Structure-Adaptive Estimation of Heterogeneous Treatment Effects with Kernel Methods

Seok-Jin Kim
Columbia University
New York, NY, USA
seok-jin.kim@columbia.edu

Abstract

We study conditional average treatment effect (CATE) estimation from observational data in regimes with potentially weak overlap, when the nuisance response functions lie in a reproducing kernel Hilbert space (RKHS), but the contrast function is structurally simpler. We consider three forms of such simplicity: (i) membership in a lower-complexity RKHS with faster eigenvalue decay, (ii) a source condition relative to the nuisance kernel, or (iii) dependence on a known low-dimensional covariate representation. We develop a unified two-stage kernel ridge regression procedure that remains valid under weak overlap and attains minimax rates determined by the complexity of the contrast function. The procedure avoids propensity-score estimation and separate nuisance convergence-rate assumptions. We also show that a simple model-selection step over candidate contrast spaces and regularization levels satisfies an oracle inequality, enabling adaptation to unknown CATE regularity.

1 Introduction

Estimating heterogeneous treatment effects is a central challenge in causal inference. In particular, conditional average treatment effects (CATEs) are fundamental for personalized decision-making in domains ranging from precision medicine to economics [Künzel et al., 2019, Kennedy, 2023]. While the Average Treatment Effect (ATE) provides a population-level summary, it can obscure substantial heterogeneity across individuals. A recurring feature in this literature is the structural asymmetry between the treatment effect of interest and the nuisance components, such as the baseline response functions. Empirically, nuisance functions may be highly complex, non-smooth, or high-dimensional even when the CATE itself is smooth, sparse, or nearly constant [Kennedy, 2023, Kennedy et al., 2024, Kato and Imaizumi, 2023]. A central theoretical question is whether one can exploit this simpler contrast structure to obtain convergence rates faster than those dictated by nuisance learning, while still achieving minimax-optimal performance.

A standard route to exploiting such structure is doubly robust (DR) estimation or double machine learning (DML). These methods typically yield an oracle term together with a product of nuisance estimation errors, but the norms involved can be quite demanding in CATE problems. For ATE, the product is often controlled in L^2, whereas for CATE it may require stronger L^4-type control, and for pointwise guarantees it can require pointwise accuracy of both the outcome regression and the propensity score. As a result, obtaining sharp adaptive guarantees is substantially harder in the heterogeneous-treatment setting. A second challenge is weak overlap. When the propensity score approaches 0 or 1, misspecification or estimation error in the propensity model can be strongly amplified through inverse-propensity weighting, obstructing optimal rates and making such procedures unstable. This has motivated a growing line of work that seeks to reduce or bypass direct reliance on propensity-score estimation in limited-overlap regimes [Hainmueller, 2012, Imai and Ratkovic, 2014, Kim and Wang, 2026].

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

At the same time, the standard DML product bound is not always rate-optimal in smoother regimes. For Hölder classes, higher-order influence-function (HOIF) methods show that one can adapt to the complexity of the CATE itself and attain rates strictly faster than the standard product-of-errors heuristic [Kennedy et al., 2024]. This fast-rate phenomenon is by now well understood in Hölder spaces, but its analogue in reproducing kernel Hilbert spaces (RKHSs) remains largely unexplored. To our knowledge, we give the first RKHS analysis of CATE with rates governed by a lower-complexity contrast class that are minimax-optimal in both sample size and overlap, for both L^2 and pointwise loss.

We also study the weak-overlap regime, indexed by a parameter $\kappa \in [0, 1]$ that quantifies how close the propensity score can be to 0 or 1. The standard pointwise overlap condition that the propensity score lies in $[\kappa, 1 - \kappa]$ is a representative example, but our analysis also accommodates more singular cases. Rather than treating κ as a fixed constant, we track its effect explicitly throughout the non-asymptotic analysis and establish optimality with respect to κ.

1.1 Contribution

We study CATE estimation from observational data when the nuisance outcome functions belong to an RKHS $\mathcal{F}$, while the CATE function $h^\star$ belongs to a strictly simpler class $\mathcal{H}$. Our goal is to match the statistical difficulty of regression over $\mathcal{H}$, while still allowing for weak overlap and without imposing structural assumptions on the propensity score beyond overlap itself. We consider three representative forms of simplicity of the CATE function $h^\star$: (i) $\mathcal{H}$ is an RKHS subspace with faster spectral decay than $\mathcal{F}$; (ii) $h^\star$ satisfies a source condition relative to $\mathcal{F}$; and (iii) $h^\star$ is represented through a known low-dimensional feature map. We study both population L^2 error and pointwise evaluation. Our central message is that, in RKHS outcome models, well-specified nuisance functions are enough to recover learning rates that adapt to the complexity of $\mathcal{H}$.

We achieve this with a unified, imputation-based two-stage kernel ridge regression (KRR) framework that applies across these three settings. It attains minimax-optimal rates for both L^2 error and pointwise evaluation, with the rate governed by the complexity of $\mathcal{H}$ rather than by the ambient nuisance space $\mathcal{F}$. In the appendix, we additionally establish an oracle inequality for a simple model-selection scheme over candidate contrast spaces and regularization levels when $\mathcal{H}$ is unknown.

Our guarantees remain valid in weak-overlap regimes, where κ may be small and the propensity score may even take the values 0 or 1 on subsets of the covariate space. In such settings, procedures that rely on propensity-score estimation can be especially fragile under weak overlap and even mild misspecification [Hainmueller, 2012, Imai and Ratkovic, 2014, Zhou et al., 2020]. By contrast, our analysis avoids propensity-score estimation altogether, requires only well-specified outcome models, and still attains rates that are optimal in both n and κ.

We summarize our contributions below:

- **Theory:** We establish optimal dependence on both n and κ, matching the oracle regression rate in $\mathcal{H}$ with effective sample size $n\kappa$. We impose no nuisance learning-rate assumption on $\mathcal{F}$ beyond well-specifiedness, including for pointwise evaluation. When $\mathcal{H}$ is unknown, we also prove a model-selection oracle inequality for the L^2 error.
- **Methodology:** We provide a unified estimation framework based on imputation and two-stage KRR that covers all three structural models. In the appendix, we complement this with a model-selection step over candidate $\mathcal{H}$ and regularization levels to adapt to unknown contrast complexity.

1.2 Related Work

RKHS in Causal Inference. Wang and Kim [2023], Mou et al. [2023] studied efficient ATE estimation within the RKHS framework. Related approaches include kernel balancing for partially conditional effects [Wang et al., 2022], kernel ridge Riesz representers [Singh, 2021], and continuous-treatment effects [Kim and Wang, 2026]. For CATE, Nie and Wager [2021], Foster and Syrgkanis [2023] analyzed KRR-based estimators under conditions requiring consistent propensity score estimation and sufficiently fast nuisance learning. Singh et al. [2024] proposed RKHS-based methods for CATE, but it remains unclear whether their rates adapt to the complexity of $h^\star$.

Weak Overlap, Bypassing Propensity Score Estimation. Our contribution is to show that, in a broad RKHS setting, the leading rate can be governed by the contrast complexity alone, without relying on propensity score consistency. This distinction is especially important under weak overlap, where propensity-based procedures can become unstable and highly sensitive to misspecification, motivating more robust weighting and overlap-diagnostic methods for such regimes [Zhou et al., 2020, Ma and Wang, 2020, Ma et al., 2025, Hainmueller, 2012, Imai and Ratkovic, 2014].

Structural Adaptivity. Kennedy [2023], Kennedy et al. [2024], Gao and Han [2020] established optimal rates for Hölder classes using DR estimators and U-statistics. More recent work has investigated particular forms of low-complexity structure, including sparse linear CATEs [Kato and Imaizumi, 2023] and CATE functions lying in an RKHS with smaller Hilbert norm than the nuisance functions [Kim et al., 2025]. Standard DML/meta-learner approaches instead emphasize orthogonalization and nuisance-rate products in semiparametric settings [Oprescu et al., 2019, Künzel et al., 2019]. We provide a more detailed comparison with existing methods in Section 4.4.

1.3 Notation

The constants $c, C, c_1, C_1, \ldots$ may vary from line to line. We use the symbol $[n]$ as a shorthand for $\{1, 2, \ldots, n\}$. For nonnegative sequences $\{a_n\}_{n=1}^{\infty}$ and $\{b_n\}_{n=1}^{\infty}$, we write $a_n \lesssim b_n$ or $a_n = \mathcal{O}(b_n)$ if there exists a positive constant C such that $a_n \leq C b_n$ for all n. We use $\tilde{\mathcal{O}}$ to denote bounds up to polylogarithmic factors. Additionally, we write $a_n \asymp b_n$ if $a_n \lesssim b_n$ and $b_n \lesssim a_n$. For a bounded linear operator A, we use $\|A\|_{\mathrm{op}}$ to denote its operator norm. For any u and v in a Hilbert space $\mathbb{H}$, their inner and outer products are denoted by $\langle u, v\rangle_{\mathbb{H}}$ and $u \otimes v$, respectively. When the context is clear, we denote the inner product as $u^\top v := \langle u, v\rangle_{\mathbb{H}}$.

2 Problem Setup

2.1 Treatment Regime and CATE

We observe n independent and identically distributed (i.i.d.) observational samples $\mathcal{D} = \{(x_i, a_i, y_i)\}_{i=1}^n$, where $x_i \in \mathcal{X}$ denotes the covariates, $a_i \in \{0, 1\}$ is a binary treatment indicator, and $y_i \in \mathbb{R}$ is the response. For a generic observation (x, a, y), let $\mathcal{P}_x$ denote the marginal distribution of x on the covariate space $\mathcal{X}$, which we assume to be a regular domain in $\mathbb{R}^d$. We write $\pi(x) := \mathbb{P}(a = 1 \mid x)$ for the propensity score.

Let $\mathcal{F}$ be an RKHS containing the nuisance functions $f_0^\star, f_1^\star$, with kernel $K_{\mathcal{F}}(\cdot, \cdot)$. Our response model is given by:

$$\text{(Nuisances)} \quad \mathbb{E}[y \mid x, a = 0] = f_0^\star(x), \quad \mathbb{E}[y \mid x, a = 1] = f_1^\star(x).$$

Throughout, we assume that the norms $\|f_0^\star\|_{\mathcal{F}}$ and $\|f_1^\star\|_{\mathcal{F}}$ are bounded by a universal constant $M > 0$. Let $\mathcal{H}$ be another function space. Our estimand of interest is the CATE function, defined as:

$$\text{(CATE)} \quad h^\star(x) := f_1^\star(x) - f_0^\star(x).$$

Our focus is the regime in which $h^\star \in \mathcal{H}$ and the hypothesis class $\mathcal{H}$ is *less complex* than the ambient nuisance space $\mathcal{F}$. In this regime, we seek sharp error bounds that adapt to the complexity of $\mathcal{H}$ and, ideally, match the minimax rates for a regression problem restricted to $\mathcal{H}$ itself.

We evaluate performance using two metrics: the (squared) $L^2(\mathcal{P}_x)$-error and the (squared) pointwise evaluation error at a fixed query point $x_0 \in \mathcal{X}$:

$$\mathcal{E}_{L^2}(\hat{h}) := \mathbb{E}_{x \sim \mathcal{P}_x} \left|\hat{h}(x) - h^\star(x)\right|^2, \qquad \mathcal{E}_{x_0}(\hat{h}) := \left|\hat{h}(x_0) - h^\star(x_0)\right|^2.$$

These are standard measures in nonparametric estimation [Wainwright, 2019, Bühlmann and Van De Geer, 2011].

2.2 Three Models of Structural Simplicity

We consider three structural scenarios in which the CATE is simpler than the nuisance components. In each case, $\mathcal{H}$ denotes a target class that is strictly less complex than $\mathcal{F}$, but the nature of this simplicity differs.

The first scenario is the most direct one: $\mathcal{H}$ is an RKHS subspace of $\mathcal{F}$ with faster eigenvalue decay.

Model 1 (Subspace)**.** $\mathcal{H}$ is an RKHS with $\mathcal{H} \subset \mathcal{F}$, and $\mathcal{H}$ exhibits faster spectral decay than $\mathcal{F}$.

Under Model 1, $\mathcal{H}$ is an RKHS; we denote its associated kernel by $K_{\mathcal{H}}$. Examples of Model 1 include:

- **Parametric vs. Nonparametric:** $\mathcal{F}$ is an infinite-dimensional RKHS (e.g., a Sobolev class), while $\mathcal{H}$ is a finite-dimensional linear subspace (e.g., polynomials).
- **Differential Sobolev Smoothness:** $\mathcal{H}$ corresponds to a smoother function class than $\mathcal{F}$. For instance, $\mathcal{F} = H^{\beta}(\mathcal{X})$ and $\mathcal{H} = H^{\gamma}(\mathcal{X})$ with $\gamma \geq \beta > \frac{d}{2}$.
- **Mixed Smoothness:** We can consider mixed Sobolev spaces [Kühn et al., 2015, Suzuki, 2018] $\mathcal{F} = H^{\beta}_{\text{mix}}(\mathcal{X})$ and $\mathcal{H} = H^{\gamma}_{\text{mix}}(\mathcal{X})$ with $\gamma \geq \beta > 1/2$, thereby relaxing the isotropic requirement $\beta > d/2$.

Relaxation via Mixed Smoothness. Isotropic Sobolev spaces $H^{\beta}(\mathcal{X})$ require $\beta > d/2$ to define an RKHS, whereas mixed Sobolev spaces $H^{\beta}_{\text{mix}}(\mathcal{X})$ relax this to $\beta > 1/2$, independently of d [Suzuki, 2018, Kühn et al., 2015]. For example, H^{1}_{mix} controls mixed first-order derivatives such as $\partial_{x_1}\partial_{x_2}\cdots\partial_{x_d} f$ without requiring higher-order derivatives in each coordinate, such as $\partial^2_{x_1} f$. This makes H^{β}_{mix} a natural model for high-dimensional CATE functions that are smooth through coordinate-wise interactions while avoiding stringent isotropic smoothness.

The second scenario formalizes simplicity through spectral regularity relative to the nuisance kernel.

Model 2 (Source Condition)**.** We assume $h^{\star}$ satisfies a source condition with parameter $0 < \nu < 1$ with respect to the kernel of $\mathcal{F}$.

Formally, let $T_{\mathcal{F}} : L^2(\mathcal{P}_x) \to L^2(\mathcal{P}_x)$ be the integral operator associated with the kernel of $\mathcal{F}$. The source condition is $h^{\star} \in \text{Range}(T_{\mathcal{F}}^{\frac{1+\nu}{2}})$. Source conditions are standard in RKHS theory [Singh et al., 2024, Jun et al., 2019, Chen et al., 2024]; they quantify smoothness by requiring the spectral coefficients of $h^{\star}$ to decay faster than those of a generic element of $\mathcal{F}$. Equivalently, $h^{\star}$ belongs to a fractional power space $[\mathcal{F}]^{\nu}$, which is strictly smaller and smoother than $\mathcal{F}$. This perspective is also central in recent analyses of overparameterized neural networks via the neural tangent kernel (NTK) [Zhang et al., 2025, Ghorbani et al., 2021].

The final scenario captures simplicity through a known low-dimensional representation.

Model 3 (Low-Dimensional Structure)**.** We assume there exists a known transformation $\tilde{x}_i$ of x_i with $\dim(\tilde{x}_i) < d$, and a function $\tilde{h}^{\star} \in \tilde{\mathcal{H}}$, where $\tilde{\mathcal{H}}$ is an RKHS, such that:

$$h^{\star}(x_i) = \tilde{h}^{\star}(\tilde{x}_i).$$

Common examples include variable selection, where $\tilde{x}_i$ is a subset of coordinates, and linear projections, where $\tilde{x}_i = \mathbf{P}x_i$ for a projection $\mathbf{P}$. For example, one may take $\mathcal{F} = H^{\beta}(\mathcal{X})$ and $\mathcal{H} = H^{\beta}(\tilde{\mathcal{X}})$, where $\tilde{\mathcal{X}}$ is the space of $\tilde{x}_i$. Because $\dim(\tilde{x}_i) < \dim(x_i)$, the statistical complexity of estimating $\tilde{h}^{\star}$ is strictly smaller, and the goal is to adapt to this lower intrinsic dimension. Under Model 3, $\tilde{\mathcal{H}}$ is an RKHS; we denote its associated kernel by $K_{\tilde{\mathcal{H}}}$. We set $\dim(\tilde{x}_i) = \tilde{d} < d$.

2.3 Assumptions

We impose the following standard assumptions common to causal inference and nonparametric regression [Künzel et al., 2019, Kennedy, 2023, Wainwright, 2019].

Assumption 1 (Consistency and Unconfoundedness)**.** Let y^0 and y^1 be the potential outcomes. We observe $y = y^a$ (consistency), and $(y^0, y^1) \perp\!\!\!\perp a \mid x$ (unconfoundedness).

Assumption 2 (Sub-Gaussian Noise)**.** Conditioned on x_i, a_i, the noise is centered and satisfies $\mathbb{E}[\exp(t\varepsilon_i) \mid x_i, a_i] \le \exp(\sigma^2 t^2/2)$ for every $t \in \mathbb{R}$, with a uniform constant upper bound σ. This permits heteroscedasticity: the conditional variance need not equal σ^2 or be constant.

Assumption 3 (Boundedness of Kernels)**.** There exists a universal constant $\xi > 0$ such that

$$\sup_{x \in \mathcal{X}} K_{\mathcal{F}}(x, x) \le \xi.$$

Assumption 4 (Weak Overlap)**.** Let $\phi(\cdot)$ denote the feature map associated with the kernel $K_{\mathcal{F}}$. There exists $\kappa \in (0, 1]$ such that

$$\mathbb{E}[\phi(x_i) \otimes \phi(x_i)\mathbf{1}(a_i = 1)] \preceq \frac{1}{\kappa}\left(\mathbb{E}[\phi(x_i) \otimes \phi(x_i)\mathbf{1}(a_i = 0)] + \frac{1}{n}\mathbf{I}\right),$$
$$\mathbb{E}[\phi(x_i) \otimes \phi(x_i)\mathbf{1}(a_i = 0)] \preceq \frac{1}{\kappa}\left(\mathbb{E}[\phi(x_i) \otimes \phi(x_i)\mathbf{1}(a_i = 1)] + \frac{1}{n}\mathbf{I}\right).$$

The parameter κ quantifies the degree of overlap. We do not treat κ as a fixed constant and instead track its effect explicitly in our non-asymptotic bounds. In this sense, our analysis covers the *weak-overlap* regime. Moreover, this condition is automatically satisfied under the usual pointwise overlap assumption $\kappa < \pi(\cdot) < 1 - \kappa$, but it is strictly weaker: it also permits more singular regimes in which $\pi(x)$ may take the values 0 or 1 on subsets of the covariate space [Wang, 2026, Kim et al., 2025].

For bounded kernels, the $n^{-1}I$ regularization ensures an admissible $\kappa_n \gtrsim n^{-1}$. The operator condition can remain informative without uniform pointwise overlap: for $X \sim \mathrm{Unif}[0, 1]$, $\pi(x) = x$, and $\mathcal{F} = H^1([0, 1])$, one has $\kappa_n \asymp n^{-1/3}$ and hence $n\kappa_n \asymp n^{2/3} \to \infty$ [Wang, 2026]. Empirical Gram matrices provide a diagnostic of overlap in the chosen kernel space, although they do not certify the population condition.

2.4 Fundamental Limits of CATE Estimation

To place our upper bounds in context, we first record the natural minimax benchmark for CATE estimation. The key comparison is with ordinary nonparametric regression over $\mathcal{H}$, which provides the appropriate oracle baseline once overlap is taken into account.

Definition 5. Let $\texttt{LB-L2}(N; \mathcal{H})$ denote the minimax squared L^2-error lower bound for regression over $\mathcal{H}$ with N samples. Similarly, let $\texttt{LB-PE}(N; \mathcal{H})$ denote the corresponding lower bound for squared pointwise evaluation error.

For instance, when $\mathcal{H} = H^\gamma(\mathcal{X})$ is a Sobolev space, these lower bounds satisfy $\texttt{LB-L2}(N; \mathcal{H}) \asymp N^{-2\gamma/(d+2\gamma)}$ and $\texttt{LB-PE}(N; \mathcal{H}) \asymp N^{-(2\gamma-d)/2\gamma}$ [Tuo and Zou, 2024, Wainwright, 2019]. With this notation in place, the following informal lemma states the benchmark relevant for CATE estimation.

Lemma 6 (Informal: Lower Bounds)**.** *The minimax squared L^2-error for estimating $h^\star$ is lower-bounded by $\texttt{LB-L2}(n\kappa; \mathcal{H})$, up to constant factors. Similarly, the minimax squared pointwise evaluation error at x_0 is lower-bounded by $\texttt{LB-PE}(n\kappa; \mathcal{H})$, up to constant factors.*

Proof. Consider an oracle setting in which the baseline function $f_0^\star$ is known exactly and the propensity score is constant, $\pi(x) \equiv \kappa$. In this case, estimating $h^\star$ is equivalent to estimating $f_1^\star$ using only the treated subsample, whose expected size is $n\kappa$. The CATE problem therefore reduces to standard regression over $\mathcal{H}$ with $n\kappa$ effective observations. Any rate faster than the corresponding regression lower bound would contradict minimax optimality for regression in $\mathcal{H}$. □

Consequently, for Sobolev spaces $\mathcal{H} = H^\gamma(\mathcal{X})$, the minimax lower bound for CATE estimation in L^2 error is $(n\kappa)^{-\frac{2\gamma}{d+2\gamma}}$, and the minimax lower bound for pointwise CATE evaluation is $(n\kappa)^{-\frac{2\gamma-d}{2\gamma}}$.

3 Methodology: A Unified Approach

This section presents a unified procedure for Models 1 to 3. The guiding idea is to decouple nuisance estimation from target estimation so that, after a first-stage correction, the second stage behaves like an ordinary regression problem for the contrast $h^\star$. This is what ultimately allows the final rate to adapt to the complexity of the contrast class.

3.1 Algorithm Structure

The procedure has two main ingredients: undersmoothed nuisance estimation and a switch-imputation regression step. The overall algorithm is summarized in Algorithm 1.

1. Nuisance Estimation via Undersmoothed KRR. We first estimate the conditional mean functions $f_0^\star$ and $f_1^\star$ from the observational data $\mathcal{D}$. The nuisance stage is intentionally undersmoothed so that its regularization bias is small enough not to dominate the final CATE error:

$$\hat{f}_a := \arg\min_{f \in \mathcal{F}} \Big\{ \frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i))^2 \mathbf{1}(a_i = a) + \bar{\lambda} \|f\|_{\mathcal{F}}^2 \Big\}, \quad a \in \{0, 1\}. \tag{1}$$

Here, $\bar{\lambda}$ is the nuisance regularization parameter. It must be small enough to suppress regularization bias, even at the cost of a larger nuisance variance; the key point is that this variance inflation will be absorbed in the second-stage analysis. For generic bounded kernels, the scaling $\bar{\lambda} \asymp \log(n)/n$ is sufficient.

2. Generating Pseudo-outcomes via Switch-Imputation. We next form pseudo-outcomes $\{m_i\}_{i=1}^n$ by imputing the unobserved arm with the first-stage estimator:

$$m_i := \begin{cases} y_i - \hat{f}_0(x_i), & \text{if } a_i = 1, \\ \hat{f}_1(x_i) - y_i, & \text{if } a_i = 0. \end{cases} \tag{2}$$

This switch-imputation construction isolates the treatment effect by subtracting the estimated baseline corresponding to the unobserved arm.

3. Regression Oracle for the Final Estimator. Finally, we feed the pseudo-pairs $\{(x_i, m_i)\}_{i=1}^n$ into a regression oracle $\mathscr{O}$, which returns

$$\hat{h} = \mathscr{O}(\{(x_i, m_i)\}_{i=1}^n).$$

The only model-dependent part of the procedure is the choice of this regression oracle, which reflects the structural assumptions imposed on $h^\star$; see Section 3.2.

Algorithm 1 Optimal CATE Learner with KRR

1: **Input:** Dataset $\mathcal{D} = \{(x_i, a_i, y_i)\}_{i=1}^n$, regression oracle $\mathscr{O}$, nuisance regularizer $\bar{\lambda} \asymp \frac{\log n}{n}$, main regularizer λ.
2: **Step 1: Nuisance Estimation**
3: Compute $\hat{f}_0, \hat{f}_1$ via Equation (1) using $\mathcal{D}$ and nuisance regularizer $\bar{\lambda}$.
4: **Step 2: Pseudo-outcome Generation**
5: For $i = 1, \ldots, n$, compute m_i via Equation (2) using $\hat{f}_0, \hat{f}_1$.
6: **Step 3: Target Estimation**
7: **Return** $\hat{h} = \mathscr{O}(\{(x_i, m_i)\}_{i=1}^n)$ with main regularizer λ.

3.2 Regression Oracle $\mathscr{O}$

We now instantiate $\mathscr{O}$ for the three structural models from Section 2. In every case, the oracle is a form of KRR applied to the pseudo-outcomes, with the second-stage function space chosen to reflect the complexity of $h^\star$.

Oracle for Model 1 (Subspace). Under Model 1, where $h^\star$ resides in a strictly simpler RKHS $\mathcal{H} \subset \mathcal{F}$, the oracle performs KRR directly within $\mathcal{H}$:

$$\hat{h} := \arg\min_{h \in \mathcal{H}} \Big\{ \frac{1}{n} \sum_{i=1}^{n} (m_i - h(x_i))^2 + \lambda \|h\|_{\mathcal{H}}^2 \Big\}.$$

The regularization parameter λ is tuned to the spectral complexity of $\mathcal{H}$, not that of the ambient nuisance space $\mathcal{F}$.

Oracle for Model 2 (Source Condition). Under Model 2, where $h^\star \in \mathcal{F}$ satisfies a source condition, the oracle performs KRR in the ambient space $\mathcal{F}$:

$$\hat{h} := \arg\min_{h \in \mathcal{F}} \Big\{ \frac{1}{n} \sum_{i=1}^{n} (m_i - h(x_i))^2 + \lambda \|h\|_{\mathcal{F}}^2 \Big\}.$$

Although the optimization is carried out in $\mathcal{F}$, the source condition permits a faster choice of λ and hence faster rates.

Oracle for Model 3 (Low-Dimensional Structure). Under Model 3, where $h^\star$ depends only on a low-dimensional feature projection $\tilde{x}_i$, the oracle performs KRR in the corresponding space $\tilde{\mathcal{H}}$:

$$\hat{h} := \arg\min_{\tilde{h}\in\tilde{\mathcal{H}}} \Big\{ \frac{1}{n}\sum_{i=1}^{n}(m_i - \tilde{h}(\tilde{x}_i))^2 + \lambda\|\tilde{h}\|^2_{\tilde{\mathcal{H}}} \Big\}.$$

This reduces the statistical problem to the intrinsic dimension of the low-dimensional representation.

In all three cases, we refer to λ as the **main regularizer**.

3.3 Adaptation to Unknown Contrast Structure

The contrast class and its regularization need not be specified in advance. Let Π be a finite library of candidate pairs $(\mathcal{H}, \lambda)$. For example, the library may combine polynomial models of different degrees, Sobolev kernels or NTKs, and models using selected coordinates. Running Algorithm 1 on the first data split for each pair gives a collection $\mathcal{M} = \{\hat{h}_1, \dots, \hat{h}_L\}$ of candidate CATE estimators, where $L = |\Pi|$.

For the selection guarantee, we truncate each candidate pointwise to $[-B, B]$ by setting $\bar{h}_j(x) = \min\{B, \max\{-B, \hat{h}_j(x)\}\}$, and write $\bar{\mathcal{M}} = \{\bar{h}_1, \dots, \bar{h}_L\}$. Here B is a constant upper bound on $\|h^\star\|_\infty$; exact knowledge of this norm is unnecessary.

Algorithm 2 in Appendix B gives the full model-selection procedure. A second split fits independent nuisance functions, and a third compares the truncated candidates against the resulting switch-imputed responses. We denote the candidate with the smallest empirical squared loss by $\bar{h}_{\mathrm{ms}} \in \bar{\mathcal{M}}$. This provides a validation criterion even though the true CATE is unobserved.

Under the assumptions of Section 2.3, for a library of polynomial size and the regularization specified in Theorem B.2, the selected estimator satisfies, with probability at least $1 - n^{-10}$,

$$\mathcal{E}_{L^2}(\bar{h}_{\mathrm{ms}}) \lesssim \min_{\bar{h}\in\bar{\mathcal{M}}} \mathcal{E}_{L^2}(\bar{h}) + \widetilde{\mathcal{O}}\left(\frac{1}{n\kappa}\right),$$

Thus, if the library contains a suitable contrast model and regularization level, selection recovers its rate up to the displayed remainder. This guarantee applies to L^2 loss.

4 Upper Bounds for L^2-Error

In this section, we present minimax-optimal upper bounds on the $L^2(\mathcal{P}_x)$-error of the estimator $\hat{h}$ produced by our meta-algorithm (Algorithm 1). Appendix D proves Theorem D.1, a single general error bound for second-stage KRR in a generic Hilbert space. All of the main corollaries in the paper are then obtained by instantiating this theorem with the appropriate choice of inputs and the evaluation metric.

4.1 Results under Model 1 (Subspace)

We define the kernel integral operator $\mathbf{T}_{\mathcal{H}} : L^2(\mathcal{P}_x) \to L^2(\mathcal{P}_x)$ associated with the kernel $K_{\mathcal{H}}$ and the marginal distribution $\mathcal{P}_x$ [Fischer and Steinwart, 2020, Zhang et al., 2023]. Concretely,

$$\mathbf{T}_{\mathcal{H}} f(x) = \int_{x'\in\mathcal{X}} K(x, x') f(x') \mathrm{d}\mathcal{P}_x(x').$$

Let $\rho_{\mathcal{H},1} \geq \rho_{\mathcal{H},2} \geq \cdots > 0$ denote the eigenvalues of $\mathbf{T}_{\mathcal{H}}$. The decay rate of these eigenvalues characterizes the complexity of the hypothesis space $\mathcal{H}$. For instance, for the Sobolev space $H^\beta(\mathcal{X})$, it is well established that $\rho_{\mathcal{H},j} \asymp j^{-2\beta/d}$ [Zhang et al., 2023]. We also assume the kernel is bounded, i.e., $\sup_{x\in\mathcal{X}} K_{\mathcal{H}}(x, x) = \mathcal{O}(1)$.

Corollary 7 (L^2-Error Bounds for Model 1)**.** *Suppose the assumptions in Section 2.3 and Model 1 hold. Then, with probability at least $1 - n^{-10}$, the estimator $\hat{h}$ from Algorithm 1 with oracle $\mathscr{O}$ for Model 1 and main regularizer λ satisfies the following bounds:*

1. ***Polynomial Decay:*** *If* $\rho_{\mathcal{H},j} \lesssim j^{-2\ell_{\mathcal{H}}}$ *for some* $\ell_{\mathcal{H}} > 1/2$*, setting* $\lambda \asymp (n\kappa)^{-\frac{2\ell_{\mathcal{H}}}{2\ell_{\mathcal{H}}+1}}$ *yields*

$$\mathcal{E}_{L^2}(\hat{h}) \lesssim (n\kappa)^{-\frac{2\ell_{\mathcal{H}}}{2\ell_{\mathcal{H}}+1}}.$$

2. ***Exponential Decay:*** *If* $\rho_{\mathcal{H},j} \lesssim \exp(-cj)$ *for some* $c > 0$*, setting* $\lambda \asymp (n\kappa)^{-1}$ *yields*

$$\mathcal{E}_{L^2}(\hat{h}) \lesssim \frac{1}{n\kappa}.$$

3. ***Finite Rank:*** *If* $\operatorname{rank}(\mathbf{T}_{\mathcal{H}}) \le D$*, setting* $\lambda \asymp (n\kappa)^{-1}$ *yields*

$$\mathcal{E}_{L^2}(\hat{h}) \lesssim \frac{D}{n\kappa}.$$

The notation $\lesssim$ *omits absolute constants and polylogarithmic factors.*

Discussion. We defer the proof to Appendix F.1. Corollary 7 shows that the estimator adapts to the spectral complexity of the contrast space $\mathcal{H}$, thereby decoupling the rate from the potentially slower complexity of the nuisance space $\mathcal{F}$. Comparing these bounds with Lemma 6, we see that the estimator is minimax optimal in both the sample size n and the overlap parameter κ. The corresponding Sobolev special cases follow immediately: when $\mathcal{F} = H^{\beta}(\mathcal{X})$ and $\mathcal{H} = H^{\gamma}(\mathcal{X})$, then our bound adapts to $n^{-\frac{2\gamma}{d+2\gamma}}$, which is much faster than the nuisance learning rate $n^{-\frac{2\beta}{d+2\beta}}$. Notably, on the nuisance side we assume only well-specification of $\mathcal{F}$; the eigendecay and rate-determining assumptions are imposed only on the target class $\mathcal{H}$.

4.2 Results under Model 2 (Source Condition)

We define the integral operator $\mathbf{T}_{\mathcal{F}} : L^2(\mathcal{P}_x) \to L^2(\mathcal{P}_x)$ associated with the kernel $K_{\mathcal{F}}$ and the marginal distribution $\mathcal{P}_x$. Let $\rho_{\mathcal{F},1} \ge \rho_{\mathcal{F},2} \ge \cdots > 0$ denote the eigenvalues of $\mathbf{T}_{\mathcal{F}}$.

Corollary 8 (L^2-Error Bounds for Model 2)**.** *Suppose the assumptions in Section 2.3 hold and* $\rho_{\mathcal{F},j} \lesssim j^{-2\ell_{\mathcal{F}}}$ *for some* $\ell_{\mathcal{F}} > 1/2$*. Under Model 2, running Algorithm 1 with oracle* $\mathscr{O}$ *for Model 2 and main regularizer* $\lambda \asymp (n\kappa)^{-\frac{2\ell_{\mathcal{F}}}{1+2\ell_{\mathcal{F}}(1+\nu)}}$ *yields*

$$\mathcal{E}_{L^2}(\hat{h}) \;\lesssim\; (n\kappa)^{-\frac{2\ell_{\mathcal{F}}(1+\nu)}{1+2\ell_{\mathcal{F}}(1+\nu)}},$$

with probability at least $1 - n^{-10}$*. The notation* $\lesssim$ *omits polylogarithmic factors.*

Discussion. We defer the proof to Appendix F.2. This result applies to general RKHSs, including settings such as NTKs where adaptivity is far from automatic. In the NTK setting of Li et al. [2024], on a bounded domain with bounded covariate density, the spectral exponent $\ell_{\mathcal{F}} = (d+1)/(2d)$ yields $\tilde{\mathcal{O}}((n\kappa)^{-\frac{(d+1)(1+\nu)}{d+(d+1)(1+\nu)}})$. This can be substantially faster than the nuisance NTK learning bound $n^{-\frac{d+1}{2d+1}}$. Moreover, combined with lower bounds under source conditions [Zhang et al., 2024], Lemma 6 shows that this rate is optimal. Fixed NTKs preserve the i.i.d. sampling setup; extending the conclusions to wide networks requires a suitable lazy-training approximation [Li et al., 2024].

4.3 Results under Model 3 (Low-Dimensional Structure)

Finally, we consider the case where $h^\star$ depends on a known low-dimensional representation of the covariates. Let $\tilde{\mathbf{T}}$ be the integral operator associated with the kernel on the lower-dimensional space $\tilde{\mathcal{H}}$, defined with respect to the distribution of $\tilde{x}_i$, with eigenvalues $\tilde{\rho}_1 \ge \tilde{\rho}_2 \ge \ldots$. We also assume that the kernel $K_{\tilde{\mathcal{H}}}$ is uniformly bounded.

Corollary 9 (L^2-Error Bounds for Model 3)**.** *Suppose the assumptions in Section 2.3 and Model 3 hold, and* $\tilde{\rho}_j \lesssim j^{-2\tilde{\ell}}$ *for some* $\tilde{\ell} > 1/2$*. Running Algorithm 1 with oracle* $\mathscr{O}$ *for Model 3 and main regularizer* $\lambda \asymp (n\kappa)^{-\frac{2\tilde{\ell}}{1+2\tilde{\ell}}}$ *yields*

$$\mathcal{E}_{L^2}(\hat{h}) \;\lesssim\; (n\kappa)^{-\frac{2\tilde{\ell}}{1+2\tilde{\ell}}}$$

with probability at least $1 - n^{-10}$*. The notation* $\lesssim$ *omits polylogarithmic factors.*

Discussion. We defer the proof to Appendix F.3. For instance, if $\tilde{\mathcal{H}} = H^\beta(\tilde{\mathcal{X}})$ for $\dim(\tilde{\mathcal{X}}) = \tilde{d}$, then $\tilde{\ell} = \beta/\tilde{d}$, and the rate simplifies to $(n\kappa)^{-\frac{2\beta}{\tilde{d}+2\beta}}$. This dependence on the intrinsic dimension $\tilde{d}$, rather than the ambient dimension d, shows that the estimator behaves as if the target were generated in the lower-dimensional space. The complexity of the nuisance functions $f_0^\star$ and $f_1^\star$, which may depend on the full dimension d, does not degrade the convergence rate of $\hat{h}$.

4.4 Comparison with Prior Work

Closest in spirit to our work is the recent study of Kim et al. [2025], which also considers CATE estimation via a two-stage KRR strategy. The adaptivity considered there, however, is different from ours. Kim et al. [2025] study a "transfer learning" regime in which the CATE $h^\star$ lies in the *same* function space as the nuisance functions ($\mathcal{H} = \mathcal{F}$), but has a substantially smaller RKHS norm ($\|h^\star\|_{\mathcal{F}} \ll \|f_0^\star\|_{\mathcal{F}}, \|f_1^\star\|_{\mathcal{F}}$). For fixed overlap and no covariate shift, the leading term in their bound is, up to logarithmic factors,

$$n^{-\alpha_{\mathcal{F}}} \|h^\star\|_{\mathcal{F}}^{2(1-\alpha_{\mathcal{F}})}, \qquad \alpha_{\mathcal{F}} = \frac{2\ell_{\mathcal{F}}}{1+2\ell_{\mathcal{F}}},$$

where $\ell_{\mathcal{F}}$ is the eigendecay exponent of the common kernel. A smaller contrast norm improves this bound, but the exponent of n remains fixed by $\mathcal{F}$. In contrast, Models 1–3 allow the exponent itself to adapt to the contrast structure. For example, under the source condition in Model 2, Corollary 8 improves the exponent from $\alpha_{\mathcal{F}}$ to $2\ell_{\mathcal{F}}(1+\nu)/(1+2\ell_{\mathcal{F}}(1+\nu))$. Under Models 1 and 3, we do not impose an eigendecay assumption on the nuisance kernel. We also establish minimax pointwise rates in both n and κ in Section 5, a guarantee not established by Kim et al. [2025].

A more detailed comparison with DR and DML-style approaches is given in Appendix A. At a high level, our method is particularly advantageous under weak overlap: it avoids propensity-score estimation entirely, and its statistical burden on the nuisance stage is limited to well-specification, while the eigendecay assumptions governing the rate are imposed only on the contrast class $\mathcal{H}$.

5 Upper Bounds for Point Evaluation

In this section, we establish minimax-optimal pointwise error bounds for our estimator. As in the L^2-analysis, the results here are obtained by applying Theorem D.1 from Appendix D under the appropriate evaluation operator.

For pointwise analysis, we focus on Sobolev-type classes, where evaluation at a point can be controlled via Sobolev embedding. The same arguments extend naturally to mixed Sobolev spaces. Throughout this section, we additionally assume that the density of x_i is bounded above and below by positive constants.

Point Evaluation under Models 1 and 2. Assume that $h^\star$ belongs to a Sobolev space $\mathcal{H} = H^\gamma(\mathcal{X})$ with smoothness $\gamma > d/2$, while the nuisance functions may lie in a rougher space $\mathcal{F}$. Under this setup, Model 2 is subsumed by Model 1, since a source condition of order ν relative to a Sobolev kernel of smoothness β implies membership in $H^{\beta(1+\nu)}$. Accordingly, we analyze Models 1 and 2 jointly by working under the Sobolev assumption $h^\star \in H^\gamma(\mathcal{X})$ and applying the Model 1 oracle.

Point Evaluation under Model 3. For Model 3, we again work with Sobolev classes and, for simplicity, assume $\mathcal{F} = H^\beta(\mathcal{X})$ and $\tilde{\mathcal{H}} = H^\beta(\tilde{\mathcal{X}})$ for some $\beta > d/2$. Recall that $\dim(\tilde{\mathcal{X}}) = \tilde{d}$.

We now present the upper bounds for point evaluation under each model.

Corollary 10 (Point Evaluation: Model 1 & 2)**.** *Suppose $\mathcal{H} = H^\gamma(\mathcal{X})$ with $\gamma > d/2$. Under the assumptions of Section 2.3 and Model 1, the estimator $\hat{h}$ produced by Algorithm 1 with main regularizer $\lambda \asymp \frac{1}{n\kappa}$ satisfies the following pointwise error bound for any fixed $x_0 \in \mathcal{X}$ with probability at least $1 - n^{-10}$:*

$$\mathcal{E}_{x_0}(\hat{h}) = |\hat{h}(x_0) - h^\star(x_0)|^2 \lesssim (n\kappa)^{-\frac{2\gamma-d}{2\gamma}}. \tag{3}$$

Here $\lesssim$ omits absolute constants and polylogarithmic factors.

Discussion. The rate in Equation (3) matches the minimax pointwise rate for estimating a function in $H^\gamma(\mathcal{X})$ from $n\kappa$ effective observations [Tuo and Zou, 2024]. Importantly, it is independent of

the smoothness of the nuisance functions $f_0^\star$ and $f_1^\star$, provided they remain in the ambient RKHS $\mathcal{F}$. The same conclusion holds for mixed Sobolev spaces $\mathcal{H} = H^\gamma_{\text{mix}}(\mathcal{X}), \gamma > \frac{1}{2}$, which yield the rate $(n\kappa)^{-\frac{2\gamma-1}{2\gamma}}$. By Lemma 6, this rate is likewise optimal. We defer the proof to Appendix F.4.

Next, we address the low-dimensional setting.

Corollary 11 (Point Evaluation: Model 3)**.** *Under the assumptions of Section 2.3 and Model 3, suppose $\tilde{\mathcal{H}} = H^\beta(\tilde{\mathcal{X}})$ and $\mathcal{F} = H^\beta(\mathcal{X})$ with $\beta > d/2$. Then the estimator $\hat{h}$ produced by Algorithm 1 with main regularizer $\lambda \asymp \frac{1}{n\kappa}$ satisfies:*

$$\mathcal{E}_{x_0}(\hat{h}) = |\hat{h}(x_0) - h^\star(x_0)|^2 \lesssim (n\kappa)^{-\frac{2\beta-\tilde{d}}{2\beta}} \tag{4}$$

with probability at least $1 - n^{-10}$. Here $\lesssim$ omits absolute constants and polylogarithmic factors.

Discussion. This bound depends on the intrinsic dimension $\tilde{d}$ rather than the ambient dimension d, thereby showing that the method adapts to low-dimensional structure in the target even when nuisance learning takes place in the full space. We defer the proof to Appendix F.5.

Comparison with DML. Relative to pointwise DR analyses such as Kennedy [2023], our pointwise guarantees avoid a substantial burden. Those results typically require sufficiently fast control of pointwise nuisance errors such as $|\hat{f}(x_0) - f^\star(x_0)|$ and $|\hat{\pi}(x_0) - \pi(x_0)|$, whereas standard learning theory more naturally provides L^2-type guarantees. Our approach bypasses propensity estimation altogether and requires only that the outcome nuisances be well specified in $\mathcal{F}$, while the leading pointwise rate is governed solely by the complexity of the target class $\mathcal{H}$.

6 Experiments

We compare Ours with plug-in KRR, the DR-learner of Kennedy [2023], and the R-learner of Nie and Wager [2021]. The R-learner fits the treatment effect through a residualized regression loss. Synthetic experiments additionally include oracle DR, which uses the true propensity score. We also use RealCause [Neal et al., 2020], which fits generative models to real data to produce treatment assignments and potential outcomes given observed covariates. Our semi-real experiments use its LaLonde benchmark, based on the National Supported Work job-training study and a nonexperimental comparison group from the Current Population Survey (CPS) [LaLonde, 1986]. The outcome is post-training earnings; RealCause uses the data version of Dehejia and Wahba [1999].

Ours remains competitive with DR, oracle DR, and R in the univariate examples, where low-dimensional nuisance estimation is relatively favorable to DR-based methods. Its advantage is clearer in the multivariate experiments: it attains the lowest mean error in all sample-size and noise comparisons, as well as all eight settings with $\kappa \in \{0.01, 0.05\}$. Noise, overlap, nuisance-kernel misspecification, and representation ablations examine the sensitivity of these findings. On 50 RealCause realizations, Ours also has the lowest average held-out mean squared error (MSE). Subtracting the common estimated outcome-noise contribution makes the excess-risk differences easier to compare and leaves the ranking unchanged. Appendix C gives the experimental setup and results in Tables 1–3. Code is available at `https://github.com/seokjinkim0428/Structure-Adaptive-CATE`.

Acknowledgments

We thank Kaizheng Wang for helpful discussions. We also thank the reviewers for their constructive feedback, which helped improve the paper.

References

Nachman Aronszajn. Theory of reproducing kernels. *Transactions of the American mathematical society*, 68(3):337–404, 1950.

Peter Bühlmann and Sara Van De Geer. *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media, 2011.

Yihang Chen, Fanghui Liu, Taiji Suzuki, and Volkan Cevher. High-dimensional kernel methods under covariate shift: Data-dependent implicit regularization. *arXiv preprint arXiv:2406.03171*, 2024.

Rajeev H. Dehejia and Sadek Wahba. Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American Statistical Association*, 94(448): 1053–1062, 1999.

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

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

Zijun Gao and Yanjun Han. Minimax optimal nonparametric estimation of heterogeneous treatment effects. *Advances in Neural Information Processing Systems*, 33:21751–21762, 2020.

Behrooz Ghorbani, Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Linearized two-layers neural networks in high dimension. *The Annals of Statistics*, 49(2), 2021.

Jens Hainmueller. Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Political analysis*, 20(1):25–46, 2012.

Kosuke Imai and Marc Ratkovic. Covariate balancing propensity score. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 76(1):243–263, 2014.

Kwang-Sung Jun, Ashok Cutkosky, and Francesco Orabona. Kernel truncated randomized ridge regression: Optimal rates and low noise acceleration. *Advances in neural information processing systems*, 32, 2019.

Masahiro Kato and Masaaki Imaizumi. Cate lasso: conditional average treatment effect estimation with high-dimensional linear regression. *arXiv preprint arXiv:2310.16819*, 2023.

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

Edward H Kennedy, Sivaraman Balakrishnan, James M Robins, and Larry Wasserman. Minimax rates for heterogeneous causal effect estimation. *Annals of statistics*, 52(2):793, 2024.

Seok-Jin Kim and Kaizheng Wang. Estimating continuous treatment effects with two-stage kernel ridge regression. *arXiv preprint arXiv:2604.13410*, 2026.

Seok-Jin Kim, Hongjie Liu, Molei Liu, and Kaizheng Wang. Transfer learning of cate with kernel ridge regression. *arXiv preprint arXiv:2502.11331*, 2025.

Thomas Kühn, Winfried Sickel, and Tino Ullrich. Approximation of mixed order sobolev functions on the d-torus: asymptotics, preasymptotics, and d-dependence. *Constructive Approximation*, 42 (3):353–398, 2015.

Sören R Künzel, Jasjeet S Sekhon, Peter J Bickel, and Bin Yu. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the national academy of sciences*, 116(10):4156–4165, 2019.

Robert J. LaLonde. Evaluating the econometric evaluations of training programs with experimental data. *The American Economic Review*, 76(4):604–620, 1986.

Yicheng Li, Zixiong Yu, Guhan Chen, and Qian Lin. On the eigenvalue decay rates of a class of neural-network related kernel functions defined on general domains. *Journal of Machine Learning Research*, 25(82):1–47, 2024.

Cong Ma, Reese Pathak, and Martin J Wainwright. Optimally tackling covariate shift in rkhs-based nonparametric regression. *The Annals of Statistics*, 51(2):738–761, 2023.

Xinwei Ma and Jingshen Wang. Robust inference using inverse probability weighting. *Journal of the American Statistical Association*, 115(532):1851–1860, 2020.

Xinwei Ma, Yuya Sasaki, and Yulong Wang. Testing limited overlap. *Econometric Theory*, 41(5): 1129–1162, 2025.

Wenlong Mou, Peng Ding, Martin J Wainwright, and Peter L Bartlett. Kernel-based off-policy estimation without overlap: Instance optimality beyond semiparametric efficiency. *arXiv preprint arXiv:2301.06240*, 2023.

Brady Neal, Chin-Wei Huang, and Sunand Raghupathi. RealCause: Realistic causal inference benchmarking. *arXiv preprint arXiv:2011.15007*, 2020.

Xinkun Nie and Stefan Wager. Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika*, 108(2):299–319, 2021.

Miruna Oprescu, Vasilis Syrgkanis, and Zhiwei Steven Wu. Orthogonal random forest for causal inference. In *International Conference on Machine Learning*, pages 4932–4941. PMLR, 2019.

Rahul Singh. Kernel ridge riesz representers: Generalization, mis-specification, and the counterfactual effective dimension. *arXiv preprint arXiv:2102.11076*, 2021. URL `https://arxiv.org/abs/2102.11076`. Revised July 2024.

Rahul Singh, Liyuan Xu, and Arthur Gretton. Kernel methods for causal functions: dose, heterogeneous and incremental response curves. *Biometrika*, 111(2):497–516, 2024.

Taiji Suzuki. Adaptivity of deep relu network for learning in besov and mixed smooth besov spaces: optimal rate and curse of dimensionality. *arXiv preprint arXiv:1810.08033*, 2018.

Rui Tuo and Lu Zou. Asymptotic theory for linear functionals of kernel ridge regression. *arXiv preprint arXiv:2403.04248*, 2024.

Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge university press, 2019.

Hengfang Wang and Jae Kwang Kim. Statistical inference using regularized m-estimation in the reproducing kernel hilbert space for handling missing data. *Annals of the Institute of Statistical Mathematics*, 75(6):911–929, 2023.

Jiayi Wang, Raymond K. W. Wong, Shu Yang, and Kwun Chuen Gary Chan. Estimation of partially conditional average treatment effect by double kernel-covariate balancing. *Electronic Journal of Statistics*, 16(2):4332–4378, 2022. doi: 10.1214/22-EJS2000.

Kaizheng Wang. Pseudo-labeling for kernel ridge regression under covariate shift. *The Annals of Statistics*, 54(1):252–276, 2026.

Haobo Zhang, Yicheng Li, Weihao Lu, and Qian Lin. On the optimality of misspecified kernel ridge regression. In *International Conference on Machine Learning*, pages 41331–41353. PMLR, 2023.

Haobo Zhang, Yicheng Li, and Qian Lin. On the optimality of misspecified spectral algorithms. *Journal of Machine Learning Research*, 25(188):1–50, 2024.

Haobo Zhang, Yicheng Li, Weihao Lu, and Qian Lin. Optimal rates of kernel ridge regression under source condition in large dimensions. *Journal of Machine Learning Research*, 26(219):1–63, 2025.

Yunji Zhou, Roland A Matsouaka, and Laine Thomas. Propensity score weighting under limited overlap and model misspecification. *Statistical methods in medical research*, 29(12):3721–3756, 2020.

Supplementary Materials

Contents

A Detailed Comparison with DML Methods **14**

B Model Selection Procedure and Adaptivity Guarantees **14**

B.1 Model Selection Procedure 14

B.2 Adaptivity Guarantees 15

B.3 Practical choices 15

C Numerical Experiments **16**

C.1 Univariate Synthetic Data 16

C.2 Multivariate Synthetic Data 17

C.3 Semi-real Data (RealCause) 19

D General Theory for the Analysis **20**

D.1 RKHS Formulation and Notation 20

D.2 Setup 20

D.3 Upper Bound Analysis 22

E Proof of the Theorem D.1 (General Theory) **22**

E.1 Notation 22

E.2 Good Event 23

E.3 Proof 24

F Proofs of the Main Corollaries from General Theory **25**

F.1 Proof of Corollary 7 (Model 1, L^2-error) 25

F.2 Proof of Corollary 8 (Model 2, L^2-error) 26

F.3 Proof of Corollary 9 (Model 3, L^2-error) 26

F.4 Proof of Corollary 10 (Models 1–2, point evaluation) 26

F.5 Proof of Corollary 11 (Model 3, point evaluation) 27

G Proof of Theorem B.2 (Oracle Inequality) **27**

G.1 Empirical target error and proxy-based selection 27

G.2 Population conversion and the role of B 27

G.3 Proof of Lemma G.1 (Oracle Inequality for Empirical L^2-Error) 28

H Technical Lemmas **30**

I Nuisance Misspecification **32**

I.1 Rates under Models 1–3 32

I.2 General bound and proof 32

J Limitations **33**

Appendix Roadmap. For convenience, we summarize the supporting material as follows. Section B presents the model-selection procedure and its oracle guarantee. Section C reports the experimental details and additional numerical results. Section D develops the general Hilbert-space error bound, followed by its proofs and technical lemmas. Section I extends the analysis to approximate outcome models and proves the additional error bound. Section J discusses limitations.

A Detailed Comparison with DML Methods

Our weak-overlap formulation is broad enough to allow singular propensity scores, thereby covering regimes in which standard DR methods are not directly applicable. Even in the classical regime $\pi(\cdot) \in [\kappa, 1-\kappa]$, however, our approach retains substantial advantages over DR methods. Avoiding propensity-score estimation is itself a substantial advantage, because propensity models are well known to be especially sensitive to weak overlap and even mild model misspecification [Hainmueller, 2012, Imai and Ratkovic, 2014, Zhou et al., 2020]. At a high level, existing DR analyses typically yield squared error bounds of the form

$$\kappa^{-A}\big((\text{Oracle rate for } \mathcal{H} \text{ with } n\kappa \text{ samples}) + \kappa^{-B}\|\hat{f}_a - f_a^\star\|_{\mathbb{X}}^2\|\hat{\pi} - \pi\|_{\mathbb{X}}^2\big).$$

For instance, Nie and Wager [2021, Eq. (12)] corresponds to $(\mathbb{X}, A, B) = (L^2, 2, 0)$, while Foster and Syrgkanis [2023, Proposition 2] corresponds to $(\mathbb{X}, A, B) = (L^4, 0, 3)$. In both cases, the resulting guarantees are highly vulnerable under weak overlap, that is, when κ is small.

- Nie and Wager [2021]: the leading term is not optimal in κ because of the prefactor κ^{-A}, even when the nuisance estimators converge sufficiently fast. By contrast, our rates are optimal in both n and κ. Their analysis also requires $\|\hat{\pi} - \pi\|_{L^2}^2 = o(1/\sqrt{n})$.
- Foster and Syrgkanis [2023]: the required L^4 nuisance error bounds are stronger than standard L^2 control, and the additional factor κ^{-3} makes the bound unstable under weak overlap. In contrast, we require only that $f^\star$ be well specified and do not estimate π. Moreover, these methods guarantee only CATE L^2 error, whereas our algorithm also yields pointwise bounds.
- Kennedy [2023]: this line of work focuses on pointwise bounds. In that setting, one may view $\mathbb{X} = L^2(\delta_{x_0})$ and $(A, B) = (0, 2)$. It requires pointwise error control for both $f^\star$ and π, whereas standard learning theory more naturally provides L^2-type guarantees. Even for point evaluation, we require only that $f^\star$ be well specified. We impose no separate nuisance-estimator error bound, and the target spectrum determines the leading rate.

B Model Selection Procedure and Adaptivity Guarantees

B.1 Model Selection Procedure

In this section, we present a model-selection procedure and its corresponding guarantee for L^2 error. In practice, the best contrast space $\mathcal{H}$ and the corresponding regularization level are rarely known a priori. While the nuisance class $\mathcal{F}$ can be tuned against observed outcomes, selecting a CATE model is harder because the target itself is not directly observed. We therefore introduce a dedicated model-selection procedure that chooses among a library of candidate estimators.

The procedure uses three disjoint data splits so that candidate generation, proxy construction, and final selection remain independent.

1. Candidate Generation ($\mathcal{D}_1$). We partition the dataset $\mathcal{D}$ into three disjoint subsets $\mathcal{D}_1, \mathcal{D}_2, \mathcal{D}_3$. For simplicity, we analyze the equal-split case $n_1 = n_2 = n_3 = n/3$; the same conclusions hold for any constant-proportion split. Using $\mathcal{D}_1$, we run Algorithm 1 over a library of configurations, for example different hypothesis classes $\{\mathcal{H}_1, \ldots, \mathcal{H}_K\}$ and regularization levels, to generate candidate estimators $\mathcal{M} = \{\hat{h}_1, \ldots, \hat{h}_L\}$. We denote the configuration library by Π, where each element is a pair $(\mathcal{H}, \lambda)$ consisting of a hypothesis class and a regularization level.

2. Truncation. To control the variance during the selection phase, we enforce a boundedness constraint. Let $B > 0$ be the truncation level. For each $\hat{h}_j \in \mathcal{M}$, we define its truncated version $\bar{h}_j(x) := \min(\max(\hat{h}_j(x), -B), B)$. Let $\bar{\mathcal{M}} = \{\bar{h}_1, \ldots, \bar{h}_L\}$ denote the set of truncated candidates.

3. Proxy Construction ($\mathcal{D}_2$). Using $\mathcal{D}_2$, we construct a high-quality proxy for the unobserved CATE by re-estimating the nuisance functions with an independent sample and undersmoothed KRR:

$$\tilde{f}_a := \arg\min_{f \in \mathcal{F}} \Big\{ \frac{1}{n_2} \sum_{i=1}^{n_2} (y_{2i} - f(x_{2i}))^2 \mathbf{1}(a_{2i} = a) + \tilde{\lambda} \|f\|_{\mathcal{F}}^2 \Big\}, \quad a \in \{0, 1\}. \tag{5}$$

These nuisance estimates depend only on $\mathcal{D}_2$ and are therefore independent of the candidate estimators generated from $\mathcal{D}_1$.

4. Empirical Risk Minimization ($\mathcal{D}_3$). Finally, using $\mathcal{D}_3$, we evaluate the candidates against the proxy induced by $\tilde{f}_0$ and $\tilde{f}_1$. For each $(x_{3i}, a_{3i}, y_{3i}) \in \mathcal{D}_3$, define

$$\tilde{m}_{3i} := \begin{cases} y_{3i} - \tilde{f}_0(x_{3i}), & \text{if } a_{3i} = 1, \\ \tilde{f}_1(x_{3i}) - y_{3i}, & \text{if } a_{3i} = 0. \end{cases} \tag{6}$$

We select the estimator that minimizes the empirical squared error with respect to these proxies:

$$\bar{h}_{\mathrm{ms}} := \arg\min_{\bar{h} \in \bar{\mathcal{M}}} \frac{1}{n_3} \sum_{i=1}^{n_3} (\bar{h}(x_{3i}) - \tilde{m}_{3i})^2.$$

Algorithm 2 CATE Model Selection

1: **Input:** Data $\mathcal{D}$, set of candidate configurations Π, truncation bound B.
2: Partition $\mathcal{D}$ into $\mathcal{D}_1, \mathcal{D}_2, \mathcal{D}_3$.
3: **Step 1:** Train candidates $\mathcal{M} = \{\hat{h}_1, \ldots, \hat{h}_L\}$ on $\mathcal{D}_1$ by running Algorithm 1 over multiple configurations $(\mathcal{H}, \lambda) \in \Pi$.
4: **Step 2:** Truncate candidates to $[-B, B]$ to obtain $\bar{\mathcal{M}} = \{\bar{h}_1, \ldots \bar{h}_L\}$.
5: **Step 3:** Estimate nuisance functions $\tilde{f}_0, \tilde{f}_1$ on $\mathcal{D}_2$ via Equation (5).
6: **Step 4:** For $(x_{3i}, a_{3i}, y_{3i}) \in \mathcal{D}_3$, compute proxies $\tilde{m}_{3i}$ using $\tilde{f}_0, \tilde{f}_1$ by Equation (6).
7: **Return** $\bar{h}_{\mathrm{ms}} = \arg\min_{\bar{h} \in \bar{\mathcal{M}}} \frac{1}{n_3} \sum_{i=1}^{n_3} (\bar{h}(x_{3i}) - \tilde{m}_{3i})^2$.

B.2 Adaptivity Guarantees

We now give the oracle guarantee for Algorithm 2.

Assumption B.1 (Bounded contrast)**.** There is a finite $B^\star$ such that $\|h^\star\|_\infty \le B^\star$.

Choose $B \ge B^\star$ in advance or using $\mathcal{D}_1$ or independent data; exact knowledge of $B^\star$ is unnecessary.

Theorem B.2 (Oracle inequality)**.** *Suppose Assumptions 1, 2, B.1, 3 and 4 hold, with bounded nuisance RKHS norms. Run Algorithm 2 with constant-proportion splits, $\tilde{\lambda} \asymp \log n / n_2$, and $B \ge B^\star$ as above. If $L = |\mathcal{M}|$ is polynomial in n, then with probability at least $1 - n^{-10}$,*

$$\mathcal{E}_{L^2}(\bar{h}_{\mathrm{ms}}) \lesssim \min_{\bar{h} \in \bar{\mathcal{M}}} \mathcal{E}_{L^2}(\bar{h}) + \widetilde{\mathcal{O}}\left(\frac{1}{n\kappa} + \frac{(B + B^\star)^2}{n} \right). \tag{7}$$

Bounded upper estimates of $B^\star$ preserve the nonparametric oracle rate whenever the parametric-order remainder is lower order. Appendix G gives the derivation, including logarithmic factors.

B.3 Practical choices

For coordinate subsets, $\log |\mathcal{M}| = \log \binom{d}{\tilde{d}} \le \tilde{d} \log(ed/\tilde{d})$ before adding kernel and regularization grids. For principal component regression, a grid of at most d leading subspaces constructed on $\mathcal{D}_1$ contributes only $\log d$. The guarantee compares supplied candidates; learning the unknown projection remains open.

For pointwise tuning, choose $\lambda \asymp (n\kappa)^{-1}$; a constant-factor approximation of κ suffices. We recommend averaging the cyclic fits.

C Numerical Experiments

We compare Ours with plug-in KRR, DR [Kennedy, 2023], oracle-propensity DR, and R [Nie and Wager, 2021], using the same datasets and test covariates for all methods. HOIF methods serve as theoretical benchmarks.

The raw ridge grid is $\mathcal{A} = \{1, 2, 4, \ldots, 512\}$ in $(K + \alpha I)^{-1} y$. In our experiments, Ours fixes the raw nuisance ridge at $c = 0.01$, corresponding to normalized regularization c/n_{stage}, and selects the target kernel and ridge by Algorithm 2. We use untruncated predictions and average the three cyclic fits. Plug-in uses three-fold cross-validation (CV) and refits each arm. DR, oracle DR, and R use three-fold CV within two swapped fits and average their predictions. R minimizes the residualized loss $(Y - \hat{m}(X) - (A - \hat{\pi}(X))h(X))^2$, with $m(x) = \mathbb{E}[Y \mid X = x]$. Estimated propensities are clipped to $[0.01, 0.99]$; oracle DR uses the true propensity.

We generate $A \mid X \sim \text{Bernoulli}(\pi_\kappa(X))$ and $Y = f_0^\star(X) + Ah^\star(X) + \varepsilon$, where $\pi_\kappa(x) = \text{Clip}(\sin(5\|x\|_2), \kappa, 1 - \kappa)$ and $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ is independent. Defaults are $\sigma = 1$ and $\kappa = 0.1$. Tables report mean integrated squared error (MISE) with Monte Carlo standard errors (SEs) on 3,000 independent uniform test points (seed 10099), excluded from tuning. Sample-size comparisons use 100 replications; sweeps and ablations use 30, with seed $100 + 10r$ for replication r.

C.1 Univariate Synthetic Data

We draw $X \sim \text{Unif}[0, 1]$ and set

$$f_0^\star(x) = 5\big(|x - 0.4| + |x - 0.8|\big), \qquad h^\star(x) = x^2.$$

The default nuisance/target kernels are unanchored Sobolev kernels of orders (H^1, H^2). We vary $n \in \{500, 1000\}$ and, at $n = 500$, noise $\sigma \in \{0.25, 0.5, 1, 2, 4\}$ and overlap $\kappa \in \{0.01, 0.02, 0.05, 0.1, 0.2, 0.3, 0.5\}$. The (H^2, H^3) ablation at $n = 1000$ misspecifies the kinked nuisance while retaining a smooth target; it changes both kernels.

This univariate setting is relatively favorable to DR-based methods, since low-dimensional outcome and propensity estimation can make their product-of-errors remainder small. Ours remains comparable to DR, oracle DR, and R (Table 1): at $n = 1000$, their MISEs are 0.0424, 0.0577, 0.0365, and 0.0335, respectively.

Table 1: Univariate synthetic results. Entries are mean MISE (SE), with the smallest mean in bold. Unless varied, $\sigma = 1$ and $\kappa = 0.1$; the $\sigma = 1$ row also supplies the $\kappa = 0.1$ reference. H^a/H^b denotes the jointly varied nuisance/target Sobolev orders; H^2 misspecifies the kinked nuisance.

Configuration	n	Reps.	Ours	Plug-in	DR	Oracle DR	R
Sample size	500	100	0.0811 (0.0056)	0.1581 (0.0087)	0.3203 (0.1599)	0.0758 (0.0057)	**0.0699** (0.0051)
Sample size	1000	100	0.0424 (0.0033)	0.0825 (0.0035)	0.0577 (0.0204)	0.0365 (0.0033)	**0.0335** (0.0031)
$\sigma = 0.25$	500	30	0.0317 (0.0061)	0.1230 (0.0104)	0.0938 (0.0631)	0.0527 (0.0076)	**0.0210** (0.0025)
$\sigma = 0.5$	500	30	0.0473 (0.0078)	0.1388 (0.0127)	0.0828 (0.0398)	0.0688 (0.0100)	**0.0368** (0.0055)
$\sigma = 1$	500	30	0.0890 (0.0098)	0.1908 (0.0201)	**0.0704** (0.0119)	0.0983 (0.0132)	0.0814 (0.0106)
$\sigma = 2$	500	30	**0.1406** (0.0141)	0.3814 (0.0451)	0.1585 (0.0237)	0.1839 (0.0227)	0.1697 (0.0162)
$\sigma = 4$	500	30	**0.2701** (0.0454)	1.0014 (0.0945)	0.5051 (0.1317)	0.4994 (0.1034)	0.3686 (0.0707)
$\kappa = 0.01$	500	30	0.2939 (0.0382)	0.6344 (0.0584)	0.3802 (0.0961)	0.3739 (0.1063)	**0.1601** (0.0143)
$\kappa = 0.02$	500	30	0.2263 (0.0319)	0.4839 (0.0480)	0.2363 (0.0416)	0.2434 (0.0423)	**0.1381** (0.0122)
$\kappa = 0.05$	500	30	0.1324 (0.0183)	0.2983 (0.0367)	0.2197 (0.0574)	0.1437 (0.0176)	**0.1058** (0.0126)
$\kappa = 0.2$	500	30	0.0634 (0.0103)	0.0970 (0.0105)	0.0678 (0.0109)	0.0676 (0.0095)	**0.0539** (0.0068)
$\kappa = 0.3$	500	30	0.0495 (0.0091)	0.0671 (0.0065)	0.0469 (0.0072)	0.0548 (0.0085)	**0.0414** (0.0067)
$\kappa = 0.5$	500	30	0.0440 (0.0065)	0.0415 (0.0039)	**0.0347** (0.0047)	0.0370 (0.0051)	0.0379 (0.0056)
H^1/H^2	1000	30	0.0459 (0.0056)	0.0790 (0.0064)	0.0440 (0.0086)	**0.0297** (0.0051)	0.0364 (0.0073)
H^2/H^3	1000	30	**0.0493** (0.0067)	0.3063 (0.0249)	6.0731 (1.6719)	0.0499 (0.0076)	0.0850 (0.0049)

C.2 Multivariate Synthetic Data

We draw $X \sim \mathrm{Unif}([-1,1]^d)$ and set

$$f_0^\star(x) = \frac{2}{d}\sum_{j=1}^{d}\sin(x_j), \qquad h^\star(x) = \frac{0.3}{p}\sum_{j=1}^{p}x_j^2.$$

We consider dense ($p = d$) and sparse ($p = 4$) targets with $d = 10, 20$. The nuisance uses a Matérn kernel with $\nu = 1.5$. Ours selects among Matérn ($\nu = 1.5, 2.5$) and radial basis function kernels on all coordinates and on $x_{1:4}$. DR, oracle DR, and R use Matérn target kernels with $\nu = 2.5$ and receive the true support in sparse settings. Sample-size comparisons use fixed kernel lengths; sweeps use lengths calibrated to correlation 0.5 at the median distance on 2,000 independent covariates (seed 42). Sweeps use $n = 1000$, $\sigma \in \{0.5, 1, 2\}$, and $\kappa \in \{0.01, 0.05, 0.1, 0.2, 0.5\}$.

The advantage of Ours is clearer in Table 2: it has the lowest mean in 4 of the 4 sample-size comparisons and all 12 noise conditions. At $\sigma = 1$, its error is approximately 77–83% below plug-in. R or DR leads in some moderate-overlap settings. Oracle DR need not dominate estimated-propensity DR in finite samples, since exact propensities can increase weighting variance.

For $(d, p) = (20, 4)$, adding irrelevant $x_{5:8}$ candidates barely changes Ours (0.0092 to 0.0094). Using only these incorrect coordinates gives 0.0065, a finite-sample result rather than evidence of support recovery. Baselines keep their true-support kernels.

Table 2: Multivariate synthetic results. Entries are mean MISE (SE), with the smallest mean in bold. Unless varied, $\sigma = 1$ and $\kappa = 0.1$; the $\sigma = 1$ row also supplies the $\kappa = 0.1$ reference. $p = d$ is dense and $p = 4$ is sparse. Sample-size rows use fixed kernel lengths; other rows use median-calibrated lengths. The $\sigma = 1$ row at $(d, p) = (20, 4)$ also gives the default representation library.

(d, p)	Configuration	n	Reps.	Ours	Plug-in	DR	Oracle DR	R
$(10, 10)$	Sample size	1000	100	**0.0099** (0.0008)	0.0478 (0.0018)	0.0184 (0.0049)	0.0201 (0.0031)	0.0127 (0.0012)
$(10, 10)$	Sample size	2000	100	**0.0062** (0.0004)	0.0295 (0.0010)	0.0175 (0.0041)	0.0099 (0.0012)	0.0076 (0.0006)
$(10, 10)$	$\sigma = 0.5$	1000	30	**0.0051** (0.0005)	0.0204 (0.0011)	0.0054 (0.0008)	0.0087 (0.0022)	0.0058 (0.0007)
$(10, 10)$	$\sigma = 1$	1000	30	**0.0081** (0.0011)	0.0479 (0.0032)	0.0119 (0.0030)	0.0222 (0.0072)	0.0134 (0.0023)
$(10, 10)$	$\sigma = 2$	1000	30	**0.0170** (0.0033)	0.1047 (0.0124)	0.0377 (0.0122)	0.0570 (0.0207)	0.0382 (0.0090)
$(10, 10)$	$\kappa = 0.01$	1000	30	**0.0088** (0.0010)	0.0511 (0.0040)	0.0102 (0.0015)	0.1019 (0.0790)	0.0128 (0.0018)
$(10, 10)$	$\kappa = 0.05$	1000	30	**0.0078** (0.0009)	0.0517 (0.0045)	0.0113 (0.0020)	0.0213 (0.0068)	0.0124 (0.0020)
$(10, 10)$	$\kappa = 0.2$	1000	30	0.0107 (0.0016)	0.0462 (0.0039)	0.0104 (0.0018)	0.0102 (0.0013)	**0.0096** (0.0011)
$(10, 10)$	$\kappa = 0.5$	1000	30	**0.0069** (0.0011)	0.0450 (0.0036)	0.0127 (0.0034)	0.0121 (0.0032)	0.0089 (0.0014)
$(10, 4)$	Sample size	1000	100	**0.0112** (0.0008)	0.0491 (0.0018)	0.0171 (0.0024)	0.0186 (0.0021)	0.0141 (0.0011)
$(10, 4)$	Sample size	2000	100	**0.0074** (0.0003)	0.0307 (0.0010)	0.0121 (0.0014)	0.0126 (0.0010)	0.0088 (0.0005)
$(10, 4)$	$\sigma = 0.5$	1000	30	**0.0062** (0.0005)	0.0212 (0.0011)	0.0066 (0.0005)	0.0085 (0.0016)	0.0079 (0.0008)
$(10, 4)$	$\sigma = 1$	1000	30	**0.0092** (0.0010)	0.0489 (0.0032)	0.0117 (0.0018)	0.0174 (0.0052)	0.0171 (0.0029)
$(10, 4)$	$\sigma = 2$	1000	30	**0.0183** (0.0033)	0.1061 (0.0124)	0.0306 (0.0067)	0.0542 (0.0220)	0.0476 (0.0103)
$(10, 4)$	$\kappa = 0.01$	1000	30	**0.0098** (0.0009)	0.0523 (0.0040)	0.0180 (0.0041)	0.0496 (0.0156)	0.0167 (0.0026)
$(10, 4)$	$\kappa = 0.05$	1000	30	**0.0090** (0.0009)	0.0527 (0.0045)	0.0131 (0.0028)	0.0261 (0.0096)	0.0156 (0.0025)
$(10, 4)$	$\kappa = 0.2$	1000	30	0.0112 (0.0014)	0.0469 (0.0039)	0.0099 (0.0011)	0.0119 (0.0017)	**0.0098** (0.0009)
$(10, 4)$	$\kappa = 0.5$	1000	30	**0.0083** (0.0011)	0.0459 (0.0036)	0.0106 (0.0017)	0.0109 (0.0016)	0.0100 (0.0014)
$(20, 20)$	$\sigma = 0.5$	1000	30	**0.0032** (0.0003)	0.0204 (0.0011)	0.0039 (0.0005)	0.0051 (0.0011)	0.0035 (0.0004)
$(20, 20)$	$\sigma = 1$	1000	30	**0.0074** (0.0009)	0.0388 (0.0031)	0.0093 (0.0016)	0.0125 (0.0038)	0.0086 (0.0009)
$(20, 20)$	$\sigma = 2$	1000	30	**0.0212** (0.0035)	0.0602 (0.0099)	0.0289 (0.0062)	0.0367 (0.0115)	0.0237 (0.0036)
$(20, 20)$	$\kappa = 0.01$	1000	30	**0.0072** (0.0009)	0.0417 (0.0036)	0.0100 (0.0017)	0.0311 (0.0101)	0.0097 (0.0015)
$(20, 20)$	$\kappa = 0.05$	1000	30	**0.0076** (0.0012)	0.0401 (0.0038)	0.0095 (0.0016)	0.0173 (0.0034)	0.0099 (0.0013)
$(20, 20)$	$\kappa = 0.2$	1000	30	**0.0077** (0.0013)	0.0388 (0.0034)	0.0114 (0.0024)	0.0104 (0.0025)	0.0092 (0.0014)
$(20, 20)$	$\kappa = 0.5$	1000	30	0.0089 (0.0012)	0.0408 (0.0028)	**0.0075** (0.0010)	0.0076 (0.0010)	0.0104 (0.0019)
$(20, 4)$	$\sigma = 0.5$	1000	30	**0.0050** (0.0004)	0.0221 (0.0011)	0.0061 (0.0004)	0.0071 (0.0009)	0.0061 (0.0004)
$(20, 4)$	$\sigma = 1$	1000	30	**0.0092** (0.0009)	0.0405 (0.0032)	0.0114 (0.0013)	0.0147 (0.0030)	0.0125 (0.0013)
$(20, 4)$	$\sigma = 2$	1000	30	**0.0232** (0.0035)	0.0620 (0.0100)	0.0314 (0.0059)	0.0433 (0.0108)	0.0339 (0.0056)
$(20, 4)$	$\kappa = 0.01$	1000	30	**0.0090** (0.0009)	0.0426 (0.0036)	0.0128 (0.0022)	0.0389 (0.0128)	0.0135 (0.0013)
$(20, 4)$	$\kappa = 0.05$	1000	30	**0.0094** (0.0012)	0.0413 (0.0039)	0.0110 (0.0017)	0.0263 (0.0048)	0.0127 (0.0013)
$(20, 4)$	$\kappa = 0.2$	1000	30	**0.0093** (0.0013)	0.0408 (0.0034)	0.0150 (0.0022)	0.0134 (0.0035)	0.0128 (0.0020)
$(20, 4)$	$\kappa = 0.5$	1000	30	**0.0108** (0.0012)	0.0425 (0.0028)	0.0138 (0.0021)	0.0131 (0.0020)	0.0147 (0.0022)
$(20, 4)$	Add $x_{5:8}$	1000	30	**0.0094** (0.0009)	0.0405 (0.0032)	0.0114 (0.0013)	0.0147 (0.0030)	0.0125 (0.0013)
$(20, 4)$	$x_{5:8}$ only	1000	30	**0.0065** (0.0006)	0.0405 (0.0032)	0.0114 (0.0013)	0.0147 (0.0030)	0.0125 (0.0013)

C.3 Semi-real Data (RealCause)

We use 50 realizations of the LaLonde CPS benchmark [LaLonde, 1986, Dehejia and Wahba, 1999] generated by RealCause [Neal et al., 2020]. Each has 16,177 observations and eight covariates. All methods use only factual (X, A, Y) for fitting, with a common $50/50$ train/test split (seed 50; 8,088 training rows). Outcome scaling is fitted on training data; in realization r, effects are divided by the training outcome range S_r. The nuisance and target Matérn kernels use $(\nu, \ell) = (1.5, 1.3)$ and $(2.5, 1.2)$, respectively, with the ridge grid and fitting schemes above. True propensities are unavailable, so oracle DR is omitted.

All methods use Nyström KRR with at most 700 uniformly sampled training landmarks per fit (seed 700); smaller fits use all training points. For N fitting rows and m landmarks, storage is $O(Nm+m^2)$ and ridge-solve cost is $O(Nm^2+m^3)$. Runs use macOS/ARM64 with single-threaded linear algebra per file; software settings and elapsed times are supplied with the code. The theoretical results concern exact KRR.

Evaluation. We compute test MSE between predicted effects and sampled differences $Y(1)-Y(0)$, both on the training-normalized outcome scale. This score includes a common outcome-noise contribution in addition to CATE estimation error.

We estimate the noise floor using the same 50 realizations. At each shared test covariate, we compute the sample variance of the unscaled effect across realizations, average over test covariates, and multiply by the average S_r^{-2} to match the outcome normalization. Subtracting this common estimate from each method's mean test MSE gives its estimated excess risk and preserves method rankings. MSE SEs are computed across realizations. For excess-risk SEs, a joint delete-one-realization jackknife recomputes both the mean MSE and the correction after omitting each realization, accounting for their dependence.

The estimated noise floor is 0.08607 (SE 0.00190). Table 3 shows the lowest mean error for Ours; its estimated excess risk is 26.6%, 33.9%, and 85.7% below plug-in, DR, and R, respectively.

Table 3: RealCause results on 50 realizations. Entries are mean normalized error (SE). Estimated excess risk subtracts a common noise floor estimated from the same files; its SE includes correction uncertainty through the joint jackknife.

Method	Sampled-effect MSE (SE)	Estimated excess risk (SE)
Ours	**0.10010** (0.00251)	**0.01402** (0.00125)
Plug-in	0.10517 (0.00262)	0.01909 (0.00114)
DR	0.10730 (0.00267)	0.02123 (0.00107)
R	0.18427 (0.00420)	0.09820 (0.00259)

D General Theory for the Analysis

We now turn to the technical core of the paper. The second stage of our method has the same form in Models 1 to 3: it is KRR applied to pseudo-outcomes produced by the nuisance stage, with only the second-stage function space changing across models. This observation allows us to analyze all three cases through a single abstract theorem in a generic Hilbert space $\mathbb{X}$. We first introduce the RKHS representation and generic notation, and then state the high-probability error bound from which all subsequent corollaries follow.

D.1 RKHS Formulation and Notation

RKHS Formulation. By standard RKHS theory [Aronszajn, 1950], there exists a Hilbert space $\mathbb{F}$ and a feature map $\phi : \mathcal{X} \to \mathbb{F}$ satisfying the reproducing identity $\langle \phi(x), \phi(x') \rangle_{\mathbb{F}} = K_{\mathcal{F}}(x, x')$. Accordingly, the hypothesis space is defined as:

$$\mathcal{F} = \{ f_\theta(\cdot) = \langle \phi(\cdot), \theta \rangle_{\mathbb{F}} \mid \theta \in \mathbb{F} \}.$$

We denote the Hilbertian element corresponding to f_θ by θ. In particular, $\theta_0^\star$ and $\theta_1^\star$ denote the Hilbertian elements corresponding to $f_0^\star$ and $f_1^\star$, respectively.

Under Model 1, the contrast class $\mathcal{H}$ is itself an RKHS. Hence there exist a Hilbert space $\mathbb{H}$ and a feature map $\psi : \mathcal{X} \to \mathbb{H}$ such that $\langle \psi(x), \psi(x') \rangle_{\mathbb{H}} = K_{\mathcal{H}}(x, x')$. Accordingly, the hypothesis space is defined as:

$$\mathcal{H} = \{ h_\eta(\cdot) = \langle \psi(\cdot), \eta \rangle_{\mathbb{H}} \mid \eta \in \mathbb{H} \}.$$

Similarly, we denote the Hilbertian element corresponding to h_η by η. We write $\eta^\star$ for the Hilbertian element corresponding to $h^\star$.

Under Model 2, we use the same notation $\eta^\star$ for the Hilbertian element corresponding to $h^\star$ in $\mathbb{F}$.

Under Model 3, the space $\tilde{\mathcal{H}}$ is again an RKHS, so there exist a Hilbert space $\tilde{\mathbb{H}}$ and a feature map $\tilde{\psi} : \mathcal{X} \to \tilde{\mathbb{H}}$ satisfying $\langle \tilde{\psi}(x), \tilde{\psi}(x') \rangle_{\tilde{\mathbb{H}}} = K_{\tilde{\mathcal{H}}}(x, x')$. Accordingly, the hypothesis space is defined as:

$$\tilde{\mathcal{H}} = \{ \tilde{h}_\eta(\cdot) = \langle \tilde{\psi}(\cdot), \eta \rangle_{\tilde{\mathbb{H}}} \mid \eta \in \tilde{\mathbb{H}} \}.$$

We denote the Hilbertian element corresponding to $\tilde{h}_\eta$ by η. We write $\eta^\star$ for the Hilbertian element corresponding to $h^\star$ in $\tilde{\mathbb{H}}$.

Design Operators. For generic elements $\{v_1, \ldots, v_N\}$ in a Hilbert space $\mathbb{X}$, define the *design operator* $\mathbf{V} : \mathbb{X} \to \mathbb{R}^N$ by

$$\mathbf{V}\theta = \left(\langle v_1, \theta \rangle, \langle v_2, \theta \rangle, \ldots, \langle v_N, \theta \rangle \right)^\top.$$

Its adjoint $\mathbf{V}^\top : \mathbb{R}^N \to \mathbb{X}$ is given by

$$\mathbf{V}^\top \mathbf{a} = \sum_{i=1}^N a_i v_i \in \mathbb{X}.$$

D.2 Setup

We now describe a generic KRR problem that covers the second stage of all three models. Across Models 1–3, the second stage is always a KRR problem in which the scalar pseudo-outcome m_i plays the role of the response and the covariate enters only through a Hilbert-space feature $w_i \in \mathbb{X}$. The only object that changes across models is therefore the choice of second-stage Hilbert space $\mathbb{X}$ and the corresponding feature representation w_i. This leads to the common estimator $\hat{\eta}$ in Equation (8), which is the unified form analyzed below. Fix a Hilbert space $\mathbb{X}$ and i.i.d. features $\{w_i\}_{i=1}^n \subset \mathbb{X}$, with associated design operator $\mathbf{W}$. For example, in Model 1 we take $\mathbb{X} = \mathbb{H}$ and $w_i = \psi(x_i)$; in Model 2, $\mathbb{X} = \mathbb{F}$ and $w_i = \phi(x_i)$; and in Model 3, $\mathbb{X} = \tilde{\mathbb{H}}$ and $w_i = \tilde{\psi}(\tilde{x}_i)$. Across Models 1–3, the regression oracle $\mathscr{O}$ applies KRR to the pseudo-outcome vector $\mathbf{M} := (m_1, \ldots, m_n)$; only the choice of $\mathbb{X}$ and w_i changes.

Regression Oracle $\mathscr{O}$ as KRR on $\mathbb{X}$. We abstract the second-stage regression in Algorithm 1 as KRR in a generic Hilbert space $\mathbb{X}$, which will be instantiated as $\mathbb{H}$, $\mathbb{F}$, or $\tilde{\mathbb{H}}$ depending on the model. Write the population second moment operator as

$$\mathbf{\Sigma} := \mathbb{E}[w_i \otimes w_i].$$

This setup specializes to our models as follows:

- Model 1: $\mathbb{X} = \mathbb{H}$, $w_i = \psi(x_i)$ (feature map of $K_{\mathcal{H}}$).
- Model 2: $\mathbb{X} = \mathbb{F}$, $w_i = \phi(x_i)$ (feature map of $K_{\mathcal{F}}$).
- Model 3: $\mathbb{X} = \tilde{\mathbb{H}}$, $w_i = \tilde{\psi}(\tilde{x}_i)$ (feature map of $K_{\tilde{\mathcal{H}}}$).

In all cases, the target function is represented by an element $\eta^\star \in \mathbb{X}$ such that

$$\langle w_i, \eta^\star \rangle_{\mathbb{X}} = h^\star(x_i).$$

Recall the pseudo-outcomes m_i from Equation (2) and their vector form $\mathbf{M}$. Then the second-stage estimator returned by $\mathscr{O}$ can be written as

$$\hat{\eta} = (\mathbf{W}^\top \mathbf{W} + n\lambda \mathbf{I})^{-1} \mathbf{W}^\top \mathbf{M}. \tag{8}$$

Here $\hat{\eta}$ denotes the Hilbertian element corresponding to the function estimator $\hat{h}$ returned by Algorithm 1.

Evaluation. To treat different loss functions simultaneously, we evaluate errors through a positive trace-class operator $\mathbf{\Sigma}_{\mathrm{ref}}$ on $\mathbb{X}$. For any estimator $\hat{\eta}$, define

$$\mathcal{E}_{\mathrm{ref}}(\hat{\eta}) := \|\hat{\eta} - \eta^\star\|^2_{\mathbf{\Sigma}_{\mathrm{ref}}} \;=\; \langle \hat{\eta} - \eta^\star, \mathbf{\Sigma}_{\mathrm{ref}}(\hat{\eta} - \eta^\star) \rangle_{\mathbb{X}}.$$

This formalism recovers the usual loss functions of interest. In Model 1, for example, $\mathbf{\Sigma}$ is the second-moment operator for $K_{\mathcal{H}}$; choosing $\mathbf{\Sigma}_{\mathrm{ref}} = \mathbf{\Sigma}$ yields the $L^2(\mathcal{P}_x)$-error, while choosing $\mathbf{\Sigma}_{\mathrm{ref}} = \psi(x_0) \otimes \psi(x_0)$ yields the pointwise error at x_0. The same interpretation applies with $\phi(x_0)$ or $\tilde{\psi}(\tilde{x}_0)$ under Models 2 and 3. For convenience, we summarize the two main cases:

- L^2-error: $\mathbf{\Sigma}_{\mathrm{ref}} = \mathbf{\Sigma}$.
- Pointwise evaluation at x_0: $\mathbf{\Sigma}_{\mathrm{ref}} = \psi(x_0) \otimes \psi(x_0)$ (or $\phi(x_0) \otimes \phi(x_0)$ / $\tilde{\psi}(\tilde{x}_0) \otimes \tilde{\psi}(\tilde{x}_0)$ in Models 2–3), so $\mathcal{E}_{\mathrm{ref}}(\hat{\eta}) = |\hat{h}(x_0) - h^\star(x_0)|^2$.

Given $\mathbf{\Sigma}_{\mathrm{ref}}$, define

$$\mathbf{S}_\lambda := \mathbf{\Sigma}_{\mathrm{ref}}^{\frac{1}{2}} (\mathbf{\Sigma} + \lambda \mathbf{I})^{-1} \mathbf{\Sigma}_{\mathrm{ref}}^{\frac{1}{2}}.$$

This operator plays the role of an effective-dimension term under the evaluation metric induced by $\mathbf{\Sigma}_{\mathrm{ref}}$.

Additional notation used only in the appendix proofs is collected in Appendix E.1 (Table 5). Table 4 summarizes the notation used in this section.

Table 4: Notation used in this section.

Symbol	Definition
Spaces and feature maps	
$\mathbb{F}$	Hilbert space associated with kernel $K_{\mathcal{F}}$ and feature map $\phi : \mathcal{X} \to \mathbb{F}$.
$\mathbb{H}$	Hilbert space associated with kernel $K_{\mathcal{H}}$ (under Model 1).
$\tilde{\mathbb{H}}$	Hilbert space associated with $K_{\tilde{\mathcal{H}}}$ (under Model 3).
$\phi, \psi, \tilde{\psi}$	Feature maps for $K_{\mathcal{F}}, K_{\mathcal{H}}, K_{\tilde{\mathcal{H}}}$, respectively.
Operators and vectors	
$\theta_0^\star, \theta_1^\star$	Hilbertian elements corresponding to nuisance functions $f_0^\star, f_1^\star$ in $\mathbb{F}$.
$\eta^\star$	Hilbertian element corresponding to the target contrast function $h^\star$ in $\mathbb{X}$.
w_i	Feature vector in $\mathbb{X}$ for the second-stage regression.
$\mathbf{W}$	design operator of $\{w_i\}_{i=1}^n$.
$\mathbf{\Sigma}$	population second moment $\mathbb{E}[w_i \otimes w_i]$.
$\mathbf{\Sigma}_{\mathrm{ref}}$	evaluation operator defining the loss.
$\mathbf{S}_\lambda$	sandwich operator $\mathbf{\Sigma}_{\mathrm{ref}}^{1/2}(\mathbf{\Sigma} + \lambda \mathbf{I})^{-1}\mathbf{\Sigma}_{\mathrm{ref}}^{1/2}$.
m_i	pseudo-outcome defined by switch imputation in Equation (2).
$\mathbf{M}$	pseudo-outcome vector $(m_i)_{i \in [n]}$.

D.3 Upper Bound Analysis

We now state a single upper bound for the unified second-stage estimator introduced above. The theorem controls the evaluation error $\mathcal{E}_{\mathrm{ref}}$ in the generic Hilbert-space formulation and therefore applies simultaneously to Models 1–3 after the appropriate instantiation of $\mathbb{X}$, w_i, and $\mathbf{\Sigma}_{\mathrm{ref}}$.

Theorem D.1 (General error bound)**.** *Suppose we run Algorithm 1 with main regularizer $\lambda \gtrsim \log n/n$ and nuisance regularizer $\bar{\lambda} \asymp \log n/n$, and with $\mathscr{O}$ instantiated as KRR on a Hilbert space $\mathbb{X}$, under the assumptions in Section 2.3. Let $\hat{h}$ be the estimator produced by the algorithm and let $\hat{\eta}$ be the corresponding Hilbertian element in $\mathbb{X}$ (defined in Equation* (8)*). Then, with probability at least $1 - n^{-10}$, we have*

$$\mathcal{E}_{\mathrm{ref}}(\hat{\eta}) \lesssim \frac{1}{n\kappa} \operatorname{Tr}(\mathbf{S}_\lambda) + \lambda^2 \|\mathbf{S}_\lambda\|_{\mathrm{op}} \|(\mathbf{\Sigma} + \lambda \mathbf{I})^{-1/2} \eta^\star\|_{\mathbb{X}}^2 + \frac{1}{n\kappa} \|\mathbf{S}_\lambda\|_{\mathrm{op}} (\|f_0^\star\|_{\mathcal{F}}^2 + \|f_1^\star\|_{\mathcal{F}}^2).$$

Discussion. The theorem recovers the standard KRR bias-variance form in $\mathbb{X}$, up to an effective noise inflation by a factor $1/\kappa$ caused by treatment imbalance. In particular, the leading rate is governed entirely by the operators $\mathbf{\Sigma}$ and $\mathbf{\Sigma}_{\mathrm{ref}}$ on $\mathbb{X}$; the nuisance space does not alter the spectral bias-variance terms. The only explicit nuisance contribution is the final term, which is of order $\tilde{\mathcal{O}}((n\kappa)^{-1})$ under the bounded-norm assumption on $f_0^\star$ and $f_1^\star$, and is therefore mild relative to the nonparametric rates of interest. Consequently, once the relevant spectral decay and trace bounds are plugged in, Theorem D.1 immediately yields the L^2-error and pointwise rates for Models 1–3.

Instantiations. For L^2-error, $\mathbf{S}_\lambda = \mathbf{\Sigma}^{\frac{1}{2}}(\mathbf{\Sigma} + \lambda \mathbf{I})^{-1}\mathbf{\Sigma}^{\frac{1}{2}}$ and $\operatorname{Tr}(\mathbf{S}_\lambda) = \sum_j \frac{\rho_j}{\rho_j + \lambda}$, where $\{\rho_j\}$ are eigenvalues of $\mathbf{\Sigma}$. For point evaluation, $\mathbf{S}_\lambda$ is rank-one with $\operatorname{Tr}(\mathbf{S}_\lambda) = \|\mathbf{S}_\lambda\|_{\mathrm{op}} = \langle \psi(x_0), (\mathbf{\Sigma} + \lambda \mathbf{I})^{-1}\psi(x_0) \rangle$ (or with ϕ / $\tilde{\psi}$ under Models 2–3), i.e., the leverage score. In Sobolev RKHS $\mathcal{H} = H^\gamma(\mathcal{X})$, Lemma H.1 yields the bound $\operatorname{Tr}(\mathbf{S}_\lambda) \lesssim \lambda^{-d/(2\gamma)}$ under Model 1.

E Proof of the Theorem D.1 (General Theory)

E.1 Notation

We collect additional notation used in the proof.

Table 5: Common notation used in the appendix proofs.

Symbol	Definition
Second moment operators	
$\mathbf{\Sigma}_{\mathcal{F}}$	$\mathbb{E}[\phi(x_i)\phi(x_i)^\top]$.
$\mathbf{\Sigma}_{\mathcal{F},0}$	$\mathbb{E}[\phi(x_i)\phi(x_i)^\top \mathbf{1}(a_i = 0)]$.
$\mathbf{\Sigma}_{\mathcal{F},1}$	$\mathbb{E}[\phi(x_i)\phi(x_i)^\top \mathbf{1}(a_i = 1)]$.
RKHS elements	
$\theta_0^\star$	Hilbertian element corresponding to $f_0^\star$ in $\mathbb{F}$.
$\theta_1^\star$	Hilbertian element corresponding to $f_1^\star$ in $\mathbb{F}$.
Design operators and vectors	
$\mathbf{X}$	design operator of $\{\phi(x_i)\}_{i=1}^n$.
$\mathbf{X}_1$	design operator of $\{\phi(x_i)\mathbf{1}(a_i = 1)\}_{i=1}^n$.
$\mathbf{X}_0$	design operator of $\{\phi(x_i)\mathbf{1}(a_i = 0)\}_{i=1}^n$.
$\mathbf{Y}_1$	vector of $\{y_i \mathbf{1}(a_i = 1)\}_{i=1}^n$.
$\mathbf{Y}_0$	vector of $\{y_i \mathbf{1}(a_i = 0)\}_{i=1}^n$.
$\boldsymbol{\varepsilon}_1$	vector of $\{\varepsilon_i \mathbf{1}(a_i = 1)\}_{i=1}^n$.
$\boldsymbol{\varepsilon}_0$	vector of $\{\varepsilon_i \mathbf{1}(a_i = 0)\}_{i=1}^n$.
m_i	pseudo-outcome defined by switch imputation in Equation (2).
$\mathbf{M}$	pseudo-outcome vector $(m_i)_{i\in[n]}$ with decomposition $\mathbf{M} = \mathbf{X}(\theta_1^\star - \theta_0^\star) + \boldsymbol{\varepsilon}' + \mathbf{X}_0(\hat{\theta}_1 - \theta_1^\star) - \mathbf{X}_1(\hat{\theta}_0 - \theta_0^\star)$.
$\boldsymbol{\varepsilon}'$	vector with entries $\varepsilon_i(-1)^{a_i+1}$, $i \in [n]$.

E.2 Good Event

By Assumption 4, for every $\lambda' \in \{\lambda, \bar{\lambda}\}$ we have

$$n\mathbf{\Sigma}_{\mathcal{F},1} \preceq \frac{1}{\kappa}(n\mathbf{\Sigma}_{\mathcal{F},0} + \mathbf{I}), \qquad n\mathbf{\Sigma}_{\mathcal{F},0} \preceq \frac{1}{\kappa}(n\mathbf{\Sigma}_{\mathcal{F},1} + \mathbf{I}).$$

Since $\lambda' \gtrsim \log n/n$, we may absorb the identity term into $n\lambda'\mathbf{I}$. Consequently, there exists an absolute constant $c > 1$ such that

$$\begin{aligned}
&\frac{\kappa}{c}(n\mathbf{\Sigma}_{\mathcal{F},0} + n\lambda'\mathbf{I}) \preceq n\mathbf{\Sigma}_{\mathcal{F},1} + n\lambda'\mathbf{I} \preceq \frac{c}{\kappa}(n\mathbf{\Sigma}_{\mathcal{F},0} + n\lambda'\mathbf{I}),\\
&\frac{\kappa}{c}(n\mathbf{\Sigma}_{\mathcal{F},0} + n\lambda'\mathbf{I}) \preceq n\mathbf{\Sigma}_{\mathcal{F}} + n\lambda'\mathbf{I} \preceq \frac{c}{\kappa}(n\mathbf{\Sigma}_{\mathcal{F},0} + n\lambda'\mathbf{I}),\\
&\frac{\kappa}{c}(n\mathbf{\Sigma}_{\mathcal{F},1} + n\lambda'\mathbf{I}) \preceq n\mathbf{\Sigma}_{\mathcal{F}} + n\lambda'\mathbf{I} \preceq \frac{c}{\kappa}(n\mathbf{\Sigma}_{\mathcal{F},1} + n\lambda'\mathbf{I}).
\end{aligned} \tag{9}$$

Applying Lemma H.2 and Remark H.3, for any $\lambda' \in \{\lambda, \bar{\lambda}\}$ and with probability at least $1 - n^{-11}$, there exists an absolute constant $c_2 > 1$ such that

$$\begin{aligned}
&\frac{1}{c_2}(n\mathbf{\Sigma}_{\mathcal{F},1} + n\lambda'\mathbf{I}) \preceq \mathbf{X}_1^\top\mathbf{X}_1 + n\lambda'\mathbf{I} \preceq c_2(n\mathbf{\Sigma}_{\mathcal{F},1} + n\lambda'\mathbf{I})\\
&\frac{1}{c_2}(n\mathbf{\Sigma}_{\mathcal{F},0} + n\lambda'\mathbf{I}) \preceq \mathbf{X}_0^\top\mathbf{X}_0 + n\lambda'\mathbf{I} \preceq c_2(n\mathbf{\Sigma}_{\mathcal{F},0} + n\lambda'\mathbf{I})\\
&\frac{1}{c_2}(n\mathbf{\Sigma}_{\mathcal{F}} + n\lambda'\mathbf{I}) \preceq \mathbf{X}^\top\mathbf{X} + n\lambda'\mathbf{I} \preceq c_2(n\mathbf{\Sigma}_{\mathcal{F}} + n\lambda'\mathbf{I}).
\end{aligned} \tag{10}$$

Moreover, for any $\lambda' \in \{\lambda, \bar{\lambda}\}$, with probability at least $1 - n^{-11}$,

$$\frac{1}{c}(n\mathbf{\Sigma} + n\lambda'\mathbf{I}) \preceq \mathbf{W}^\top\mathbf{W} + n\lambda'\mathbf{I} \preceq c(n\mathbf{\Sigma} + n\lambda'\mathbf{I}) \tag{11}$$

for some absolute constant $c > 1$.

Let $\mathscr{E}_0$ be the event on which the above concentration inequalities hold. Then

$$\mathbb{P}[\mathscr{E}_0] \geq 1 - 3n^{-11}.$$

Combining Equation (9) and Equation (10), on the event $\mathscr{E}_0$ we obtain

$$\begin{aligned}
\frac{\kappa}{c}(\mathbf{X}_0^\top \mathbf{X}_0 + n\lambda' \mathbf{I}) &\preceq \mathbf{X}_1^\top \mathbf{X}_1 + n\lambda' \mathbf{I} \preceq \frac{c}{\kappa}(\mathbf{X}_0^\top \mathbf{X}_0 + n\lambda' \mathbf{I}) \\
\frac{\kappa}{c}(\mathbf{X}_0^\top \mathbf{X}_0 + n\lambda' \mathbf{I}) &\preceq \mathbf{X}^\top \mathbf{X} + n\lambda' \mathbf{I} \preceq \frac{c}{\kappa}(\mathbf{X}_0^\top \mathbf{X}_0 + n\lambda' \mathbf{I}) \\
\frac{\kappa}{c}(\mathbf{X}_1^\top \mathbf{X}_1 + n\lambda' \mathbf{I}) &\preceq \mathbf{X}^\top \mathbf{X} + n\lambda' \mathbf{I} \preceq \frac{c}{\kappa}(\mathbf{X}_1^\top \mathbf{X}_1 + n\lambda' \mathbf{I})
\end{aligned} \tag{12}$$

for some constant $c > 1$.

E.3 Proof

We work on the event $\mathscr{E}_0$, which controls $\mathbf{X}_0, \mathbf{X}_1, \mathbf{X}$ as defined above.

Error Decomposition. Recalling the pseudo-outcome decomposition,

$$\hat{\eta} = (\mathbf{W}^\top \mathbf{W} + n\lambda \mathbf{I})^{-1} \mathbf{W}^\top \big(\mathbf{W}\eta^\star + \boldsymbol{\varepsilon}' + \mathbf{X}_0(\hat{\theta}_1 - \theta_1^\star) - \mathbf{X}_1(\hat{\theta}_0 - \theta_0^\star) \big).$$

The estimation error can be written as

$$\begin{aligned}
&\hat{\eta} - \eta^\star \\
&= (\mathbf{W}^\top \mathbf{W} + n\lambda \mathbf{I})^{-1} \big(\mathbf{W}^\top \mathbf{X}_0(\hat{\theta}_1 - \theta_1^\star) - \mathbf{W}^\top \mathbf{X}_1(\hat{\theta}_0 - \theta_0^\star) - n\lambda \eta^\star \big) \\
&\quad + (\mathbf{W}^\top \mathbf{W} + n\lambda \mathbf{I})^{-1} \mathbf{W}^\top \boldsymbol{\varepsilon}'.
\end{aligned}$$

Accordingly, the error decomposes as

$$\begin{aligned}
&\|\boldsymbol{\Sigma}_{\mathrm{ref}}^{\frac{1}{2}}(\hat{\eta} - \eta^\star)\|_{\mathbb{X}} \\
&\lesssim \|\boldsymbol{\Sigma}_{\mathrm{ref}}^{\frac{1}{2}}(\mathbf{W}^\top \mathbf{W} + n\lambda \mathbf{I})^{-1} \mathbf{W}^\top \mathbf{X}_1(\theta_0^\star - \hat{\theta}_0)\|_{\mathbb{X}} + \|\boldsymbol{\Sigma}_{\mathrm{ref}}^{\frac{1}{2}}(\mathbf{W}^\top \mathbf{W} + n\lambda \mathbf{I})^{-1} \mathbf{W}^\top \mathbf{X}_0(\hat{\theta}_1 - \theta_1^\star)\|_{\mathbb{X}} \\
&\quad + \|\boldsymbol{\Sigma}_{\mathrm{ref}}^{\frac{1}{2}}(\mathbf{W}^\top \mathbf{W} + n\lambda \mathbf{I})^{-1} n\lambda \eta^\star\|_{\mathbb{X}} + \|\boldsymbol{\Sigma}_{\mathrm{ref}}^{\frac{1}{2}}(\mathbf{W}^\top \mathbf{W} + n\lambda \mathbf{I})^{-1} \mathbf{W}^\top \boldsymbol{\varepsilon}'\|_{\mathbb{X}}.
\end{aligned}$$

We first bound the term

$$\begin{aligned}
&\|\boldsymbol{\Sigma}_{\mathrm{ref}}^{\frac{1}{2}}(\mathbf{W}^\top \mathbf{W} + n\lambda \mathbf{I})^{-1} \mathbf{W}^\top \mathbf{X}_0(\hat{\theta}_1 - \theta_1^\star)\|_{\mathbb{X}} \\
&= \|\boldsymbol{\Sigma}_{\mathrm{ref}}^{\frac{1}{2}}(\mathbf{W}^\top \mathbf{W} + n\lambda \mathbf{I})^{-1} \mathbf{W}^\top \mathbf{X}_0(\mathbf{X}_1^\top \mathbf{X}_1 + n\bar{\lambda} \mathbf{I})^{-1}(\mathbf{X}_1^\top \boldsymbol{\varepsilon}_1 - n\bar{\lambda}\theta_1^\star)\|_{\mathbb{X}} \\
&\leq \|\boldsymbol{\Sigma}_{\mathrm{ref}}^{\frac{1}{2}}(\mathbf{W}^\top \mathbf{W} + n\lambda \mathbf{I})^{-1} \mathbf{W}^\top \mathbf{X}_0(\mathbf{X}_1^\top \mathbf{X}_1 + n\bar{\lambda} \mathbf{I})^{-1} \mathbf{X}_1^\top \boldsymbol{\varepsilon}_1\|_{\mathbb{X}} \\
&\quad + \|\boldsymbol{\Sigma}_{\mathrm{ref}}^{\frac{1}{2}}(\mathbf{W}^\top \mathbf{W} + n\lambda \mathbf{I})^{-1} \mathbf{W}^\top \mathbf{X}_0(\mathbf{X}_1^\top \mathbf{X}_1 + n\bar{\lambda} \mathbf{I})^{-1} n\bar{\lambda}\theta_1^\star\|_{\mathbb{X}}
\end{aligned}$$

Bias Term. Under the event $\mathscr{E}_0$, we have

$$\begin{aligned}
\mathsf{B}_1 &:= \|\boldsymbol{\Sigma}_{\mathrm{ref}}^{\frac{1}{2}}(\mathbf{W}^\top \mathbf{W} + n\lambda \mathbf{I})^{-1} \mathbf{W}^\top \mathbf{X}_0(\mathbf{X}_1^\top \mathbf{X}_1 + n\bar{\lambda} \mathbf{I})^{-1}(n\bar{\lambda}\theta_1^\star)\|_{\mathbb{X}} \\
&\lesssim \|\boldsymbol{\Sigma}_{\mathrm{ref}}^{\frac{1}{2}}(\mathbf{W}^\top \mathbf{W} + n\lambda \mathbf{I})^{-\frac{1}{2}}\|_{\mathrm{op}} \|(\mathbf{W}^\top \mathbf{W} + n\lambda \mathbf{I})^{-\frac{1}{2}} \mathbf{W}^\top\|_{\mathrm{op}} \\
&\quad \times \|\mathbf{X}_0(\mathbf{X}_1^\top \mathbf{X}_1 + n\bar{\lambda} \mathbf{I})^{-\frac{1}{2}}\|_{\mathrm{op}} \|(\mathbf{X}_1^\top \mathbf{X}_1 + n\bar{\lambda} \mathbf{I})^{-\frac{1}{2}}(n\bar{\lambda}\theta_1^\star)\|_{\mathbb{F}} \\
&\overset{(i)}{\lesssim} \|\boldsymbol{\Sigma}_{\mathrm{ref}}^{\frac{1}{2}}(n\boldsymbol{\Sigma} + n\lambda \mathbf{I})^{-\frac{1}{2}}\|_{\mathrm{op}} \times 1 \times \frac{1}{\sqrt{\kappa}} \times \sqrt{n\bar{\lambda}} \|\theta_1^\star\|_{\mathbb{F}} \\
&\lesssim \frac{1}{\sqrt{n}} \|\mathbf{S}_\lambda\|_{\mathrm{op}}^{\frac{1}{2}} \times 1 \times \frac{1}{\sqrt{\kappa}} \times \sqrt{n\bar{\lambda}} \|\theta_1^\star\|_{\mathbb{F}} \\
&\lesssim \frac{1}{\sqrt{\kappa}} \|\mathbf{S}_\lambda\|_{\mathrm{op}}^{\frac{1}{2}} \sqrt{\bar{\lambda}} \|\theta_1^\star\|_{\mathbb{F}} \\
&\lesssim \frac{\sqrt{\log n}}{\sqrt{n\kappa}} \|\mathbf{S}_\lambda\|_{\mathrm{op}}^{\frac{1}{2}} \|\theta_1^\star\|_{\mathbb{F}}.
\end{aligned}$$

In step (i), we used the definition of the good event $\mathscr{E}_0$ (specifically Equation (12) and Equation (11)).

Variance Term. By the Hanson–Wright inequality, on the event $\mathscr{E}_0$, we obtain the following upper bound with probability at least $1 - n^{-11}$:

$$
\begin{aligned}
&\mathsf{V}_1^2 \\
&:= \|\mathbf{\Sigma}_{\mathrm{ref}}^{\frac{1}{2}}(\mathbf{W}^\top\mathbf{W} + n\lambda\mathbf{I})^{-1}\mathbf{W}^\top\mathbf{X}_0(\mathbf{X}_1^\top\mathbf{X}_1 + n\bar{\lambda}\mathbf{I})^{-1}\mathbf{X}_1^\top\boldsymbol{\varepsilon}_1\|_{\mathbb{X}}^2 \\
&\lesssim \log n \cdot \mathrm{Tr}\Big(\mathbf{\Sigma}_{\mathrm{ref}}^{\frac{1}{2}}(\mathbf{W}^\top\mathbf{W} + n\lambda\mathbf{I})^{-1}\mathbf{W}^\top\mathbf{X}_0(\mathbf{X}_1^\top\mathbf{X}_1 + n\bar{\lambda}\mathbf{I})^{-1}\mathbf{X}_1^\top\mathbf{X}_1 \\
&\qquad \times (\mathbf{X}_1^\top\mathbf{X}_1 + n\bar{\lambda}\mathbf{I})^{-1}\mathbf{X}_0^\top\mathbf{W}(\mathbf{W}^\top\mathbf{W} + n\lambda\mathbf{I})^{-1}\mathbf{\Sigma}_{\mathrm{ref}}^{\frac{1}{2}}\Big) \\
&\lesssim \mathrm{Tr}\Big(\mathbf{\Sigma}_{\mathrm{ref}}^{\frac{1}{2}}(\mathbf{W}^\top\mathbf{W} + n\lambda\mathbf{I})^{-1}\mathbf{W}^\top\mathbf{X}(\mathbf{X}_1^\top\mathbf{X}_1 + n\bar{\lambda}\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{W} \\
&\qquad \times (\mathbf{W}^\top\mathbf{W} + n\lambda\mathbf{I})^{-1}\mathbf{\Sigma}_{\mathrm{ref}}^{\frac{1}{2}}\Big)\log n \\
&\overset{(i)}{\lesssim} \frac{1}{\kappa}\,\mathrm{Tr}\Big(\mathbf{\Sigma}_{\mathrm{ref}}^{\frac{1}{2}}(\mathbf{W}^\top\mathbf{W} + n\lambda\mathbf{I})^{-1}\mathbf{W}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X} + n\bar{\lambda}\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{W} \\
&\qquad \times (\mathbf{W}^\top\mathbf{W} + n\lambda\mathbf{I})^{-1}\mathbf{\Sigma}_{\mathrm{ref}}^{\frac{1}{2}}\Big)\log n \\
&\lesssim \frac{1}{\kappa}\,\mathrm{Tr}\Big(\mathbf{\Sigma}_{\mathrm{ref}}^{\frac{1}{2}}(\mathbf{W}^\top\mathbf{W} + n\lambda\mathbf{I})^{-1}\mathbf{W}^\top\mathbf{W}(\mathbf{W}^\top\mathbf{W} + n\lambda\mathbf{I})^{-1}\mathbf{\Sigma}_{\mathrm{ref}}^{\frac{1}{2}}\Big)\log n \\
&\lesssim \frac{1}{\kappa}\,\mathrm{Tr}(\mathbf{\Sigma}_{\mathrm{ref}}^{\frac{1}{2}}(\mathbf{W}^\top\mathbf{W} + n\lambda\mathbf{I})^{-1}\mathbf{\Sigma}_{\mathrm{ref}}^{\frac{1}{2}})\log n \\
&\overset{(ii)}{\lesssim} \frac{\log n}{n\kappa}\,\mathrm{Tr}(\mathbf{\Sigma}_{\mathrm{ref}}(\mathbf{\Sigma} + \lambda\mathbf{I})^{-1}) \\
&= \frac{\log n}{n\kappa}\,\mathrm{Tr}(\mathbf{S}_\lambda).
\end{aligned}
$$

Here, in step (i) we used Equation (12), and in step (ii) we used Equation (11).

The remaining propagated term is bounded similarly:

$$
\|\mathbf{\Sigma}_{\mathrm{ref}}^{\frac{1}{2}}(\mathbf{W}^\top\mathbf{W} + n\lambda\mathbf{I})^{-1}\mathbf{W}^\top\mathbf{X}(\theta_0^\star - \hat{\theta}_0)\|_{\mathbb{X}}^2 \lesssim \frac{\log n}{n\kappa}\,\mathrm{Tr}(\mathbf{S}_\lambda) + \frac{\log n}{n\kappa}\|\theta_0^\star\|_{\mathbb{F}}^2.
$$

Ridge Bias Term. Finally, we bound the ridge bias term:

$$
\begin{aligned}
\|\mathbf{\Sigma}_{\mathrm{ref}}^{\frac{1}{2}}(\mathbf{W}^\top\mathbf{W} + n\lambda\mathbf{I})^{-1}n\lambda\eta^\star\|_{\mathbb{X}} &\lesssim \|\mathbf{\Sigma}_{\mathrm{ref}}^{\frac{1}{2}}(\mathbf{W}^\top\mathbf{W} + n\lambda\mathbf{I})^{-\frac{1}{2}}\|_{\mathrm{op}} \cdot \|(\mathbf{W}^\top\mathbf{W} + n\lambda\mathbf{I})^{-\frac{1}{2}}n\lambda\eta^\star\|_{\mathbb{X}} \\
&\overset{(i)}{\lesssim} \|\mathbf{\Sigma}_{\mathrm{ref}}^{\frac{1}{2}}(n\mathbf{\Sigma} + n\lambda\mathbf{I})^{-\frac{1}{2}}\|_{\mathrm{op}} \cdot \|(n\mathbf{\Sigma} + n\lambda\mathbf{I})^{-\frac{1}{2}}n\lambda\eta^\star\|_{\mathbb{X}} \\
&\lesssim \frac{1}{\sqrt{n}}\|\mathbf{S}_\lambda\|_{\mathrm{op}}^{\frac{1}{2}} \times \sqrt{n}\lambda \times \|(\mathbf{\Sigma} + \lambda\mathbf{I})^{-\frac{1}{2}}\eta^\star\|_{\mathbb{X}} \\
&= \lambda\|\mathbf{S}_\lambda\|_{\mathrm{op}}^{\frac{1}{2}}\|(\mathbf{\Sigma} + \lambda\mathbf{I})^{-\frac{1}{2}}\eta^\star\|_{\mathbb{X}}
\end{aligned}
$$

where in step (i) we used Equation (11).

Combining the variance and bias bounds above with the decomposition yields the stated bound in Theorem D.1.

F Proofs of the Main Corollaries from General Theory

We now show how each main corollary follows from Theorem D.1 by specializing $\mathbf{\Sigma}_{\mathrm{ref}}, \mathbf{\Sigma}$, and $\mathbb{X}$, bounding $\|\mathbf{S}_\lambda\|_{\mathrm{op}}$ and $\mathrm{Tr}(\mathbf{S}_\lambda)$, and choosing λ. Polylogarithmic factors are absorbed into $\lesssim$. Throughout this section, w_i, $\mathbb{X}$, $\mathbf{\Sigma}$, and $\mathbf{\Sigma}_{\mathrm{ref}}$ are as defined in Section D and Table 4.

F.1 Proof of Corollary 7 (Model 1, L^2-error)

Set $\mathbf{\Sigma}_{\mathrm{ref}} = \mathbf{\Sigma}_{\mathcal{H}}$ and $\mathbf{\Sigma} = \mathbf{\Sigma}_{\mathcal{H}}$, where $\mathbf{\Sigma}_{\mathcal{H}} = \mathbb{E}[\psi(x) \otimes \psi(x)]$. Let $\mathbf{\Sigma}_{\mathcal{H}} = \sum_{j\geq 1}\rho_{\mathcal{H},j}\,u_j \otimes u_j$ be the eigendecomposition. Then

$$
\mathbf{S}_\lambda = \mathbf{\Sigma}_{\mathcal{H}}^{1/2}(\mathbf{\Sigma}_{\mathcal{H}} + \lambda\mathbf{I})^{-1}\mathbf{\Sigma}_{\mathcal{H}}^{1/2}
$$

has eigenvalues $\rho_{\mathcal{H},j}/(\rho_{\mathcal{H},j}+\lambda)$, so

$$\|\mathbf{S}_\lambda\|_{\mathrm{op}} = \max_j \frac{\rho_{\mathcal{H},j}}{\rho_{\mathcal{H},j}+\lambda} \le 1, \qquad \mathrm{Tr}(\mathbf{S}_\lambda) = \sum_j \frac{\rho_{\mathcal{H},j}}{\rho_{\mathcal{H},j}+\lambda}.$$

For the ridge bias, note that

$$\lambda^2 \|(\mathbf{\Sigma}_{\mathcal{H}}+\lambda\mathbf{I})^{-1/2}\eta^\star\|_{\mathbb{H}}^2 \le \lambda \|\eta^\star\|_{\mathbb{H}}^2,$$

so the bias term is $O(\lambda)$. Plugging these into Theorem D.1 yields

$$\mathcal{E}_{L^2}(\hat{h}) \lesssim \frac{1}{n\kappa}\sum_j \frac{\rho_{\mathcal{H},j}}{\rho_{\mathcal{H},j}+\lambda} + \lambda.$$

If $\rho_{\mathcal{H},j} \lesssim j^{-2\ell_{\mathcal{H}}}$, the effective-dimension bound $\sum_j \frac{\rho_{\mathcal{H},j}}{\rho_{\mathcal{H},j}+\lambda} \lesssim \lambda^{-1/(2\ell_{\mathcal{H}})}$ holds [Zhang et al., 2023, Ma et al., 2023], and choosing $\lambda \asymp (n\kappa)^{-\frac{2\ell_{\mathcal{H}}}{2\ell_{\mathcal{H}}+1}}$ gives the rate in Corollary 7. The finite-rank case follows by the bound $\sum_j \frac{\rho_{\mathcal{H},j}}{\rho_{\mathcal{H},j}+\lambda} \le D$.

F.2 Proof of Corollary 8 (Model 2, L^2-error)

Set $\mathbf{\Sigma}_{\mathrm{ref}} = \mathbf{\Sigma}_{\mathcal{F}}$ and $\mathbf{\Sigma} = \mathbf{\Sigma}_{\mathcal{F}}$, where $\mathbf{\Sigma}_{\mathcal{F}} = \mathbb{E}[\phi(x)\otimes\phi(x)]$. Write $\mathbf{\Sigma}_{\mathcal{F}} = \sum_{j\ge1}\rho_{\mathcal{F},j}\, u_j \otimes u_j$. Then $\|\mathbf{S}_\lambda\|_{\mathrm{op}} \le 1$ and $\mathrm{Tr}(\mathbf{S}_\lambda) = \sum_j \frac{\rho_{\mathcal{F},j}}{\rho_{\mathcal{F},j}+\lambda}$.

For the ridge bias, we use the source condition in its spectral form: there exists $\gamma \in \mathbb{F}$ such that $\gamma = \mathbf{\Sigma}_{\mathcal{F}}^{-\nu/2}\eta^\star$. Since $\mathbf{\Sigma}_{\mathrm{ref}} = \mathbf{\Sigma}_{\mathcal{F}}$, we can sharpen the generic bound by tracking the $\mathbf{\Sigma}_{\mathcal{F}}$ factor

$$\begin{aligned}\lambda^2\|\mathbf{\Sigma}_{\mathcal{F}}^{1/2}(\mathbf{\Sigma}_{\mathcal{F}}+\lambda\mathbf{I})^{-1}\eta^\star\|^2 &\lesssim \lambda^2\|(\mathbf{\Sigma}_{\mathcal{F}}+\lambda\mathbf{I})^{-\frac{1-\nu}{2}}(\mathbf{\Sigma}_{\mathcal{F}}+\lambda\mathbf{I})^{-\frac{\nu}{2}}\eta^\star\|^2 \\ &\lesssim \lambda^{1+\nu}.\end{aligned}$$

Thus the ridge bias term is $O(\lambda^{1+\nu})$ under the source condition. Therefore,

$$\mathcal{E}_{L^2}(\hat{h}) \lesssim \frac{1}{n\kappa}\sum_j \frac{\rho_{\mathcal{F},j}}{\rho_{\mathcal{F},j}+\lambda} + \lambda^{1+\nu}.$$

If $\rho_{\mathcal{F},j} \lesssim j^{-2\ell_{\mathcal{F}}}$, then $\sum_j \frac{\rho_{\mathcal{F},j}}{\rho_{\mathcal{F},j}+\lambda} \lesssim \lambda^{-1/(2\ell_{\mathcal{F}})}$ [Ma et al., 2023], and choosing $\lambda \asymp (n\kappa)^{-\frac{2\ell_{\mathcal{F}}}{1+2\ell_{\mathcal{F}}(1+\nu)}}$ yields the rate in Corollary 8.

F.3 Proof of Corollary 9 (Model 3, L^2-error)

Set $\mathbf{\Sigma}_{\mathrm{ref}} = \mathbf{\Sigma}_{\tilde{\mathcal{H}}}$ and $\mathbf{\Sigma} = \mathbf{\Sigma}_{\tilde{\mathcal{H}}}$, where $\mathbf{\Sigma}_{\tilde{\mathcal{H}}} = \mathbb{E}[\tilde{\psi}(x)\otimes\tilde{\psi}(x)]$, and write $\mathbf{\Sigma}_{\tilde{\mathcal{H}}} = \sum_{j\ge1}\tilde{\rho}_j\,\tilde{u}_j\otimes\tilde{u}_j$. Then $\|\mathbf{S}_\lambda\|_{\mathrm{op}} \le 1$ and $\mathrm{Tr}(\mathbf{S}_\lambda) = \sum_j \frac{\tilde{\rho}_j}{\tilde{\rho}_j+\lambda}$. As in Model 1, the ridge bias satisfies $\lambda^2\|(\mathbf{\Sigma}_{\tilde{\mathcal{H}}} + \lambda\mathbf{I})^{-1/2}\tilde{\eta}^\star\|^2 \le \lambda\|\tilde{\eta}^\star\|_{\tilde{\mathbb{H}}}^2$, so it is $O(\lambda)$ when $\|\tilde{\eta}^\star\|_{\tilde{\mathbb{H}}}$ is bounded. Hence,

$$\mathcal{E}_{L^2}(\hat{h}) \lesssim \frac{1}{n\kappa}\sum_j \frac{\tilde{\rho}_j}{\tilde{\rho}_j+\lambda} + \lambda.$$

If $\tilde{\rho}_j \lesssim j^{-2\tilde{\ell}}$, then $\sum_j \frac{\tilde{\rho}_j}{\tilde{\rho}_j+\lambda} \lesssim \lambda^{-1/(2\tilde{\ell})}$ [Ma et al., 2023], and choosing $\lambda \asymp (n\kappa)^{-\frac{2\tilde{\ell}}{1+2\tilde{\ell}}}$ gives the rate in Corollary 9.

F.4 Proof of Corollary 10 (Models 1–2, point evaluation)

For a fixed x_0, let $\mathbf{\Sigma}_{\mathrm{ref}} = \psi(x_0)\otimes\psi(x_0)$ and $\mathbf{\Sigma} = \mathbf{\Sigma}_{\mathcal{H}}$, where $\mathbf{\Sigma}_{\mathcal{H}} = \mathbb{E}[\psi(x)\otimes\psi(x)]$; then

$$\mathbf{S}_\lambda = \big((\mathbf{\Sigma}+\lambda\mathbf{I})^{-1/2}\psi(x_0)\big)\otimes\big((\mathbf{\Sigma}+\lambda\mathbf{I})^{-1/2}\psi(x_0)\big)$$

is rank-one with

$$\|\mathbf{S}_\lambda\|_{\mathrm{op}} = \mathrm{Tr}(\mathbf{S}_\lambda) = \mathcal{Q}_\lambda^2 := \langle\psi(x_0), (\mathbf{\Sigma}+\lambda\mathbf{I})^{-1}\psi(x_0)\rangle,$$

the standard leverage score. Theorem D.1 yields

$$\mathcal{E}_{x_0}(\hat{h}) \lesssim \left(\frac{1}{n\kappa} + \lambda\right) \mathcal{Q}_\lambda^2.$$

When $\mathcal{H} = H^\gamma(\mathcal{X})$ with $\gamma > d/2$, Lemma H.1 (Sobolev embedding / leverage score control) gives $\mathcal{Q}_\lambda^2 \lesssim \lambda^{-d/(2\gamma)}$ [Tuo and Zou, 2024]. Taking $\lambda \asymp (n\kappa)^{-1}$ yields $\mathcal{E}_{x_0}(\hat{h}) \lesssim (n\kappa)^{-(2\gamma-d)/(2\gamma)}$, as in Corollary 10.

F.5 Proof of Corollary 11 (Model 3, point evaluation)

Let $\mathbf{\Sigma}_{\mathrm{ref}} = \tilde{\psi}(\tilde{x}_0) \otimes \tilde{\psi}(\tilde{x}_0)$ and $\mathbf{\Sigma} = \mathbf{\Sigma}_{\tilde{\mathcal{H}}}$, where $\mathbf{\Sigma}_{\tilde{\mathcal{H}}} = \mathbb{E}[\tilde{\psi}(x) \otimes \tilde{\psi}(x)]$, so that

$$\mathbf{S}_\lambda = \left((\mathbf{\Sigma}_{\tilde{\mathcal{H}}} + \lambda \mathbf{I})^{-1/2} \tilde{\psi}(\tilde{x}_0)\right) \otimes \left((\mathbf{\Sigma}_{\tilde{\mathcal{H}}} + \lambda \mathbf{I})^{-1/2} \tilde{\psi}(\tilde{x}_0)\right)$$

and $\|\mathbf{S}_\lambda\|_{\mathrm{op}} = \mathrm{Tr}(\mathbf{S}_\lambda) = \tilde{\mathcal{Q}}_\lambda^2 := \langle \tilde{\psi}(\tilde{x}_0), (\mathbf{\Sigma}_{\tilde{\mathcal{H}}} + \lambda \mathbf{I})^{-1} \tilde{\psi}(\tilde{x}_0) \rangle$. Then

$$\mathcal{E}_{x_0}(\hat{h}) \lesssim \left(\frac{1}{n\kappa} + \lambda\right) \tilde{\mathcal{Q}}_\lambda^2.$$

If $\tilde{\mathcal{H}} = H^\beta(\tilde{\mathcal{X}})$ with $\beta > \tilde{d}/2$, Sobolev embedding Lemma H.1 yields $\tilde{\mathcal{Q}}_\lambda^2 \lesssim \lambda^{-\tilde{d}/(2\beta)}$ [Tuo and Zou, 2024]. Choosing $\lambda \asymp (n\kappa)^{-1}$ gives $\mathcal{E}_{x_0}(\hat{h}) \lesssim (n\kappa)^{-(2\beta-\tilde{d})/(2\beta)}$, matching Corollary 11.

G Proof of Theorem B.2 (Oracle Inequality)

G.1 Empirical target error and proxy-based selection

For a candidate g, define its *true empirical error*

$$R_3(g) = \frac{1}{n_3} \sum_{i=1}^{n_3} (g(x_{3i}) - h^\star(x_{3i}))^2.$$

This differs from the observable proxy loss minimized by Algorithm 2. The latter includes outcome noise and cannot be substituted for R_3 in a concentration argument around population CATE error.

Lemma G.1 (Empirical target-error oracle inequality)**.** *Under the assumptions of Theorem B.2, on the covariance concentration event used below, condition on $\mathcal{D}_1$ and on the covariates and treatments in $\mathcal{D}_2, \mathcal{D}_3$. With conditional probability at least $1 - \delta$ over the remaining outcome noise,*

$$R_3(\bar{h}_{\mathrm{ms}}) \le C \min_{g \in \bar{\mathcal{M}}} R_3(g) + C \left[\frac{\log n}{n_2 \kappa} + \left(\frac{1}{n_3} + \frac{1}{n_2 \kappa}\right) \log(L/\delta)\right].$$

The candidate set and B are fixed under this conditioning, but the proxy-training outcomes in $\mathcal{D}_2$ are not conditioned away. As shown below, the conditional proxy-noise vector has sub-Gaussian norm at most $C(1 + \sqrt{n_3/(n_2\kappa)})$, and its squared mean bias is at most $Cn_3 \log n/(n_2\kappa)$. Lemma G.2, with $\gamma = 1$, then proves the display. Combining the design event with $\delta = n^{-12}$ gives the required empirical comparison with overall failure probability of order n^{-11}, after increasing constants.

G.2 Population conversion and the role of B

Condition now only on $\mathcal{D}_1$ (and any independent data used to choose B). Put $M = B + B^\star$, $R(g) = \mathcal{E}_{L^2}(g)$, and $Z_i(g) = (g(x_{3i}) - h^\star(x_{3i}))^2$. For each truncated candidate,

$$0 \le Z_i(g) \le M^2, \qquad \mathrm{Var}(Z_i(g)) \le \mathbb{E} Z_i(g)^2 \le M^2 R(g).$$

Bernstein's inequality, a union bound over L candidates, and $\sqrt{uv} \le \eta u + C_\eta v$ give, simultaneously for all candidates,

$$\begin{aligned} |R_3(g) - R(g)| &\le \sqrt{2M^2 R(g) t/n_3} + 2M^2 t/(3n_3) \\ &\le \tfrac{1}{2} R(g) + CM^2 t/n_3, \qquad t = \log(2Ln^{12}), \end{aligned}$$

with probability at least $1 - n^{-12}$. In particular,

$$R(g) \le 2R_3(g) + CM^2 t/n_3, \qquad R_3(g) \le \tfrac{3}{2}R(g) + CM^2 t/n_3.$$

Apply the first inequality to $\bar{h}_{\mathrm{ms}}$, the empirical oracle lemma, and the second inequality to a population-optimal candidate. Since $n_1 \asymp n_2 \asymp n_3 \asymp n$, their intersection yields (7). No independence between these events is needed. The variance bound proportional to $R(g)$ is essential: a uniform $O(M^2/\sqrt{n_3})$ deviation would not establish the stated fast remainder. If a random B fails to bound $h^\star$ with probability δ_B, a further union bound adds δ_B.

G.3 Proof of Lemma G.1 (Oracle Inequality for Empirical L^2-Error)

Write $n_j = |\mathcal{D}_j|$ for $j = 1, 2, 3$, with $n = n_1 + n_2 + n_3$, and use $\tilde{\lambda} \asymp \log n / n_2$ as in (5). Expectations below condition on $\mathcal{D}_1$ and the covariates and treatments in $\mathcal{D}_2, \mathcal{D}_3$, leaving their outcome noises random. Let $\mathbf{X}_j$ have rows $\phi(x_{ji})^\top$, and let $\mathbf{X}_{j,a}$ have rows $\mathbf{1}(a_{ji} = a)\phi(x_{ji})^\top$, for $i \in [n_j]$ and $j = 2, 3$. Write $\boldsymbol{\varepsilon}_{j,a} = (\mathbf{1}(a_{ji} = a)\varepsilon_{ji})_{i=1}^{n_j}$ and $\boldsymbol{\varepsilon}'_3 = \boldsymbol{\varepsilon}_{3,1} - \boldsymbol{\varepsilon}_{3,0}$. Let $\mathbf{M}_3 := (\tilde{m}_{3i})_{i=1}^{n_3}$ and $\mathbf{M}_3^\star := (h^\star(x_{3i}))_{i=1}^{n_3}$. Let $\tilde{\theta}_a$ denote the Hilbertian element corresponding to $\tilde{f}_a$ for $a \in \{0, 1\}$.

$$\begin{aligned}\mathbf{M}_3 &= (\mathbf{X}_{3,0}\tilde{\theta}_1 - \mathbf{X}_{3,0}\theta_0^\star - \boldsymbol{\varepsilon}_{3,0}) + (\mathbf{X}_{3,1}\theta_1^\star + \boldsymbol{\varepsilon}_{3,1} - \mathbf{X}_{3,1}\tilde{\theta}_0) \\ &= \mathbf{X}_3(\theta_1^\star - \theta_0^\star) + \boldsymbol{\varepsilon}'_3 + \mathbf{X}_{3,0}(\tilde{\theta}_1 - \theta_1^\star) - \mathbf{X}_{3,1}(\tilde{\theta}_0 - \theta_0^\star). \end{aligned} \tag{13}$$

Thus $\mathbf{M}_3 - \mathbf{M}_3^\star = \mathsf{V}_{\mathrm{ms}} + \mathsf{B}_{\mathrm{ms}}$, where

$$\begin{aligned}\mathsf{V}_{\mathrm{ms}} &:= \boldsymbol{\varepsilon}'_3 + \mathbf{X}_{3,0}(\mathbf{X}_{2,1}^\top \mathbf{X}_{2,1} + n_2\tilde{\lambda}\mathbf{I})^{-1}\mathbf{X}_{2,1}^\top \boldsymbol{\varepsilon}_{2,1} \\ &\quad - \mathbf{X}_{3,1}(\mathbf{X}_{2,0}^\top \mathbf{X}_{2,0} + n_2\tilde{\lambda}\mathbf{I})^{-1}\mathbf{X}_{2,0}^\top \boldsymbol{\varepsilon}_{2,0}, \\ \mathsf{B}_{\mathrm{ms}} &:= n_2\tilde{\lambda}\mathbf{X}_{3,1}(\mathbf{X}_{2,0}^\top \mathbf{X}_{2,0} + n_2\tilde{\lambda}\mathbf{I})^{-1}\theta_0^\star \\ &\quad - n_2\tilde{\lambda}\mathbf{X}_{3,0}(\mathbf{X}_{2,1}^\top \mathbf{X}_{2,1} + n_2\tilde{\lambda}\mathbf{I})^{-1}\theta_1^\star. \end{aligned}$$

Good Event. Apply Lemma H.2 with n_j observations on $\mathcal{D}_j$, for $j = 2, 3$, at $\lambda' = \tilde{\lambda}$. Since $n_2 \asymp n_3 \asymp n$, a sufficiently large constant in $\tilde{\lambda} \asymp \log n / n_2$ gives the following bounds with probability at least $1 - n^{-11}$ for each comparison, for some constant $c_2 > 1$:

$$\begin{aligned}&\frac{1}{c_2}(n_2\boldsymbol{\Sigma}_{\mathcal{F},1} + n_2\lambda'\mathbf{I}) \preceq \mathbf{X}_{2,1}^\top \mathbf{X}_{2,1} + n_2\lambda'\mathbf{I} \preceq c_2(n_2\boldsymbol{\Sigma}_{\mathcal{F},1} + n_2\lambda'\mathbf{I}), \\ &\frac{1}{c_2}(n_2\boldsymbol{\Sigma}_{\mathcal{F},0} + n_2\lambda'\mathbf{I}) \preceq \mathbf{X}_{2,0}^\top \mathbf{X}_{2,0} + n_2\lambda'\mathbf{I} \preceq c_2(n_2\boldsymbol{\Sigma}_{\mathcal{F},0} + n_2\lambda'\mathbf{I}), \\ &\frac{1}{c_2}(n_2\boldsymbol{\Sigma}_{\mathcal{F}} + n_2\lambda'\mathbf{I}) \preceq \mathbf{X}_2^\top \mathbf{X}_2 + n_2\lambda'\mathbf{I} \preceq c_2(n_2\boldsymbol{\Sigma}_{\mathcal{F}} + n_2\lambda'\mathbf{I}). \end{aligned} \tag{14}$$

$$\begin{aligned}&\frac{1}{c_2}(n_3\boldsymbol{\Sigma}_{\mathcal{F},1} + n_3\lambda'\mathbf{I}) \preceq \mathbf{X}_{3,1}^\top \mathbf{X}_{3,1} + n_3\lambda'\mathbf{I} \preceq c_2(n_3\boldsymbol{\Sigma}_{\mathcal{F},1} + n_3\lambda'\mathbf{I}), \\ &\frac{1}{c_2}(n_3\boldsymbol{\Sigma}_{\mathcal{F},0} + n_3\lambda'\mathbf{I}) \preceq \mathbf{X}_{3,0}^\top \mathbf{X}_{3,0} + n_3\lambda'\mathbf{I} \preceq c_2(n_3\boldsymbol{\Sigma}_{\mathcal{F},0} + n_3\lambda'\mathbf{I}), \\ &\frac{1}{c_2}(n_3\boldsymbol{\Sigma}_{\mathcal{F}} + n_3\lambda'\mathbf{I}) \preceq \mathbf{X}_3^\top \mathbf{X}_3 + n_3\lambda'\mathbf{I} \preceq c_2(n_3\boldsymbol{\Sigma}_{\mathcal{F}} + n_3\lambda'\mathbf{I}). \end{aligned} \tag{15}$$

Let $\mathscr{E}_{\mathrm{ms}}$ be the event on which these inequalities hold. Then

$$\mathbb{P}[\mathscr{E}_{\mathrm{ms}}] \ge 1 - 6n^{-11}.$$

Combining Equations (14) and (15) with the operator-overlap assumption, on $\mathscr{E}_{\mathrm{ms}}$ we obtain, for $a \in \{0, 1\}$,

$$\begin{aligned}\frac{\kappa n_3}{c n_2}(\mathbf{X}_{2,a}^\top \mathbf{X}_{2,a} + n_2\lambda'\mathbf{I}) &\preceq \mathbf{X}_{3,1-a}^\top \mathbf{X}_{3,1-a} + n_3\lambda'\mathbf{I} \\ &\preceq \frac{c n_3}{\kappa n_2}(\mathbf{X}_{2,a}^\top \mathbf{X}_{2,a} + n_2\lambda'\mathbf{I}), \end{aligned} \tag{16}$$

for some constant $c > 1$.

To apply Lemma G.2, it suffices to bound

$$\|\mathbf{M}_3 - \mathbb{E}[\mathbf{M}_3]\|_{\psi_2} = \|\mathsf{V}_{\mathrm{ms}}\|_{\psi_2}, \qquad \|\mathbb{E}[\mathbf{M}_3] - \mathbf{M}_3^\star\|_2 = \|\mathsf{B}_{\mathrm{ms}}\|_2.$$

Bounding $\|\mathsf{V}_{\mathrm{ms}}\|_{\psi_2}$. Observe that

$$\begin{aligned}\|\mathsf{V}_{\mathrm{ms}}\|_{\psi_2} \lesssim 1 &+ \|\mathbf{X}_{3,0}(\mathbf{X}_{2,1}^\top \mathbf{X}_{2,1} + n_2\tilde{\lambda}\mathbf{I})^{-1}\mathbf{X}_{2,1}^\top\|_{\mathrm{op}} \\ &+ \|\mathbf{X}_{3,1}(\mathbf{X}_{2,0}^\top \mathbf{X}_{2,0} + n_2\tilde{\lambda}\mathbf{I})^{-1}\mathbf{X}_{2,0}^\top\|_{\mathrm{op}}.\end{aligned}$$

We bound one term; the other is analogous:

$$\begin{aligned}&\|\mathbf{X}_{3,0}(\mathbf{X}_{2,1}^\top \mathbf{X}_{2,1} + n_2\tilde{\lambda}\mathbf{I})^{-1}\mathbf{X}_{2,1}^\top\|_{\mathrm{op}}^2 \\ &= \|\mathbf{X}_{2,1}(\mathbf{X}_{2,1}^\top \mathbf{X}_{2,1} + n_2\tilde{\lambda}\mathbf{I})^{-1}\mathbf{X}_{3,0}^\top\mathbf{X}_{3,0}(\mathbf{X}_{2,1}^\top \mathbf{X}_{2,1} + n_2\tilde{\lambda}\mathbf{I})^{-1}\mathbf{X}_{2,1}^\top\|_{\mathrm{op}} \\ &\overset{(i)}{\lesssim} \frac{n_3}{\kappa n_2}\|\mathbf{X}_{2,1}(\mathbf{X}_{2,1}^\top \mathbf{X}_{2,1} + n_2\tilde{\lambda}\mathbf{I})^{-1}\mathbf{X}_{2,1}^\top\|_{\mathrm{op}} \\ &\leq \frac{n_3}{\kappa n_2},\end{aligned}$$

where step (i) follows from Equation (16). Hence

$$\|\mathsf{V}_{\mathrm{ms}}\|_{\psi_2} \lesssim 1 + \sqrt{\frac{n_3}{\kappa n_2}}.$$

Bounding $\|\mathsf{B}_{\mathrm{ms}}\|_2$. We have

$$\begin{aligned}\|\mathsf{B}_{\mathrm{ms}}\|_2 \leq\ & n_2\tilde{\lambda}\|\mathbf{X}_{3,1}(\mathbf{X}_{2,0}^\top \mathbf{X}_{2,0} + n_2\tilde{\lambda}\mathbf{I})^{-1}\theta_0^\star\|_2 \\ &+ n_2\tilde{\lambda}\|\mathbf{X}_{3,0}(\mathbf{X}_{2,1}^\top \mathbf{X}_{2,1} + n_2\tilde{\lambda}\mathbf{I})^{-1}\theta_1^\star\|_2.\end{aligned}$$

We bound the first term as

$$\begin{aligned}&n_2\tilde{\lambda}\|\mathbf{X}_{3,1}(\mathbf{X}_{2,0}^\top \mathbf{X}_{2,0} + n_2\tilde{\lambda}\mathbf{I})^{-1}\theta_0^\star\|_2 \\ &\leq n_2\tilde{\lambda}\|\mathbf{X}_{3,1}(\mathbf{X}_{2,0}^\top \mathbf{X}_{2,0} + n_2\tilde{\lambda}\mathbf{I})^{-1/2}\|_{\mathrm{op}}\|(\mathbf{X}_{2,0}^\top \mathbf{X}_{2,0} + n_2\tilde{\lambda}\mathbf{I})^{-1/2}\theta_0^\star\|_{\mathbb{F}} \\ &\overset{(i)}{\lesssim} n_2\tilde{\lambda}\sqrt{\frac{n_3}{\kappa n_2}}\frac{\|\theta_0^\star\|_{\mathbb{F}}}{\sqrt{n_2\tilde{\lambda}}} \lesssim \sqrt{\frac{n_3\tilde{\lambda}}{\kappa}},\end{aligned}$$

where step (i) follows from Equation (16) and the last inequality uses the bounded nuisance RKHS norm. The other term satisfies the same bound, so

$$\frac{\|\mathsf{B}_{\mathrm{ms}}\|_2^2}{n_3} \lesssim \frac{\tilde{\lambda}}{\kappa} \asymp \frac{\log n}{n_2\kappa}.$$

Applying Lemma G.2 to the n_3 validation points with $\gamma = 1$, the remainder is bounded by

$$\frac{\|\mathsf{B}_{\mathrm{ms}}\|_2^2}{n_3} + \frac{\|\mathsf{V}_{\mathrm{ms}}\|_{\psi_2}^2 \log(L/\delta)}{n_3} \lesssim \frac{\log n}{n_2\kappa} + \left(\frac{1}{n_3} + \frac{1}{n_2\kappa}\right)\log(L/\delta).$$

This proves Lemma G.1.

Lemma G.2 (Theorem 5.2 from Wang 2026)**.** *Let $\{\boldsymbol{z}_i\}_{i=1}^n$ be deterministic elements in a set $\mathcal{Z}$, let $g^\star$ and $\{g_j\}_{j=1}^m$ be deterministic functions on $\mathcal{Z}$, and let $\widetilde{g}$ be a random function on $\mathcal{Z}$. Define*

$$\mathcal{L}(g) = \frac{1}{n}\sum_{i=1}^n \left|g\left(\boldsymbol{z}_i\right) - g^\star\left(\boldsymbol{z}_i\right)\right|^2$$

for any function g on $\mathcal{Z}$. Assume that the random vector $\widetilde{\boldsymbol{y}} = \left(\widetilde{g}\left(\boldsymbol{z}_1\right), \widetilde{g}\left(\boldsymbol{z}_2\right), \cdots, \widetilde{g}\left(\boldsymbol{z}_n\right)\right)^\top$ satisfies $\|\widetilde{\boldsymbol{y}} - \mathbb{E}\widetilde{\boldsymbol{y}}\|_{\psi_2} \leq V < \infty$. Choose any

$$\widehat{j} \in \operatorname*{argmin}_{j\in[m]}\left\{\frac{1}{n}\sum_{i=1}^n \left|g_j\left(\boldsymbol{z}_i\right) - \widetilde{g}\left(\boldsymbol{z}_i\right)\right|^2\right\}.$$

There exists a universal constant C such that for any $\delta \in (0, 1]$, with probability at least $1 - \delta$ we have

$$\mathcal{L}\left(g_{\widehat{j}}\right) \leq \inf_{\gamma>0}\left\{(1+\gamma)\min_{j\in[m]}\mathcal{L}\left(g_j\right) + C\left(1+\gamma^{-1}\right)\left(\mathcal{L}(\mathbb{E}\widetilde{g}) + \frac{V^2\log(m/\delta)}{n}\right)\right\}.$$

Consequently,

$$\mathbb{E}\mathcal{L}\left(g_{\widehat{j}}\right) \leq \inf_{\gamma>0}\left\{(1+\gamma)\min_{j\in[m]}\mathcal{L}\left(g_j\right) + C\left(1+\gamma^{-1}\right)\left(\mathcal{L}(\mathbb{E}\widetilde{g}) + \frac{V^2(1+\log m)}{n}\right)\right\}.$$

H Technical Lemmas

Lemma H.1 (Key Inequality for Point Evaluation: Sobolev Space)**.** *Let K be a Sobolev kernel on $H^\beta(\mathcal{X})$ with feature map $\phi(\cdot)$, and let $\mathbf{\Sigma} = \int_{x\in\mathcal{X}} \phi(x)\phi(x)^\top \mathrm{d}x$. Suppose $\mathcal{X} \subset \mathbb{R}^d$ and $\beta > d/2$. For any $x_0 \in \mathcal{X}$ and any $\lambda > 0$, we have*

$$\phi(x_0)^\top(\mathbf{\Sigma} + \lambda\mathbf{I})^{-1}\phi(x_0) \lesssim \lambda^{-\frac{d}{2\beta}}.$$

Hence,

$$\phi(x_0)\phi(x_0)^\top \lesssim \lambda^{-\frac{d}{2\beta}}(\mathbf{\Sigma} + \lambda\mathbf{I}).$$

Proof. Let $Q = \phi(x_0)^\top(\mathbf{\Sigma} + \lambda\mathbf{I})^{-1}\phi(x_0)$. Our goal is to find an upper bound for Q.

Let $\theta \in \mathbb{H}$ be defined as $\theta = (\mathbf{\Sigma} + \lambda\mathbf{I})^{-1}\phi(x_0)$. The quantity Q can be written as an inner product in the RKHS $\mathbb{H}$:

$$Q = \left\langle \phi(x_0), (\mathbf{\Sigma} + \lambda\mathbf{I})^{-1}\phi(x_0)\right\rangle_{\mathbb{H}} = \langle \phi(x_0), \theta\rangle_{\mathbb{H}}.$$

From the definition of θ, we have $(\mathbf{\Sigma} + \lambda\mathbf{I})\theta = \phi(x_0)$. Taking the inner product with θ on both sides gives:

$$\langle(\mathbf{\Sigma} + \lambda\mathbf{I})\theta, \theta\rangle_{\mathbb{H}} = \langle \phi(x_0), \theta\rangle_{\mathbb{H}}.$$

This simplifies to:

$$\langle\mathbf{\Sigma}\theta, \theta\rangle_{\mathbb{H}} + \lambda\langle\theta, \theta\rangle_{\mathbb{H}} = Q.$$

Let $g_\theta(x) = \langle\phi(x), \theta\rangle_{\mathbb{H}}$ be the function in the Sobolev space $H^\beta(\mathcal{X})$ corresponding to $\theta \in \mathbb{H}$. By the reproducing property, $g_\theta(x_0) = \langle\phi(x_0), \theta\rangle_{\mathbb{H}} = Q$.

The norms of g_θ can be related to the expression for Q:

1. The squared H^β norm is: $\|g_\theta\|^2_{H^\beta(\mathcal{X})} = \|\theta\|^2_{\mathbb{H}} = \langle\theta, \theta\rangle_{\mathbb{H}}$.
2. The squared L^2 norm is: $\|g_\theta\|^2_{L^2(\mathcal{X})} = \int_{\mathcal{X}} |g_\theta(x)|^2 dx = \int_{\mathcal{X}} \langle\phi(x), \theta\rangle^2_{\mathbb{H}}\, dx$
   $= \left\langle \left(\int_{\mathcal{X}} \phi(x)\phi(x)^\top dx\right)\theta, \theta\right\rangle_{\mathbb{H}} = \langle\mathbf{\Sigma}\theta, \theta\rangle_{\mathbb{H}}$.

Substituting these into the expression for Q, we get:

$$\|g_\theta\|^2_{L^2(\mathcal{X})} + \lambda\|g_\theta\|^2_{H^\beta(\mathcal{X})} = Q.$$

Since both norm terms are non-negative, we can establish bounds for each:

$$\|g_\theta\|^2_{L^2(\mathcal{X})} \le Q \implies \|g_\theta\|_{L^2(\mathcal{X})} \le \sqrt{Q}$$

$$\lambda\|g_\theta\|^2_{H^\beta(\mathcal{X})} \le Q \implies \|g_\theta\|_{H^\beta(\mathcal{X})} \le \sqrt{\frac{Q}{\lambda}}$$

Now apply the interpolation inequality to g_θ. Let $q = \frac{d}{2\beta}$.

$$Q = g_\theta(x_0) \le \|g_\theta\|_{L^\infty(\mathcal{X})} \le C\|g_\theta\|^{1-q}_{L^2(\mathcal{X})}\|g_\theta\|^q_{H^\beta(\mathcal{X})}.$$

Substitute the bounds we found for the norms:

$$Q \le C\left(\sqrt{Q}\right)^{1-q}\left(\sqrt{\frac{Q}{\lambda}}\right)^q$$

and hence

$$Q^{\frac{1}{2}} \le C \cdot \lambda^{-\frac{q}{2}}$$

which completes the first inequality.

$$\phi(x_0)^\top(\mathbf{\Sigma} + \lambda\mathbf{I})^{-1}\phi(x_0) \lesssim \lambda^{-\frac{d}{2\beta}}.$$

For the second inequality, let $v \in \mathbb{H}$. Then

$$\langle\phi(x_0)\phi(x_0)^\top v, v\rangle_{\mathbb{H}} = |\langle\phi(x_0), v\rangle_{\mathbb{H}}|^2$$

$$
\begin{aligned}
&= \left| \left\langle (\mathbf{\Sigma} + \lambda \mathbf{I})^{-1/2} \phi(x_0), (\mathbf{\Sigma} + \lambda \mathbf{I})^{1/2} v \right\rangle_{\mathbb{H}} \right|^2 \\
&\leq \phi(x_0)^\top (\mathbf{\Sigma} + \lambda \mathbf{I})^{-1} \phi(x_0) \cdot \langle (\mathbf{\Sigma} + \lambda \mathbf{I}) v, v \rangle_{\mathbb{H}} \\
&\lesssim \lambda^{-d/(2\beta)} \langle (\mathbf{\Sigma} + \lambda \mathbf{I}) v, v \rangle_{\mathbb{H}}.
\end{aligned}
$$

Since this holds for all v, we conclude

$$
\phi(x_0)\phi(x_0)^\top \lesssim \lambda^{-d/(2\beta)} (\mathbf{\Sigma} + \lambda \mathbf{I}).
$$

□

Lemma H.2 (Corollary E.1 from Wang 2026)**.** *Let $\{\boldsymbol{x}_i\}_{i=1}^n$ be i.i.d. random elements in a separable Hilbert space $\mathbb{H}$ with $\mathbf{\Sigma} := \mathbb{E}(\boldsymbol{x}_i \otimes \boldsymbol{x}_i)$ trace class. Define $\hat{\mathbf{\Sigma}} = \frac{1}{n} \sum_{i=1}^n \boldsymbol{x}_i \otimes \boldsymbol{x}_i$. Choose any constant $\gamma \in (0,1)$ and define the event $\mathcal{A} = \{(1-\gamma)(\mathbf{\Sigma} + \lambda \boldsymbol{I}) \preceq \hat{\mathbf{\Sigma}} + \lambda \boldsymbol{I} \preceq (1+\gamma)(\mathbf{\Sigma} + \lambda \boldsymbol{I})\}$. 1. If $\|\boldsymbol{x}_i\|_{\mathbb{H}} \leq \xi$ holds almost surely for some constant ξ, then there exists a constant $C \geq 1$ determined by γ such that $\mathbb{P}(\mathcal{A}) \geq 1 - \delta$ holds so long as $\delta \in (0, 1/14]$ and $\lambda \geq \frac{C\xi \log(n/\delta)}{n}$. Hence, with probability at least $1 - \delta$, we have*

$$
\frac{1}{C\xi} \left(\widehat{\mathbf{\Sigma}} + \frac{\log(n/\delta)}{n} \mathbf{I} \right) \preceq \mathbf{\Sigma} + \frac{\log(n/\delta)}{n} \mathbf{I} \preceq C\xi \left(\widehat{\mathbf{\Sigma}} + \frac{\log(n/\delta)}{n} \mathbf{I} \right).
$$

Remark H.3 (A useful observation)**.** For any two trace-class operators $\mathbf{S}_1, \mathbf{S}_2$, suppose that

$$
\mathbf{S}_1 \preceq c_1 (\mathbf{S}_2 + c_2 \log n \mathbf{I})
$$

for some absolute constant $c_1 > 1, c_2 > 0$. Then, for any $c_3 > 0$, for $c = \max(1, c_2/c_3)$, we have

$$
\mathbf{S}_1 \preceq c(\mathbf{S}_2 + c_3 \log n \mathbf{I}).
$$

Lemma H.4 (Lemma E.1 from Wang 2026)**.** *Suppose that $\boldsymbol{x} \in \mathbb{R}^d$ is a zero-mean random vector with $\|\boldsymbol{x}\|_{\psi_2} \leq 1$. There exists a universal constant $C > 0$ such that for any symmetric and positive semi-definite matrix $\mathbf{\Sigma} \in \mathbb{R}^{d \times d}$,*

$$
\mathbb{P}\left(\boldsymbol{x}^\top \mathbf{\Sigma} \boldsymbol{x} \leq C \operatorname{Tr}(\mathbf{\Sigma}) t \right) \geq 1 - e^{-r(\mathbf{\Sigma})t}, \quad \forall t \geq 1.
$$

Here $r(\mathbf{\Sigma}) = \operatorname{Tr}(\mathbf{\Sigma}) / \|\mathbf{\Sigma}\|_2$ is the effective rank of $\mathbf{\Sigma}$.

I Nuisance Misspecification

We extend Theorem D.1 by allowing approximate outcome models. Write

$$f_a^\star = q_a^\star + r_a, \qquad q_a^\star \in \mathcal{F}, \quad a \in \{0,1\},$$

where $q_a^\star$ has bounded RKHS norm and r_a is the approximation residual. The actual contrast $h^\star = f_1^\star - f_0^\star$ retains its assumed structure under Model 1, 2, or 3; we do not require $q_1^\star - q_0^\star = h^\star$.

I.1 Rates under Models 1–3

Corollary I.1 (Bounded nuisance misspecification)**.** *Suppose* $\max_{a\in\{0,1\}} \|r_a\|_{L^\infty(\mathcal{P}_x)} \le \varepsilon$*. Under the remaining assumptions and regularization choices of Corollaries 7–9, the squared* $L^2(\mathcal{P}_x)$*-error satisfies, with probability at least* $1 - n^{-10}$*,*

$$\mathcal{E}_{L^2}(\hat{h}) \lesssim \frac{\varepsilon^2}{\kappa} + \begin{cases} (n\kappa)^{-\frac{2\ell_{\mathcal{H}}}{1+2\ell_{\mathcal{H}}}}, & \textit{Model 1}, \\ (n\kappa)^{-\frac{2\ell_{\mathcal{F}}(1+\nu)}{1+2\ell_{\mathcal{F}}(1+\nu)}}, & \textit{Model 2}, \\ (n\kappa)^{-\frac{2\tilde{\ell}}{1+2\tilde{\ell}}}, & \textit{Model 3}. \end{cases} \tag{17}$$

Here the spectral exponents are as in the corresponding corollaries, and logarithmic factors are suppressed. For Model 1 with exponential decay or rank at most D*, the first term inside the cases is replaced by* $(n\kappa)^{-1}$ *or* $D/(n\kappa)$*, respectively.*

Proof. For L^2-error, $\|\mathbf{S}_\lambda\|_{\mathrm{op}} \le 1$, while $n^{-1}(\|\boldsymbol{r}_0\|_2^2 + \|\boldsymbol{r}_1\|_2^2) \le 2\varepsilon^2$. Proposition I.2 below therefore contributes at most ε^2/κ, up to constants. The remaining terms follow the calculations in Appendices F.1–F.3, including the sharper source-condition bias bound for Model 2. □

Thus a well-specified rate $(n\kappa)^{-\alpha}$ is preserved whenever $\varepsilon^2 \lesssim \kappa(n\kappa)^{-\alpha}$.

I.2 General bound and proof

We use the notation of Appendices D and E, with $h^\star(x) = \langle w(x), \eta^\star\rangle_{\mathbb{X}}$ and $\mathbf{\Sigma}_{\mathrm{ref}} = \mathbf{\Sigma}$ for L^2-error, and keep the kernels and regularizers fixed in the following decomposition.

Let $\theta_{a,q}^\star \in \mathbb{F}$ represent $q_a^\star$. Define

$$\boldsymbol{r}_a = (r_a(x_i))_{i\in[n]}, \quad \mathbf{D}_a = \mathrm{diag}\big(\mathbf{1}(a_i = a)\big)_{i\in[n]}, \quad \mathbf{A}_a = (\mathbf{X}_a^\top \mathbf{X}_a + n\bar{\lambda}\mathbf{I})^{-1},$$

and let $\mathbf{A}_\lambda = (\mathbf{W}^\top\mathbf{W} + n\lambda\mathbf{I})^{-1}$. The residual contribution to the pseudo-outcomes is

$$\boldsymbol{d} = \mathbf{D}_1\boldsymbol{r}_0 - \mathbf{D}_0\boldsymbol{r}_1 + \mathbf{X}_0\mathbf{A}_1\mathbf{X}_1^\top\boldsymbol{r}_1 - \mathbf{X}_1\mathbf{A}_0\mathbf{X}_0^\top\boldsymbol{r}_0. \tag{18}$$

Write the bound from Theorem D.1, with $q_a^\star$ in the nuisance norms, as

$$\mathcal{R}_{n,\kappa}(q) := \frac{1}{n\kappa}\mathrm{Tr}(\mathbf{S}_\lambda) + \lambda^2\|\mathbf{S}_\lambda\|_{\mathrm{op}}\|(\mathbf{\Sigma}+\lambda\mathbf{I})^{-1/2}\eta^\star\|_{\mathbb{X}}^2 + \frac{\|\mathbf{S}_\lambda\|_{\mathrm{op}}}{n\kappa}\big(\|q_0^\star\|_{\mathcal{F}}^2 + \|q_1^\star\|_{\mathcal{F}}^2\big).$$

Proposition I.2 (Approximate outcome models)**.** *Under the conditions of Theorem D.1, with nuisance well-specification replaced by the decomposition above, the estimator satisfies, with probability at least* $1 - n^{-10}$*,*

$$\mathcal{E}_{\mathrm{ref}}(\hat{\eta}) \lesssim \mathcal{R}_{n,\kappa}(q) + \Delta_{\mathrm{mis}}, \qquad \Delta_{\mathrm{mis}} := \|\mathbf{\Sigma}_{\mathrm{ref}}^{1/2}\mathbf{A}_\lambda\mathbf{W}^\top\boldsymbol{d}\|_{\mathbb{X}}^2, \tag{19}$$

where logarithmic factors are suppressed as in the general bound. On the same event,

$$\Delta_{\mathrm{mis}} \lesssim \frac{\|\mathbf{S}_\lambda\|_{\mathrm{op}}}{n}\|\boldsymbol{d}\|_2^2 \lesssim \frac{\|\mathbf{S}_\lambda\|_{\mathrm{op}}}{\kappa}\left(\frac{\|\boldsymbol{r}_0\|_2^2 + \|\boldsymbol{r}_1\|_2^2}{n}\right). \tag{20}$$

Proof. We follow the error decomposition in Appendix E. The first-stage estimator satisfies

$$\begin{aligned}\hat{\theta}_a - \theta^\star_{a,q} &= \mathbf{A}_a\big(\mathbf{X}_a^\top \boldsymbol{\varepsilon}_a - n\bar{\lambda}\theta^\star_{a,q}\big) + \mathbf{A}_a\mathbf{X}_a^\top \boldsymbol{r}_a \\ &=: (\hat{\theta}_{a,q} - \theta^\star_{a,q}) + \mathbf{A}_a\mathbf{X}_a^\top \boldsymbol{r}_a.\end{aligned}$$

Thus $\hat{\theta}_{a,q}$ is the KRR fit to the responses $q_a^\star(x_i) + \varepsilon_i$ in arm a. Substituting into the switch-imputation formula gives

$$\mathbf{M} = \mathbf{W}\eta^\star + \boldsymbol{\varepsilon}' + \mathbf{X}_0(\hat{\theta}_{1,q} - \theta^\star_{1,q}) - \mathbf{X}_1(\hat{\theta}_{0,q} - \theta^\star_{0,q}) + \boldsymbol{d}.$$

Consequently,

$$\begin{aligned}\hat{\eta} - \eta^\star = \mathbf{A}_\lambda\Big[\mathbf{W}^\top\big\{\boldsymbol{\varepsilon}' + \mathbf{X}_0(\hat{\theta}_{1,q} - \theta^\star_{1,q}) - \mathbf{X}_1(\hat{\theta}_{0,q} - \theta^\star_{0,q})\big\} - n\lambda\eta^\star\Big] \\ + \mathbf{A}_\lambda\mathbf{W}^\top \boldsymbol{d}.\end{aligned}$$

The first line has exactly the nuisance-bias, centered-noise, and target-ridge terms bounded in the proof of Theorem D.1, with $\theta^\star_{a,q}$ in place of $\theta^\star_a$. Applying those bounds and the squared triangle inequality proves (19). This argument only uses the representation of the actual contrast in $\mathbb{X}$, not the identity $q_1^\star - q_0^\star = h^\star$.

To bound the new term, work on the same good event $\mathscr{E}_0$. By (11),

$$\|\boldsymbol{\Sigma}_{\mathrm{ref}}^{1/2}\mathbf{A}_\lambda\mathbf{W}^\top\|_{\mathrm{op}}^2 \le \|\boldsymbol{\Sigma}_{\mathrm{ref}}^{1/2}\mathbf{A}_\lambda\boldsymbol{\Sigma}_{\mathrm{ref}}^{1/2}\|_{\mathrm{op}} \lesssim \frac{\|\mathbf{S}_\lambda\|_{\mathrm{op}}}{n}.$$

Moreover, (12) gives, for $a \in \{0,1\}$,

$$\|\mathbf{X}_{1-a}\mathbf{A}_a\mathbf{X}_a^\top\|_{\mathrm{op}} \le \|\mathbf{X}_{1-a}\mathbf{A}_a^{1/2}\|_{\mathrm{op}}\|\mathbf{A}_a^{1/2}\mathbf{X}_a^\top\|_{\mathrm{op}} \lesssim \kappa^{-1/2}.$$

Since $\|\mathbf{D}_a\|_{\mathrm{op}} \le 1$ and $\kappa \le 1$, equation (18) implies $\|\boldsymbol{d}\|_2^2 \lesssim \kappa^{-1}(\|\boldsymbol{r}_0\|_2^2 + \|\boldsymbol{r}_1\|_2^2)$. Combining these inequalities proves (20). □

L^2 control of the residuals. More generally, (20) uses only the empirical residual energy $\widehat{\varepsilon}_n^2 := n^{-1}\sum_i\{r_0(x_i)^2 + r_1(x_i)^2\}$. If the fixed residual functions satisfy $\|r_0\|_{L^2(\mathcal{P}_x)}^2 + \|r_1\|_{L^2(\mathcal{P}_x)}^2 \le \varepsilon^2$, Markov's inequality gives $\widehat{\varepsilon}_n^2 \le \varepsilon^2/\delta$ with probability at least $1-\delta$. Thus (17) holds with the additional term $\varepsilon^2/(\kappa\delta)$ and probability at least $1 - n^{-10} - \delta$. The uniform residual bound in Corollary I.1 avoids this extra factor.

J Limitations

Our theory concerns binary treatments and outcome-based imputation, without double robustness. Combining undersmoothed outcome regression with a DR estimator is left to future work. Approximate outcome models introduce an additional error term, quantified and proved in Appendix I; the contrast structure itself is still assumed to be correctly specified. The low-dimensional model assumes a supplied representation; learning an unknown projection remains open. Our model-selection guarantee is limited to L^2 loss and does not establish pointwise adaptation over unknown contrast classes. The selection guarantee also requires a valid upper bound B on $\|h^\star\|_\infty$. Our experiments use a fixed nuisance regularization constant; we do not study noise-adaptive tuning. We leave extensions to noise with only finite moments to future work.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state the main contributions and scope, and these are reflected in the theory and experiments.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Appendix J discusses the limitations; Appendices B and I give details on adaptation and nuisance misspecification.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: The assumptions are stated in the setup and theorem statements, and full proofs are provided in the appendix.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper specifies the data-generating processes, baselines, kernels, tuning choices, splits, and evaluation metrics used in the experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

Justification: We include an anonymized supplementary package with the experiment code, the RealCause CSV files used in the paper, and instructions to reproduce the main results. Synthetic data are generated by the provided scripts, and Appendix C documents the evaluation setup.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Appendix C describes the train/test protocol, hyperparameters, regularization grids, kernel choices, and benchmark details.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The experimental tables report standard errors of the mean over repeated runs or dataset realizations, and the number of repetitions is stated.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   Justification: Appendix C.3 reports the Nyström implementation and CPU platform; the saved per-file results contain elapsed times, and the supplied manifests record software versions and thread settings. We did not log peak memory or a complete hardware resource profile for every experiment.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The work is methodological and uses standard synthetic and semi-synthetic benchmarks, and we are not aware of any deviation from the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: There is no societal impact of the work performed.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release high-risk models or scraped datasets.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Appendix C cites the RealCause benchmark, states that the official repository is MIT-licensed, and explains that the anonymized supplementary package includes the specific RealCause CSV files used in the experiments.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [N/A]

Justification: The submission does not introduce or release new datasets, models, or code assets.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: The paper does not involve crowdsourcing or research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: The paper does not involve crowdsourcing or research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: LLMs are not part of the core methodology or experiments in this work.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.