Estimation of causal dose-response functions under data fusion

Jaewon Lim[1], and Alex Luedtke[2,3]

[1]Department of Biostatistics, University of Washington
[2]Department of Health Care Policy, Harvard University
[3]Department of Statistics, University of Washington

Abstract

Estimating the causal dose-response function is challenging, particularly when data from a single source are insufficient to estimate responses precisely across all exposure levels. To overcome this limitation, we propose a data fusion framework that leverages multiple data sources that are partially aligned with the target distribution. Specifically, we derive a Neyman-orthogonal loss function tailored for estimating the dose-response function within data fusion settings. To improve computational efficiency, we propose a stochastic approximation that retains orthogonality. We apply kernel ridge regression with this approximation, which provides closed-form estimators. Our theoretical analysis demonstrates that incorporating additional data sources yields tighter finite-sample regret bounds and improved worst-case performance, as confirmed via minimax lower bound comparison. Simulation studies validate the practical advantages of our approach, showing improved estimation accuracy when employing data fusion. This study highlights the potential of data fusion for estimating non-smooth parameters such as causal dose-response functions.

Keywords: causal-dose response function; data fusion; orthogonal statistical learning; minimax; reproducing kernel Hilbert space

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. (2025), 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.

2 Data Fusion Framework and Estimation Algorithm

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

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

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 prespecified 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

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

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 (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 for some $Q^0 \in \mathcal{Q}$. Condition 1 can be relaxed in the sense that it suffices for there to exist at least one $Q \in \mathcal{Q}$ satisfying the condition, even if the true Q^0 fails to meet Condition 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(a) guarantees that the probabilities of correctly and partially aligned sources are uniformly bounded away from zero across $\mathcal{P}$. Conditions 1(b) and 1(c) are used to identify the risk through the observed distribution P^0. Specifically, Condition 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(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(b) to define a Neyman–orthogonal loss for the risk in the next section.

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

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

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

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

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 (Bickel, 1982). This estimator adjusts for the bias induced by plug-in estimation of estimator $\widehat{P}_n$ of P, where $D^*_P(\theta)$ denotes a gradient of $P \mapsto L_P(\theta)$ at P. 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

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

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

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

with nuisance parameters defined as

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

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), for the remainder of the paper.

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

Condition 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 $\left|\theta_0(a)\right| = \left|\mathbb{E}_{Q_X}\left[m_0(x,a)\right]\right| \leq 1$ 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 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 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 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(b) and 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 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 3(a) requires that the source probability estimators are uniformly bounded away from zero over $\mathcal{P}$. Condition 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 $M_\lambda := M_w^{-1}(M_\eta + M_w + 2)$.

Algorithm 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:

$$\widehat{\xi} := \frac{1}{\widehat{P}_n(S \in \mathcal{S}_X)}, \qquad \widehat{\eta} := \frac{1}{\widehat{P}_n(S \in \mathcal{S}_Y)}, \qquad \widehat{w}(x,a) := \text{direct estimator of } w_0(x,a),$$
$$\widehat{m}(x,a) := \mathbb{E}_{\widehat{P}_n}[Y \mid X = x, A = a, S \in \mathcal{S}_Y], \qquad \widehat{\tau}_0(a) := \mathbb{E}_{\widehat{P}_n}[\widehat{m}(X,a) \mid S \in \mathcal{S}_X].$$

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.

The loss function defined in (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:

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

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:

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

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 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 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

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

Algorithm 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, 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 S3) 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

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

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 $\mathcal{H}_c$.

3 Excess Risk Guarantees and Worst-Case Efficiency Gains from Using Data Fusion

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 summarizes these relationships between excess risk bounds and worst-case guarantees under data fusion and no fusion.

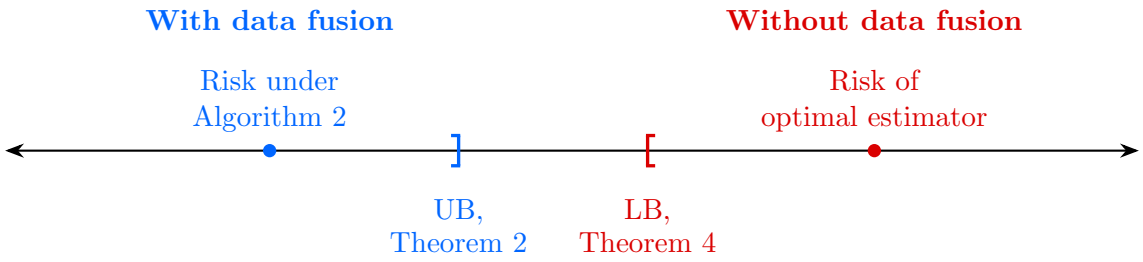

Figure 1: Number line demonstrating approach for establishing data fusion necessarily improves worst-case performance. The key idea is to show that a **high-probability worst-case upper bound (UB)** of the risk for kernel ridge regression with data fusion lies below a **high-probability worst-case lower bound (LB)** for any estimator without it.

3.2 Excess Risk Bound

This subsection establishes oracle excess risk bounds for the estimators obtained from Algorithms 1 and 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 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 $1 \le p < \infty$, we define

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

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.

Lemma 1 (Oracle Excess Risk Bound in Terms of Target and Nuisance Errors)**.** *Assume Conditions 1 to 3 hold. If $\widehat{\theta}$ and $\widehat{g}$ are the target estimator and nuisance estimators from Algorithm 1, then*

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

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 S3. 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 and 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 $\mathcal{D}$, defined as

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

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

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

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

We now present the oracle excess risk bound for the estimator obtained from Algorithm 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

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

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

Theorem 1 (Excess Risk Bound for Algorithm 1)**.** *Suppose Conditions 1 to 3 hold. Fix $\delta\in(0,1)$, and let $(\widehat{\theta},\widehat{g})$ be as in Algorithm 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 $\delta_n^2=\Omega\big(n^{-1}\log\log n\big)$, 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 $\theta_1,\theta_2\in\Theta$,

$$\big|\ell_{P^0}(\theta_1,\widehat{g};z)-\ell_{P^0}(\theta_2,\widehat{g};z)\big|\le B_{P^0}(\delta)\,\big\|\big(\theta_1(b)-\theta_2(b),\ \theta_1(a)-\theta_2(a)\big)\big\|_2. \tag{2}$$

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 $s \geq 0$, 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, $0 < s_1 < s_2$, $\mathcal{H}^{s_2} \subseteq \mathcal{H}^{s_1} \subseteq \mathcal{H}^0 \subseteq L^2(\mu)$. 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 the $\mathcal{H}$ and its associated kernel $\mathcal{K}$.

Condition 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 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 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 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 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 2 and an additional smoothness condition-and also is large enough so that estimating θ_0 is challenging.

Condition 5 (Characterization of $\mathcal{Q}$). $\mathcal{Q}$ consists of all distributions satisfying Condition 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 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 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 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 2. For clarity, we focus on this constrained formulation throughout the remainder of the paper.

Theorem 2 (Excess Risk Bound for Algorithm 2)**.** *Assume that Conditions 1 to 5 hold. Fix $\delta \in (0,1)$, and let $(\widehat{\theta}, \widehat{g})$ be as in Algorithm 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

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

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

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

Theorem 2 is a special case of Theorem 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). 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 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 4—also yields a bound uniform over $\mathcal{P}$, obtained by taking the supremum over $P \in \mathcal{P}$.

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:

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

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

$$\begin{aligned}&\underline{\xi}_0 := \frac{1}{P^0(S \in \mathcal{S}_X \cap \mathcal{S}_Y)}, \quad \underline{\eta}_0 := \frac{1}{P^0(S \in \mathcal{S}_X \cap \mathcal{S}_Y)}, \quad \underline{w}_0(x,a) := \frac{1}{p^0(a \mid x, S \in \mathcal{S}_X \cap \mathcal{S}_Y)}\,, \\ &\underline{m}_0(x,a) := \mathbb{E}_{P^0}\left[Y \mid X = x, A = a, S \in \mathcal{S}_X \cap \mathcal{S}_Y\right], \quad \underline{\tau}_0(a) := \mathbb{E}_{P^0}\left[\underline{m}_0(X,a) \mid S \in \mathcal{S}_X \cap \mathcal{S}_Y\right],\end{aligned}$$

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 2 with data fusion, and $(\underline{\theta}, \underline{g})$ the corresponding estimators without data fusion, where $\underline{g}\colon (a,x) \mapsto (\widehat{\underline{\xi}}, \widehat{\underline{\eta}}, \widehat{\underline{w}}(a,x), \widehat{\underline{m}}(a,x), \widehat{\underline{\tau}}(a))$.

For Theorem 3 below, we impose the same nuisance estimation and cross-validation conditions as in the second statement of Theorem 2. Specifically, suppose the nuisance estimators satisfy

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

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 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 to 5 hold, and that Conditions 1 and 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 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 6 (Eigenvalue Decay at Exactly Polynomial Rate). There exists $p \in (0,1)$ 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 2 (Minimax Lower Bound without Data Fusion)**.** *Assume that Conditions 1 to 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* (S6) *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* $1 - 2\delta$,

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

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 2 with the uniform excess risk upper bound for the data-fusion estimator from Theorem 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 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 2): for any $\varepsilon > 0$, there exists $M_g < \infty$ such that

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

We further suppose the regularization parameter is chosen as

$$c \propto (n\log(n)^{-2})^{\frac{\alpha}{p+(1-2\alpha)}}. \tag{5}$$

Theorem 4 (Sufficient Conditions for Data Fusion to Improve Worst-Case Performance)**.** *Suppose $(\widehat{\theta}, \widehat{g})$ satisfy the conditions above, c satisfies (5), Conditions 1 to 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 (S10),*

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

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

$$\sup_{P\in\mathcal{P}_{XY}} P^n\Big\{L_P(\underline{\theta}, g_0) \ge r(\delta, n)\Big\} - \sup_{P\in\mathcal{P}_{XY}} P^n\Big\{L_P(\widehat{\theta}, g_0) > r(\delta, n)/\log(n)^{\frac{h(1-2\alpha)}{2[p+(1-2\alpha)]}}\Big\} \;\ge\; 1 - 2\delta. \tag{6}$$

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 2, and that, by taking the supremum of the excess-risk bound in Theorem 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 (6), it follows that

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

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

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 aligning datasets via data fusion improves predictive performance, even when the reference measure is different than the data-

generating measure or 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 CDRFs, 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 $\text{Beta}(5, 5)$, a bell-shaped distribution emphasizing the center of the range, thereby making estimation accuracy near $a = 0.5$ more important. The third is $\text{Beta}(0.5, 0.5)$, 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 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 a SuperLearner package (van der Laan et al., 2007), combining mean regression, random forests,

Table 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 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

(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

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 S3.

Given the estimated nuisance components and the selected λ, we computed the closed-form estimator of the CDRF as in Algorithm 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. Across all settings, incorporating data fusion improves performance. This includes the discontinuous CDRF case in Table 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.

5 Discussion

Our data fusion framework can also be applied to the estimation of non-pathwise differentiable function-valued parameters such as 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 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).

Acknowledgements

This work was supported by the National Science Foundation and National Institutes of Health under award numbers DMS-2210216 and DP2-LM013340.

Supplementary Materials

R code for the CDRF estimation function Algorithm 2 and simulation codes for Section 4 are available at https://github.com/imjaewon07/CDRF_data_fusion.

References

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.

Bahadori, T., Tchetgen, E. T., & 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*.

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.

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. (2025). Towards a unified theory for semiparametric data fusion with individual-level data. *arXiv preprint arXiv:2409.09973*.

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*, 1–14.

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.

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

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.

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.

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*, 1–11.

Luedtke, A. (2025). Simplifying debiased inference via automatic differentiation and probabilistic programming. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, qkaf052.

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

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.

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.

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.

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

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

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.

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.

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.

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

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.

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

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

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. In *Proceedings of the 40th international conference on machine learning* (pp. 41331–41353, Vol. 202). PMLR.

Appendices

Algorithm S3 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.
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.

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

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

A Cross Validation Algorithm in Kernel Ridge Regression

B Proofs

B.1 Derivation of Neyman-Orthogonal Loss and a Stochastic Approximation Thereof (Lemma S1)

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 not locally nonparametric, standard calculations or a chain rule (Kennedy, 2023; Luedtke,

2025) 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

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

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

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

where

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

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 S1 (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*

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

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

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

$$+ 2\eta_0 w_0(x,a)\mathbf{1}(s \in \mathcal{S}_Y)\,(\theta(a) - \theta_0(a))\,(m_0(x,a) - y)\,. \tag{S1}$$

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

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

where

$$\begin{aligned}\underline{\ell}_{P^0}(\theta, g_0; x, a, y, s, b) := {}& (\theta(b) - \theta_0(b))^2 \\ &+ 2\underline{\xi}_0\mathbf{1}(s \in \mathcal{S}_X \cap \mathcal{S}_Y)\,(\theta(b) - \underline{\tau}_0(b))\,(\underline{\tau}_0(b) - \underline{m}_0(b,x)) \\ &+ 2\underline{\eta}_0\underline{w}_0(x,a)\mathbf{1}(s \in \mathcal{S}_X \cap \mathcal{S}_Y)\,(\theta(a) - \underline{\tau}_0(a))\,(\underline{m}_0(x,a) - y)\,.\end{aligned} \tag{S2}$$

Proof of Lemma S1. 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

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

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

B.2 Closed-Form Solution of Kernel Ridge Regression (Lemma S2)

Lemma S2 (Closed-form solution of kernel ridge regression)**.** *The closed-form estimator $\widehat{\theta}$ obtained in Algorithm 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 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 2.

Proof of Lemma S2. 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

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

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 4(a), the block matrix $\mathbf{K}$ is invertible, we obtain that the minimizer $(\widehat{\beta},\widehat{\gamma})$ satisfies

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

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

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

$$\mathcal{D}_\theta \mathcal{D}_g L(\theta^*, g_0)[h, k] := \frac{\partial^2}{\partial t\, \partial s} L(\theta^* + th,\, g_0 + sk)\Big|_{t=0,\, s=0}.$$

B.3 Properties of the Population Loss (Lemma S3)

Lemma S3 (Properties of the population loss)**.** *The loss function in (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:*

$$\mathcal{D}_\theta L_{P^0}(\theta^*, g_0)[\theta - \theta^*] = 2\,\mathbb{E}_\mu\left[\big(\theta(A) - \theta^*(A)\big)\big(\theta^*(A) - \theta_0(A)\big)\right] \geq 0,$$

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:*

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

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 S3. 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

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

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

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

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,

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

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 $g \in \mathcal{G}$,

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

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

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

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

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

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

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

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

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

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

□

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.

B.4 Oracle Excess Risk Bound in Terms of Target and Nuisance Errors (Proof of Lemma 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 $\hat{g}$:

$$L_{P^0}(\hat{\theta}, \hat{g}) - L_{P^0}(\theta^*, \hat{g}) = \mathcal{D}_\theta L_{P^0}(\theta^*, \hat{g})[\hat{\theta} - \theta^*] + \|\hat{\theta} - \theta^*\|^2_{L^2(\mu)}. \tag{S3}$$

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 (S3) 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

$$\|\hat{\theta} - \theta^*\|^2_{L^2(\mu)} \leq 2\left(L_{P^0}(\hat{\theta}, \hat{g}) - L_{P^0}(\theta^*, \hat{g})\right) - 2\mathcal{D}_\theta L_{P^0}(\theta^*, g_0)[\hat{\theta} - \theta^*] + M_\lambda^2 \|\hat{g} - g_0\|^4_{L^4(Q_X^0 \times \mu; \ell^2)}. \tag{S4}$$

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 (S4) 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^*. □

Based on Lemma 1, we can derive the regret bound stated in Theorem 1. To prove Theorem 1, we employ the Talagrand concentration inequality for Lipschitz loss functions, allowing for the target parameter to be vector-valued.

B.5 Excess Risk Bound for Algorithm 1 (Proof of Theorem 1)

Proof of Theorem 1. In our setting, the loss depends on the target function evaluated at two possibly distinct inputs a and b. To directly leverage the framework in Foster and Syrgkanis (2023), we reformulate the scalar-valued target $\theta : \mathbb{R} \to \mathbb{R}$ as a vector-valued function

$$\boldsymbol{\theta} : \mathbb{R}^2 \to \mathbb{R}^2, \quad \boldsymbol{\theta}(a, b) = \big(\theta(a), \theta(b)\big).$$

This representation allows us to treat $\theta(a)$ and $\theta(b)$ jointly. Note that estimating $\boldsymbol{\theta}$ is equivalent to estimating θ, but the vector-valued formulation streamlines the notation and aligns our setting with existing results on Lipschitz losses of vector-valued target parameters. We define the loss at $\boldsymbol{\theta}$ as

$$\begin{aligned}\ell_{P^0}(\boldsymbol{\theta}, g_0; z) &:= \ell_{P^0}(\theta, g_0; z),\\ L_{P^0}(\boldsymbol{\theta}, g_0; z) &:= L_{P^0}(\theta, g_0; z),\end{aligned}$$

and similarly denote $\boldsymbol{\theta}^*(a, b) = (\theta^*(a), \theta^*(b))$.

It is straightforward to verify that the multidimensional analogues of all conditions in Lemma S3 hold for $\boldsymbol{\theta}$ as well. In particular:

(a) **First-order optimality** For all $\theta \in \Theta \subset L^2(\mu)$,

$$\mathcal{D}_{\boldsymbol{\theta}} L_{P^0}(\boldsymbol{\theta}^*, g_0)\big[\boldsymbol{\theta} - \boldsymbol{\theta}^*\big] = 2\mathbb{E}_\mu\big[\big(\theta(A) - \theta^*(A)\big)\big(\theta^*(A) - \theta_0(A)\big)\big] \geq 0.$$

(b) **Neyman orthogonality** For all $\theta \in L^2(\mu)$ and $g \in \mathcal{G}$,

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

(c) **Second-order smoothness and strong convexity** For all $\theta \in \Theta$, $\overline{\theta} \in L^2(\mu)$, and $g \in \mathcal{G}$,

$$\begin{aligned}\mathcal{D}^2_{\boldsymbol{\theta}} L_P(\overline{\boldsymbol{\theta}}, g)\big[\boldsymbol{\theta} - \boldsymbol{\theta}^*, \boldsymbol{\theta} - \boldsymbol{\theta}^*\big] &= 2\mathbb{E}_\mu\big[(\theta(A) - \theta^*(A))^2\big]\\ &= 2\mathbb{E}_{\mu\times\mu}\big[(\theta(A) - \theta^*(A))(\theta(B) - \theta^*(B))\big]\\ &= \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|^2_{L^2(\mu\times\mu;\ell^2)}.\end{aligned}$$

(d) **Higher-order smoothness** For all $\theta \in L^2(\mu)$, $g \in \mathcal{G}$, and $\overline{g} \in \mathrm{star}(\mathcal{G}, g_0)$,

$$\Big|\mathcal{D}^2_g \mathcal{D}_{\boldsymbol{\theta}} L_P(\boldsymbol{\theta}^*, \overline{g})\big[\boldsymbol{\theta} - \boldsymbol{\theta}^*, g - g_0, g - g_0\big]\Big| \leq \sqrt{2} M_\lambda \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_{L^2(\mu\times\mu;\ell^2)} \|g - g_0\|^2_{L^4(Q^0_X\times\mu;\ell^2)}.$$

We next establish that the loss under data fusion is Lipschitz continuous in its first argument, when

evaluated at the nuisance estimators $\widehat{g}$ from Algorithm 1. To see this, note that

$$\begin{aligned}
&\left|\ell_{P^0}(\boldsymbol{\theta}_1, \widehat{g}; z) - \ell_{P^0}(\boldsymbol{\theta}_2, \widehat{g}; z)\right| = \left|\ell_{P^0}(\theta_1, \widehat{g}; z) - \ell_{P^0}(\theta_2, \widehat{g}; z)\right| \\
&= \Big|\big(\theta_1(b) - \theta_2(b)\big)\Big(\theta_1(b) + \theta_2(b) - 2\widehat{\tau}(b) + 2\widehat{\xi}\,\mathbf{1}(s \in \mathcal{S}_X)\big(\widehat{\tau}(b) - \widehat{m}(b,x)\big)\Big) \\
&\quad + \big(\theta_1(a) - \theta_2(a)\big)\Big(2\widehat{\eta}\,\widehat{w}(x,a)\,\mathbf{1}(s \in \mathcal{S}_Y)\big(\widehat{m}(x,a) - y\big)\Big)\Big|
\end{aligned}$$

This expression can be written as an inner product and bounded by

$$\left\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\right\|_2 \cdot B(z),$$

where $B(z)$ is defined as

$$B(z) := \left\| \begin{pmatrix} \theta_1(b) + \theta_2(b) - 2\widehat{\tau}(b) + 2\widehat{\xi}\,\mathbf{1}(s \in \mathcal{S}_X)(\widehat{\tau}(b) - \widehat{m}(b,x)) \\ 2\widehat{\eta}\,\widehat{w}(x,a)\mathbf{1}(s \in \mathcal{S}_Y)(\widehat{m}(x,a) - y) \end{pmatrix} \right\|_2.$$

We next derive a uniform bound on $B(z)$. First, observe that

$$\begin{aligned}
B(z) &\le \Big|\theta_1(b) + \theta_2(b) - 2\widehat{\tau}(b) + 2\widehat{\xi}\mathbf{1}(s \in \mathcal{S}_X)\big(\widehat{\tau}(b) - \widehat{m}(b,x)\big)\Big| \\
&\quad + \left|2\widehat{\eta}\,\widehat{w}(x,a)\mathbf{1}(s \in \mathcal{S}_Y)\big(\widehat{m}(x,a) - y\big)\right| \\
&\le 2\|\theta\|_\infty + 2|\widehat{\tau}(b)| + 2\widehat{\xi} \cdot \left|\widehat{\tau}(b) - \widehat{m}(x,b)\right| + 2\widehat{\eta} \cdot \widehat{w}(x,a)\mathbf{1}(s \in \mathcal{S}_Y)\big(|\widehat{m}(x,a)| + |y|\big) \\
&\le 4 + 4\widehat{\xi} + 2\widehat{\eta} \cdot \widehat{w}(x,a)\mathbf{1}(s \in \mathcal{S}_Y)\big(1 + |y|\big), \quad Q_X^0 \times \mu\text{-a.e.},
\end{aligned}$$

by $\|\theta\|_\infty \le 1$ and Condition 3(b), and, under the consistency assumptions on $\widehat{\xi}, \widehat{\eta}$, and $\widehat{w}$, the following hold with high probability:

1. Since $|\widehat{\xi} - \xi_0|^2 \xrightarrow{P^0} 0$, there exists $N(\delta)$ such that, for all $n \ge N(\delta)$,

$$P^0\Big(|\widehat{\xi} - \xi_0| \le \delta\xi_0\Big) \ge 1 - \delta/12, \quad \text{and thus} \quad P^0\Big(\widehat{\xi} \le (1+\delta)\xi_0\Big) \ge 1 - \delta/12.$$

2. Similarly, since $|\widehat{\eta} - \eta_0|^2 \xrightarrow{P^0} 0$, there exists $N(\delta)$ such that, for all $n \ge N(\delta)$,

$$P^0\Big(|\widehat{\eta} - \eta_0| \le \delta\eta_0\Big) \ge 1 - \delta/12, \quad \text{and thus} \quad P^0\Big(\widehat{\eta} \le (1+\delta)\eta_0\Big) \ge 1 - \delta/12.$$

3. Since $\|\widehat{w} - w_0\|_{L^\infty(Q_X^0 \times \mu)} \xrightarrow{P^0} 0$, there exists $N(\delta)$ such that, for all $n \ge N(\delta)$,

$$P^0\left(|\widehat{w}(x,a)| \le (1+\delta) \cdot \|w_0\|_\infty \text{ for } Q_X^0 \times \mu\text{-almost every } (x,a)\right) \ge 1 - \delta/12.$$

Combining these results, we conclude that

$$B(z) \le 4 + 4(1+\delta)\xi_0 + 2(1+\delta)^2\eta_0\,\|w_0\|_\infty\,\mathbf{1}(s \in \mathcal{S}_Y)(1 + |y|),$$

$Q_X^0 \times \mu$-a.e. with probability at least $1 - \delta/4$.

We now use that Y is sub-exponential, as stated in Condition 2(b). This moment condition yields the following sub-exponential tail bound (see, e.g., Vershynin (2018)): for any $t > 0$,

$$\begin{aligned}
&P^0(\mathbf{1}(S \in \mathcal{S}_Y)\ c_w|Y - \theta_0(a)| \geq t \mid X = x, A = a) \\
&= P^0(S \in \mathcal{S}_Y,\ c_w|Y - \theta_0(a)| \geq t \,|\, X = x, A = a) \\
&= P^0(S \in \mathcal{S}_Y \mid X = x, A = a)\ P^0\Big(|Y - \theta_0(a)| \geq \tfrac{t}{c_w} \,\Big|\, X = x, A = a, S \in \mathcal{S}_Y\Big) \\
&\leq 2\exp\Big(-\tfrac{1}{2}\min\Big\{\tfrac{(t/c_w)^2}{\sigma^2},\ \tfrac{t/c_w}{L}\Big\}\Big),
\end{aligned}$$

where $c_w := \eta_0 \|w_0\|_{L^\infty(Q_X^0 \times \mu)}$.
To derive a high-probability bound, set the right-hand side less than or equal to $\delta/4$, i.e.,

$$2\exp\Big(-\tfrac{1}{2}\min\Big\{\tfrac{(t/c_w)^2}{\sigma^2},\ \tfrac{t/c_w}{L}\Big\}\Big)\ \leq\ \frac{\delta}{4}.$$

This inequality holds whenever

$$t\ \geq\ c_w \cdot \max\Big\{\sigma\sqrt{2\log\tfrac{8}{\delta}},\ \ 2L\log\tfrac{8}{\delta}\Big\}.$$

Consequently, with probability at least $1 - \delta/4$,

$$\mathbf{1}\{s \in \mathcal{S}_Y\}\,\eta_0\|w_0\|_\infty\,|y - \theta_0(a)|\ \leq\ \eta_0\|w_0\|_\infty \cdot \max\Big\{\sigma\sqrt{2\log\tfrac{8}{\delta}},\ \ 2L\log\tfrac{8}{\delta}\Big\}.$$

In turn, this bound implies that, with probability at least $1 - \delta/4$,

$$\begin{aligned}
\mathbf{1}\{s \in \mathcal{S}_Y\}\,\eta_0\|w_0\|_\infty\,|y| &\leq \mathbf{1}\{s \in \mathcal{S}_Y\}\,\eta_0\|w_0\|_\infty\,|\theta_0(a)| + \mathbf{1}\{s \in \mathcal{S}_Y\}\,\eta_0\|w_0\|_\infty\,|y - \theta_0(a)| \\
&\leq C_{\sigma,\delta}\ \eta_0\|w_0\|_\infty,
\end{aligned}$$

where $C_{\sigma,\delta} := 1 + \max\Big\{\sigma\sqrt{2\log(8/\delta)},\ 2L\log(8/\delta)\Big\}$. Putting everything together, we obtain

$$\begin{aligned}
B(z) &\leq 4 + 4(1+\delta)\xi_0 + 2(1+\delta)^2\eta_0\|w_0\|_\infty(1 + C_{\sigma,\delta}) \\
&\leq 4(1+\delta)^2\big(1 + \xi_0 + C_{\sigma,\delta}\ \eta_0\|w_0\|_\infty\big) =: B_{P^0}(\delta),
\end{aligned}$$

$Q_X^0 \times \mu$-a.e. with probability at least $1 - \delta/2$ for sufficiently large n. Therefore,

$$|\ell_{P^0}(\boldsymbol{\theta}_1, \widehat{g}; z) - \ell_{P^0}(\boldsymbol{\theta}_2, \widehat{g}; z)| \leq B_{P^0}(\delta)\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|_2,$$

with probability at least $1 - \delta/2$ for sufficiently large n. This further entails

$$\begin{aligned}
\|\ell_{P^0}(\boldsymbol{\theta}_1, \widehat{g}; z) - \ell_{P^0}(\boldsymbol{\theta}_2, \widehat{g}; z)\|_{L^\infty(Q_X^0 \times \mu)} &\leq B_{P^0}(\delta)\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|_{L^\infty(\mu;\ell^2)}, \\
\|\ell_{P^0}(\boldsymbol{\theta}_1, \widehat{g}; z) - \ell_{P^0}(\boldsymbol{\theta}_2, \widehat{g}; z)\|_{L^2(Q_X^0 \times \mu)} &\leq B_{P^0}(\delta)\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|_{L^2(\mu;\ell^2)},
\end{aligned}$$

again with probability at least $1 - \delta/2$ for sufficiently large n.

Next, we show that the excess risk bound can be expressed as the sum of the squared critical radius multiplied by the Lipschitz constant, and the fourth power of the nuisance estimation error. Applying Theorem 3 of Foster and Syrgkanis (2023) with the parameters

$$K_2 = 2, \quad R = \sqrt{2}, \quad \beta_1 = 2, \quad \beta_2 = 2M_\lambda, \quad r = 0, \quad \lambda = 2, \quad \kappa = 0, \quad B_1 = B_{P^0}(\delta)^2, \quad B_2 = M_\lambda^2,$$

we obtain

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

for some universal constant $C > 0$ (as in the proof of Theorem 3 in Foster and Syrgkanis (2023)). This bound holds for finite samples n with probability at least $1 - \delta/2$. Since we have shown that the loss is Lipschitz in its first argument at $\widehat{g}$ with probability at least $1 - \delta/2$ for large n, we can combine both results and apply a union bound to conclude that the excess risk bound holds with probability at least $1 - \delta$ for large n. □

We have established the excess risk bound for the estimators in Algorithm 1. We now turn to Algorithm 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.

B.6 Excess Risk Bound for Algorithm 2 (Proof of Theorem 2)

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

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

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 3, we revisit the higher-order smoothness condition (d) in the proof of Theorem 1 to adjust it for the kernel ridge regression setting. By Lemma 5.1 of Mendelson and Neeman (2010), for all $\theta \in \Theta$,

$$\|\theta - \theta^*\|_{L^\infty(\mu)} \leq C_p \|\theta - \theta^*\|_{\mathcal{H}}^p \|\theta - \theta^*\|_{L^2(\mu)}^{1-p}, \tag{S5}$$

where C_p is a constant depending on A, p, G in Condition 4(b). By Hölder's inequality and (S5),

$$\begin{aligned} \left|\mathcal{D}_g^2 \mathcal{D}_{\boldsymbol{\theta}} L_P(\boldsymbol{\theta}^*, \overline{g})\left[\boldsymbol{\theta} - \boldsymbol{\theta}^*, g - g_0, g - g_0\right]\right| &\leq 2\mathbb{E}_{Q_X^0 \times \mu}\left[\left|(\theta - \theta^*)(g - g_0)^T A (g - g_0)\right|\right] \\ &\leq 2\|(\theta - \theta^*)\|_{L^\infty(\mu)} \mathbb{E}_{Q_X^0 \times \mu}\left[\left|(g - g_0)^T A (g - g_0)\right|\right] \\ &\leq 2M_\lambda \|(\theta - \theta^*)\|_{L^\infty(\mu)} \|g - g_0\|_{L^2(Q_X^0 \times \mu, \ell^2)}^2 \\ &\leq 2M_\lambda C_p \|\theta - \theta^*\|_{\mathcal{H}}^p \|\theta - \theta^*\|_{L^2(\mu)}^{1-p} \|g - g_0\|_{L^2(Q_X^0 \times \mu, \ell^2)}^2 \end{aligned}$$

$$\leq 2M_\lambda C_p c^p \|\theta - \theta^*\|^{1-p}_{L^2(\mu)} \|g - g_0\|^2_{L^2(Q^0_X \times \mu, \ell^2)}$$
$$\leq 2M_\lambda C_p c^p \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|^{1-p}_{L^2(\mu)} \|g - g_0\|^2_{L^2(Q^0_X \times \mu, \ell^2)}.$$

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:

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

with C is an universal constant and probability at least $1 - \delta$ 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, $\delta_n = K c^{\frac{p}{1+p}} \log(n)^{\frac{1}{1+p}} n^{-\frac{1}{2(1+p)}}$ is a valid critical radius and if we plug this critical radius into the sample bound above,

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

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

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

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

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

B.7 Lipschitz Constant Decreases in Data Fusion (Lemma S4)

Lemma S4 (Lipschitz constant decreases in data fusion)**.** *Suppose the conditions of Theorem 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 S4. In the proof of Theorem 1, we showed that there exists $N(\delta)$ such that, for all $n \geq N(\delta)$, with probability at least $1 - \delta/2$,

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

By a similar argument as in the proof of Theorem 1, for all $n \geq N(\delta)$,

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

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,

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

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

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

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

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

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 2.

B.8 Data Fusion Improves the Excess Risk Upper Bound (Proof of Theorem 3)

Proof. Under the assumptions of Theorem 2, there exists a universal constant $C > 0$ such that

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

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

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

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

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

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. □

B.9 Minimax Lower Bound without Data Fusion (Proof of Lemma 2)

Define

$$N_\delta := \frac{1}{4} \left(\sqrt{3 \log \frac{2}{\delta} \;+\; 4 \left(\frac{\delta}{32} \right)^{-1-\frac{(1-2\alpha)}{p}}} \;-\; \sqrt{3 \log \frac{2}{\delta}} \right)^2. \tag{S6}$$

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

Proof of Lemma 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 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 $r \geq 0$,

$$\begin{aligned} P^n\{\|\underline{\theta}_\tau - \theta_Q\|^2 \geq t\} &\geq P^n\{\|\underline{\theta}_\tau - \theta_Q\|^2 \geq t, K \leq r\} \\ &= \sum_{k=1}^{r} Q^k\{\|\tau_k - \theta_Q\|^2 \geq t\} P^n\{K = k\} \\ &\geq \min_{k:0 \leq k \leq r} Q^k\{\|\tau_k - \theta_Q\|^2 \geq t\} \; P^n\{K \leq r\}. \end{aligned}$$

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

$$\begin{aligned}
&\inf_{\tau} \sup_{Q \in \mathcal{Q}} \sup_{P \in \mathcal{P}(Q)} P^n\{\|\underline{\theta}_\tau - \theta_Q\|^2 \geq t\} \\
&\geq (1 - \delta/2) \inf_{\tau} \sup_{Q \in \mathcal{Q}} \min_{k:0 \leq k \leq r} Q^k\{\|\tau_k - \theta_Q\|^2 \geq t\} \\
&\geq (1 - \delta/2) \sup_{\Pi} \inf_{\tau} \min_{k:0 \leq k \leq r} \int Q^k\{\|\tau_k - \theta_Q\|^2 \geq t\} \Pi(dQ), \qquad \text{(S7)}
\end{aligned}$$

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,

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

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 $j^\star := \arg\min_{j:0 \leq j \leq r} \int Q^j\{\|\tau_j - \theta_Q\|^2 \geq t\} \Pi(dQ)$. This construction then implies that

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

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 (S7), this shows that

$$\inf_{\tau} \sup_{Q \in \mathcal{Q}} \sup_{P \in \mathcal{P}(Q)} P^n\{\|\underline{\theta}_\tau - \theta_Q\|^2 \geq t\} \geq (1 - \delta/2) \sup_{\Pi} \inf_{\tau} \int Q^r\{\|\tau_r - \theta_Q\|^2 \geq t\} \Pi(dQ). \qquad \text{(S8)}$$

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 $M \geq 2^{r^{p/[p+(1-2\alpha)]}/8}$ 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:* $\|\theta_j - \theta_k\| \geq c_\delta \kappa \, r^{-\frac{1-2\alpha}{p+(1-2\alpha)}}$ for all $0 \leq j < k \leq M$.

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

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

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

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 (S8) lower bounds as

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

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 $M \geq 2^{n/8}$ 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 (S6), $r := C_{\delta,v} n_{XY} \geq (\delta/32)^{-[p+(1-2\alpha)]/p}$, and so the right-hand side is lower bounded by $1 - \delta/2$. Plugging in our chosen t and recalling (S8) 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 (S9) 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 (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 $Y \mid A = X, A = x \sim \mathcal{N}(\theta_i(a), \overline{\sigma}^2)$ where $\overline{\sigma} := \min(\sigma, L)$, where σ and L are defined in Condition 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 2(b), (ii) $\mathbb{E}_{Q_i}[Y|A, X]$ 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 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 $i \neq j$,

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

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

$$c_\delta/8 = \min \left\{ 2^{-2(1-2\alpha)/p} R/(16\delta G_1), \ (\overline{\sigma}^2 \log 2) 2^{-(1-2\alpha)/p}/256, \ \frac{2^{-(1-2\alpha)/p}}{16\delta \left(\sup_l \|e_l\|_\infty\right)^2} \right\}$$

$$\geq \min\left\{2^{-2(1-2\alpha)/p}R/(16G_1),\ (\overline{\sigma}^2\log 2)2^{-(1-2\alpha)/p}/256,\ \frac{2^{-(1-2\alpha)/p}}{16\big(\sup_l \|e_l\|_\infty\big)^2}\right\} =: c_{\mathcal{Q}}. \tag{S9}$$

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

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

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

□

B.10 Sufficient Conditions for Data Fusion to Improve Worst-Case Performance (Proof of Theorem 4)

Proof. We will show that, for all n, n_{XY} large enough, the first term in (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 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 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

$$\begin{aligned} B(z) &\leq \Big|\theta_1(b) + \theta_2(b) - 2\widehat{\tau}(b) + 2\widehat{\xi}\,\mathbf{1}(s \in \mathcal{S}_X)\big(\widehat{\tau}(b) - \widehat{m}(b,x)\big)\Big| \\ &\quad + \Big|2\,\widehat{\eta}\,\widehat{w}(x,a)\,\mathbf{1}(s \in \mathcal{S}_Y)\big(\widehat{m}(x,a) - y\big)\Big| \\ &\leq 2\|\theta\|_\infty + 2|\widehat{\tau}(b)| + 2\widehat{\xi}\,\big|\widehat{\tau}(b) - \widehat{m}(b,x)\big| + 2\widehat{\eta}\,\widehat{w}(x,a)\,\mathbf{1}(s \in \mathcal{S}_Y)\big(|\widehat{m}(x,a)| + |y|\big) \\ &\leq 4 + 4\widehat{\xi} + 2\widehat{\eta}\,\widehat{w}(x,a)\,\mathbf{1}(s \in \mathcal{S}_Y)\big(1 + |y|\big) \\ &\leq 4 + 4M_\xi + 2M_\eta M_w\,\mathbf{1}(s \in \mathcal{S}_Y)\,(1 + |y|) \qquad Q_X^0 \times \mu\text{-a.e.}, \end{aligned}$$

by $\|\theta\|_\infty \leq 1$ and Condition 3. By Condition 2(b) and the same argument as in the proof of Theorem 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 $1 - 3/4\delta$. By (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 an 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 $n \geq N_4(\delta)$,

$$C' \, (n \log(n)^{-2})^{-\rho} \leq \tfrac{c_Q}{4} \, \delta \, n_{XY}^{-\rho} \, (\log n)^{-h\rho/2} = r(\delta, n) \, (\log n)^{-h\rho/2}.$$

Letting

$$N_\delta := \max\{N_1(\delta), N_2(\delta), N_3(\delta), N_4(\delta)\}, \tag{S10}$$

we conclude

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

□