(\delta)\}$,

$$\sup_{P \in \mathcal{P}_{XY}} P^n\Big(L_P(\underline{\theta}, g_0) \;\geq\; \frac{c_{\mathcal{Q}}}{2}\,\delta(1-\delta)\,n_{XY}^{-\rho}\Big) \;\geq\; 1 - \delta,$$

and combining this with the fact that $\delta(1-\delta) \geq \delta/2$ for $\delta \in (0, 1/2]$ yields that, for all such n, n_{XY},

$$\sup_{P \in \mathcal{P}_{XY}} P^n\big(L_P(\underline{\theta}, g_0) \;\geq\; \tfrac{c_{\mathcal{Q}}}{4}\,\delta\,n_{XY}^{-\rho}\big) \;\geq\; 1 - \delta.$$

Step 2. Show $\sup_{P \in \mathcal{P}_{XY}} P^n\Big(L_P(\widehat{\theta}, g_0) > r(\delta, n)/\log(n)^{h\rho/2}\Big) < \delta$. To prove the result, we will apply Theorem 3 of Foster and Syrgkanis (2023) as we did in the proof of Theorem 1.2. We first establish that the loss is Lipschitz in its first argument with a Lipschitz constant uniform over $P \in \mathcal{P}_{XY}$. To see this, for any $P \in \mathcal{P}_{XY}$, $\theta_1, \theta_2 \in \Theta$, and realization $z := (x, a, b, y, s)$ of $Z \sim P \times \mu$,

$$\big|\ell_P(\theta_1, \widehat{g}; z) - \ell_P(\theta_2, \widehat{g}; z)\big| = \big\|\,(\theta_1(a) - \theta_2(a),\; \theta_1(b) - \theta_2(b))\,\big\|_2 \cdot B(z),$$

where $B(z)$ is defined as

$$B(z) := \left\| \begin{pmatrix} \theta_1(b) + \theta_2(b) - 2\widehat{\tau}(b) + 2\widehat{\xi}\,\mathbf{1}(s \in \mathcal{S}_X)\big(\widehat{\tau}(b) - \widehat{m}(b,x)\big) \\ 2\,\widehat{\eta}\,\widehat{w}(x,a)\,\mathbf{1}(s \in \mathcal{S}_Y)\big(\widehat{m}(x,a) - y\big) \end{pmatrix} \right\|_2.$$

We next derive a uniform bound on $B(z)$. First, observe that

$$B(z) \leq \Big|\theta_1(b) + \theta_2(b) - 2\widehat{\tau}(b) + 2\widehat{\xi}\,\mathbf{1}(s \in \mathcal{S}_X)\big(\widehat{\tau}(b) - \widehat{m}(b,x)\big)\Big|$$

$$
\begin{aligned}
&+ \left| 2\,\widehat{\eta}\,\widehat{w}(x,a)\,\mathbf{1}(s \in \mathcal{S}_Y)\big(\widehat{m}(x,a) - y\big) \right| \\
&\leq 2\|\theta\|_\infty + 2|\widehat{\tau}(b)| + 2\widehat{\xi}\,\big|\widehat{\tau}(b) - \widehat{m}(b,x)\big| + 2\widehat{\eta}\,\widehat{w}(x,a)\,\mathbf{1}(s \in \mathcal{S}_Y)\big(|\widehat{m}(x,a)| + |y|\big) \\
&\leq 4 + 4\widehat{\xi} + 2\widehat{\eta}\,\widehat{w}(x,a)\,\mathbf{1}(s \in \mathcal{S}_Y)\big(1 + |y|\big) \\
&\leq 4 + 4M_\xi + 2M_\eta M_w\,\mathbf{1}(s \in \mathcal{S}_Y)\,(1 + |y|) \qquad Q_X^0 \times \mu\text{-a.e.},
\end{aligned}
$$

by $\|\theta\|_\infty \leq 1$ and Condition 1.3. By Condition 1.2(b) and the same argument as in the proof of Theorem 1.1, with probability at least $1 - \delta/4$,

$$
\mathbf{1}(s \in \mathcal{S}_Y)\,\eta_0\,\|w_0\|_\infty\,|y| \;\leq\; C_{\sigma,\delta}\,\eta_0\,\|w_0\|_\infty.
$$

Since $\eta_0\,\|w_0\|_\infty \geq (M_\eta M_w)^{-1} > 0$, it follows that $\mathbf{1}(s \in \mathcal{S}_Y)\,|y| \leq C_{\sigma,\delta}$, which further implies

$$
B(z) \leq 4\big(1 + M_\xi + C_\delta\,M_\eta M_w\big) =: B_{\mathcal{P}} \qquad Q_X^0 \times \mu\text{-a.e.}
$$

Then we apply Theorem 3 of Foster and Syrgkanis (2023) with the parameters

$$
K_2 = 2, \quad R = \sqrt{2}, \quad \beta_1 = 2, \quad r = p, \quad \lambda = 2, \quad \kappa = 0, \quad B_1 = B_{\mathcal{P}}^2, \quad B_2 = (M_\lambda C_p c^p)^{\frac{2}{1+p}},
$$

and plug in the valid critical radius $\delta_n \propto c^{\frac{p}{1+p}}(n\log(n)^{-2})^{-\frac{1}{2+p}}$ resulting in the following bound:

$$
L_P(\widehat{\theta}, g_0) - L_P(\theta^*, g_0) \lesssim c^{\frac{2p}{1+p}}(n\log(n)^{-2})^{-\frac{1}{1+p}} + \frac{\log(\delta^{-1})}{n} + c^{\frac{2p}{1+p}}\|\widehat{g} - g_0\|_{L^2(Q_X\times\mu,\ell^2)}^{\frac{4}{1+p}}
$$

with probability at least $1 - 3\delta/4$. By (1.4), there exists $N_3(\delta) < \infty$ such that, for all $n \geq N_3(\delta)$, for all $P \in \mathcal{P}_{XY}$,

$$
\|\widehat{g} - g_0\|_{L^2(Q_X^0\times\mu;\ell^2)} \leq M_g\;(n\log(n)^{-2})^{-1/4},
$$

with probability at least $1 - \delta/4$. Then, the nuisance estimation error can be absorbed into the target estimation error, yielding, for all $n \geq N_3(\delta)$,

$$
L_P(\widehat{\theta}, g_0) - L_P(\theta^*, g_0) \lesssim c^{\frac{2p}{1+p}}(n\log(n)^{-2})^{-\frac{1}{1+p}}
$$

with probability at least $1-\delta$. Then for any $P \in \mathcal{P}_{XY}$, the excess risk with respect to the true CDRF θ_0 is bounded by

$$L_P(\widehat{\theta}, g_0) \leq C'(n \log(n)^{-2})^{-\rho},$$

by plugging in $c \propto (n \log(n)^{-2})^{\rho}$ and C' is a universal constant. Hence, for $n \geq N_3(\delta)$,

$$\sup_{P \in \mathcal{P}_{XY}} P^n(L_P(\widehat{\theta}, g_0) \leq C'(n \log(n)^{-2})^{-\rho}) \geq 1-\delta.$$

Next, using $n_{XY} \leq n/(\log n)^{2+h}$, we can rewrite the right-hand side in terms of n_{XY}: for some $N_4(\delta) < \infty$ and all $n \geq N_4(\delta)$,

$$C' \, (n \log(n)^{-2})^{-\rho} \leq \tfrac{c_{\mathcal{Q}}}{4} \, \delta \, n_{XY}^{-\rho} \, (\log n)^{-h\rho/2} = r(\delta, n) \, (\log n)^{-h\rho/2}.$$

Letting

$$N_\delta := \max\{N_1(\delta), N_2(\delta), N_3(\delta), N_4(\delta)\}, \tag{A.10}$$

we conclude

$$\sup_{P \in \mathcal{P}_{XY}} P^n\Big\{ L_P(\underline{\theta}, g_0) \geq r(\delta, n) \Big\} - \sup_{P \in \mathcal{P}_{XY}} P^n\Big\{ L_P(\widehat{\theta}, g_0) > r(\delta, n)/\log(n)^{\frac{h(1-2\alpha)}{2[p+(1-2\alpha)]}} \Big\} \; \geq \; 1 - 2\delta.$$

□

Appendix B

SUPPLEMENTARY MATERIALS FOR CHAPTER 2

B.1 Efficient influence operators for the examples of Section 2.3

B.1.1 Conditional variance (Item 1)

Derivation of the EIO. This is the conditional variance primitive of Appendix C.2.4 in Luedtke (2026), specialized to $u(z) = y$ and $Q = P^0$ (so that $Q_X = P^0_X$). □

B.1.2 Conditional causal effect functions (Example 2)

Derivation of the EIO. Writing $z = (x, a, y)$, for each treatment level $j \in \{0, 1\}$ let $\mu_{\cdot,j} : P \mapsto \big(x \mapsto \mu_{P,j}(x)\big)$ denote the arm-j outcome-regression function, viewed as a map into $L^2(P^0_X)$. As in the regression example (Example 5 / Appendix B.6 of Luedtke (2026)), $\mu_{\cdot,j}$ depends on P only through the conditional law $P_{Y|A,X}(\cdot \mid A = j, \cdot)$, so its local parameter picks up only the $Y \mid A, X$ score. The adjoint computation, carried out with the inverse-propensity change of measure $\mathbb{1}\{a = j\}/g_P(j \mid x)$, gives

$$\dot{\mu}^*_{P,j}(w)(z) = \frac{dP^0_X}{dP_X}(x)\, w(x)\, \frac{\mathbb{1}\{a = j\}}{g_P(j \mid x)}\big\{y - \mu_{P,j}(x)\big\}, \qquad w \in L^2(P^0_X),\ j \in \{0, 1\}.$$

Since m is continuously differentiable and $\nu(P) = m\big(\mu_{P,1}, \mu_{P,0}\big)$ pointwise, the chain rule gives, for any score s,

$$\dot{\nu}_P(s) = M_{\partial_b m}\, \dot{\mu}_{P,1}(s) + M_{\partial_c m}\, \dot{\mu}_{P,0}(s),$$

where, for $\varphi \in \{\partial_b m, \partial_c m\}$, the operator $M_\varphi : L^2(P^0_X) \to L^2(P^0_X)$ denotes multiplication by $x \mapsto \varphi\big(\mu_{P,1}(x), \mu_{P,0}(x)\big)$, bounded provided $\varphi\big(\mu_{P,1}, \mu_{P,0}\big) \in L^\infty(P^0_X)$. Taking adjoints and using that multiplication by a real-valued function is self-adjoint, $M^*_\varphi = M_\varphi$, we obtain

$$\dot{\nu}^*_P(w) = \dot{\mu}^*_{P,1}\big(\partial_b m \cdot w\big) + \dot{\mu}^*_{P,0}\big(\partial_c m \cdot w\big),$$

with $\partial_b m$ and $\partial_c m$ evaluated at $\big(\mu_{P,1}(\cdot), \mu_{P,0}(\cdot)\big)$. Substituting the two operators above yields

$$\dot{\nu}_P^*(w)(z) = w(x)\,\frac{dP_X^0}{dP_X}(x)\Bigg[\partial_b m\big(\mu_{P,1}(x), \mu_{P,0}(x)\big)\,\frac{\mathbb{1}\{a=1\}}{g_P(1 \mid x)}\{y - \mu_{P,1}(x)\} \\ + \partial_c m\big(\mu_{P,1}(x), \mu_{P,0}(x)\big)\,\frac{\mathbb{1}\{a=0\}}{g_P(0 \mid x)}\{y - \mu_{P,0}(x)\}\Bigg],$$

as claimed. The derivation requires the strong positivity condition $\operatorname{ess\,inf}_x g_P(a \mid x) > 0$ for $a \in \{0,1\}$ and the conditional second-moment condition $\operatorname{ess\,sup}_x \mathrm{E}_P[Y^2 \mid A = a, X = x] < \infty$ ensuring each $\dot{\mu}_{P,j}^*$ is well defined. □

B.1.3 Longitudinal modified treatment policies (Example 2.2)

Consider longitudinal data $Z = (L_1, A_1, L_2, A_2, \ldots, L_\tau, A_\tau, Y) \sim P \in \mathcal{P}$, as introduced in Díaz et al. (2023), where L_t denotes time-varying covariates, A_t the treatment at time t, and $Y = L_{\tau+1}$ the end-of-study outcome. Let H_t denote the history just before A_t, so that $H_1 = L_1$ and $H_{t+1} = (H_t, A_t, L_{t+1})$. For each t, let d_t be a known policy that may depend on the current natural value of treatment a_t and on the history h_t, but not on past natural treatment values, as in Díaz et al. (2023), and write $a_t^d := d_t(a_t, h_t)$ for the shifted treatment. Throughout, we assume that, for each t and every $P \in \mathcal{P}$:

(L1) the conditional law of A_t^d given H_t under P is absolutely continuous with respect to that of A_t given H_t, with density ratio

$$r_{P,t}(a_t, h_t) := \frac{g_{P,t}^d(a_t \mid h_t)}{g_{P,t}(a_t \mid h_t)} \quad \text{satisfying} \quad \operatorname{ess\,sup} r_{P,t} < \infty,$$

where $g_{P,t}(\cdot \mid h_t)$ and $g_{P,t}^d(\cdot \mid h_t)$ denote the conditional densities of A_t and A_t^d given $H_t = h_t$ under P; this is a quantitative form of the positivity condition of Díaz et al. (2023);

(L2) $P_{A_t,H_t}^0 \ll P_{A_t,H_t}$ with $\rho_t := dP_{A_t,H_t}^0 / dP_{A_t,H_t}$ satisfying $\operatorname{ess\,sup} \rho_t < \infty$;

(L3) $\operatorname{ess\,sup} \mathrm{E}_P[Y^2 \mid A_\tau, H_\tau] < \infty$.

These conditions ensure that the maps below are well defined on the indicated L^2 spaces. The function-valued sequential regressions $m_t : \mathcal{P} \to L^2(P^0_{A_t,H_t})$ are defined by the base case

$$m_\tau(P)(a_\tau, h_\tau) := \mathrm{E}_P\left[Y \mid A_\tau = a_\tau, H_\tau = h_\tau\right],$$

and the backward recursion, for $t = \tau - 1, \ldots, 1$,

$$m_t(P)(a_t, h_t) := \mathrm{E}_P\left[m_{t+1}(P)\left(A^d_{t+1}, H_{t+1}\right) \middle| A_t = a_t, H_t = h_t\right].$$

Equivalently, $m_t(P)$ is the composition of primitive maps

$$m_\tau(P) = \eta_\tau(P, 0), \qquad m_t(P) = \eta_t(P,\, \xi_{t+1}(P,\, m_{t+1}(P))) \quad (t < \tau),$$

where

$$\begin{aligned}
&\eta_\tau : \mathcal{P} \times \{0\} \to L^2(P^0_{A_\tau,H_\tau}),\ (P, 0) \mapsto \mathrm{E}_P[\, Y \mid (A_\tau, H_\tau) = \cdot\,],\\
&\xi_t : \mathcal{P} \times L^2(P^0_{A_t,H_t}) \to L^2(P^0_{A_t,H_t}),\ (P, u) \mapsto \left[(a_t, h_t) \mapsto u(a^d_t, h_t)\right],\\
&\eta_t : \mathcal{P} \times L^2(P^0_{A_{t+1},H_{t+1}}) \to L^2(P^0_{A_t,H_t}),\ (P, u) \mapsto \mathrm{E}_P[\, u(A_{t+1}, H_{t+1}) \mid (A_t, H_t) = \cdot\,].
\end{aligned}$$

For a fixed time point $t \in \{1, \ldots, \tau\}$, we consider the function-valued parameter $\nu_t : \mathcal{P} \to L^2(P^0_{A_t,H_t})$ defined by

$$\nu_t(P)(a_t, h_t) := m_t(P)\big(a^d_t,\, h_t\big) = \xi_t(P,\, m_t(P))\,(a_t, h_t),$$

so that ν_t admits the fully recursive primitive representation

$$\nu_t(P) = \xi_t\Big(P,\, \eta_t\Big(P,\, \xi_{t+1}\Big(P,\, \cdots \eta_{\tau-1}\big(P,\, \xi_\tau(P,\, \eta_\tau(P, 0))\,\big) \cdots\Big)\Big)\Big).$$

Under the identification conditions of Díaz et al. (2023) (sequential randomization and positivity), $\nu_t(P)(a_t, h_t)$ equals the mean counterfactual outcome had treatment followed the policy $(d_s)_{s=t}^{\tau}$ from time t onward, among units with natural treatment and history $(A_t, H_t) = (a_t, h_t)$. In particular, for $t = 1$, the marginal parameter $\mathrm{E}_P[\nu_1(P)(A_1, L_1)]$ recovers the scalar LMTP parameter of Díaz et al. (2023). Since the policies d_s are known and do not depend on P, each shift primitive ξ_s carries no score and contributes to the local parameter only through its function argument.

Lemma B.1 (Efficient influence operator of the LMTP curve)**.** *Suppose* (L1)–(L3) *hold, and adopt the convention* $m_{\tau+1}(P)\big(A^d_{\tau+1}, H_{\tau+1}\big) := Y$*. Then* ν_t *is pathwise differentiable at every* $P \in \mathcal{P}$*, and its efficient influence operator* $\dot{\nu}^*_{t,P} : L^2(P^0_{A_t,H_t}) \to L^2_0(P)$ *is, for any* $w \in L^2(P^0_{A_t,H_t})$*,*

$$\dot{\nu}^*_{t,P}(w)(z) = \omega_{P,w}(a_t, h_t) \sum_{k=t}^{\tau} \left(\prod_{j=t+1}^{k} r_{P,j}(a_j, h_j) \right) \Big\{ m_{k+1}(P)\big(a^d_{k+1}, h_{k+1}\big) - m_k(P)(a_k, h_k) \Big\},$$

where the empty product is one and

$$\omega_{P,w}(a_t, h_t) := r_{P,t}(a_t, h_t)\, \mathrm{E}_P\big[\rho_t(A_t, H_t)\, w(A_t, H_t) \;\big|\; A^d_t = a_t,\ H_t = h_t\big]\,.$$

Proof. Throughout, fix $P \in \mathcal{P}$ and suppress it from $m_k := m_k(P)$ where convenient.

Step 1: local parameter. Each primitive is totally pathwise differentiable in the sense of Luedtke (2026). Since the policies d_k are known, the shift $\xi_k(P, \cdot)$ does not depend on P and is a bounded linear map by (L1); it therefore carries no score, and its derivative in the function argument is the map itself. The conditional mean primitive η_k is covered by Table 1 and Appendix B.6 of Luedtke (2026). In its function argument, $\eta_k(P, \cdot)$ is again bounded and linear. To compute its partial local parameter in P, we freeze the function argument at its value $u = \xi_{k+1}(P, m_{k+1})$, so that $u(A_{k+1}, H_{k+1}) = m_{k+1}(A^d_{k+1}, H_{k+1})$ and $\eta_k(P, u) = m_k$. The map $P' \mapsto \eta_k(P', u)$ is then the outcome regression of the fixed variable $V := m_{k+1}(A^d_{k+1}, H_{k+1})$ on (A_k, H_k), and its local parameter takes the classical regression form. Indeed, along a submodel $\{P_\epsilon\}$ with score s, the conditional distribution of Z given $(A_k, H_k) = (a_k, h_k)$ has score $s(Z) - \mathrm{E}_P[s(Z) \mid a_k, h_k]$, and consequently

$$\begin{aligned} \frac{\partial}{\partial \epsilon}\bigg|_0 \mathrm{E}_{P_\epsilon}[V \mid a_k, h_k] &= \mathrm{E}_P\big[V\big\{s(Z) - \mathrm{E}_P[s(Z) \mid a_k, h_k]\big\} \,\big|\, a_k, h_k\big] \\ &= \mathrm{E}_P\big[\big\{V - m_k(a_k, h_k)\big\}\, s(Z) \,\big|\, a_k, h_k\big] =: D_{k,P}(s)(a_k, h_k), \end{aligned}$$

where the second equality holds because both centerings subtract the same quantity inside the conditional expectation. Under the convention $m_{\tau+1}(A^d_{\tau+1}, H_{\tau+1}) = Y$, the same display covers the base case $k = \tau$. Applying the chain rule (Lemma S3 of Luedtke, 2026) to the

composition $m_k = \eta_k\big(P, \xi_{k+1}(P, m_{k+1})\big)$, the total local parameter decomposes into the direct contribution $D_{k,P}$ and the contribution propagated through the two linear maps, which yields the recursion

$$\dot{m}_{k,P}(s) = D_{k,P}(s) + \big[\eta_k(P,\cdot) \circ \xi_{k+1}(P,\cdot)\big]\dot{m}_{k+1,P}(s), \qquad \dot{m}_{\tau+1,P}(s) := 0,$$

and $\dot{\nu}_{t,P}(s) = \xi_t\big(P, \dot{m}_{t,P}(s)\big)$.

Step 2: a change-of-variables identity. The adjoint computation in the next step repeatedly encounters expectations in which a function is evaluated at the shifted pair (A_k^d, H_k), whereas the building blocks from Step 1 are indexed by the natural pair (A_k, H_k). The following identity converts between the two at the cost of a density-ratio weight. For any P-integrable u and any bounded $\sigma(A_k, H_k)$-measurable random variable V,

$$\mathrm{E}_P\big[V\, u\big(A_k^d, H_k\big)\big] = \mathrm{E}_P\big[r_{P,k}(A_k, H_k)\, \bar{V}_k(A_k, H_k)\, u(A_k, H_k)\big]\,, \tag{B.1}$$

where $\bar{V}_k(a,h) := \mathrm{E}_P\big[V \mid A_k^d = a,\, H_k = h\big]$. To see this, first condition on $\sigma(A_k^d, H_k)$, which replaces V by $\bar{V}_k(A_k^d, H_k)$ and leaves an expectation of a function of (A_k^d, H_k) alone. Rewriting that expectation as an integral over the conditional law of A_k^d given H_k, and then changing to the conditional law of A_k given H_k, introduces precisely the density ratio $r_{P,k}$ and yields the right-hand side. Two special cases will be used below. When V is $\sigma(H_k)$-measurable, conditioning on the shifted pair leaves it unchanged, so $\bar{V}_k = V$ and (B.1) simplifies to $\mathrm{E}_P[V\, u(A_k^d, H_k)] = \mathrm{E}_P[r_{P,k}(A_k, H_k)\, V\, u(A_k, H_k)]$. In general, however, the conditional expectation in $\bar{V}_k$ does not disappear; applied at the initial time point with $V = \rho_t(A_t, H_t)\, w(A_t, H_t)$, it produces the leading weight $\omega_{P,w}$ of the lemma, and is precisely the adjoint of the treatment-shift primitive ξ_t.

Step 3: adjoint computation. Fix $w \in L^2(P^0_{A_t,H_t})$. Changing measure via (L2) and then applying (B.1) at $k = t$ with $V = \rho_t(A_t, H_t)\, w(A_t, H_t)$—the shifted evaluation $\dot{m}_{t,P}(s)(A_t^d, H_t)$ being well defined by (L1)—we obtain

$$\big\langle \dot{\nu}_{t,P}(s), w \big\rangle_{L^2(P^0_{A_t,H_t})} = \mathrm{E}_P\big[\rho_t(A_t, H_t)\, w(A_t, H_t)\, \dot{m}_{t,P}(s)\big(A_t^d, H_t\big)\big]$$

$$= \mathrm{E}_P[\omega_{P,w}(A_t, H_t)\, \dot{m}_{t,P}(s)(A_t, H_t)]\,,$$

with $\omega_{P,w} = r_{P,t} \cdot \bar{V}_t$ as displayed in the lemma. We now claim that, for every $k \in \{t, \ldots, \tau\}$ and every bounded $\sigma(A_k, H_k)$-measurable G,

$$\mathrm{E}_P[G\, \dot{m}_{k,P}(s)(A_k, H_k)] = \sum_{j=k}^{\tau} \mathrm{E}_P\left[G \prod_{i=k+1}^{j} r_{P,i}(A_i, H_i) \left\{ m_{j+1}\big(A_{j+1}^d, H_{j+1}\big) - m_j(A_j, H_j) \right\} s(Z) \right]. \tag{B.2}$$

The claim follows by backward induction on k. For the $D_{k,P}$ term of the recursion, the tower property gives $\mathrm{E}_P[G\, D_{k,P}(s)(A_k, H_k)] = \mathrm{E}_P[G\{m_{k+1}(A_{k+1}^d, H_{k+1}) - m_k(A_k, H_k)\}s(Z)]$, the $j = k$ summand. For the remaining term, the tower property gives

$$\mathrm{E}_P\big[G\, \big[\eta_k(P, \cdot) \circ \xi_{k+1}(P, \cdot)\big] \dot{m}_{k+1,P}(s)\, (A_k, H_k)\big] = \mathrm{E}_P\big[G\, \dot{m}_{k+1,P}(s)\big(A_{k+1}^d, H_{k+1}\big)\big]\,,$$

and since G is a function of H_{k+1}, identity (B.1) at $k+1$ applies in its simplified form, yielding $\mathrm{E}_P[G\, r_{P,k+1}(A_{k+1}, H_{k+1})\, \dot{m}_{k+1,P}(s)(A_{k+1}, H_{k+1})]$; the induction hypothesis with $G' := G \cdot r_{P,k+1}(A_{k+1}, H_{k+1})$, which is $\sigma(A_{k+1}, H_{k+1})$-measurable and bounded by (L1), completes the step. Applying (B.2) with $k = t$ and $G = \omega_{P,w}(A_t, H_t)$—noting that the accumulated weights $\omega_{P,w}(A_t, H_t) \prod_{i=t+1}^{j} r_{P,i}(A_i, H_i)$ are $\sigma(A_j, H_j)$-measurable—gives

$$\big\langle \dot{\nu}_{t,P}(s), w \big\rangle_{L^2(P^0_{A_t,H_t})} = \mathrm{E}_P\big[\dot{\nu}^*_{t,P}(w)(Z)\, s(Z)\big] \qquad \text{for all scores } s,$$

with $\dot{\nu}^*_{t,P}(w)$ as displayed in the lemma. Each summand of $\dot{\nu}^*_{t,P}(w)$ is conditionally mean zero given (A_k, H_k), since its weight is $\sigma(A_k, H_k)$-measurable and $\mathrm{E}_P[m_{k+1}(A_{k+1}^d, H_{k+1}) \mid A_k, H_k] = m_k(A_k, H_k)$; hence $\dot{\nu}^*_{t,P}(w) \in L^2_0(P)$.

Step 4: boundedness. It remains to verify that $\dot{\nu}^*_{t,P}$ is a bounded operator into $L^2(P)$. The second-moment condition (L3) propagates through the recursion: conditional means preserve essential bounds by Jensen's inequality, and shift evaluations do so by (L1), so each m_k, and hence each residual $m_{k+1}(A_{k+1}^d, H_{k+1}) - m_k(A_k, H_k)$, is essentially bounded. By Jensen's inequality and changes of measure using (L1) and (L2), $\|\omega_{P,w}\|_{L^2(P_{A_t,H_t})} \le C\, \|w\|_{L^2(P^0_{A_t,H_t})}$ for a constant C depending only on $\operatorname{ess\,sup} r_{P,t}$ and $\operatorname{ess\,sup} \rho_t$, and the accumulated ratios

$\prod_j r_{P,j}$ are essentially bounded. Hence each summand of $\dot{\nu}^*_{t,P}(w)$ has $L^2(P)$ norm bounded by a constant multiple of $\|w\|_{L^2(P^0_{A_t,H_t})}$, as required. □

B.2 Efficient influence operator for the $\mathcal{H}$-embedded parameter (proof of Lemma 2.1)

Proof. We apply Theorem 1 of Luedtke (2026) to establish pathwise differentiability of γ and derive its efficient influence operator. To this end, we apply the chain rule in Lemma S3 in Luedtke (2026) for our primitives. For the first primitive, we view $\theta_1(P,0) = \nu(P)$ as taking values in the fixed ambient space $L^2(P^0_X)$. Hence, when the adjoint of the differential of θ_1 is represented using the usual $L^2(P_X)$-efficient influence operator, a change-of-measure factor appears. Indeed, for $w \in L^2(P^0_X)$,

$$\langle \dot{\nu}_P(s), w\rangle_{L^2(P^0_X)} = \left\langle \dot{\nu}_P(s), w\frac{dP^0_X}{dP_X} \right\rangle_{L^2(P_X)}.$$

Therefore, whenever $w\, dP^0_X/dP_X \in L^2(P_X)$, the adjoint of the differential of θ_1 at $(P,0)$ is

$$\dot{\theta}^*_{1,P,0}(w) = \left(\dot{\nu}^*_P\left[w\frac{dP^0_X}{dP_X}\right], 0\right).$$

In particular, at $P_X = P^0_X$, $\dot{\theta}^*_{1,P^0,0}(w) = \big(\dot{\nu}^*_0(w), 0\big)$.

We now study the total pathwise differentiability of the projection primitive $\theta_2(P,u) := \Pi^{\perp}_P(u)$, where $\Pi^{\perp}_P$ denotes the $L^2(P_X)$-orthogonal projection onto $\mathcal{S}^{\perp}_P$. Equivalently, Π_P denotes the $L^2(P_X)$-orthogonal projection onto $\mathcal{S}$ and $\Pi^{\perp}_P(u) = u - \Pi_P(u)$. Although $\Pi^{\perp}_P(u)$ belongs to $\mathcal{S}^{\perp}_P$ as an element of $L^2(P_X)$, we view it as an element of the fixed ambient space $L^2(P^0_X)$.

Fix (P,u) and write $a := \Pi_P(u), v := \Pi^{\perp}_P(u) = u - a$. The element a is characterized by the normal equation

$$\langle u - a, q\rangle_{L^2(P_X)} = 0 \qquad \text{for all } q \in \mathcal{S}.$$

For generic $(P', u', a') \in \mathcal{P} \times \mathcal{B}(P_X) \times \mathcal{B}(P_X)$, define $F(P', u', a')(q) := \langle u' - a', q\rangle_{L^2(P'_X)}$ for any $q \in \mathcal{S}$. Then $F(P', u', \Pi_{P'}(u')) \equiv 0$. The map F is continuously Fréchet differentiable in (P', u', a') under the regularity conditions imposed below. Its derivative with respect to a'

at (P, u, a) is

$$D_a F(P, u, a)[h](q) = -\langle h, q \rangle_{L^2(P_X)}, \qquad h, q \in \mathcal{S}.$$

Since $\mathcal{S}$ is a Hilbert space with inner product inherited from $L^2(P_X)$, the map $h \mapsto \{q \mapsto \langle h, q \rangle_{L^2(P_X)}\}$ is a topological isomorphism from $\mathcal{S}$ onto $\mathcal{S}^*$. Hence $D_a F(P, u, a)$ is a Banach space isomorphism. By the implicit function theorem in Banach spaces (Lang, 1996, Theorem 1.5.9), the map $(P', u') \mapsto \Pi_{P'}(u')$ is Fréchet differentiable in a neighborhood of (P, u).

Let $(s, t) \in \dot{\mathcal{P}}_P \oplus \dot{\mathcal{U}}_u$, where s is a score direction at P and t is a perturbation of u in the fixed ambient space $L^2(P_X^0)$. We identify t with its P_X-equivalence-class representative whenever it appears inside an $L^2(P_X)$ inner product. Since θ_2 depends on P only through the induced marginal distribution P_X, the score enters through the marginal score

$$s_X(x) := E_P\{s(Z) \mid X = x\}.$$

Indeed, if $\{P_\epsilon : \epsilon \geq 0\}$ is a quadratic-mean differentiable path through P with score s, then the induced marginal path $\{(P_\epsilon)_X : \epsilon \geq 0\}$ is differentiable at P_X with score s_X. Hence, for every square-integrable function φ,

$$\left.\frac{d}{d\epsilon}\right|_{\epsilon=0} \int \varphi(x)\, d(P_\epsilon)_X(x) = \int \varphi(x) s_X(x)\, dP_X(x).$$

Now fix a when taking the derivative with respect to (P, u). For a path u_ϵ satisfying

$$\|u_\epsilon - u - \epsilon t\|_{L^2(P_X^0)} = o(\epsilon),$$

write

$$r_\epsilon := u_\epsilon - u - \epsilon t.$$

Then $\|r_\epsilon\|_{L^2(P_X^0)} = o(\epsilon)$, and hence, with $v := u - a = \Pi_P^\perp(u)$,

$$u_\epsilon - a = v + \epsilon t + r_\epsilon.$$

Under the standing local equivalence and bounded density-ratio conditions between P_X and P_X^0, this also implies $\|r_\epsilon\|_{L^2(P_X)} = o(\epsilon)$. Thus, in $L^2(P_X)$, $u_\epsilon - a = v + \epsilon t + o(\epsilon)$. Therefore,

for each $q \in \mathcal{S}$,

$$F(P_\epsilon, u_\epsilon, a)(q) = \int (u_\epsilon - a)(x) q(x)\, d(P_\epsilon)_X(x) = \int \{v(x) + \epsilon t(x) + o(\epsilon)\} q(x)\, d(P_\epsilon)_X(x).$$

Using the first-order expansion of the marginal measure,

$$d(P_\epsilon)_X(x) = \{1 + \epsilon s_X(x)\}\, dP_X(x) + o(\epsilon),$$

in the usual pathwise sense, we obtain

$$\begin{aligned} F(P_\epsilon, u_\epsilon, a)(q) &= \int \{v(x) + \epsilon t(x)\} q(x) \{1 + \epsilon s_X(x)\}\, dP_X(x) + o(\epsilon) \\ &= \int v(x) q(x)\, dP_X(x) + \epsilon \int t(x) q(x)\, dP_X(x) + \epsilon \int v(x) q(x) s_X(x)\, dP_X(x) + o(\epsilon). \end{aligned}$$

Since

$$F(P, u, a)(q) = \int v(x) q(x)\, dP_X(x),$$

it follows that

$$\begin{aligned} D_{(P,u)} F(P, u, a)[(s,t)](q) &= \left. \frac{d}{d\epsilon} \right|_{\epsilon=0} F(P_\epsilon, u_\epsilon, a)(q) \\ &= \int \{t(x) + v(x) s_X(x)\} q(x)\, dP_X(x) \\ &= E_{P_X}\left[\{t + v s_X\} q\right]. \end{aligned}$$

Thus the derivative with respect to (P, u) contains two first-order contributions: the perturbation of the argument u, which gives $E_{P_X}(tq)$, and the perturbation of the integration measure P_X, which gives $E_{P_X}(vqs_X)$. On the other hand, the derivative with respect to a is

$$D_a F(P, u, a)[h](q) = -E_{P_X}[hq].$$

Since the projection satisfies $F(P_\epsilon, u_\epsilon, \Pi_{P_\epsilon}(u_\epsilon))(q) = 0$ for all $q \in \mathcal{S}$, differentiating this identity at $\epsilon = 0$ gives

$$D_{(P,u)} F(P, u, a)[(s,t)](q) + D_a F(P, u, a)[\dot{a}](q) = 0,$$

where $\dot{a} := D\{\Pi_P u\}_{P,u}[(s,t)]$ is the total pathwise derivative of a. Therefore,

$$E_{P_X}\left[\{t + v s_X\} q\right] - E_{P_X}[\dot{a}\, q] = 0 \qquad \text{for all } q \in \mathcal{S}.$$

Equivalently,

$$\langle \dot{a}, q\rangle_{L^2(P_X)} = \langle t + vs_X, q\rangle_{L^2(P_X)} \qquad \text{for all } q \in \mathcal{S}.$$

Since $\dot{a} \in \mathcal{S}$, this characterizes $\dot{a}$ as the $L^2(P_X)$-orthogonal projection of $t + vs_X$ onto $\mathcal{S}$. Hence

$$D\{\Pi_P u\}_{P,u}[(s,t)] = \Pi_P\{t + vs_X\}.$$

Consequently, since $\theta_2(P,u) = u - \Pi_P u$,

$$\dot{\theta}_{2,P,u}(s,t) = t - \Pi_P\{t + vs_X\} = \Pi_P^{\perp} t - \Pi_P\{vs_X\}.$$

We next compute the Hermitian adjoint of $\dot{\theta}_{2,P,u}$ under the fixed ambient inner product on $L^2(P_X^0)$. Let $\rho_P(x) := \frac{dP_X^0}{dP_X}(x)$. For $w \in L^2(P_X^0)$, we have

$$\begin{aligned}\left\langle \dot{\theta}_{2,P,u}(s,t), w \right\rangle_{L^2(P_X^0)} &= E_{P_X^0}\left[\{t - \Pi_P(t + vs_X)\}w\right] \\ &= E_{P_X}\left[\{t - \Pi_P(t + vs_X)\}\rho_P w\right].\end{aligned}$$

Using the self-adjointness of Π_P in $L^2(P_X)$, this becomes

$$\begin{aligned}\left\langle \dot{\theta}_{2,P,u}(s,t), w \right\rangle_{L^2(P_X^0)} &= E_{P_X}\left[t\{\rho_P w - \Pi_P(\rho_P w)\}\right] - E_{P_X}\left[vs_X \Pi_P(\rho_P w)\right] \\ &= E_{P_X^0}\left[t\left\{w - \rho_P^{-1}\Pi_P(\rho_P w)\right\}\right] - E_P\left[s(Z)v(X)\Pi_P(\rho_P w)(X)\right].\end{aligned}$$

Since scores belong to the mean-zero tangent space, the score component of the adjoint must be represented by a mean-zero function. Hence we write

$$-E_P\left[s(Z)v(X)\Pi_P(\rho_P w)(X)\right] = E_P\left[s(Z)\left\{-\left[v(X)\Pi_P(\rho_P w)(X) - E_P\{v(X)\Pi_P(\rho_P w)(X)\}\right]\right\}\right].$$

Therefore, the Hermitian adjoint of $\dot{\theta}_{2,P,u}$, viewed as an operator $\dot{\mathcal{P}}_P \oplus L^2(P_X^0) \to L^2(P_X^0)$, is

$$\dot{\theta}_{2,P,u}^*(w) = \left(-\left[v(X)\Pi_P(\rho_P w)(X) - E_P\{v(X)\Pi_P(\rho_P w)(X)\}\right],\ \rho_P^{-1}\Pi_P^{\perp}(\rho_P w)\right).$$

Here $v = \Pi_P^{\perp}(u)$. Since $v \in \mathcal{S}_P^{\perp}$ and $\Pi_P(\rho_P w) \in \mathcal{S}$, we have

$$E_P\{v(X)\Pi_P(\rho_P w)(X)\} = \langle v, \Pi_P(\rho_P w)\rangle_{L^2(P_X)} = 0.$$

Nevertheless, we retain the centering term to emphasize that the first component of the adjoint belongs to the mean-zero tangent space.

We next derive the total pathwise derivative of θ_3. Recall that

$$\theta_3(P,u) = \left[x' \mapsto \int K(x',x)u(x)\,dP_X(x)\right] \in \mathcal{H}.$$

Let $s \in \dot{\mathcal{P}}_P$ be a score direction and let $t \in L^2(P_X^0)$ be a perturbation of u. Then, along any path with score s and perturbation t,

$$\dot{\theta}_{3,P,u}(s,t) = \left[x' \mapsto \int K(x',x)\{t(x) + u(x)s_X(x)\}\,dP_X(x)\right].$$

Equivalently,

$$\dot{\theta}_{3,P,u}(s,t) = \int K(\cdot,x)\{t(x) + u(x)s_X(x)\}\,dP_X(x).$$

For $w \in \mathcal{H}$, the reproducing property gives

$$\begin{aligned}\left\langle \dot{\theta}_{3,P,u}(s,t), w\right\rangle_{\mathcal{H}} &= E_{P_X}\left[\{t(X) + u(X)s_X(X)\}w(X)\right] \\ &= E_{P_X}\{t(X)w(X)\} + E_P\{s(Z)u(X)w(X)\}.\end{aligned}$$

Since the perturbation t is represented in the fixed ambient space $L^2(P_X^0)$,

$$E_{P_X}\{t(X)w(X)\} = E_{P_X^0}\left[t(X)\rho_P(X)^{-1}w(X)\right].$$

Therefore, the Hermitian adjoint of $\dot{\theta}_{3,P,u}$, viewed as an operator $\dot{\theta}_{3,P,u} : \dot{\mathcal{P}}_P \oplus L^2(P_X^0) \to \mathcal{H}$, is

$$\dot{\theta}^*_{3,P,u}(w) = \left(uw - E_P\{u(X)w(X)\},\ \rho_P^{-1}w\right).$$

The first component is centered so that it belongs to the mean-zero tangent space. Applying the chain rule for total pathwise differentiability, we now derive the efficient influence operator of $\gamma(P) = \theta_3(P, \theta_2(P, \theta_1(P,0)))$. For notational convenience, define

$$h_1 := \theta_1(P,0) = \nu(P), \qquad h_2 := \theta_2(P,h_1) = \Pi_P^{\perp}h_1,$$

so that $\gamma(P) = \theta_3(P,h_2)$. Let $w \in \mathcal{H}$, and write $[w]_{P_X}$ simply as w when no confusion can arise. By the adjoint form of the chain rule,

$$\dot{\gamma}^*_P(w) = \dot{\theta}^*_{3,P,h_2}(w)_0 + \dot{\theta}^*_{2,P,h_1}\left(\dot{\theta}^*_{3,P,h_2}(w)_1\right)_0 + \dot{\theta}^*_{1,P,0}\left[\dot{\theta}^*_{2,P,h_1}\left(\dot{\theta}^*_{3,P,h_2}(w)_1\right)_1\right].$$

Here the subscripts 0 and 1 denote, respectively, the score component and the second-argument component of the adjoint. We first apply the adjoint of the kernel embedding primitive. Since

$$\dot{\theta}^*_{3,P,h_2}(w) = \big(h_2(X)w(X) - E_P\{h_2(X)w(X)\}, \rho_P^{-1}w\big)\,,$$

the first contribution to $\dot{\gamma}^*_P(w)$ is $h_2(X)w(X) - E_P\{h_2(X)w(X)\}$. Next, apply the adjoint of the projection primitive to $\rho_P^{-1}w$. Since $h_2 = \Pi_P^{\perp}(h_1)$,

$$\dot{\theta}^*_{2,P,h_1}(\rho_P^{-1}w) = \big(-\left[h_2(X)\Pi_P(w)(X) - E_P\{h_2(X)\Pi_P(w)(X)\}\right], \rho_P^{-1}\Pi_P^{\perp}(w)\big)\,.$$

The first component contributes

$$-h_2(X)\Pi_P(w)(X) + E_P\{h_2(X)\Pi_P(w)(X)\}.$$

The second component is passed through the adjoint of θ_1. Because $\theta_1(P,0) = \nu(P)$ is viewed as taking values in the fixed ambient space $L^2(P_X^0)$, $\dot{\theta}^*_{1,P,0}(r) = \dot{\nu}^*_P(\rho_P r)$. Thus, with $r = \rho_P^{-1}\Pi_P^{\perp}(w)$, $\dot{\theta}^*_{1,P,0}\left\{\rho_P^{-1}\Pi_P^{\perp}(w)\right\} = \dot{\nu}^*_P\{\Pi_P^{\perp}(w)\}$. Combining the three contributions yields

$$\dot{\gamma}^*_P(w)(z) = \Pi_P^{\perp}\{\nu(P)\}(x)\,\Pi_P^{\perp}(w)(x) - E_{P_X}\left[\Pi_P^{\perp}\{\nu(P)\}(X)\,\Pi_P^{\perp}(w)(X)\right] + \dot{\nu}^*_P\{\Pi_P^{\perp}(w)\}(z),$$

using $E_{P_X}\{h_2(X)\Pi_P(w)(X)\} = 0$. □

B.3 Efficient influence function for the $\mathcal{H}$-embedded parameter (proof of Theorem 2.1)

Proof. By Lemma 2.1, γ is pathwise differentiable at P and has efficient influence operator $\dot{\gamma}^*_P : \mathcal{H} \to \dot{\mathcal{P}}_P$. If an $\mathcal{H}$-valued EIF ϕ_P exists, then it must satisfy

$$\dot{\gamma}^*_P(w)(z) = \langle w, \phi_P(z)\rangle_{\mathcal{H}} \qquad \text{for every } w \in \mathcal{H}.$$

Taking $w = K_{x'}$ and using the reproducing property gives

$$\phi_P(z)(x') = \langle K_{x'}, \phi_P(z)\rangle_{\mathcal{H}} = \dot{\gamma}^*_P(K_{x'})(z).$$

Substituting $w = K_{x'}$ into the expression for $\dot{\gamma}_P^*$ in Lemma 2.1 gives (2.5). The assumed square-integrability condition ensures that the pointwise representer belongs to $L_0^2(P; \mathcal{H})$. Therefore, by the Hilbert-valued influence-function characterization in Theorem 1 of Luedtke and Chung (2024), ϕ_P is the efficient influence function for γ. □

B.4 Efficient influence operator for the $L^2(P_X^0)$-embedded parameter (proof of Lemma 2.2)

Proof. We first compute the total pathwise derivative of θ_4. Since θ_4 acts on its second argument as the fixed inclusion map $\zeta_0 : \mathcal{H} \to L^2(P_X^0)$, $\zeta_0(u) = [u]_{P_X^0}$, its first-order variation along a direction $(s, t) \in \dot{\mathcal{P}}_P \oplus \mathcal{H}$ is

$$\dot{\theta}_{4,P,u}(s, t) = [t]_{P_X^0}.$$

Thus the score direction s does not contribute to the derivative of θ_4 under this fixed-ambient-space representation. The adjoint of $\dot{\theta}_{4,P,u}$ is obtained as follows. For $w \in L^2(P_X^0)$,

$$\begin{aligned}\left\langle \dot{\theta}_{4,P,u}(s, t), w \right\rangle_{L^2(P_X^0)} &= \left\langle [t]_{P_X^0}, w \right\rangle_{L^2(P_X^0)} = \int t(x) w(x)\, dP_X^0(x) \\ &= \left\langle t, \int K(\cdot, x) w(x)\, dP_X^0(x) \right\rangle_{\mathcal{H}}.\end{aligned}$$

Since there is no term involving s, the Hermitian adjoint is $\dot{\theta}_{4,P,u}^*(w) = (0, T_{K,P^0} w)$. Now apply the adjoint form of the chain rule to $\Gamma(P) = \theta_4\{P, \gamma(P)\}$. For $w \in L^2(P_X^0)$,

$$\dot{\Gamma}_P^*(w) = \dot{\theta}_{4,P,\gamma(P)}^*(w)_0 + \dot{\gamma}_P^* \left\{ \dot{\theta}_{4,P,\gamma(P)}^*(w)_1 \right\} = \dot{\gamma}_P^*(T_{K,P^0} w).$$

Substituting $T_{K,P^0} w$ into the expression for $\dot{\gamma}_P^*$ from Lemma 2.1 yields

$$\begin{aligned}\dot{\Gamma}_P^*(w)(z) = \Pi_P^\perp\{\nu(P)\}(x)\, \Pi_P^\perp([T_{K,P^0} w]_{P_X})(x) - E_{P_X}\left[\Pi_P^\perp\{\nu(P)\}(X)\, \Pi_P^\perp([T_{K,P^0} w]_{P_X})(X)\right] \\ + \dot{\nu}_P^* \left\{ \Pi_P^\perp([T_{K,P^0} w]_{P_X}) \right\}(z).\end{aligned}$$

□

B.5 *Efficient influence function for the $L^2(P_X^0)$-embedded parameter (proof of Theorem 2.2)*

Proof. By Lemma 2.2, for every $w \in L^2(P_X^0)$,

$$\dot{\Gamma}_P^*(w)(z) = \dot{\Gamma}_{P,12}^*(w)(z) + \dot{\Gamma}_{P,3}^*(w)(z),$$

where

$$\dot{\Gamma}_{P,12}^*(w)(z) = \Pi_P^\perp\{\nu(P)\}(x)\,\Pi_P^\perp\big([T_{K,P^0}w]_{P_X}\big)(x) - E_{P_X}\left[\Pi_P^\perp\{\nu(P)\}(X)\,\Pi_P^\perp\big([T_{K,P^0}w]_{P_X}\big)(X)\right],$$
$$\dot{\Gamma}_{P,3}^*(w)(z) = \dot{\nu}_P^*\big(\Pi_P^\perp([T_{K,P^0}w]_{P_X})\big)(z).$$

Step 1 (form of the representer). The element $T_{K,P^0}w = \int K(\cdot,x)\,w(x)\,dP_X^0(x)$ belongs to $\mathcal{H}$: under the bounded-kernel condition, $\int \|K(\cdot,x)\|_{\mathcal{H}}\,|w(x)|\,dP_X^0(x) = \int \sqrt{K(x,x)}\,|w(x)|\,dP_X^0(x) < \infty$ by Cauchy–Schwarz, so the integrand is Bochner integrable in $\mathcal{H}$. Viewing this element in $L^2(P_X)$ through the inclusion $\mathcal{H} \hookrightarrow L^2(P_X)$ (valid since $P_X \sim P_X^0$), the same integral is Bochner integrable in $L^2(P_X)$, because $\int \big\|[K(\cdot,x)]_{P_X}\big\|_{L^2(P_X)}\,|w(x)|\,dP_X^0(x) < \infty$. Since $\Pi_P^\perp$ is a continuous linear operator on $L^2(P_X)$, it commutes with this $L^2(P_X)$-valued Bochner integral, giving

$$\Pi_P^\perp\big([T_{K,P^0}w]_{P_X}\big) = \int \Pi_P^\perp\{[K(\cdot,x')]_{P_X}\}\,w(x')\,dP_X^0(x') \qquad \text{in } L^2(P_X).$$

Substituting this identity and applying Fubini's theorem yields

$$\dot{\Gamma}_{P,12}^*(w)(z) = \langle \Phi_{P,12}(z), w\rangle_{L^2(P_X^0)}.$$

For the third term, define the bounded linear operator $A_P : L^2(P_X^0) \to \mathcal{S}_P^\perp$ by $A_P w := \Pi_P^\perp([T_{K,P^0}w]_{P_X})$, so that $\psi_{P,k} = A_P e_k$. Under the bounded kernel and bounded density ratio conditions of Lemma 2.2, A_P is bounded; since $\dot{\nu}_P^*$ is bounded and linear, so is $w \mapsto \dot{\nu}_P^*(A_P w)$. Expanding $w = \sum_k \langle w, e_k\rangle_{L^2(P_X^0)} e_k$ and using continuity,

$$\dot{\Gamma}_{P,3}^*(w)(\cdot) = \sum_{k=1}^{\infty} \langle w, e_k\rangle_{L^2(P_X^0)}\,\dot{\nu}_P^*(\psi_{P,k})(\cdot) \qquad \text{in } L^2(P). \tag{B.3}$$

Step 2 (existence and integrability). Write $c_k(z) := \dot{\nu}_P^*(\psi_{P,k})(z)$. By Tonelli's theorem, (2.7) implies $\sum_k c_k(z)^2 < \infty$ for P-almost every z, so $\Phi_{P,3}(z) = \sum_k c_k(z) e_k$ converges in $L^2(P_X^0)$ for P-almost every z, and by Parseval's identity

$$\|\Phi_{P,3}(z)\|^2_{L^2(P_X^0)} = \sum_{k=1}^{\infty} c_k(z)^2, \qquad E_P\|\Phi_{P,3}(Z)\|^2_{L^2(P_X^0)} = E_P\left[\sum_{k=1}^{\infty} c_k(Z)^2\right] < \infty.$$

For such z, continuity of the inner product gives $\langle \Phi_{P,3}(z), w\rangle_{L^2(P_X^0)} = \sum_k \langle w, e_k\rangle_{L^2(P_X^0)} c_k(z)$, and by Cauchy–Schwarz

$$\left|\dot{\Gamma}^*_{P,3}(w)(z)\right| \le \|\Phi_{P,3}(z)\|_{L^2(P_X^0)}\, \|w\|_{L^2(P_X^0)}.$$

Hence $w \mapsto \dot{\Gamma}^*_{P,3}(w)(z)$ is a bounded linear functional with Riesz representation $\Phi_{P,3}(z)$; matching with (B.3) shows $\dot{\Gamma}^*_{P,3}(w) = \langle \Phi_{P,3}(\cdot), w\rangle_{L^2(P_X^0)}$ in $L^2(P)$.

Finally, $\Phi_{P,12}(z)$ depends on z only through $x = f(z)$, is manifestly $L^2(P_X^0)$-valued, and under the bounded-kernel condition together with $\nu(P) \in \mathcal{B}(P_X^0)$ satisfies $E_P\|\Phi_{P,12}(Z)\|^2_{L^2(P_X^0)} < \infty$ and $\dot{\Gamma}^*_{P,12}(w) = \langle \Phi_{P,12}(\cdot), w\rangle_{L^2(P_X^0)}$ in $L^2(P)$. Therefore $\Phi_P = \Phi_{P,12} + \Phi_{P,3} \in L_0^2(P; L^2(P_X^0))$ represents $\dot{\Gamma}^*_P$, i.e. $\dot{\Gamma}^*_P(w) = \langle \Phi_P(\cdot), w\rangle_{L^2(P_X^0)}$ in $L^2(P)$ for every w, so Φ_P is the efficient influence function of Γ at P. □

B.6 Weak convergence of $\widehat{\gamma}_n$ (proof of Theorem 2.3)

Proof. By the decomposition preceding Condition 2.1,

$$\widehat{\gamma}_n - \gamma(P^0) - P_n\phi_0 = \mathcal{R}_n^\gamma + \mathcal{D}_n^\gamma,$$

where $\mathcal{R}_n^\gamma = \gamma(\widehat{P}_n) + P^0\phi_n - \gamma(P^0)$ and $\mathcal{D}_n^\gamma = (P_n - P^0)(\phi_n - \phi_0)$. Condition (i)–(ii) give $\|\mathcal{R}_n^\gamma\|_{\mathcal{H}} = o_p(n^{-1/2})$, as the remainder is second-order (see the discussion following the condition), and Condition (iii) with sample splitting gives $\|\mathcal{D}_n^\gamma\|_{\mathcal{H}} = o_p(n^{-1/2})$. Hence $\widehat{\gamma}_n - \gamma(P^0) = P_n\phi_0 + o_p(n^{-1/2})$, establishing asymptotic linearity and, by the argument of Theorem 2 of Luedtke and Chung (2024), regularity. Since $\phi_0 \in L_0^2(P^0; \mathcal{H})$, the $\mathcal{H}$-valued central limit theorem yields the stated tight Gaussian limit $\mathbb{G}$ with the given marginal variances. □

B.7 Weak convergence of $\widehat{\Gamma}_n$ (proof of Theorem 2.4)

Proof. Identical to that of Theorem 2.3, replacing $(\gamma, \phi, \mathcal{H})$ by $(\Gamma, \Phi, L^2(P_X^0))$ and invoking Condition 2.2; the square-integrability $\Phi_0 \in L_0^2(P^0; L^2(P_X^0))$ from Theorem 2.2 ensures the $L^2(P_X^0)$-valued central limit theorem applies. □

B.8 Confidence sets of hypothesis tests (proof of Theorem 2.5)

Proof. The claims follow from Theorems 4 and 3(i) of Luedtke and Chung (2024), applied with $(\bar{\nu}_n, \nu(P^0), \mathcal{H}, \phi)$ there replaced by $(\widehat{\eta}_n, \eta(P^0), \mathcal{W}, \Psi)$ here. Their proofs use only the weak convergence $\sqrt{n}\{\widehat{\eta}_n - \eta(P^0)\} \rightsquigarrow \mathbb{G}_\bullet$, supplied by Theorem 2.3 (resp. 2.4); the stated conditions on Ω_n, Ω_0, Ψ_0, and Ψ_n^j; and the Hilbert-valued bootstrap central limit theorem of Giné and Zinn — none of which is specific to the codomain or the parameter studied there. In particular, no step requires $\eta(P^0) = 0$. □

B.9 Validity of the test (proof of Theorem 2.6)

Proof. (i) Under H_0, $\eta(P^0) = 0$, so $\{T_n > \zeta_n\} = \{\eta(P^0) \notin C_n\}$, and the claim is immediate from Theorem 2.5.

(ii) If $\nu(P^0) \notin \mathcal{S}$, then $\eta(P^0) \neq 0$ by the injectivity of the kernel embedding, so $c_\star := W(\eta(P^0); \Omega_0) > 0$ by positive-definiteness of Ω_0. Asymptotic linearity (Theorem 2.3, resp. 2.4) gives $\widehat{\eta}_n \xrightarrow{p} \eta(P^0)$, and combining this with $\|\Omega_n - \Omega_0\|_{\mathrm{op}} = o_p(1)$ and the continuity of $(h, \Omega) \mapsto W(h; \Omega)$ yields $n^{-1}T_n = W(\widehat{\eta}_n; \Omega_n) \xrightarrow{p} c_\star > 0$, whence $T_n \to \infty$ in probability. Since Theorem 2.5 gives $\zeta_n \xrightarrow{p} \zeta_{1-\alpha} < \infty$ regardless of whether H_0 holds, the claim follows; compare the analogous argument for a point null in Theorem 3.3 of Jain and Luedtke (2026). □

B.10 Simulation details and additional results

B.10.1 Data-generating process

Covariates are generated from a Gaussian copula with equicorrelation: $W \sim N(0, \Sigma_\rho)$ with $\Sigma_\rho = (1-\rho)I_3 + \rho \mathbf{1}\mathbf{1}^\top$, and $X_i = 2\Phi(W_i) - 1$ for $i = 1, 2, 3$, so that each coordinate is exactly Unif$[-1, 1]$ for every ρ while the coordinates are dependent for $\rho \neq 0$; the main grid uses $\rho = 0.5$. The structural functions are

$$g_0(x) = \text{expit}(0.5x_1 - 0.25x_2 + 0.25x_3), \qquad \mu_0(x) = \sin(\pi x_1) + x_2^2 + x_1 x_2 + 0.5x_3,$$

with $A \mid X \sim \text{Bern}\{g_0(X)\}$ and $Y = \mu_0(X) + A\tau_0(X) + \varepsilon$, $\varepsilon \sim N(0, 1)$. The null CATEs are those of Table 2.1; each lies in its own $\mathcal{S}$ and violates the other two nulls. Alternatives are $\tau_0 = \tau_{\text{null}} + \delta c_\perp$ with raw directions, per null, $c = x_1 x_2$ (smooth) and $\cos(3\pi x_1)$ (rough) for the linear null, $x_1 x_2$ and $\cos(3\pi x_1)x_2$ for the additive null, and $x_2 x_3$ and $\cos(3\pi x_3)$ for the subset null. Each direction is projected onto $\mathcal{S}^\perp$ and unit-normalized in $L^2(P_X^0)$, $c_\perp = \Pi_0^\perp c / \|\Pi_0^\perp c\|_{L^2(P_X^0)}$, on a single large reference sample generated once and reused across all cells, so that δ equals the $L^2(P_X^0)$ distance from τ_0 to $\mathcal{S}$ up to Monte Carlo error in the reference fit.

B.10.2 Implementation

Projection estimation. The projections $\widehat{\Pi}^\perp$ are computed on the evaluation sample from spline sieve bases whose number of interior knots grows with n. For the linear null the basis is $\{1, x_1, x_2, x_3\}$ and the projection is exact least squares. For the subset null the basis is a tensor-product B-spline basis in (x_1, x_2); for the additive null it is the union of centered univariate B-spline blocks, and the projection is computed by backfitting iterated to tolerance 10^{-8}, with the closed-form solution via the stacked normal equations used as a cross-check. Basis matrices are orthonormalized with a rank guard, and condition numbers are monitored.

Nuisance estimation. Nuisances are estimated by two-fold cross-fitting. The propensity

score is estimated by a SuperLearner (van der Laan et al., 2007) that forms a convex combination of logistic regression, LightGBM under two configurations, and a shallow random forest, with weights selected by 5-fold cross-validated log-loss. Estimated propensities are clipped to $[10^{-3}, 1 - 10^{-3}]$. Outcome regressions μ_a, $a \in \{0, 1\}$, are RBF kernel ridge regressions with ridge penalty 10^{-2} and the median-heuristic bandwidth, computed on a subsample of at most 500 training points, fit separately by treatment arm.

Test statistics and thresholds. The embedded parameter uses the augmented ANOVA kernel of Section 2.4.1. The covariance operator Σ_n is the empirical covariance of the estimated influence functions, $\Omega_n = [(1 - \lambda)\Sigma_n + \lambda I]^{-1}$ with $\lambda = 0.5$ in the main grid, and thresholds use $B = 999$ multiplier bootstrap replicates (Algorithm 2.1).

B.10.3 Additional results

Coverage of the confidence sets. Table B.1 reports empirical coverage of the 95% confidence sets of Theorem 2.5 for the Γ-path with estimated nuisances, aggregated over the δ grid (5000 replications per cell); coverage is computed at the true embedded parameter $\Gamma(P^0)$, which is nonzero under the alternatives, so the table exercises the "whether or not H_0 holds" guarantee. Spherical coverage is close to nominal in every cell (range 0.926–0.953). The Wald-type sets undercover in the subset cell (0.902–0.926 across n, with per-δ minima as low as 0.74), reflecting the sensitivity of the regularized inverse Ω_n to estimation error of Σ_n in the low-variance directions that this weighting upweights; the spherical sets, which require no covariance estimation, are unaffected.

Table B.1: Empirical coverage of 95% confidence sets (Γ-path, estimated nuisances, 5000 replications; standard errors in parentheses).

$\mathcal{S}$	n	sph.	Wald
linear	500	0.944 (0.003)	0.941 (0.003)
	1000	0.947 (0.003)	0.936 (0.003)
	2000	0.938 (0.003)	0.940 (0.003)
	4000	0.942 (0.003)	0.940 (0.003)
additive	500	0.949 (0.003)	0.935 (0.003)
	1000	0.939 (0.003)	0.936 (0.003)
	2000	0.953 (0.003)	0.944 (0.003)
	4000	0.941 (0.003)	0.937 (0.003)
subset	500	0.929 (0.004)	0.903 (0.004)
	1000	0.949 (0.003)	0.919 (0.004)
	2000	0.935 (0.003)	0.926 (0.004)
	4000	0.926 (0.004)	0.902 (0.004)

Appendix C

SUPPLEMENTARY MATERIALS FOR CHAPTER 3

C.1 Regularity conditions of corrected loss (proof of Lemma 3.1)

Throughout, write $h := \theta - \theta^*$ and, for a function $f(v', x, y)$ and a kernel $m(v' \mid w, x, y)$,

$$b_f^m(w, x, y) \;:=\; \int f(v', x, y)\, m(v' \mid w, x, y)\, \mu(dv').$$

All interchanges of derivative, expectation, and integral are justified by dominated convergence, using the bounds defining $\mathcal{Q}$ and $\mathcal{R}$ together with Condition 3.2. We first show that, for every $(\theta, g, \rho, q) \in \Theta \times \mathcal{G} \times \mathcal{R} \times \mathcal{Q}$,

$$\tilde{L}_{P^0}(\theta, g, \rho, q) \;=\; L_{P^0}(\theta, g) \;+\; \mathrm{E}_{P^0}\Big[\frac{\rho - \rho_0}{\rho}(W, X, Y)\; b^{q-q_0}_{\ell(\theta,g;\,\cdot)}(W, X, Y)\Big]. \tag{C.1}$$

Since ρ and q depend only on (W, X, Y), conditioning on (W, X, Y) in the definition (3.3) gives

$$\mathrm{E}_{P^0}\big[\tilde{\ell}(\theta, g, \rho, q;\, Z)\,\big|\, W, X, Y\big] = \frac{\mathrm{E}_{P^0}[S\,\ell(\theta, g;\, (V, X, Y)) \mid W, X, Y]}{\rho} + \Big(1 - \frac{\rho_0}{\rho}\Big) b^q_{\ell(\theta,g;\,\cdot)},$$

using $\mathrm{E}_{P^0}[S \mid W, X, Y] = \rho_0$ in the second term. For the numerator of the first term, (3.1) gives $\mathrm{E}_{P^0}[S\,\ell \mid W, X, Y] = \rho_0\, \mathrm{E}_{P^0}[\ell \mid W, X, Y]$, and disintegration of the conditional law of V gives $\mathrm{E}_{P^0}[\ell \mid W, X, Y] = b^{q_0}_{\ell(\theta,g;\,\cdot)}$. Hence

$$\mathrm{E}_{P^0}\big[\tilde{\ell}\,\big|\, W, X, Y\big] = \frac{\rho_0}{\rho}\, b^{q_0}_\ell + \Big(1 - \frac{\rho_0}{\rho}\Big) b^q_\ell = b^{q_0}_\ell + \frac{\rho - \rho_0}{\rho}\big(b^q_\ell - b^{q_0}_\ell\big).$$

Taking expectations and noting $\mathrm{E}_{P^0}[b^{q_0}_\ell] = L_{P^0}(\theta, g)$ by the tower property yields (C.1). We record two consequences. First, the deviation term vanishes whenever $\rho = \rho_0$ or $q = q_0$, identically in (θ, g); in particular,

$$\tilde{L}_{P^0}(\cdot, \eta_0) = L_{P^0}(\cdot, g_0) \quad \text{as an equality of functions on } \Theta. \tag{C.2}$$

Second, since ℓ enters the deviation term linearly and the (ρ, q)-factors do not involve θ, θ-derivatives of the deviation term are obtained by replacing ℓ with the corresponding pointwise derivative.

We next show that, for any $f = f(v', x, y)$ with $\mathrm{E}_{P^0}[f(V, X, Y)^2] < \infty$, any $\rho \in \mathcal{R}$, and any $q \in \mathcal{Q}$,

$$\left|\mathrm{E}_{P^0}\left[\frac{\rho - \rho_0}{\rho}\, b_f^{q-q_0}\right]\right| \;\le\; e_\rho(\rho)\, e_q(q)\, \mathrm{E}_{P^0}\big[f(V, X, Y)^2\big]^{1/2}. \tag{C.3}$$

Writing $q - q_0 = \{(q-q_0)/\sqrt{q_0}\}\sqrt{q_0}$ and applying the Cauchy–Schwarz inequality with respect to μ,

$$\big|b_f^{q-q_0}(w, x, y)\big| \le \Big(\int f^2\, q_0\, d\mu\Big)^{1/2}\Big(\int \frac{(q - q_0)^2}{q_0}\, d\mu\Big)^{1/2} = \mathrm{E}_{P^0}\big[f^2 \,\big|\, w, x, y\big]^{1/2}\chi_q(w, x, y),$$

where the equality uses disintegration and the definition of χ_q. Substituting and applying the generalized Hölder inequality with exponents $(4, 4, 2)$,

$$\left|\mathrm{E}_{P^0}\left[\frac{\rho - \rho_0}{\rho}\, b_f^{q-q_0}\right]\right| \le \mathrm{E}_{P^0}\left[\left(\frac{\rho - \rho_0}{\rho}\right)^4\right]^{1/4}\mathrm{E}_{P^0}\big[\chi_q^4\big]^{1/4}\mathrm{E}_{P^0}\Big[\mathrm{E}_{P^0}\big[f^2 \,\big|\, W, X, Y\big]\Big]^{1/2},$$

which equals the right-hand side of (C.3) by the definitions of e_ρ, e_q and the tower property.

We first establish (b) and (c). Both sides of (C.2) agree on all of Θ, hence so do their θ derivatives of any order at any point of Θ; since Θ is convex, $\mathrm{star}(\Theta, \theta^*) \subseteq \Theta$, so this includes every $\bar\theta$ appearing below. Therefore

$$D_\theta \tilde L_{P^0}(\theta^*, \eta_0)[h] = D_\theta L_{P^0}(\theta^*, g_0)[h], \qquad D_\theta^2 \tilde L_{P^0}(\bar\theta, \eta_0)[h, h] = D_\theta^2 L_{P^0}(\bar\theta, g_0)[h, h],$$

and (b) and (c) follow from Condition 3.1(b) and Condition 3.1(c), respectively. For (a), differentiating $D_\theta \tilde\ell(\theta^*, \cdot\,;\, z)[h]$ at η_0 in each nuisance direction,

$$\begin{aligned} D_\rho D_\theta \tilde\ell(\theta^*, \eta_0;\, z)[h,\, \rho - \rho_0] &= \frac{s\,(\rho - \rho_0)}{\rho_0^2}\Big(b^{q_0}_{D_\theta \ell(\theta^*, g_0;\, \cdot)[h]} - D_\theta \ell(\theta^*, g_0;\, (v, x, y))[h]\Big), \\ D_q D_\theta \tilde\ell(\theta^*, \eta_0;\, z)[h,\, q - q_0] &= \Big(1 - \frac{s}{\rho_0}\Big) b^{q-q_0}_{D_\theta \ell(\theta^*, g_0;\, \cdot)[h]}, \end{aligned}$$

and both have mean zero after conditioning on (W, X, Y): the first since $\mathrm{E}_{P^0}[S\, D_\theta \ell \mid W, X, Y] = \rho_0\, b^{q_0}_{D_\theta \ell[h]}$ by (3.1) and disintegration, the second since $\mathrm{E}_{P^0}[S \mid W, X, Y] =$

ρ_0. In the g direction, $\tilde{L}_{P^0}(\cdot,\cdot,\rho_0,q_0) = L_{P^0}$ by (C.1), so $D_g D_\theta \tilde{L}_{P^0}(\theta^*,\eta_0)[h,\, g-g_0] = D_g D_\theta L_{P^0}(\theta^*,g_0)[h,\, g-g_0] = 0$ by Condition 3.1(a). For (d), differentiating (C.1) in θ, with $f := D_\theta \ell(\theta^*, g;\, \cdot)[h]$,

$$D_\theta \tilde{L}_{P^0}(\theta^*,g,\rho,q)[h] - D_\theta L_{P^0}(\theta^*,g)[h] = \mathrm{E}_{P^0}\Big[\frac{\rho-\rho_0}{\rho}\, b_f^{q-q_0}\Big].$$

Applying (C.3) and then Condition 3.2(b), the right-hand side is at most $e_\rho(\rho)\, e_q(q)\, \beta_4 \,\|h\|_\Theta$ in absolute value. For (e), differentiating (C.1) twice in θ, with $f := D_\theta^2 \ell(\bar\theta, g;\, \cdot)[h,h]$,

$$D_\theta^2 \tilde{L}_{P^0}(\bar\theta,\eta)[h,h] = D_\theta^2 L_{P^0}(\bar\theta,g)[h,h] + \mathrm{E}_{P^0}\Big[\frac{\rho-\rho_0}{\rho}\, b_f^{q-q_0}\Big].$$

For the first term, Condition 3.1(e) gives

$$D_\theta^2 L_{P^0}(\bar\theta,g)[h,h] \geq \lambda\,\|h\|_\Theta^2 - \kappa\,\|g-g_0\|_{\mathcal{G}}^{4/(1+r)}.$$

For the second, (C.3) and Condition 3.2(a) give

$$\Big|\mathrm{E}_{P^0}\Big[\frac{\rho-\rho_0}{\rho}\, b_f^{q-q_0}\Big]\Big| \leq \beta_3\, e_\rho(\rho)\, e_q(q)\,\|h\|_\Theta \leq \frac{\lambda}{4}\,\|h\|_\Theta^2 + \frac{\beta_3^2}{\lambda}\, e_\rho(\rho)^2 e_q(q)^2,$$

the last step by Young's inequality $ab \leq \frac{\lambda}{4}a^2 + \frac{1}{\lambda}b^2$ with $a = \|h\|_\Theta$ and $b = \beta_3\, e_\rho e_q$. Combining the last two bounds with the decomposition above,

$$D_\theta^2 \tilde{L}_{P^0}(\bar\theta,\eta)[h,h] \geq \frac{3\lambda}{4}\,\|h\|_\Theta^2 - \kappa\,\|g-g_0\|_{\mathcal{G}}^{4/(1+r)} - \frac{\beta_3^2}{\lambda}\, e_\rho(\rho)^2 e_q(q)^2.$$

C.2 Oracle excess risk bound (proof of Lemma 3.2)

The argument uses only $\widehat\theta \in \Theta$ and $\widehat\eta = (\widehat g, \widehat\rho, \widehat q) \in \mathcal{G}\times\mathcal{R}\times\mathcal{Q}$, so the resulting inequalities are deterministic and hold for arbitrary such pairs. All scalar functions of a line-segment parameter appearing below, namely $t \mapsto \tilde{L}_{P^0}(\theta^* + t(\widehat\theta-\theta^*), \widehat\eta)$, $t \mapsto D_\theta L_{P^0}(\theta^*,\, g_0 + t(\widehat g - g_0))[\widehat\theta - \theta^*]$, and $t \mapsto L_{P^0}(\theta^* + t(\widehat\theta-\theta^*), g_0)$, are twice continuously differentiable on $[0,1]$, by the dominated-convergence argument recorded at the start of the proof of Lemma 3.1. Hence, Taylor's theorem with Lagrange remainder therefore applies to each.

We also use a weighted Young inequality twice below: for $x, y \geq 0$, $r \in [0,1)$, and any $\eta > 0$,

$$x^{1-r}\, y \;\leq\; \frac{\eta}{4}\, x^2 + \Big(\frac{2}{\eta}\Big)^{\frac{1-r}{1+r}}\, y^{\frac{2}{1+r}}. \tag{C.4}$$

This is Young's inequality with parameter, $ab \leq \frac{\varepsilon}{p}\, a^p + \frac{\varepsilon^{-p'/p}}{p'}\, b^{p'}$ for conjugate exponents (p, p') and $\varepsilon > 0$, applied with $p = 2/(1-r)$, $a = x^{1-r}$, $b = y$, and $\varepsilon = \eta/2$, after bounding $1/p \leq 1/2$ and $1/p' \leq 1$.

By Taylor's theorem applied to $t \mapsto \tilde{L}_{P^0}(\theta^* + t(\widehat{\theta} - \theta^*), \widehat{\eta})$ there exists $\bar{\theta} \in \mathrm{star}(\Theta, \theta^*)$ such that

$$\tilde{L}_{P^0}(\widehat{\theta}, \widehat{\eta}) - \tilde{L}_{P^0}(\theta^*, \widehat{\eta}) \;=\; D_\theta \tilde{L}_{P^0}(\theta^*, \widehat{\eta})[\widehat{\theta} - \theta^*] \;+\; \tfrac{1}{2}\, D_\theta^2 \tilde{L}_{P^0}(\bar{\theta}, \widehat{\eta})[\widehat{\theta} - \theta^*,\, \widehat{\theta} - \theta^*].$$

Applying Lemma 3.1(e) to the second term and rearranging,

$$\frac{3\lambda}{8}\, \big\|\widehat{\theta} - \theta^*\big\|_\Theta^2 \;\leq\; \tilde{L}_{P^0}(\widehat{\theta}, \widehat{\eta}) - \tilde{L}_{P^0}(\theta^*, \widehat{\eta}) - D_\theta \tilde{L}_{P^0}(\theta^*, \widehat{\eta})[\widehat{\theta} - \theta^*] + \frac{\kappa}{2}\, \|\widehat{g} - g_0\|_{\mathcal{G}}^{4/(1+r)} + \frac{\beta_3^2}{2\lambda}\, e_\rho(\widehat{\rho})^2\, e_q(\widehat{q})^2. \tag{C.5}$$

We now bound the first-order term in (C.5). By Lemma 3.1(d),

$$-\, D_\theta \tilde{L}_{P^0}(\theta^*, \widehat{\eta})[\widehat{\theta} - \theta^*] \;\leq\; -\, D_\theta L_{P^0}(\theta^*, \widehat{g})[\widehat{\theta} - \theta^*] \;+\; \beta_4\, e_\rho(\widehat{\rho})\, e_q(\widehat{q})\, \big\|\widehat{\theta} - \theta^*\big\|_\Theta. \tag{C.6}$$

Next, by Taylor's theorem applied to $t \mapsto D_\theta L_{P^0}(\theta^*, g_0 + t(\widehat{g} - g_0))[\widehat{\theta} - \theta^*]$ there exists $\bar{g} \in \mathrm{star}(\mathcal{G}, g_0)$ such that

$$\begin{aligned} D_\theta L_{P^0}(\theta^*, \widehat{g})[\widehat{\theta} - \theta^*] \;&= D_\theta L_{P^0}(\theta^*, g_0)[\widehat{\theta} - \theta^*] \;+\; D_g D_\theta L_{P^0}(\theta^*, g_0)[\widehat{\theta} - \theta^*,\, \widehat{g} - g_0] \\ &\quad +\; \tfrac{1}{2}\, D_g^2 D_\theta L_{P^0}(\theta^*, \bar{g})[\widehat{\theta} - \theta^*,\, \widehat{g} - g_0,\, \widehat{g} - g_0]. \end{aligned}$$

The middle term vanishes by Neyman orthogonality, Condition 3.1(a), and the last term is bounded in absolute value by $\frac{\beta_2}{2} \|\widehat{\theta} - \theta^*\|_\Theta^{1-r}\, \|\widehat{g} - g_0\|_{\mathcal{G}}^2$ by Condition 3.1(d). Hence

$$-\, D_\theta L_{P^0}(\theta^*, \widehat{g})[\widehat{\theta} - \theta^*] \;\leq\; -\, D_\theta L_{P^0}(\theta^*, g_0)[\widehat{\theta} - \theta^*] \;+\; \frac{\beta_2}{2}\, \big\|\widehat{\theta} - \theta^*\big\|_\Theta^{1-r}\, \|\widehat{g} - g_0\|_{\mathcal{G}}^2. \tag{C.7}$$

The two cross terms produced by (C.6) and (C.7) are now absorbed into the curvature, spending $\lambda/16$ on each. For the β_2 term, apply (C.4) with $x = \|\widehat{\theta} - \theta^*\|_\Theta$ and $y = \|\widehat{g} - g_0\|_{\mathcal{G}}^2$,

multiply by $\beta_2/2$, and choose $\eta = \lambda/(2\beta_2)$:

$$\frac{\beta_2}{2}\left\|\widehat{\theta}-\theta^*\right\|_\Theta^{1-r}\left\|\widehat{g}-g_0\right\|_{\mathcal{G}}^2 \;\le\; \frac{\lambda}{16}\left\|\widehat{\theta}-\theta^*\right\|_\Theta^2 \;+\; \frac{\beta_2}{2}\Big(\frac{4\beta_2}{\lambda}\Big)^{\frac{1-r}{1+r}}\left\|\widehat{g}-g_0\right\|_{\mathcal{G}}^{4/(1+r)}. \tag{C.8}$$

For the β_4 term, the quadratic case of Young's inequality, $ab \le \frac{\lambda}{16}a^2 + \frac{4}{\lambda}b^2$, gives

$$\beta_4\, e_\rho(\widehat{\rho})\, e_q(\widehat{q})\left\|\widehat{\theta}-\theta^*\right\|_\Theta \;\le\; \frac{\lambda}{16}\left\|\widehat{\theta}-\theta^*\right\|_\Theta^2 \;+\; \frac{4\beta_4^2}{\lambda}\, e_\rho(\widehat{\rho})^2\, e_q(\widehat{q})^2. \tag{C.9}$$

Substituting (C.6)-(C.9) into (C.5) and moving the two $\frac{\lambda}{16}\|\widehat{\theta}-\theta^*\|_\Theta^2$ terms to the left,

$$\begin{aligned}\frac{\lambda}{4}\left\|\widehat{\theta}-\theta^*\right\|_\Theta^2 \;&\le \tilde{L}_{P^0}(\widehat{\theta},\widehat{\eta}) - \tilde{L}_{P^0}(\theta^*,\widehat{\eta}) \;-\; D_\theta L_{P^0}(\theta^*,g_0)[\widehat{\theta}-\theta^*] \\ &\quad+\; \Big[\frac{\kappa}{2}+\frac{\beta_2}{2}\Big(\frac{4\beta_2}{\lambda}\Big)^{\frac{1-r}{1+r}}\Big]\left\|\widehat{g}-g_0\right\|_{\mathcal{G}}^{4/(1+r)} \;+\; \Big[\frac{\beta_3^2}{2\lambda}+\frac{4\beta_4^2}{\lambda}\Big]e_\rho(\widehat{\rho})^2\, e_q(\widehat{q})^2,\end{aligned}$$

and multiplying by $4/\lambda$,

$$\left\|\widehat{\theta}-\theta^*\right\|_\Theta^2 \;\le\; \frac{4}{\lambda}\Big(\tilde{L}_{P^0}(\widehat{\theta},\widehat{\eta})-\tilde{L}_{P^0}(\theta^*,\widehat{\eta})-D_\theta L_{P^0}(\theta^*,g_0)[\widehat{\theta}-\theta^*]\Big) + C_g\,\left\|\widehat{g}-g_0\right\|_{\mathcal{G}}^{4/(1+r)} + C_{\rho q}\, e_\rho(\widehat{\rho})^2\, e_q(\widehat{q})^2, \tag{C.10}$$

where

$$C_g \;:=\; \frac{2\kappa}{\lambda}+\frac{2\beta_2}{\lambda}\Big(\frac{4\beta_2}{\lambda}\Big)^{\frac{1-r}{1+r}}, \qquad C_{\rho q} \;:=\; \frac{2\beta_3^2+16\beta_4^2}{\lambda^2},$$

which depend only on $(\lambda,\kappa,\beta_2,\beta_3,\beta_4,r)$; for $r=0$ these evaluate to $C_g = \frac{8\beta_2^2}{\lambda^2}+\frac{2\kappa}{\lambda}$ and $C_{\rho q} = \frac{2\beta_3^2+16\beta_4^2}{\lambda^2}$. Since $D_\theta L_{P^0}(\theta^*,g_0)[\widehat{\theta}-\theta^*] \ge 0$ by Condition 3.1(b), dropping it from (C.10) establishes (3.4). By Taylor's theorem applied to $t \mapsto L_{P^0}(\theta^*+t(\widehat{\theta}-\theta^*),g_0)$ there exists $\bar{\theta}' \in \mathrm{star}(\Theta,\theta^*)$ such that, using Condition 3.1(c) for the second-order term,

$$\begin{aligned}L_{P^0}(\widehat{\theta},g_0)-L_{P^0}(\theta^*,g_0) &= D_\theta L_{P^0}(\theta^*,g_0)[\widehat{\theta}-\theta^*] + \tfrac{1}{2}\,D_\theta^2 L_{P^0}(\bar{\theta}',g_0)[\widehat{\theta}-\theta^*,\,\widehat{\theta}-\theta^*] \\ &\le D_\theta L_{P^0}(\theta^*,g_0)[\widehat{\theta}-\theta^*] + \frac{\beta_1}{2}\left\|\widehat{\theta}-\theta^*\right\|_\Theta^2.\end{aligned} \tag{C.11}$$

We first note that $\beta_1 \ge \lambda$ unless $\Theta = \{\theta^*\}$, in which degenerate case both assertions of the lemma hold trivially: taking $g = g_0$ in Condition 3.1(e), so that the κ term vanishes, and comparing with Condition 3.1(c) at the same $\bar{\theta}$ gives $\lambda\|\theta-\theta^*\|_\Theta^2 \le D_\theta^2 L_{P^0}(\bar{\theta},g_0)[\theta-\theta^*,\,\theta-\theta^*] \le \beta_1\|\theta-\theta^*\|_\Theta^2$ for every $\theta\in\Theta$, and any θ with $\|\theta-\theta^*\|_\Theta \ne 0$ forces $\lambda \le \beta_1$. Now bound $\|\widehat{\theta}-\theta^*\|_\Theta^2$ in (C.11) by the right-hand side of (C.10), retaining the first-order term:

$$L_{P^0}(\widehat{\theta},g_0)-L_{P^0}(\theta^*,g_0) \le \frac{\beta_1}{2}\Big[\frac{4}{\lambda}\Big(\tilde{L}_{P^0}(\widehat{\theta},\widehat{\eta})-\tilde{L}_{P^0}(\theta^*,\widehat{\eta})\Big) + C_g\left\|\widehat{g}-g_0\right\|_{\mathcal{G}}^{4/(1+r)} + C_{\rho q}e_\rho(\widehat{\rho})^2 e_q(\widehat{q})^2\Big]$$

$$+ \left(1 - \frac{2\beta_1}{\lambda}\right) D_\theta L_{P^0}(\theta^*, g_0)[\widehat{\theta} - \theta^*].$$

Since $\beta_1 \geq \lambda$, the coefficient $1 - 2\beta_1/\lambda \leq -1$ is negative, and the first-order term is nonnegative by Condition 3.1(b), so the last term may be dropped. This establishes (3.5) and completes the proof.

C.3 Excess risk of the corrected empirical risk minimizer (proof of Theorem 3.1)

Throughout this section we condition on $\widehat{\eta} = (\widehat{g}, \widehat{\rho}, \widehat{q})$, which the sample splitting of Algorithm 3.1 renders independent of the target-stage sample; all probabilities and expectations over data refer to the target-stage sample, whose empirical distribution is $\mathbb{P}_n^{(2)}$ and whose size we denote by n. Since the resulting bound holds with conditional probability at least $1-\delta$ for every realization of $\widehat{\eta}$, it holds with unconditional probability at least $1-\delta$, which is the form stated in the theorem. We assume Θ is separable, so that the suprema below are measurable. We use the notation of Section 3.3.2: $f_\theta = \tilde{\ell}(\theta, \widehat{\eta}; \cdot) - \tilde{\ell}(\theta^*, \widehat{\eta}; \cdot)$, $\Delta_\theta = \ell(\theta, \widehat{g}; \cdot) - \ell(\theta^*, \widehat{g}; \cdot)$, the localized classes $\tilde{\mathcal{F}}_{\widehat{\eta}}(\delta)$, the constant $\tilde{b} = 4B/\rho_{\min}$, and the critical radius r_ψ; and, from the proof of Lemma 3.1, the kernel average $b_f^m(w,x,y) = \int f(v',x,y)\, m(v' \mid w,x,y)\, \mu(dv')$. Unpacking the definition (3.3),

$$f_\theta(z) \;=\; \frac{s}{\widehat{\rho}(w,x,y)}\, \Delta_\theta(v,x,y) \;+\; \left(1 - \frac{s}{\widehat{\rho}(w,x,y)}\right) b^{\widehat{q}}_{\Delta_\theta}(w,x,y), \tag{C.12}$$

and $f_{\theta^*} = 0$. We record two facts about sub-root functions. First, any sub-root ψ is continuous on $(0,\infty)$: at a jump point δ_0, monotonicity of $\psi(\delta)/\delta$ across δ_0 forces the jump to be nonpositive. Consequently $\psi(r_\psi) \leq r_\psi^2/\tilde{b}$, and for every $\delta \geq r_\psi$,

$$\psi(\delta) \;\leq\; \frac{\delta}{r_\psi}\, \psi(r_\psi) \;\leq\; \frac{\delta\, r_\psi}{\tilde{b}} \;\leq\; \frac{\delta^2}{\tilde{b}}, \tag{C.13}$$

using that $\psi(\delta)/\delta$ is nonincreasing. Second, for $a \geq 1$ and $\delta > 0$, the same monotonicity gives $\psi(a\delta) \leq a\, \psi(\delta)$. Finally, r_ψ is finite: $\psi(\delta) \leq \psi(1)\, \delta$ for $\delta \geq 1$, so the defining inequality holds once $\delta \geq \max(1, \tilde{b}\, \psi(1))$.

We first show that f_θ is bounded in supremum norm and in variance; both bounds enter the concentration inequality below. For every $\theta \in \Theta$,

$$\|f_\theta\|_{L^\infty(P^0)} \;\le\; \tilde{b}, \qquad \mathrm{E}_{P^0}\big[f_\theta(Z)^2\big] \;\le\; \tilde{L}^2\,\|\theta-\theta^*\|_\Theta^2\,, \qquad \text{where } \tilde{L} := \frac{B\sqrt{2(1+C/c_0)}}{\rho_{\min}}. \tag{C.14}$$

By Condition 3.3 and the class bound $\sup_{\theta\in\Theta}\|\theta\|_{L^\infty(P_V)} \le 1$,

$$|\Delta_\theta(v,x,y)| \le B\;|\theta(v)-\theta^*(v)| \le 2B, \qquad \big|b^{\widehat{q}}_{\Delta_\theta}(w,x,y)\big| \le b^{\widehat{q}}_{|\Delta_\theta|}(w,x,y) \le 2B, \tag{C.15}$$

the last inequality because the integrand is at most $2B$ and $\widehat{q}(\cdot \mid w,x,y)$ integrates to one. Since $\widehat{\rho} \ge \rho_{\min}$, the absolute values of the two coefficients in (C.12) sum to at most $2/\rho_{\min}-1$ on $\{s=1\}$ and to 1 on $\{s=0\}$, so $\|f_\theta\|_{L^\infty(P^0)} \le 2B(2/\rho_{\min}-1) \le \tilde{b}$. For the variance bound, apply $(a+b)^2 \le 2a^2+2b^2$ to (C.12). For the first term, $s^2=s$, and conditioning on (W,X,Y) together with (3.1) gives

$$\mathrm{E}_{P^0}\Big[\frac{S}{\widehat{\rho}^2}\,\Delta_\theta^2\Big] = \mathrm{E}_{P^0}\Big[\frac{\rho_0}{\widehat{\rho}^2}\,\mathrm{E}_{P^0}\big[\Delta_\theta^2 \,\big|\, W,X,Y\big]\Big] \le \frac{1}{\rho_{\min}^2}\,\mathrm{E}_{P^0}\big[\Delta_\theta(V,X,Y)^2\big].$$

For the second term, $|1-s/\widehat{\rho}| \le 1/\rho_{\min}$, and the Cauchy–Schwarz inequality with respect to $\widehat{q}\,d\mu$ followed by the density-ratio bound $\widehat{q}/q_0 \le C/c_0$ and disintegration give

$$\big(b^{\widehat{q}}_{\Delta_\theta}\big)^2 \le b^{\widehat{q}}_{\Delta_\theta^2} \le \frac{C}{c_0}\,b^{q_0}_{\Delta_\theta^2} = \frac{C}{c_0}\,\mathrm{E}_{P^0}\big[\Delta_\theta^2 \,\big|\, W,X,Y\big].$$

Taking expectations, summing, and bounding $\mathrm{E}_{P^0}[\Delta_\theta(V,X,Y)^2] \le B^2\,\|\theta-\theta^*\|_\Theta^2$ by (C.15) yields (C.14).

Next we show that the deviations $(\mathbb{P}_n^{(2)}-P^0)f_\theta$ are uniformly small with high probability, by a peeling argument. Fix $t>0$ and set

$$\delta_t \;:=\; r_\psi + (\tilde{b}+\tilde{L})\sqrt{\frac{t+1}{n}}, \tag{C.16}$$

so that $\delta_t \ge r_\psi$ and $n^2\delta_t^2 \ge n\,(\tilde{b}+\tilde{L})^2(t+1)$. We claim that the event

$$\mathcal{E}_t \;:=\; \Big\{\,\big|\big(\mathbb{P}_n^{(2)}-P^0\big)f_\theta\big| \;\le\; \frac{\lambda}{8}\,\|\theta-\theta^*\|_\Theta^2 \;+\; c_1\Big(\delta_t^2+\frac{t+1}{n}\Big) \quad \text{for all } \theta\in\Theta\,\Big\} \tag{C.17}$$

has conditional probability at least $1-e^{-t}$, for a constant c_1 depending only on $(B,\lambda,\rho_{\min},C/c_0)$.

For $k \geq 0$ let $Z_k := \sup\{|(\mathbb{P}_n^{(2)} - P^0)f| : f \in \tilde{\mathcal{F}}_{\tilde{\eta}}(2^k\delta_t)\}$. By symmetrization, the inclusion of the class in its star hull, the majorization by ψ, the sub-root scaling with $a = 2^k$, and (C.13) at $\delta_t \geq r_\psi$,

$$\mathrm{E}[Z_k] \;\leq\; 2\,\mathfrak{R}_n\big(\tilde{\mathcal{F}}_{\tilde{\eta}}(2^k\delta_t)\big) \;\leq\; 2\,\psi(2^k\delta_t) \;\leq\; 2^{k+1}\,\psi(\delta_t) \;\leq\; \frac{2^{k+1}\delta_t^2}{\tilde{b}}. \tag{C.18}$$

By (C.14), the centered functions $\pm(f - P^0 f)$ with $f \in \tilde{\mathcal{F}}_{\tilde{\eta}}(2^k\delta_t)$ have mean zero, supremum norm at most $2\tilde{b}$, and second moments $\mathrm{E}_{P^0}[(f - P^0 f)^2] \leq \mathrm{E}_{P^0}[f^2] \leq \tilde{L}^2\,4^k\delta_t^2$, and the supremum of the resulting empirical process over this centered class is Z_k. Talagrand's concentration inequality (Bousquet, 2003), applied to this class, gives, for any $u > 0$, with conditional probability at least $1 - e^{-u}$,

$$Z_k \;\leq\; \mathrm{E}[Z_k] + \sqrt{\frac{2u\big(\tilde{L}^2 4^k\delta_t^2 + 4\tilde{b}\,\mathrm{E}[Z_k]\big)}{n}} + \frac{2\tilde{b}\,u}{3n} \;\leq\; 2\,\mathrm{E}[Z_k] + \tilde{L}\,2^k\delta_t\sqrt{\frac{2u}{n}} + \frac{3\tilde{b}\,u}{n},$$

where the second inequality uses $\sqrt{8u\tilde{b}\,\mathrm{E}[Z_k]/n} \leq \mathrm{E}[Z_k] + 2u\tilde{b}/n$. Apply this with $u = t_k := t + (k+1)\log 2$; since $\sum_{k\geq 0} e^{-t_k} = e^{-t}\sum_{k\geq 0} 2^{-(k+1)} = e^{-t}$, with conditional probability at least $1 - e^{-t}$, simultaneously for every $k \geq 0$,

$$Z_k \;\leq\; \frac{2^{k+2}\delta_t^2}{\tilde{b}} \;+\; \tilde{L}\,2^k\delta_t\sqrt{\frac{2t_k}{n}} \;+\; \frac{3\tilde{b}\,t_k}{n}, \tag{C.19}$$

where the first term uses (C.18). It remains to show that (C.19) implies the bound in (C.17). Fix $\theta \in \Theta$. If $\|\theta - \theta^*\|_\Theta \leq \delta_t$, then $|(\mathbb{P}_n^{(2)} - P^0)f_\theta| \leq Z_0$, and (C.19) with $k = 0$, together with $\sqrt{2t_0/n} \leq \sqrt{2t/n} + \sqrt{2/n}$ and $ab \leq (a^2 + b^2)/2$, gives

$$\big|\big(\mathbb{P}_n^{(2)} - P^0\big)f_\theta\big| \;\leq\; \frac{4\delta_t^2}{\tilde{b}} + \delta_t^2 + \frac{\tilde{L}^2(t+1)}{n} + \frac{3\tilde{b}(t+1)}{n},$$

which is of the form required by (C.17).

Otherwise $\|\theta - \theta^*\|_\Theta \in (2^{k-1}\delta_t,\, 2^k\delta_t]$ for a unique $k \geq 1$, so that $|(\mathbb{P}_n^{(2)} - P^0)f_\theta| \leq Z_k$ and $2^k\delta_t \leq 2\,\|\theta - \theta^*\|_\Theta$, and, since $\|\theta - \theta^*\|_\Theta > 2^{k-1}\delta_t$ and $\log_2 a \leq a$ for $a \geq 1$,

$$k + 1 \;\leq\; 2 + \log_2\frac{\|\theta - \theta^*\|_\Theta}{\delta_t} \;\leq\; 2 + \frac{\|\theta - \theta^*\|_\Theta}{\delta_t}. \tag{C.20}$$

We bound the three terms of (C.19) in turn, using the weighted inequality $ab \le \frac{\lambda}{48}a^2 + \frac{12}{\lambda}b^2$ throughout. For the first term,

$$\frac{2^{k+2}\delta_t^2}{\tilde{b}} \;\le\; \frac{8\,\delta_t}{\tilde{b}}\,\|\theta-\theta^*\|_\Theta \;\le\; \frac{\lambda}{48}\,\|\theta-\theta^*\|_\Theta^2 + \frac{768}{\lambda\tilde{b}^2}\,\delta_t^2.$$

For the second term, split $\sqrt{2t_k} \le \sqrt{2t} + \sqrt{2(k+1)}$ using $\log 2 \le 1$. The t-part satisfies

$$\tilde{L}\,2^k\delta_t\sqrt{\frac{2t}{n}} \;\le\; 2\tilde{L}\,\|\theta-\theta^*\|_\Theta\sqrt{\frac{2t}{n}} \;\le\; \frac{\lambda}{48}\,\|\theta-\theta^*\|_\Theta^2 + \frac{96\,\tilde{L}^2\,t}{\lambda\,n}.$$

For the $(k+1)$-part, $\sqrt{k+1} \le \sqrt{2}\cdot 2^{k/2}$ gives $\tilde{L}\,2^k\delta_t\sqrt{2(k+1)/n} \le 2\tilde{L}\,2^{3k/2}\delta_t\,n^{-1/2}$, and writing $2^{3k/2}\delta_t = (2^k\delta_t)^{3/2}\delta_t^{-1/2} \le (2\,\|\theta-\theta^*\|_\Theta)^{3/2}\delta_t^{-1/2}$, then applying $ab \le \varepsilon a^{4/3} + \varepsilon^{-3}b^4$ with $a = \|\theta-\theta^*\|_\Theta^{3/2}$, $b = 6\tilde{L}\,\delta_t^{-1/2}n^{-1/2}$, and $\varepsilon = \lambda/48$,

$$\tilde{L}\,2^k\delta_t\sqrt{\frac{2(k+1)}{n}} \;\le\; \frac{\lambda}{48}\,\|\theta-\theta^*\|_\Theta^2 + \Big(\frac{48}{\lambda}\Big)^3\frac{1296\,\tilde{L}^4}{\delta_t^2\,n^2} \;\le\; \frac{\lambda}{48}\,\|\theta-\theta^*\|_\Theta^2 + \Big(\frac{48}{\lambda}\Big)^3\frac{1296\,\tilde{L}^2}{n},$$

the last step by $\delta_t^2 n^2 \ge \tilde{L}^2 n$ from (C.16). For the third term, (C.20) gives

$$\frac{3\tilde{b}\,t_k}{n} \;\le\; \frac{3\tilde{b}\,t}{n} + \frac{6\tilde{b}}{n} + \frac{3\tilde{b}}{n\,\delta_t}\,\|\theta-\theta^*\|_\Theta \;\le\; \frac{3\tilde{b}(t+2)}{n} + \frac{\lambda}{48}\,\|\theta-\theta^*\|_\Theta^2 + \frac{108\,\tilde{b}^2}{\lambda\,n^2\delta_t^2},$$

and $n^2\delta_t^2 \ge \tilde{b}^2(t+1)\,n \ge \tilde{b}^2 n$ bounds the last summand by $108/(\lambda n)$. Summing the four displays, the coefficient of $\|\theta-\theta^*\|_\Theta^2$ is $4\cdot\lambda/48 = \lambda/12 \le \lambda/8$, and every remaining summand is a multiple, by a constant depending only on $(B, \lambda, \rho_{\min}, C/c_0)$, of δ_t^2 or of $(t+1)/n$. This proves the claim (C.17).

Since $\theta^* \in \Theta$ and $\widehat{\theta}$ minimizes the corrected empirical risk, $\mathbb{P}_n^{(2)} f_{\widehat{\theta}} \le 0$. Hence, on $\mathcal{E}_t$,

$$\tilde{L}_{P^0}(\widehat{\theta},\widehat{\eta}) - \tilde{L}_{P^0}(\theta^*,\widehat{\eta}) \;=\; P^0 f_{\widehat{\theta}} \;\le\; \big|\big(\mathbb{P}_n^{(2)} - P^0\big) f_{\widehat{\theta}}\big| \;\le\; \frac{\lambda}{8}\,\big\|\widehat{\theta}-\theta^*\big\|_\Theta^2 \;+\; c_1\Big(\delta_t^2 + \frac{t+1}{n}\Big). \quad \text{(C.21)}$$

Substituting (C.21) into (3.4) of Lemma 3.2,

$$\big\|\widehat{\theta}-\theta^*\big\|_\Theta^2 \;\le\; \frac{1}{2}\,\big\|\widehat{\theta}-\theta^*\big\|_\Theta^2 \;+\; \frac{4c_1}{\lambda}\Big(\delta_t^2 + \frac{t+1}{n}\Big) \;+\; C_g\,\|\widehat{g}-g_0\|_{\mathcal{G}}^{4/(1+r)} \;+\; C_{\rho q}\,e_\rho(\widehat{\rho})^2\,e_q(\widehat{q})^2,$$

and rearranging,

$$\big\|\widehat{\theta}-\theta^*\big\|_\Theta^2 \;\le\; \frac{8c_1}{\lambda}\Big(\delta_t^2 + \frac{t+1}{n}\Big) \;+\; 2C_g\,\|\widehat{g}-g_0\|_{\mathcal{G}}^{4/(1+r)} \;+\; 2C_{\rho q}\,e_\rho(\widehat{\rho})^2\,e_q(\widehat{q})^2. \qquad \text{(C.22)}$$

By (C.16) and $(a+b)^2 \le 2a^2+2b^2$, $\delta_t^2 \le 2r_\psi^2 + 2(\tilde{b}+\tilde{L})^2(t+1)/n$. Take $t = \log(1/\delta)$, so that $\mathcal{E}_t$ has probability at least $1-\delta$; since $\delta \le 1/2$ gives $t \ge \log 2 > 1/2$ and hence $t+1 \le 3t$, the first term of (C.22) is at most $c\,(r_\psi^2 + \log(1/\delta)/n)$ for a constant c depending only on $(B, \lambda, \rho_{\min}, C/c_0)$, which establishes the first assertion of the theorem. For the risk bound, substitute (C.21) and then (C.22) into (3.5) of Lemma 3.2.