Biometrics 64, 1–81 DOI: 10.1111/j.1541-0420.2005.00454.x
May 2025

Supplementary materials for "Proximal Causal Inference for Modified Treatment Policies" by Antonio Olivas-Martinez, Peter B. Gilbert, and Andrea Rotnitzky

Antonio Olivas-Martinez[1], Peter Gilbert[1,2,3], and Andrea Rotnitzky[1]

[1]Department of Biostatistics, University of Washington, Seattle, WA, U.S.A.

[2]Vaccine and Infectious Disease Division, Fred Hutchinson Cancer Center, Seattle, WA, U.S.A

[3]Public Health Science Division, Fred Hutchinson Cancer Center, Seattle, WA, U.S.A.

Summary:

This paper has been submitted for consideration for publication in *Biometrics*

Web Appendix A. Proximal Identification and Estimation in a potentially restricted population

Here, we outline our methodology for estimating the counterfactual MTP mean within a potentially restricted population.

Recall that $O = (L, Z, W, X, Y)$ is the vector of observed data randomly sampled from a distribution P_0, with L representing observed confounders, Z and W denoting the negative control treatment and outcome, and X and Y corresponding to the exposure and outcome of interest. For a given function $q : \text{supp}\,(X, L) \to \mathbb{R}$, let $\mathcal{S}$ be a subset of $\text{supp}\,(X, L)$ such that $P\{(X, L) \in \mathcal{S}\} > 0$ and that for every $(x, l) \in \mathcal{S}$, $(q(x, l), l)$ belongs to $\text{supp}(X, L)$. The target parameter is

$$\psi_0 = \mathbb{E}\left[Y\left\{q\left(X, L\right)\right\} | (X, L) \in \mathcal{S}\right], \tag{1}$$

which represents the counterfactual MTP mean within the population $\{O : (X, L) \in \mathcal{S}\}$.

The target estimand in Example 1 of the manuscript, i.e., $\mathbb{E}\{Y(X + \delta)|(X, L) \in \mathcal{S}\}$ with $\mathcal{S} = \{(x, l) : l \in \text{supp}(L), x \in [c(l), d(l) - \delta]\}$, has the form of expression (1). Notably, $\mathcal{S}$ is the largest subset of $\text{supp}(X, L)$ satisfying $q\left(\mathcal{S}\right) \subset \text{supp}\,(X, L)$. Therefore, the subpopulation $\{O : (X, L) \in \mathcal{S}\}$ is the largest population for which the counterfactual MTP mean, under the shift intervention $q(x, l) = x + \delta$, can be estimated using observed data. However, subpopulations $\{O : X \in \mathcal{S}'\}$, where $\mathcal{S}'$ is a strict subset of $\mathcal{S}$, may also be of interest, particularly for comparing the counterfactual MTP mean across different policies within a common population $\{O : (X, L) \in \mathcal{S}'\}$.

The estimand considered in Example 2 also follows the structure of (1), with $\mathcal{S} = \text{supp}\,(X, L)$. More generally, the target estimands discussed in the main manuscript have the form in (1) with $\mathcal{S} = \text{supp}\,(X, L)$.

For completeness, we restate the assumptions and theorems introduced in the main manuscript, adjusting them to apply to the restricted population $\{O : (X, L) \in \mathcal{S}\}$, where appropriate.

A.1 *Proximal Identifying Assumptions*

For each $l \in \text{supp}\,(L)$, define $\mathcal{S}(l) := \{x \in \text{supp}\,(X) : (x, l) \in \mathcal{S}\}$.

ASSUMPTION 1 (Consistency): $Y(X) = Y$.

ASSUMPTION 2 (Randomization of policy related populations): For all $(x, l, u) \in \text{supp}(X, L, U)$ such that $x \in \mathcal{S}(l)$ and $x' = q(x, l)$, the conditional distribution of $Y(x')|X = x, L = l, U = u$ is equal to the conditional distribution of $Y(x')|X = x', L = l, U = u$.

ASSUMPTION 3 (Common support): For all $(l, u) \in \text{supp}\,(L, U)$, $\text{supp}\,(X|L = l, U = u) \equiv \text{supp}\,(X|L = l)$.

ASSUMPTION 4 (Negative control treatment): $Z \perp Y|X, L, U$.

ASSUMPTION 5 (Negative control outcome): $(X, Z) \perp W|U, L$.

ASSUMPTION 6 (Monotone and smooth policy): For each $l \in \text{supp}\,(L)$, the map $q(\cdot, l) : \mathcal{S}(l) \to \text{supp}\,(X)$ is strictly monotone and differentiable almost everywhere with respect to the distribution of X given $L = l$.

For each $l \in \text{supp}\,(L)$, given that $q\,(\mathcal{S}(l), l) \subset \text{supp}\,(X|L = l)$, assumption 6 ensures that the policy map $q : x \mapsto q(x, l)$ has an inverse $q^{-1} : q\,(\mathcal{S}(l), l) \to \mathcal{S}(l)$ that is well-defined and almost everywhere differentiable with respect to the distribution of X given $L = l$.

ASSUMPTION 7 (Outcome bridge): There exists $h_0 \in \mathcal{L}^2(X, L, W)$ solving for all $(x, l, u) \in \text{supp}\,(X, L, U)$ such that $x \in \mathcal{S}(l) \cup q\,(\mathcal{S}(l), l)$, the integral equation

$$\mathbb{E}\,(Y|X = x, L = l, U = u) = \int h\,(x, l, w)\, p_{W|X,L,U}\,(w|x, l, u)\, dw. \tag{2}$$

ASSUMPTION 8 (Treatment bridge): There exists $g_0 \in \mathcal{L}^2(X, L, Z)$ such that for all $(x, l, u) \in$

$\text{supp}\,(X, L, U)$ satisfying that $x \in \mathcal{S}(l) \cup q\,(\mathcal{S}(l), l)$ and that the map $x' \mapsto q(x', l)$ is differentiable at $x' = x$, g_0 is the solution to the integral equation

$$\alpha_0\,(x, l, u) = \int g\,(x, l, z)\, p_{Z|X,L,U}\,(z|x, l, u)\, dz, \tag{3}$$

where

$$\alpha_0\,(x, l, u) = I\big\{x \in q\,(\mathcal{S}(l), l)\,\big\}\frac{dq^{-1}(x, l)}{dx}\frac{p_{X|L,U}\,\{q^{-1}\,(x, l)\,|l, u\}}{p_{X|L,U}\,(x|l, u)}.$$

Note that the bridge equations (2) and (3) must hold for almost every $(x, l, u) \in \text{supp}\,(X, L, U)$ such that $x \in \mathcal{S}(l) \cup q(\mathcal{S}(l), l)$. This is because, within the strata defined by the potential confounders, the bridge functions allow us to infer the hypothetical outcome for an individual with observed exposure x, had they instead been exposed to the value prescribed by the policy $q(x, l)$.

A.2 *Conditions for Existence of the Latent Bridge Functions*

The existence of solutions to equations (2) and (3) becomes more intuitive when the variables Z, W, and U are discrete. In such cases, detailed conditions for the existence of solutions to equation (2) are provided by Miao et al. (2018); Tchetgen Tchetgen et al. (2020), and Shi et al. (2023). Solutions to both the treatment bridge equation for the average treatment effect (ATE) and generalized ATE are discussed in Cui et al. (2023) and Kallus et al. (2022), respectively.

Suppose W, Z, and U take a finite number of values, denoted by w_i, z_j, and u_k for $i = 1, \ldots, n_W$, $j = 1, \ldots, n_Z$, and $k = 1, \ldots, n_U$, respectively. For any $(x, l, u) \in \text{supp}\,(X, L, U)$ such that $x \in \mathcal{S}(l) \cup q(\mathcal{S}(l), l)$, let $P\,(\mathbf{W}|\mathbf{U}, x, l)$ denote the $n_W \times n_U$ matrix whose (i, k)-entry is $\mathbb{P}(W = w_i|U = u_k, X = x, L = l)$. Similarly, let $P\,(\mathbf{Z}|\mathbf{U}, x, l)$ be the $n_Z \times n_U$ matrix with (j, k)-entry $\mathbb{P}(Z = z_j|U = u_k, X = x, L = l)$. Define $R_0(x, l, \mathbf{U})$ and $A_0(x, l, \mathbf{U})$ as the n_U-dimensional vectors whose k-th entries are $\mathbb{E}\,(Y|X = x, L = l, U = u_k)$ and $\alpha_0(x, l, u_k)$,

respectively. With this notation, the bridge equations can be written as:

$$R_0(x,l,\mathbf{U}) = P\left(\mathbf{W}|\mathbf{U},x,l\right)^T h(x,l,\mathbf{W}), \quad \text{and}$$

$$A_0(x,l,\mathbf{U}) = P\left(\mathbf{Z}|\mathbf{U},x,l\right)^T g(x,l,\mathbf{Z}).$$

From this formulation, it is evident that solutions to the bridge equations exist if only if $R_0(x,l,\mathbf{U})$ and $A_0(x,l,\mathbf{U})$ lie in the vector spaces spanned by the row vectors of $P\left(\mathbf{W}|\mathbf{U},x,l\right)$ and $P\left(\mathbf{Z}|\mathbf{U},x,l\right)$, respectively. If $n_W, n_Z \geqslant n_U$ and the matrices $P\left(\mathbf{W}|\mathbf{U},x,l\right)$ and $P\left(\mathbf{Z}|\mathbf{U},x,l\right)$ are full row-rank, then the bridge functions always exist. However, the solutions are non-unique unless $n_W = n_U = n_Z$.

In the continuous case, the existence of solutions can be guaranteed by following the results of Miao et al. (2018) and Cui et al. (2023), which are based on the following completeness and regularity conditions.

Assumption 9 (Completeness conditions):

a) $\mathbb{E}\{\eta(X,L,U)|X=x,L=l,W\} = 0 \Rightarrow \eta(x,l,U) = 0$ almost everywhere.

b) $\mathbb{E}\{\eta(X,L,U)|X=x,L=l,Z\} = 0 \Rightarrow \eta(x,l,U) = 0$ almost everywhere.

We focus on the existence of solutions to equation (2), with similar reasoning applying to the treatment bridge equation (3). For each $(x,l) \in \text{supp}\,(X,L)$, let $\mathcal{T}^W_{x,l} : \mathcal{L}^2(W|X=x,L=l) \longrightarrow \mathcal{L}^2(U|X=x,L=l)$ denote the conditional expectation operator defined by $\mathcal{T}^W_{x,l} h(x,l,U) = \mathbb{E}\{h(X,L,W)|X=x,L=l,U\}$. Following Example 2.3 in Carrasco et al. (2007), the following assumption ensures that $\mathcal{T}^W_{x,l}$ is a compact operator.

Assumption 10 (Compactness of $\mathcal{T}^W_{x,l}$): For almost every $(x,l) \in \text{supp}\,(X,L)$, the following integral is finite: $\iint p_{W|X,L,U}(w|x,l,u)p_{U|X,L,W}(u|x,l,w)dwdu < +\infty$.

Under assumption 10, the operator $\mathcal{T}^W_{x,l}$ admits a singular value decomposition $(\sigma_{x,l,j}, \varphi_{x,l,j}, \phi_{x,l,j})_{j=1}^{\infty}$ (see Theorem 15.16 in Kress (2010)). Using this notation, we now state the regularity conditions.

ASSUMPTION 11 (Regularity conditions):

a) $\mathbb{E}(Y|X=x, L=l, U) \in \mathcal{L}^2\left(U|X=x, L=l\right)$.

b) $\sum_{j=1}^{\infty} \sigma_{x,l,j}^{-2} \left\{\int \mathbb{E}(Y|X=x, L=l, U=u)\phi_{x,l,j}(x,l,u)p_{U|X,L}(u|x,l)du\right\}^2 < +\infty$.

With these conditions in place, by directly applying Picard's theorem (Theorem 15.18 in Kress (2010)), we establish the following lemma, which guarantees the existence of solutions to equation (2).

LEMMA 2 (Miao, Geng, and Tchetgen Tchetgen (2018)): *Under assumptions 9a, 10, and 11, there exists a solution to the integral equation (2).*

A.3 *Proximal Identification Results*

The result linking solutions to the latent equations (2) and (3) to the solutions of integral equations that depend solely on observed data is as follows:

THEOREM 1 (Observed Bridge Equations):

i) [Kallus, Mao, and Uehara (2022)] Under assumptions 4 and 5, any solution $h_0\left(x,l,w\right)$ to equation (2) also solves, for all $(x,l,z) \in supp(X,L,Z)$ such that $x \in \mathcal{S}(l) \cup q\left(\mathcal{S}(l),l\right)$, the following equation

$$\mathbb{E}\left(Y|X=x, L=l, Z=z\right) = \int h\left(x,l,w\right) p_{W|X,L,Z}\left(w|x,l,z\right) dw. \tag{4}$$

ii) Under assumption 5, any solution to equation (3) also solves, for all $(x,l,w) \in supp(X,L,W)$ such that $x \in \mathcal{S}(l) \cup q\left(\mathcal{S}(l),l\right)$ and the map $x' \mapsto q(x',l)$ is differentiable at $x'=x$, the following equation

$$\alpha_0\left(x,l,w\right) = \int g\left(x,l,z\right) p_{Z|X,L,W}\left(z|x,l,w\right) dz, \tag{5}$$

where, with a slight abuse of notation, α_0 is defined as in equation (3) but with u replaced by w.

Next, we present the identification result for the counterfactual MTP mean within the population $\{O : (X, L) \in \mathcal{S}\}$.

THEOREM 2 (Identification): *Suppose assumptions 1 - 6 hold. If, in addition, either assumption 7 holds and equation (5) has at least one solution, or assumption 8 holds and equation (4) has at least one solution, then ψ_0 satisfies all of the following:*

i) (Outcome bridge representation)

$$\psi_0 = \frac{\mathbb{E}\left[h^\dagger \{q(X, L), L, W\} \cdot I\{(X, L) \in \mathcal{S}\}\right]}{P\{(X, L) \in \mathcal{S}\}},$$

where $h^\dagger$ is any solution to equation (4).

ii) (Treatment bridge representation)

$$\psi_0 = \frac{\mathbb{E}\left\{Y g^\dagger(X, L, Z)\right\}}{P\{(X, L) \in \mathcal{S}\}},$$

where $g^\dagger$ is any solution of the equation (5).

iii) (Double robust representation)

$$\psi_0 = \frac{\mathbb{E}\left\{\phi(O; h^\dagger, g^\dagger)\right\}}{P\{(X, L) \in \mathcal{S}\}},$$

where at least one of $h^\dagger$ and $g^\dagger$ solves the corresponding integral equations (4) and (5) respectively, and for any $h \in \mathcal{L}^2(X, L, W)$ and $g \in \mathcal{L}^2(X, L, Z)$,

$$\phi(O; h, g) := h\{q(X, L), L, W\} \cdot I\{(X, L) \in \mathcal{S}\} + g(X, L, Z)\{Y - h(X, L, W)\}.$$

In the outcome bridge and double robust representations of Theorem 2, the outcome regression function $h^\dagger$ is evaluated at the exposure level determined by the policy function q. This highlights the necessity for equation (4) to hold for all $(x, l, z) \in \text{supp}(X, L, Z)$ such that $x \in \mathcal{S}(l) \cup q(\mathcal{S}(l), l)$. Conversely, while the policy function q does not appear explicitly in the treatment bridge representation, it is implicit in the definition of α_0 in equation (5). Consequently, while our objective is to estimate the counterfactual MTP mean within the population $\{O : (X, L) \in \mathcal{S}\}$, we leverage data from the potentially larger population $\{O : L \in \text{supp}(L), X \in \mathcal{S}(L) \cup q(\mathcal{S}(L), L)\}$ to achieve this goal.

In addition, while our identification strategy is motivated by the approach proposed by Kallus et al. (2022), it introduces a key difference. The identification strategy by Kallus et al. (2022) requires assumption 7 to hold and equation (5) to have at least one solution for the outcome bridge representation, with analogous assumptions for the treatment bridge representation of the generalized ATE. In contrast, our approach requires either of these combinations of assumptions to obtain both representations, thereby offering more flexibility.

A.4 *Alternative Identification Strategy*

An alternative identification strategy, which was first proposed in proximal inference by Miao et al. (2018); Tchetgen Tchetgen et al. (2020) and later adopted by Cui et al. (2023), relies on different assumptions to achieve parametric identification. This approach either assumes a solution to the observed outcome bridge equation (4) along with the completeness condition in assumption 9b, or a solution to the observed treatment bridge equation (5) alongside the completeness condition in assumption 9a. The first combination of assumptions guarantees the existence of a solution to the outcome bridge equation (2), while the second implies the existence of a solution to the treatment bridge equation (3). This result is captured in the following result.

Lemma 3 (Miao, Geng, and Tchetgen Tchetgen (2018); and Cui et al. (2023)):

i) Under assumptions 4, 5, and 9b, any solution to the integral equation (4) also solves equation (2).

ii) Under assumptions 5 and 9a, any solution to the integral equation (5) also solves equation (3).

In conjunction with Theorem 1, the identification strategy by Miao et al. (2018); Tchetgen Tchetgen et al. (2020) and Cui et al. (2023), implies that equations (2) and (4) share the

same solutions, with z replaced by u, and vice versa. Similarly, it implies that equations (3) and (5) share the same solutions, with w replaced by u, and vice versa.

In the discrete setting, Kallus et al. (2022) demonstrated that the identification strategy by Miao et al. (2018); Tchetgen Tchetgen et al. (2020) and Cui et al. (2023) is strictly stronger than the identification strategy outlined in assertion iii) of Theorem 2. However, in the continuous case, these strategies are not directly comparable. Kallus et al. (2022) strategy avoids completeness conditions but requires the existence of both an outcome bridge and a treatment bridge function, with one latent and one observed. In contrast, the strategy of Miao et al. (2018); Tchetgen Tchetgen et al. (2020), and Cui et al. (2023) relies on an assumption for only one bridge function (either outcome or treatment), while enforcing that the latent and observed integral equations share the same solutions. We built on the strategy proposed by Kallus et al. (2022) because we focus on a non-parametric estimator that uses both bridge functions to achieve a $\sqrt{n}$-convergence rate.

A.5 *Estimation*

The following result establishes the pathwise differentiability of the target parameter ψ_0 in a nonparametric model restricted to distributions that satisfy the assumptions of Theorem 2. Under these restrictions, ψ_0 admits a regular and asymptotically linear estimator with an influence function given below. This influence function will be used to construct a one-step estimator for ψ_0.

LEMMA 4: *Suppose that the assumptions required in Theorem 2 are satisfied and let h_0 and g_0 be the minimum norm solutions to the integral equations (4) and (5). Then, the target parameter defined in (1) is pathwise differentiable with an influence function given by*

$$\psi_P^1(O) = \frac{\phi(O; h_0, g_0)}{P\{(X, L) \in \mathcal{S}\}} - \psi_0 \frac{I\{(X, L) \in \mathcal{S}\}}{P\{(X, L) \in \mathcal{S}\}}.$$

The identification result for ψ_0 suggests a plug-in estimator of the form $\hat{\psi} = \hat{\phi}/\hat{p}$, where $\hat{\phi}$ is an estimator of $\phi_0 := \mathbb{E}\left[\phi(O; h^\dagger, g^\dagger)\right] = \mathbb{E}\left[h\left\{q(X, L), L, W\right\} \cdot I\left\{(X, L) \in \mathcal{S}\right\}\right]$ and $\hat{p}$ is an estimator of $p_0 := P\left\{I(X, L) \in \mathcal{S}\right\}$. To construct $\hat{\phi}$, we partition the sample into K equal-sized folds $\{I_1, \ldots, I_K\}$, as described in Section 4 of the manuscript. For each $k \in \{1, \ldots, K\}$, we use data from all parts except I_k to obtain estimators $\hat{h}^{(-k)}$ and $\hat{g}^{(-k)}$ for the minimum-norm solutions h_0 and g_0, respectively. Within each fold k, we compute an estimator $\hat{\phi}_k$ for ϕ_0 as

$$\hat{\phi}_k = \frac{1}{|I_k|} \sum_{i: O_i \in I_k} \hat{h}^{(-k)}\left\{q(X_i, L_i), L_i, W_i\right\} \cdot I\left\{(X_i, L_i) \in \mathcal{S}\right\},$$

and aggregate these to obtain the cross-fitted estimator:

$$\hat{\phi}_{CF} = \frac{1}{K} \sum_{k=1}^{K} \hat{\phi}_k.$$

We define $\hat{p} = \frac{1}{n}\sum_{i=1}^{n} I\left\{(X_i, L_i) \in \mathcal{S}\right\}$, and the initial (plug-in) estimator for ψ_0 as $\tilde{\psi} = \hat{\phi}_{CF}/\hat{p}$. The one-step cross-fitted estimator is then defined as

$$\begin{aligned}
\hat{\psi}_{CF} &= \tilde{\psi} + \frac{1}{n}\sum_{i=1}^{n} \psi^1_{\widehat{P}}(O_i) \\
&= \tilde{\psi} + \frac{1}{n}\sum_{k=1}^{K} \sum_{i: O_i \in I_k} \left[\frac{\phi\left\{O_i; \hat{h}^{(-k)}, \hat{g}^{(-k)}\right\}}{\hat{p}} - \frac{\tilde{\psi}}{\hat{p}} \cdot I\left\{(X_i, L_i) \in \mathcal{S}\right\} \right] \\
&= \tilde{\psi} + \frac{1}{\hat{p}} \cdot \frac{1}{K}\sum_{k=1}^{K} \frac{1}{|I_k|} \sum_{i: O_i \in I_k} \phi\left\{O_i; \hat{h}^{(-k)}, \hat{g}^{(-k)}\right\} - \frac{\tilde{\psi}}{\hat{p}} \cdot \frac{1}{n}\sum_{i=1}^{n} I\left\{(X_i, L_i) \in \mathcal{S}\right\} \\
&= \frac{1}{\hat{p}} \cdot \frac{1}{K}\sum_{k=1}^{K} \frac{1}{|I_k|} \sum_{i: O_i \in I_k} \phi\left\{O_i; \hat{h}^{(-k)}, \hat{g}^{(-k)}\right\}.
\end{aligned}$$

Note that when $\mathcal{S} = \text{supp}\,(X, L)$, i.e., when $I\left\{(X, L) \in \mathcal{S}\right\} \equiv 1$, the one-step cross-fitted estimator $\hat{\psi}_{CF}$ aligns with the estimator proposed in the main manuscript.

For a given function $f \in \mathcal{L}^2(V)$, let $\|f\|_\infty$ denote its supreme norm $\|f\|_\infty := \sup_{v \in \text{supp}\,(V)} |f(v)|$. The asymptotic behavior of our doubly-robust cross-fitted estimator is provided in the following Theorem.

THEOREM 3: *Let h_0 and g_0 denote the minimum-norm solutions of the integral equations*

(4) and (5), respectively. If $|\alpha_0(X,L,W)| \leqslant B$, $|Y| \leqslant B$, *either* $\|h_0\|_\infty + \|\hat{g}\|_\infty \leqslant B$ *or* $\|\hat{h}\|_\infty + \|g_0\|_\infty \leqslant B$ *for some* B, *and for all* $k \in \{1,\ldots,K\}$, *the estimators* $\hat{h}^{(-k)}$ *and* $\hat{g}^{(-k)}$ *are norm consistent in the sense that* $\left\|\hat{h}^{(-k)} - h_0\right\|_2 = o_p(1)$ *and* $\left\|\hat{g}^{(-k)} - g_0\right\|_2 = o_p(1)$, *then the estimator* $\hat{\psi}_{CF}$ *satisfies*

$$\sqrt{n}\left(\hat{\psi}_{CF} - \psi_0\right) = \sqrt{n}\mathbb{E}_n\left[\frac{\phi(O;h_0,g_0)}{P\{(X,L)\in\mathcal{S}\}} - \psi_0\frac{I\{(X,L)\in\mathcal{S}\}}{P\{(X,L)\in\mathcal{S}\}}\right] + \sqrt{n}R_n + o_p(1),$$

where

$$R_n = \frac{1}{P\{(X,L)\in\mathcal{S}\}}\cdot\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}\left[\left\{\hat{h}^{(-k)} - h_0\right\}\left\{g_0 - \hat{g}^{(-k)}\right\}\right].$$

In particular, if $R_n = o_p\left(n^{-1/2}\right)$ *and* $\mathbb{E}\{\phi(O;h_0,g_0)^2\} < \infty$, *then*

$$\sqrt{n}\left(\psi_{CF} - \psi_0\right) \xrightarrow{d} \mathcal{N}(0,\tau^2), \tag{6}$$

where $\tau^2 = \mathbb{E}\left\{\left[\frac{\phi(O;h_0,g_0)}{P\{(X,L)\in\mathcal{S}\}} - \psi_0\frac{I\{(X,L)\in\mathcal{S}\}}{P\{(X,L)\in\mathcal{S}\}}\right]^2\right\}$.

In Web Appendix B, we provide norm-consistent estimators for the bridge functions, which, under certain assumptions, ensure that R_n is $o_p\left(n^{-1/2}\right)$.

Web Appendix B. Estimation of the Bridge Functions

In this section, we provide the convergence analysis of the estimators of the bridge functions and their closed-form expressions using reproducing kernel Hilbert spaces.

B.1 *Convergence analysis*

We present convergence guarantees for the estimator $\hat{h}^{(-k)}$ of h_0, with analogous results applying for the estimator $\hat{g}^{(-k)}$ of g_0. To simplify the notation, we omit the superscript $(-k)$ from $\hat{h}^{(-k)}$ in the following discussion.

Let $\mathcal{T} : \mathcal{L}^2(X,L,W) \to \mathcal{L}^2(X,L,Z)$ and $\mathcal{T}^* : \mathcal{L}^2(X,L,Z) \to \mathcal{L}^2(X,L,W)$ denote the conditional expectation operators defined for any $h \in \mathcal{L}^2(X,L,W)$ and $g \in \mathcal{L}^2(X,L,Z)$ as:

$$\mathcal{T}h(X,L,Z) := \mathbb{E}\left[h(X,L,W)|X,L,Z\right] \quad \text{and} \quad \mathcal{T}^*g(X,L,Z) := \mathbb{E}\left[g(X,L,Z)|X,L,W\right].$$

Note that $\mathcal{T}^*$ is the adjoint operator of $\mathcal{T}$: for any $h \in \mathcal{L}^2(X,L,W)$ and $g \in \mathcal{L}^2(X,L,Z)$, we

have

$$\langle \mathcal{T}h, g\rangle_2 = \mathbb{E}\left[h(X,L,W)g(X,L,Z)\right] = \langle h, \mathcal{T}^* g\rangle_2,$$

where $\langle\cdot,\cdot,\rangle_2$ denotes the inner product in the $\mathcal{L}^2$ space defined by the variables involved.

As indicated before, we assume $\mathcal{H}$ and $\mathcal{G}'$ are reproducing kernel Hilbert spaces with kernels denoted $K_{\mathcal{H}}$ and $K_{\mathcal{G}'}$. We let $\mathcal{T}_{K_{\mathcal{H}}} : \mathcal{L}^2(X,L,W) \to \mathcal{L}^2(X,L,W)$ denote the integral operator associated with $K_{\mathcal{H}}$, i.e.,

$$\mathcal{T}_{K_{\mathcal{H}}} h(x,l,w) := \int K_{\mathcal{H}}\left(x,l,w;x',l',w'\right) h(x',l',w') p_{X,L,W}(x',l',w') dx'dl'dw'.$$

We make the following assumptions:

ASSUMPTION 12: (Compactness and injectivity of $\mathcal{T}_{K_{\mathcal{H}}}$)

a) (Compact support) $\text{supp}\,(X,L,W)$ are compact.

b) (Square-integrability) The kernel function $K_{\mathcal{H}} : \text{supp}\,(X,L,W) \times \text{supp}\,(X,L,W) \to \mathbb{R}$ is continuous and satisfies

$$\int \left\{K_{\mathcal{H}}\left[(x,l,w);(x',l',w')\right]\right\}^2 p_{X,L,W}(x,l,w) p_{X,L,W}(x',l',w') dxdldwdx'dl'dw' < \infty.$$

c) (Injectivity of the integral operator) The integral operator $\mathcal{T}_{K_{\mathcal{H}}} : \mathcal{L}^2(X,L,W) \to \mathcal{L}^2(X,L,W)$ is injective, i.e., its null space $\{h \in \mathcal{L}^2(X,L,W) : \mathcal{T}_{K_{\mathcal{H}}} h = 0\} = \{0\}$ is trivial.

REMARK 1: Assumption 12b is the assumption that the operator $\mathcal{T}_{K_{\mathcal{H}}}$ is Hilbert-Schmidt (see Theorem 2.34 of Carrasco et al. (2007)). Under assumptions 12a and b, the operator $\mathcal{T}_{K_{\mathcal{H}}}$ admits a singular value decomposition consisting of a sequence of eigenfunctions $(\varphi_j)_{j=1}$ that forms an orthonormal sequence of $\mathcal{L}^2(X,L,W)$ with a corresponding sequence of non-negative eigenvalues $(\eta_j)_{j=1}^{\infty}$ converging to 0 (see Theorems 2.39, 2.41, and 2.42 of Carrasco et al. (2007)):

$$\mathcal{T}_{K_{\mathcal{H}}}(\varphi_j) = \eta_j \varphi_j \quad \text{for } j = 1, 2, \ldots,$$

and

$$K_{\mathcal{H}}(v, v') = \sum_{j=1}^{\infty} \eta_j \varphi_j(v)\varphi_j(v') \quad \forall v, v' \in \text{supp}\,(X, L, W).$$

Additionally, assumption 12c ensures that the eigenvalues $(\eta_j)_{j=1}^{\infty}$ are strictly positive and that the eigenfunctions $(\varphi_j)_{j=1}$ form an orthonormal basis of $\mathcal{L}^2(X, L, W)$. Consequently, $\mathcal{H}$ is a separable[1] Hilbert space with the following representation (see Corollary 12.26 of Wainwright (2019))

$$\mathcal{H} \equiv \left\{ h = \sum_{j=1}^{\infty} \gamma_j \varphi_j \Big| \text{ for some } (\gamma_j)_{j=1}^{\infty} \text{ with } \sum_{j=1}^{\infty} \frac{\gamma_j^2}{\eta_j} < \infty \right\}, \tag{7}$$

and equipped with the inner product:

$$\langle h, h' \rangle_{\mathcal{H}} := \sum_{j=1}^{\infty} \frac{\langle h, \varphi_j \rangle_2 \langle h', \varphi_j \rangle_2}{\eta_j}.$$

Remark 2: Assumption 12c holds when the kernel $K_{\mathcal{H}} : \text{supp}\,(X, L, W) \times \text{supp}\,(X, L, W) \to \mathbb{R}$ is a radial function, i.e., for any $v_1, v_2 \in \text{supp}\,(X, L, W)$, $K(v_1, v_2) = \rho(\|v_1 - v_2\|_2)$, where $\rho : [0, \infty) \to \mathbb{R}$, and the Fourier transform of ρ is strictly positive. In Web Appendix F, we demonstrate that the Gaussian kernel satisfies these conditions. Other Matérn kernels also satisfy these conditions.

Defining the linear operator $T_{\mathcal{H}_2}^{-1/2} : \mathcal{H}_2 \to \mathcal{L}^2(X, L, W)$ as

$$T_{\mathcal{H}}^{-1/2} h := \sum_{j=1}^{\infty} \frac{\langle h, \varphi_j \rangle_2}{\sqrt{\eta_j}} \varphi_j,$$

we have, for any $h \in \mathcal{H}_2$,

$$\|h\|_{\mathcal{H}}^2 = \left\| T_{\mathcal{H}}^{-1/2} h \right\|_{\mathcal{L}^2(X,L,W)}^2.$$

Let $\tilde{\eta} := \max_j \eta_j < \infty$. Since $\sum_{j=1}^{\infty} \eta_j \langle h, \varphi_j \rangle_2^2 \leqslant \tilde{\eta} \sum_{j=1}^{\infty} \langle h, \varphi_j \rangle_2^2 = \tilde{\eta} \|h\|_2^2$, the operator $T_{\mathcal{H}}^{-1/2}$ is invertible, and its inverse $T_{\mathcal{H}}^{1/2} : \mathcal{L}^2(X, L, W) \to \mathcal{H}$ is defined for any $h \in \mathcal{L}^2(X, L, W)$ as

$$T_H^{1/2} h := \sum_{j=1}^{\infty} \sqrt{\eta_j} \langle h, \varphi_j \rangle_2 \varphi_j.$$

[1] A separable space is a metric space that contains a countable dense subset.

Notably, $T_H^{1/2}$ is a bounded linear operator with $\left\|T_{\mathcal{H}}^{1/2}\right\| \leqslant \sqrt{\widetilde{\eta}}$, where the norm of any operator A is defined as $\|A\| := \sup_{x\neq 0} \frac{\|Ax\|_2}{\|x\|_2}$.

Consider the operator $\mathcal{T} \circ T_{\mathcal{H}}^{1/2}$ which maps $\mathcal{L}^2(X, LW)$ onto $\mathcal{T}(\mathcal{H}_2) \subset \mathcal{L}^2(X, L, Z)$. Its adjoint is the operator $T_{\mathcal{H}}^{1/2} \circ \mathcal{T}^*$ which maps $\mathcal{L}^2(X, L, Z)$ onto $T_{\mathcal{H}}^{1/2}(\mathcal{T}^*[\mathcal{L}^2(X, L, Z)]) \subset T_{\mathcal{H}}^{1/2}[\mathcal{L}^2(X, L, W)] = \mathcal{H}_2$ since for any $h \in \mathcal{L}^2(X, L, W)$ and $g \in \mathcal{L}^2(X, L, Z)$, we have

$$\begin{aligned}
\left\langle \mathcal{T} \circ T_{\mathcal{H}}^{1/2} h, g \right\rangle_{\mathcal{L}^2(X,L,Z)} &= \left\langle T_{\mathcal{H}}^{1/2} h, \mathcal{T}^* g \right\rangle_{\mathcal{L}^2(X,L,W)} \\
&= \left\langle \sum_{j=1}^{\infty} \sqrt{\eta_j} \langle h, \varphi_j \rangle_{\mathcal{L}^2(X,L,W)} \varphi_j, \mathcal{T}^* g \right\rangle_{\mathcal{L}^2(X,L,W)} \\
&= \sum_{j=1}^{\infty} \sqrt{\eta_j} \langle h, \varphi_j \rangle_{\mathcal{L}^2(X,L,W)} \langle \varphi_j, \mathcal{T}^* g \rangle_{\mathcal{L}^2(X,L,W)} \\
&= \left\langle h, \sum_{j=1}^{\infty} \sqrt{\eta_j} \langle \varphi_j, \mathcal{T}^* g \rangle_{\mathcal{L}^2(X,L,W)} \varphi_j \right\rangle_{\mathcal{L}^2(X,L,W)} \\
&= \left\langle h, T_{\mathcal{H}}^{1/2} \circ \mathcal{T}^* g \right\rangle_{\mathcal{L}^2(X,L,W)}.
\end{aligned}$$

Now, define $\widetilde{\mathcal{T}} := \frac{1}{2}\mathcal{T} \circ T_{\mathcal{H}}^{1/2}$. Its adjoint satisfies $\widetilde{\mathcal{T}}^* = \frac{1}{2}\mathcal{T}_{\mathcal{H}}^{1/2} \circ \mathcal{T}^*$. To ensure that our minimax approach—regularizing the outer minimization with a penalty based on the norm of the function class $\mathcal{H}$—is suitable for estimating the minimum-norm solution h_0, we impose the following condition on h_0:

Assumption 13 ($\beta-$source condition): There exists $w_0 \in \mathcal{L}^2(X, L, W)$ such that the minimum-norm solution h_0 satisfies

$$h_0 = T_{\mathcal{H}}^{1/2} \circ \left(\widetilde{\mathcal{T}}^* \circ \widetilde{\mathcal{T}}\right)^{\beta/2} w_0.$$

Assumption 13 implies that $h_0 \in \mathcal{H}_2$. Consequently, we can apply the operator $T_{\mathcal{H}}^{-1/2}$ to h_0, and so, the following regularized optimization problem is well defined:

$$h_* = \arg \min_{h \in \mathcal{L}^2(X,L,W)} \left\{ \left\| \widetilde{\mathcal{T}}(T_{\mathcal{H}}^{-1/2} h_0 - h) \right\|_2^2 + \lambda_{\mathcal{H}} \|h\|_2^2 \right\}.$$

Since $\widetilde{\mathcal{T}}$ is a bounded linear operator, Theorem 16.4 of Kress (2010) guarantees that the

solution h_* to this optimization problem exists and is unique. Furthermore, this optimization problem can be equivalently reformulated as:

$$h_* = \arg \min_{h \in \mathcal{L}^2(X,L,W)} \left\{ \frac{1}{4} \left\| \mathcal{T}(h_0 - T_{\mathcal{H}}^{1/2} h) \right\|_2^2 + \lambda_{\mathcal{H}} \| T_{\mathcal{H}}^{-1/2} \circ T_{\mathcal{H}}^{1/2} h \|_2^2 \right\}.$$

REMARK 3: Since $T_{\mathcal{H}}^{1/2}$ is a bijective map from $\mathcal{L}^2(X, L, W)$ onto $\mathcal{H}_2$, i.e., $T_{\mathcal{H}}^{1/2}\left(\mathcal{L}^2(X, L, W)\right) = \mathcal{H}_2$, the uniqueness of solutions for regularized optimization problems implies that $T_{\mathcal{H}}^{1/2} h_*$ solves the following optimization problem:

$$T_{\mathcal{H}}^{1/2} h_* = \arg \min_{h \in \mathcal{H}} \left\{ \frac{1}{4} \| \mathcal{T}(h_0 - h) \|_2^2 + \lambda_{\mathcal{H}} \| h \|_{\mathcal{H}}^2 \right\}.$$

In addition, to address potential challenges in the inner maximization step of our minimax approach, we impose the following assumption on the relationship between the function classes $\mathcal{H}$ and $\mathcal{G}'$:

ASSUMPTION 14 (Closedness of $\mathcal{H}$ with respect to $\mathcal{G}'$): For any $h \in \mathcal{H}$, $\mathcal{T}(h_0 - h) \in \mathcal{G}'$.

REMARK 4: Under assumption 14, Lemma 1 ensures that $T_{\mathcal{H}}^{1/2} h_*$ solves the optimization problem:

$$T_{\mathcal{H}}^{1/2} h_* = \arg \min_{h \in \mathcal{H}} \max_{g \in \mathcal{G}'} \left\{ \mathbb{E}\left\{ g(X, L, Z) \left[Y - h(X, L, W) \right] - g(X, L, Z)^2 \right\} + \lambda_{\mathcal{H}} \| h \|_{\mathcal{H}}^2 \right\}. \quad (8)$$

This shows that $T_{\mathcal{H}}^{1/2} h_*$ solves a population optimization problem analogous to the one addressed by our estimator, except that it does not include the inner penalty term $\lambda_{\mathcal{G}'} \| g \|_{\mathcal{G}'}^2$.

By leveraging standard results on regularization bias under source conditions (see Proposition 3.11 of Carrasco et al. (2007), Theorem 1.4 of Cavalier (2011), and Lemma 3 of Bennett et al. (2023)), we derive the following bounds for $\left\| h_* - T_{\mathcal{H}}^{-1/2} h_0 \right\|_2$ and $\left\| \mathcal{T}\left(T_{\mathcal{H}}^{1/2} h_* - h_0 \right) \right\|_2$.

THEOREM 4 (Regularization Bias): *Under assumptions 12-13, we have*

$$\left\| h_* - T_{\mathcal{H}}^{-1/2} h_0 \right\|_2^2 \leqslant \| w_0 \|_2^2 \left(\frac{\tilde{\eta}}{4} \right)^{\max\{0, \beta - 2\}} \lambda_{\mathcal{H}}^{\min\{\beta, 2\}} \quad \textit{and}$$

$$\left\| \mathcal{T}\left(T_{\mathcal{H}}^{1/2} h_* - h_0 \right) \right\|_2^2 \leqslant 4 \| w_0 \|_2^2 \left(\frac{\tilde{\eta}}{4} \right)^{\max\{0, \beta - 1\}} \lambda_{\mathcal{H}}^{\min\{\beta + 1, 2\}}.$$

To proceed toward the main result, we introduce additional notation and concepts. Let $\mathcal{F}$ denote a class of real-valued functions $f : \mathcal{X} \to \mathbb{R}$. The function class $\mathcal{F}$ is said to be uniformly bounded by b if, for all $f \in \mathcal{F}$, $\|f\|_\infty \leqslant b$. For any $B > 0$, we define $\mathcal{F}_B := \{f \in \mathcal{F} : \|f\|_{\mathcal{F}} \leqslant B\}$, where $\|\cdot\|_{\mathcal{F}}$ denotes the norm associated with the class $\mathcal{F}$. The class $\mathcal{F}$ is called *star-shaped* if for any $f \in \mathcal{F}$ and $\alpha \in [0, 1]$, the scaled function αf also belongs to $\mathcal{F}$. For any function class $\mathcal{F}$, we define $star\,(\mathcal{F}) := \{\alpha f | f \in \mathcal{F}, \alpha \in [0, 1]\}$, which is a star-shaped class that contains all function in $\mathcal{F}$. For any function $f^* \in \mathcal{F}$, we defined the class $\mathcal{F} - f^*$ centered at f^* as $\mathcal{F} - f^* := \{f - f^* | f \in \mathcal{F}\}$.

For a given radius $\delta > 0$, the *localized Rademacher complexity* of a function class $\mathcal{F}$ is defined as

$$\mathcal{R}_n\left(\mathcal{F}, \delta\right) := \mathbb{E}_{\varepsilon, X_1, \ldots, X_n}\left[\sup_{f \in \mathcal{F}: \|f\|_2 \leqslant \delta}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i f(X_i)\right|\right],$$

where $\varepsilon_1, \ldots, \varepsilon_n$ are independent Rademacher random variables. Localized Rademacher complexity, a concept from statistical learning theory, quantifies the complexity of a class of functions within a neighborhood of the target function. This measure focuses on functions that are *close* to the true target, thereby avoiding unnecessary increases in complexity by ignoring functions that are far from the true solution.

The *critical radius* of a function class $\mathcal{F}$ is defined as the smallest positive radius δ for which the following inequality holds:

$$\mathcal{R}\left(\mathcal{F}, \delta\right) \leqslant \delta^2.$$

Intuitively, the critical radius identifies the smallest scale at which a learning algorithm can reliably estimate functions in $\mathcal{F}$, balancing between the complexity of the function class and the inherent uncertainty in the data.

Additionally, in line with assumption 14 and similar to Dikkala et al. (2020) and Ghassami

et al. (2022), we assume a Lipschitz relationship between the norms of the function classes $\mathcal{H}$ and $\mathcal{G}'$ under the operator $\mathcal{T}$:

ASSUMPTION 15 (Lipschitz condition between $\mathcal{H}$ and $\mathcal{G}'$): There exists $L > 0$ such that for all $h \in \mathcal{H}$, $\|\mathcal{T}h\|_{\mathcal{G}'} \leqslant L \|h\|_{\mathcal{H}}$.

REMARK 5: It is well known (see Chapter 12 of Wainwright (2019)) that, for a reproducing kernel Hilbert space $\mathcal{F}$ generated by a Gaussian kernel, and for any $B > 0$, the class $\mathcal{F}_B$ is uniformly bounded by B, i.e., $\sup_{x\in\mathcal{X}} |f(x)| \leqslant B$ for all $f \in \mathcal{F}_B$. For completeness, this result is demonstrated in Web Appendix F.2. A similar property, with an appropriate scaling factor, also holds for other kernels, such as Matérn kernels.

We now present our main result:

THEOREM 5: *Suppose assumptions 12-15 hold, and that* $|Y| \leqslant B$ *and* $\|h_0\|_{\mathcal{H}} \leqslant B$ *for some* B. *Let* $B_1 = B + 1$ *and* $B_2 = \frac{1}{2}L(2B + 1)$, *and define the function classes*

$$\mathcal{H}_{B_1} \cdot \mathcal{G}'_{B_2} := \left\{h \cdot g | h \in \mathcal{H}_{B_1}, g \in \mathcal{G}'_{B_2}\right\} \quad \textit{and} \quad Y \cdot \mathcal{G}'_{B_2} = \left\{Y \cdot g | g \in \mathcal{G}'_{B_2}\right\}.$$

Let $n \geqslant 3$, *and let* $\delta_n^2 \geqslant c_1 \frac{\log(\log(n))}{n}$ *be an upper bound on the critical radius of* $star(\mathcal{H}_{B_1} \cdot \mathcal{G}'_{B_2})$, $star(Y \cdot \mathcal{G}'_{B_2})$, $star(\mathcal{H}_{B_1})$, *and* $star(\mathcal{G}'_{B_2})$. *If* $\hat{h} \in \mathcal{H}_{B_1}$, *then the estimator* $\hat{h}$ *satisfies, with high probability,*

$$\left\|\hat{h} - h_0\right\|_2^2 = O\left(\frac{\delta_n^2}{\lambda_{\mathcal{H}}} + \frac{\lambda_{\mathcal{G}'}}{\lambda_{\mathcal{H}}} + \|w_0\|_2^2 \lambda_{\mathcal{H}}^{\min\{\beta,1\}}\right),$$

and

$$\left\|\mathcal{T}\left(\hat{h} - h_0\right)\right\|_2^2 = O\left(\delta_n^2 + \lambda_{\mathcal{G}'} + \|w_0\|_2^2 \lambda_{\mathcal{H}}^{\min\{\beta+1,2\}}\right).$$

REMARK 6: Theorem 5 establishes that the bias of $\hat{h}$ can be controlled by appropriately selecting the regularization parameters $\lambda_{\mathcal{H}}$ and $\lambda_{\mathcal{G}'}$. Specifically, if $\beta \geqslant 1$, setting $\lambda_{\mathcal{H}} = O(\delta_n)$ and $\lambda_{\mathcal{G}'} = O\left(\delta_n^2\right)$ ensures that $\left\|\hat{h} - h_0\right\|_2 = O(\delta_n^{1/2})$ and $\left\|\mathcal{T}\left(\hat{h} - h_0\right)\right\|_2 = O(\delta_n)$.

REMARK 7: In our numerical experiments and application, we choose $\mathcal{H}$ and $\mathcal{G}'$ as reproducing kernel Hilbert spaces generated by a Gaussian kernel. This choice results in a critical radius of $\delta_n = O\left(\sqrt{\frac{\log(n)}{n}}\right)$. Consequently, the remainder term R_n introduced in Theorem 3 satisfies $R_n = o_p(n^{-1/2})$.

To obtain the convergences guarantees for the estimator $\hat{g}$ of g_0, we adapt the previous discussion as follows. We consider $\mathcal{T}_{K_\mathcal{G}} : \mathcal{L}^2(X, L, Z) \to \mathcal{L}^2(X, L, Z)$, the integral operator associated with $K_\mathcal{G}$, i.e.,

$$\mathcal{T}_{K_\mathcal{G}} g(x, l, z) := \int K_\mathcal{G}\left(x, l, z; x', l', z'\right) g(x', l', z') p_{X,L,Z}(x', l', z') dx' dl' dz',$$

and impose an assumption analogue to assumption 12 to ensure that the operator $\mathcal{T}_{K_\mathcal{G}}$ admits a singular value decompositions consisting of a sequence of eigenfunctions $(\phi_j)_{j=1}$ that forms an orthonormal basis of $\mathcal{L}^2(X, L, Z)$ with a sequence of positive eigenvalues $(\mu_j)_{j=1}^\infty$ converging to 0. Therefore, $\mathcal{G}$ is a separable Hilbert space with the following representation

$$\mathcal{G} \equiv \left\{ g = \sum_{j=1}^\infty \gamma_j \phi_j \Big| \text{ for some } (\gamma_j)_{j=1}^\infty \text{ with } \sum_{j=1}^\infty \frac{\gamma_j^2}{\mu_j} < \infty \right\}, \tag{9}$$

and equipped with the inner product:

$$\langle g, g' \rangle_\mathcal{G} := \sum_{j=1}^\infty \frac{\langle g, \phi_j \rangle_2 \langle g', \phi_j \rangle_2}{\mu_j}.$$

Let $\mathcal{G}_2$ denote the space consisting of all the functions in $\mathcal{G}$ but equipped with the inner product inhered from $\mathcal{L}^2(X, L, Z)$, that is, $\mathcal{G}_2$ is a subspace of $\mathcal{L}^2(X, L, Z)$. We then define the linear operator $T_\mathcal{G}^{-1/2} : \mathcal{G}_2 \to \mathscr{L}^2(X, L, Z)$ as

$$T_\mathcal{G}^{-1/2} g := \sum_{j=1}^\infty \frac{\langle g, \phi_j \rangle_2}{\sqrt{\mu_j}} \phi_j,$$

to have, for any $g \in \mathcal{G}_2$,

$$\|g\|_\mathcal{G}^2 = \left\| T_\mathcal{G}^{-1/2} g \right\|_{\mathcal{L}^2(X,L,Z)}^2.$$

The operator $T_\mathcal{G}^{1/2} : \mathcal{L}^2(X, L, Z) \to \mathcal{G}_2$ defined for any $g \in \mathcal{L}^2(X, L, Z)$ by

$$T_\mathcal{G}^{1/2} g := \sum_{j=1}^\infty \sqrt{\mu_j} \langle g, \phi_j \rangle_2 \phi_j$$

is the inverse of $T_{\mathcal{G}}^{-1/2}$. In addition, let $I_q : \mathcal{L}^2(X,L,Z) \to \mathcal{L}^2(X,L,Z)$ be the operator defined for any $g \in \mathcal{L}^2(X,L,Z)$ by

$$(I_q g)(x,l,z) = I\{x \in q(\mathcal{S}(l),l)\} \cdot g(x,l,z).$$

It is clear that I_q is a bounded linear operator. Now, define $\widetilde{\mathcal{T}}_G := \frac{1}{2}\mathcal{T}^* \circ I_q \circ T_{\mathcal{G}}^{1/2}$. Its adjoint satisfies $\widetilde{\mathcal{T}}_G^* = \frac{1}{2} T_{\mathcal{G}}^{1/2} \circ I_q \circ \mathcal{T}$ since for any $h \in \mathcal{L}^2(X,L,W)$ and $g \in \mathcal{L}^2(X,L,Z)$, we have

$$\begin{aligned}
\left\langle \mathcal{T}^* \circ I_q \circ T_{\mathcal{G}}^{1/2} g, h \right\rangle_{\mathcal{L}^2(W)} &= \left\langle I_q \circ T_{\mathcal{G}}^{1/2} g, \mathcal{T} h \right\rangle_{\mathcal{L}^2(Z)} \\
&= \left\langle I_q \left(\sum_{j=1}^{\infty} \sqrt{\mu_j} \langle g, \phi_j \rangle_{\mathcal{L}^2(Z)} \phi_j \right), \mathcal{T} h \right\rangle_{\mathcal{L}^2(Z)} \\
&= \sum_{j=1}^{\infty} \sqrt{\mu_j} \langle g, \phi_j \rangle_{\mathcal{L}^2(Z)} \left\langle I_q \circ \phi_j, \mathcal{T} h \right\rangle_{\mathcal{L}^2(Z)} \\
&\overset{(i)}{=} \sum_{j=1}^{\infty} \sqrt{\mu_j} \langle g, \phi_j \rangle_{\mathcal{L}^2(Z)} \left\langle \phi_j, I_q \circ \mathcal{T} h \right\rangle_{\mathcal{L}^2(Z)} \\
&= \left\langle g, \sum_{j=1}^{\infty} \sqrt{\mu_j} \left\langle \phi_j, I_q \circ \mathcal{T}^* g \right\rangle_{\mathcal{L}^2(W)} \phi_j \right\rangle_{\mathcal{L}^2(Z)} \\
&= \left\langle g, T_{\mathcal{G}}^{1/2} \circ I_q \circ \mathcal{T} h \right\rangle_{\mathcal{L}^2(Z)},
\end{aligned}$$

where the equality in (i) follows from

$$\begin{aligned}
\langle I_q \circ \phi_j, \mathcal{T} h \rangle_{\mathcal{L}^2(X,L,Z)} &= \int I\{x \in q(\mathcal{S}(l),l)\} \phi_j(x,l,z) (\mathcal{T} h)(x,l,z) p_{X,L,Z} dx dl dz \\
&= \langle \phi_j, I_q \circ \mathcal{T} h \rangle_{\mathcal{L}^2(X,L,Z)}.
\end{aligned}$$

It is clear that the minimum-norm solution g_0 to the integral equation (5) satisfies $g_0(x,l,z) = 0$ for all $(x,l,z) \in \operatorname{supp}(X,L,Z)$ such that $x \in q(\mathcal{S}(l),l)$. Hence, to guarantee that our minimax approach can be used to estimate a solution to the integral equation (5), we impose the following condition:

Assumption 16 ($\tilde{\beta}-$source condition): There exists $z_0 \in \mathcal{L}^2(X,L,Z)$ such that the minimum-norm solution g_0 satisfies

$$g_0 = I_q \circ T_{\mathcal{G}}^{1/2} \circ \left(\widetilde{\mathcal{T}}_G^* \circ \widetilde{\mathcal{T}}_G \right)^{\tilde{\beta}/2} z_0.$$

Assumption 16 implies that g_0 can be written as $g_0 = I_q \tilde{g}_0$ with $\tilde{g}_0 = T_{\mathcal{G}}^{1/2} \circ \left(\widetilde{\mathcal{T}}_G^* \circ \widetilde{\mathcal{T}}_G\right)^{\tilde{\beta}/2} z_0$. That is, $\tilde{g}_0 \in \mathcal{G}_2$ and satisfies a $\tilde{\beta}$-source condition of the same form as the one that was assumed for h_0. Hence, since our estimator $\hat{g}$ has the form $\hat{g}(x,l,z) = I\{x \in q(\mathcal{S}(l),l)\}\,\hat{\Gamma}(x,l,z)$ where $\hat{\Gamma}(x,l,z)$ is an estimator for $\tilde{g}_0$, a convergence analysis similar to the one employed for $\hat{h}$ provides bounds for $\left\|\hat{\Gamma} - \tilde{g}_0\right\|_2^2$ and $\left\|\tilde{I}_q \circ \mathcal{T}^*\left(\hat{\Gamma} - \tilde{g}_0\right)\right\|_2^2$, where $\tilde{I}_q : \mathcal{L}^2(X,L,W) \to \mathcal{L}^2(X,L,W)$ is the operator defined for any $h \in \mathcal{L}^2(X,L,W)$ by

$$(\tilde{I}_q h)(x,l,w) = I\{x \in q(\mathcal{S}(l),l)\}\,h(x,l,w).$$

The same bounds hold for $\|\hat{g} - g_0\|_2^2$ and $\|\mathcal{T}^*(\hat{g} - g_0)\|_2^2$ since $\hat{g} - g_0 = I_q \circ \left(\hat{\Gamma} - \tilde{g}_0\right)$, $\|I_q\| \leqslant 1$, and $\mathcal{T}^* \circ I_q = \tilde{I}_q \circ \mathcal{T}^*$.

Here are the details. Consider the following regularized optimization problem

$$g_* = \arg\min_{g \in \mathcal{L}^2(X,L,Z)} \left\{ \left\| \widetilde{\mathcal{T}}_G (T_{\mathcal{G}}^{-1/2} \tilde{g}_0 - g) \right\|_2^2 + \lambda_{\mathcal{G}} \|g\|_2^2 \right\}.$$

This problem is well-defined since $\tilde{g}_0 \in \mathcal{G}_2$, ensuring that the term $T_{\mathcal{G}}^{-1/2}\tilde{g}_0$ is well-defined. Moreover, since $\widetilde{\mathcal{T}}_G$ is a bounded linear operator, the solution g_* exists and is unique. In addition, since $g_0 = I_q \tilde{g}_0$, this optimization program can be reformulated as:

$$g_* = \arg\min_{g \in \mathcal{L}^2(X,L,Z)} \left\{ \frac{1}{4} \left\| \mathcal{T}^*(g_0 - I_q \circ T_{\mathcal{G}}^{1/2} g) \right\|_2^2 + \lambda_{\mathcal{G}} \|T_{\mathcal{G}}^{-1/2} \circ T_{\mathcal{G}}^{1/2} g\|_2^2 \right\}.$$

Using that $T_{\mathcal{G}}^{1/2}$ is a bijective map between $\mathcal{L}^2(X,L,Z)$ and $\mathcal{G}_2$, it follows that $T_{\mathcal{G}}^{1/2} g_*$ solves the following optimization problem:

$$T_{\mathcal{G}}^{1/2} g_* = \arg\min_{g \in \mathcal{G}} \left\{ \frac{1}{4} \|\mathcal{T}^*(g_0 - I_q \circ g)\|_2^2 + \lambda_{\mathcal{G}} \|g\|_{\mathcal{G}}^2 \right\}.$$

Since

$$\|\mathcal{T}^*(g_0 - I_q \circ g)\|_2^2 = \|\mathcal{T}^*(I_q \circ \tilde{g}_0 - I_q \circ g)\|_2^2 = \left\| \tilde{I}_q \circ \mathcal{T}^*(\tilde{g}_0 - g) \right\|_2^2,$$

we impose the following assumption on the relationship between the function classes $\mathcal{G}$ and $\mathcal{H}'$:

Assumption 17 (Closedness of $\mathcal{G}$ with respect to $\mathcal{H}'$): For any $g \in \mathcal{G}$, $\mathcal{T}^*(\tilde{g}_0 - g) \in \mathcal{H}'$.

Under assumption 17, Lemma 1 guarantees that $T_{\mathcal{G}}^{1/2} g_*$ solves the optimization problem:

$$T_{\mathcal{G}}^{1/2} g_* = \arg\min_{g\in\mathcal{G}} \max_{h\in\mathcal{H}'} \mathbb{E}_n \Big[h\left\{q(X,L),L,W\right\} I\left\{(X,L)\in\mathcal{S}\right\} \\ - I\left\{X \in q(\mathcal{S}(L),L)\right\} h(X,L,W) g(X,L,Z) \\ - I\left\{X \in q(\mathcal{S}(L),L)\right\} h(X,L,W)^2 \Big] + \lambda_{\mathcal{G}} \|g\|_{\mathcal{G}}. \tag{10}$$

Again, we first provide bounds for $\left\| g_* - T_{\mathcal{G}}^{-1/2} \tilde{g}_0 \right\|_2^2$ and $\left\| \mathcal{T}^* \left(T_{\mathcal{G}}^{1/2} g_* - \tilde{g}_0 \right) \right\|_2^2$:

THEOREM 6 (Regularization bias for g_0): *Under an assumption analogous to 12 for the integral operator $T_{K_{\mathcal{G}}}$ and assumption 16, we have:*

$$\left\| g_* - T_{\mathcal{G}}^{-1/2} \tilde{g}_0 \right\|_2^2 \leqslant \|z_0\|_2^2 \left(\frac{\tilde{\mu}}{4} \right)^{\max\left\{0,\tilde{\beta}-2\right\}} \lambda_{\mathcal{G}}^{\min\left\{\tilde{\beta},2\right\}}$$

and

$$\left\| \mathcal{T}^* \left(T_{\mathcal{G}}^{1/2} g_* - \tilde{g}_0 \right) \right\|_2^2 \leqslant 4 \|z_0\|_2^2 \left(\frac{\tilde{\mu}}{4} \right)^{\max\left\{0,\tilde{\beta}-1\right\}} \lambda_{\mathcal{G}}^{\min\left\{\tilde{\beta}+1,2\right\}},$$

where $\tilde{\mu} = \sup_j \mu_j < \infty$.

To obtain the other convergent result, we need the following assumption:

ASSUMPTION 18 (Lipschitz condition between $\mathcal{G}$ and $\mathcal{H}'$): There exists $L > 0$ such that for all $g \in \mathcal{G}$, $\|\mathcal{T}^* g\|_{\mathcal{H}'} \leqslant L \|g\|_{\mathcal{G}}$.

THEOREM 7: *Suppose the integral operator $T_{K_{\mathcal{G}}}$ satisfies an assumption analogous to 12, that assumptions 16-18 hold, and that $|\alpha_0(X,L,W)| \leqslant B$ and $\|\tilde{g}_0\|_{\mathcal{G}} \leqslant B$ for some B. Let $B_1 = B+1$ and $B_2 = \frac{1}{2}L(2B+1)$, and define the function classes*

$$\mathcal{G}_{B_1} \cdot \mathcal{H}'_{B_2} := \left\{ g \cdot h | g \in \mathcal{G}_{B_1}, h \in \mathcal{H}'_{B_2} \right\} \quad \text{and} \quad \mathcal{H}'_{B_2} \circ q = \left\{ h(q(x,l),l,w) \cdot I\left\{(x,l)\in\mathcal{S}\right\} | h \in \mathcal{H}'_{B_2} \right\}.$$

Let $n \geqslant 3$, and let $\delta_n^2 \geqslant c_1 \frac{\log(\log(n))}{n}$ be an upper bound on the critical radius of star$(\mathcal{G}_{B_1} \cdot \mathcal{H}'_{B_2})$, star$(\mathcal{H}'_{B_2} \circ q)$, star$(\mathcal{G}_{B_1})$, and star$(\mathcal{H}'_{B_2})$. If $\hat{\Gamma} \in \mathcal{G}_{B_1}$, then the estimator $\hat{\Gamma}$ satisfies, with high probability,

$$\left\| \hat{\Gamma} - \tilde{g}_0 \right\|_2^2 = O\left(\frac{\delta_n^2}{\lambda_{\mathcal{G}}} + \frac{\lambda_{\mathcal{H}'}}{\lambda_{\mathcal{G}}} + \|z_0\|_2^2 \lambda_{\mathcal{G}}^{\min\{\tilde{\beta},1\}} \right),$$

and

$$\left\| \tilde{I}_q \circ \mathcal{T}^* \left(\hat{\Gamma} - \tilde{g}_0 \right) \right\|_2^2 = O\left(\delta_n^2 + \lambda_{\mathcal{H}'} + \|z_0\|_2^2 \lambda_{\mathcal{G}}^{\min\left\{\tilde{\beta}+1,2\right\}} \right).$$

B.2 *Estimators Using Reproducing Kernel Hilbert Spaces*

In this section, we derive closed-form expressions for the estimators of the bridge functions using reproducing kernel Hilbert spaces. Recall that the estimator $\hat{h}$ for h_0 is defined as the solution to the regularized minimax problem

$$\hat{h} = \arg\min_{h \in \mathcal{H}} \max_{g \in \mathcal{G}'} \mathbb{E}_n \left[g(X, L, Z) \left\{ Y - h(X, L, W) \right\} - g(X, L, Z)^2 \right] - \lambda_{\mathcal{G}'} \|g\|_{\mathcal{G}'} + \lambda_{\mathcal{H}} \|h\|_{\mathcal{H}}^2.$$

We begin by analyzing the inner maximization. Specifically, for any function $h \in \mathcal{H}$, we show that the inner problem satisfies

$$\begin{aligned} &\max_{g \in \mathcal{G}'} \mathbb{E}_n \left[g(X, L, Z) \left\{ Y - h(X, L, W) \right\} - g(X, L, Z)^2 \right] - \lambda_{\mathcal{G}'} \|g\|_{\mathcal{G}'}^2 \\ &\qquad = \frac{1}{4} \left\{ \xi_n(h) \right\}^T \left[\frac{1}{n} K_{\mathcal{G}',n} + \lambda_{\mathcal{G}'} I_n \right]^{-1} K_{\mathcal{G}',n} \left\{ \xi_n(h) \right\}, \end{aligned} \tag{11}$$

where $K_{\mathcal{G}',n}$ is the $n \times n$ matrix with (i,j)-th entry given by $K_{\mathcal{G}'}\left((x_i, l_i, z_i), (x_j, l_j, z_j)\right)$, $\xi_n(h) = \frac{1}{n}\left[\tilde{Y}_n - \tilde{h}_n\right]$, $\tilde{Y}_n = (y_1, \ldots, y_n)^T$ and $\tilde{h}_n = [h(x_1, l_1, w_1), \ldots, h(x_n, l_n, w_n)]^T$.

By the generalized representer theorem (see Schölkopf, Herbrich, and Smola (2001)), the solution to the maximization problem (11) has the form

$$g(x, l, z) = \sum_{j=1}^{n} \alpha_j K_{\mathcal{G}'}[(x, l, z), (x_j, l_j, z_j)],$$

where $\alpha = (\alpha_1, \ldots, \alpha_n)^T \in \mathbb{R}^n$. This result ensures that the solution is fully characterized by the observed data through the kernel evaluations. For $i = 1, \ldots, n$, we can write $g(x_i, l_i, z_i)$ as:

$$g(x_i, l_i, z_i) = e_i^T K_{\mathcal{G},n} \alpha$$

where e_i is the i-th standard basis vector in $\mathbb{R}^n$. Using this representation, the terms in (11)

can be written as follows:

$$\mathbb{E}_n\left[Yg(X,L,Z)\right] = \frac{1}{n}\sum_{i=1}^{n} y_i e_i^T K_{\mathcal{G}',n}\alpha = \frac{1}{n}\tilde{Y}_n^T K_{\mathcal{G}',n}\alpha,$$

$$\mathbb{E}_n\left[g(X,L,Z)h(X,L,W)\right] = \frac{1}{n}\sum_{i=1}^{n} h(x_i,l_i,w_i)e_i^T K_{\mathcal{G}',n}\alpha = \frac{1}{n}\tilde{h}_n^T K_{\mathcal{G}',n}\alpha,$$

$$\mathbb{E}_n\left[g(X,L,Z)^2\right] = \frac{1}{n}\sum_{i=1}^{n}\left[\alpha^T K_{\mathcal{G}',n} e_i e_i^T K_{\mathcal{G}',n}\alpha\right] = \frac{1}{n}\alpha^T K_{\mathcal{G}',n}^2\alpha,$$

$$\|g\|_{\mathcal{G}'}^2 = \sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j K_{\mathcal{G}'}((x_i,l_i,z_i),(x_j,l_j,z_j)) = \alpha^T K_{\mathcal{G}',n}\alpha.$$

Substituting these expressions into (11), we obtain

$$\begin{aligned}&\mathbb{E}_n\left[g(X,L,Z)\left\{Y-h(X,L,W)\right\} - g(X,L,Z)^2\right] - \lambda_{\mathcal{G}'}\|g\|_{\mathcal{G}'}^2\\ &= \frac{1}{n}\alpha^T K_{\mathcal{G}',n}\left[\tilde{Y}_n - \tilde{h}_n\right] - \alpha^T\left[\frac{1}{n}K_{\mathcal{G}',n}^2 + \lambda_{\mathcal{G}'}K_{\mathcal{G}',n}\right]\alpha.\end{aligned}$$

This is a convex quadratic problem with a unique maximum. Taking the derivative with respect to α and setting it to zero, the solution is

$$\alpha^* = \frac{1}{2n}\left[\frac{1}{n}K_{\mathcal{G}',n} + \lambda_{\mathcal{G}'}I_n\right]^{-1}\left[\tilde{Y}_n - \tilde{h}_n\right].$$

Substituting α^* back into the objective, we have

$$\begin{aligned}&\max_{g\in\mathcal{G}'}\mathbb{E}_n\left[g(X,L,Z)\left\{Y-h(X,L,W)\right\} - g(X,L,Z)^2\right] - \lambda_{\mathcal{G}'}\|g\|_{\mathcal{G}'}^2\\ &= \frac{1}{2n^2}\left[\tilde{Y}_n - \tilde{h}_n\right]^T\left[\frac{1}{n}K_{\mathcal{G}',n} + \lambda_{\mathcal{G}'}I_n\right]^{-1}K_{\mathcal{G}',n}\left[\tilde{Y}_n - \tilde{h}\right]\\ &\quad - \frac{1}{4n^2}\left\{\left[\frac{1}{n}K_{\mathcal{G}',n} + \lambda_{\mathcal{G}'}I_n\right]^{-1}\left[\tilde{Y}_n - \tilde{h}_n\right]\right\}^T K_{\mathcal{G}',n}\left[\tilde{Y}_n - \tilde{h}_n\right]\\ &= \left\{\xi_n(h)\right\}^T\Gamma_{\mathcal{G}'}K_{\mathcal{G}',n}\left\{\xi_n(h)\right\},\end{aligned}$$

with $\Gamma_{\mathcal{G}'} = \frac{1}{4}\left[\frac{1}{n}K_{\mathcal{G}',n} + \lambda_{\mathcal{G}'}I_n\right]^{-1}$.

The original optimization problem reduces to the following:

$$\hat{h} = \arg\min_{h\in\mathcal{H}}\left\{\xi_n(h)\right\}^T\Gamma_{\mathcal{G}'}K_{\mathcal{G}',n}\left\{\xi_n(h)\right\} + \lambda_{\mathcal{H}}\|h\|_{\mathcal{H}}^2.$$

By the representer theorem, the solution to this minimization problem takes the form

$$\hat{h}(x,l,w) = \sum_{j=1}^{n}\gamma_j K_{\mathcal{H}}((x,l,w),(x_j,l_j,w_j)),$$

where $\gamma = (\gamma_1, \ldots, \gamma_n)^T$ is the coefficient vector to be determined. Substituting this representation into the objective function, we obtain:

$$\frac{1}{n^2}\tilde{Y}_n^T \Gamma_{\mathcal{G}'} K_{\mathcal{G}',n} \tilde{Y}_n - \frac{2}{n^2}\gamma^T K_{\mathcal{H},n} \Gamma_{\mathcal{G}'} K_{\mathcal{G}',n} \tilde{Y}_n + \gamma^T \left[\frac{1}{n^2} K_{\mathcal{H},n} \Gamma_{\mathcal{G}'} K_{\mathcal{G}',n} K_{\mathcal{H},n} + \lambda_{\mathcal{H}} K_{\mathcal{H},n}\right] \gamma.$$

This problem is convex, ensuring a unique minimizer. Taking the derivative with respect to γ and setting it to zero yields the optimal solution:

$$\gamma^* = \left(K_{\mathcal{H},n} \Gamma_{\mathcal{G}'} K_{\mathcal{G}',n} K_{\mathcal{H},n} + n^2 \lambda_{\mathcal{H}} K_{\mathcal{H},n}\right)^{\dagger} K_{\mathcal{H},n} \Gamma_{\mathcal{G}'} K_{\mathcal{G}',n} \tilde{Y}_n,$$

where $\dagger$ denotes the Moore-Penrose pseudoinverse. It remains to ensure that $\left\|\hat{h}\right\|_{\mathcal{H}} \leqslant B_1$ for some pre-specified B_1. If $(\gamma^*)^T K_{\mathcal{H},n} \gamma^* \leqslant B_1^2$, the condition is satisfied. Otherwise, by Exercise 4.22 of Boyd (2004), the optimal γ^* is given by

$$\gamma^* = \left(P + \frac{\tilde{u}}{B_1^2} K_{\mathcal{H},n}\right)^{-1} \tilde{q},$$

where $P = \frac{1}{n^2} K_{\mathcal{H},n} \Gamma_{\mathcal{G}'} K_{\mathcal{G}',n} K_{\mathcal{H},n} + \lambda_{\mathcal{H}} K_{\mathcal{H},n}$, $\tilde{q} = \frac{1}{n^2} K_{\mathcal{H},n} \Gamma_{\mathcal{G}'} K_{\mathcal{G}',n} \tilde{Y}_n$, and $\tilde{u}$ is the largest solution to the nonlinear equation:

$$\tilde{q}^T K_{\mathcal{H},n}^{-1/2} \left(K_{\mathcal{H},n}^{-1/2} P K_{\mathcal{H},n}^{-1/2} + \frac{u}{B_1^2} I_n\right)^{-2} K_{\mathcal{H},n}^{-1/2} \tilde{q} = B_1^2.$$

For the estimator $\hat{g}$ for g_0, recall that

$$\hat{g}(x, l, z) = I\left\{x \in q(\mathcal{S}(l), l)\right\} \cdot \hat{\Gamma}(x, l, z)$$

where $\hat{\Gamma}$ is defined as the solution to the optimization problem

$$\begin{aligned}\hat{\Gamma} = \arg\min_{g \in \mathcal{G}} \max_{h \in \mathcal{H}'} \mathbb{E}_n \Big[& h\left\{q(X, L), L, W\right\} I\left\{(X, L) \in \mathcal{S}\right\} \\ & - I\left\{X \in q(\mathcal{S}(L), L)\right\} h(X, L, W) g(X, L, Z) \\ & - I\left\{X \in q(\mathcal{S}(L), L)\right\} h(X, L, W)^2 \Big] - \lambda_{\mathcal{H}'} \|h\|_{\mathcal{H}'} + \lambda_{\mathcal{G}} \|g\|_{\mathcal{G}}^2.\end{aligned}$$

We first show that for any $g \in \mathcal{G}$, we have

$$\begin{aligned}\max_{h \in \mathcal{H}'} \mathbb{E}_n \Big[& h\left\{q(X, L), L, W\right\} I\left\{(X, L) \in \mathcal{S}\right\} - I\left\{X \in q(\mathcal{S}(L), L)\right\} h(X, L, W) g(X, L, Z) \\ & - I\left\{X \in q(\mathcal{S}(L), L)\right\} h(X, L, W)^2 \Big] - \lambda_{\mathcal{H}'} \|h\|_{\mathcal{H}'} \\ = & \frac{1}{4} \left\{\zeta_n(g)\right\}^T \left[\frac{1}{n} K_{\mathcal{H}',n} R_n K_{\mathcal{H}',n} + \lambda_{\mathcal{H}'} K_{\mathcal{H}',n}\right]^{-1} \left\{\zeta_n(g)\right\},\end{aligned} \tag{12}$$

where $K_{\mathcal{H}',n}$ is the $n \times n$ matrix with (i,j)-th entry given by $K_{\mathcal{H}'}\left((x_i, l_i, w_i), (x_j, l_j, w_j)\right)$. The term $\zeta_n(g)$ is given by $\zeta_n(g) = \frac{1}{n}\left[\tilde{K}_{\mathcal{H},n}^T \tilde{d}_n - K_{\mathcal{H}',n} R_n \tilde{g}_n\right]$, where $\tilde{g}_n = [g(x_1, l_1, z_1), \ldots, g(x_n, l_n, z_n)]^T$, $\tilde{d}_n$ is the n-dimensional vector with entries $d_i = I\left\{(x_i, l_i) \in \mathcal{S}\right\}$ for $i = 1, \ldots, n$, and R_n is the diagonal matrix with values $r_i = I\left\{x_i \in q\left(\mathcal{S}(l_i), l_i\right)\right\}$ for $i = 1, \ldots, n$ in its diagonal. $\tilde{K}_{\mathcal{H}',n}$ is the $n \times n$ matrix with (i,j)-th entry given by $K_{\mathcal{H}'}[(q(x_i, l_i), l_i, w_i), (x_j, l_j, w_j)]$.

The generalized representer theorem ensures that the solution to (12) has the form

$$h(x, l, w) = \sum_{j=1}^{n} \alpha_j K_{\mathcal{H}'}[(x, l, w), (x_j, l_j, w_j)],$$

where $\alpha = (\alpha_1, \ldots, \alpha_n)^T \in \mathbb{R}^n$. The terms in the optimization problem can now be written as follows:

$$\begin{aligned}
\mathbb{E}_n\left[h\left\{q(X, L), L, W\right\} \cdot I\left\{(X, L) \in \mathcal{S}\right\}\right] &= \frac{1}{n}\sum_{i=1}^{n} d_i \tilde{K}_{\mathcal{H}',n}\alpha = \frac{1}{n}\tilde{d}_n^T \tilde{K}_{\mathcal{H}',n}\alpha, \\
\mathbb{E}_n\left[I\left\{X \in q(\mathcal{S}(L), L)\right\} h(X, L, W) g(X, L, Z)\right] &= \frac{1}{n}\sum_{i=1}^{n} g(x_i, l_i, z_i) r_i e_i^T K_{\mathcal{H}',n}\alpha = \frac{1}{n}\tilde{g}_n^T R_n K_{\mathcal{H}',n}\alpha, \\
\mathbb{E}_n\left[I\left\{X \in q(\mathcal{S}(L), L)\right\} h(X, L, W)^2\right] &= \frac{1}{n}\sum_{i=1}^{n}\left[\alpha^T K_{\mathcal{H}',n} e_i r_i e_i^T K_{\mathcal{H}',n}\alpha\right] \\
&= \frac{1}{n}\alpha^T K_{\mathcal{H}',n} R_n K_{\mathcal{H}',n}\alpha, \\
\|h\|_{\mathcal{H}'}^2 &= \alpha^T K_{\mathcal{H}',n}\alpha.
\end{aligned}$$

Substituting these expressions into the original optimization problem, we obtain

$$\begin{aligned}
&\max_{h \in \mathcal{H}'} \mathbb{E}_n\Big[h\left\{q(X, L), L, W\right\} I\left\{(X, L) \in \mathcal{S}\right\} - I\left\{X \in q(\mathcal{S}(L), L)\right\} h(X, L, W) g(X, L, Z) \\
&\qquad - I\left\{X \in q(\mathcal{S}(L), L)\right\} h(X, L, W)^2\Big] - \lambda_{\mathcal{H}'}\|h\|_{\mathcal{H}'} \\
&= \frac{1}{n}\alpha^T\left[\tilde{K}_{\mathcal{H}',n}^T \tilde{d}_n - K_{\mathcal{H}',n} R_n \tilde{g}_n\right] - \alpha^T\left[\frac{1}{n} K_{\mathcal{H}',n} R_n K_{\mathcal{H}',n} + \lambda_{\mathcal{H}'} K_{\mathcal{H}',n}\right]\alpha.
\end{aligned}$$

The solution to this convex quadratic problem is

$$\alpha^* = \frac{1}{2n}\left[\frac{1}{n} K_{\mathcal{H}',n} R_n K_{\mathcal{H}',n} + \lambda_{\mathcal{H}'} K_{\mathcal{H}',n}\right]^{-1}\left[\tilde{K}_{\mathcal{H}',n}^T \tilde{d}_n - \tilde{K}_{\mathcal{H}',n} R_n \tilde{g}_n\right].$$

Substituting α^* back into the objective function, we get

$$\begin{aligned}
&\max_{h\in\mathcal{H}'}\mathbb{E}_n\Big[h\left\{q(X,L),L,W\right\}I\left\{(X,L)\in\mathcal{S}\right\}-I\left\{X\in q(\mathcal{S}(L),L)\right\}h(X,L,W)g(X,L,Z)\\
&\qquad -I\left\{X\in q(\mathcal{S}(L),L)\right\}h(X,L,W)^2\Big]-\lambda_{\mathcal{H}'}\|h\|_{\mathcal{H}'}\\
&\quad=\frac{1}{4n^2}\left[\tilde{K}_{\mathcal{H}',n}^T\tilde{d}_n-K_{\mathcal{H}',n}R_n\tilde{g}_n\right]^T\left[\frac{1}{n}K_{\mathcal{H}',n}R_nK_{\mathcal{H}',n}+\lambda_{\mathcal{H}'}K_{\mathcal{H}',n}\right]^{-1}\left[\tilde{K}_{\mathcal{H}',n}^T\tilde{d}_n-K_{\mathcal{H}',n}R_n\tilde{g}_n\right]\\
&\quad=\frac{1}{4}\left\{\zeta_n(g)\right\}^T\left[\frac{1}{n}K_{\mathcal{H}',n}R_nK_{\mathcal{H}',n}+\lambda_{\mathcal{H}'}K_{\mathcal{H}',n}\right]^{-1}\left\{\zeta_n(g)\right\}.
\end{aligned}$$

The original optimization problem simplifies to:

$$\hat{\Gamma}=\arg\min_{g\in\mathcal{G}}\frac{1}{4}\left\{\zeta_n(g)\right\}^T\left[\frac{1}{n}K_{\mathcal{H}',n}R_nK_{\mathcal{H}',n}+\lambda_{\mathcal{H}'}K_{\mathcal{H}',n}\right]^{-1}\left\{\zeta_n(g)\right\}+\lambda_{\mathcal{H}}\|h\|_{\mathcal{H}}^2.$$

By the representer theorem, the solution to this optimization problem can be expressed as:

$$\hat{\Gamma}(x,l,z)=\sum_{j=1}^{n}\theta_jK_{\mathcal{G}}((x,l,z),(x_j,l_j,z_j)),$$

with $\theta=(\theta_1,\ldots,\theta_n)^T\in\mathbb{R}^n$. Let $\Gamma_{\mathcal{H}'}=\frac{1}{4}\left[\frac{1}{n}K_{\mathcal{H}',n}R_n+\lambda_{\mathcal{H}'}I_n\right]^{-1}$. Then, the objective function becomes:

$$\begin{aligned}
&\frac{1}{4n^2}\tilde{d}_n^T\tilde{K}_{\mathcal{H}',n}\left[\frac{1}{n}K_{\mathcal{H}',n}R_nK_{\mathcal{H}',n}+\lambda_{\mathcal{H}'}K_{\mathcal{H}',n}\right]^{-1}K_{\mathcal{H}',n}^T\tilde{d}_n\\
&\qquad-\frac{2}{n^2}\theta^TK_{\mathcal{G},n}R_n\Gamma_{\mathcal{H}'}\tilde{K}_{\mathcal{H}',n}^T\tilde{d}_n+\theta^T\left[\frac{1}{n^2}K_{\mathcal{G},n}R_n\Gamma_{\mathcal{H}'}K_{\mathcal{H}',n}R_nK_{\mathcal{G},n}+\lambda_{\mathcal{G}}K_{\mathcal{G},n}\right]\theta.
\end{aligned}$$

This problem is convex, and its solution is:

$$\theta^*=\left(K_{\mathcal{G},n}R_n\Gamma_{\mathcal{H}'}K_{\mathcal{H}',n}R_nK_{\mathcal{G},n}+n^2\lambda_{\mathcal{G}}K_{\mathcal{G},n}\right)^{\dagger}K_{\mathcal{G},n}R_n\Gamma_{\mathcal{H}'}\tilde{K}_{\mathcal{H}',n}^T\tilde{d}_n.$$

If $\left(\theta^*\right)^TK_{\mathcal{G},n}\theta^*>B_1^2$, by Exercise 4.22 of Boyd (2004), the optimal θ^* that ensures $\left\|\hat{\Gamma}\right\|_{\mathcal{G}}\leqslant B_1$ is given by

$$\theta^*=\left(P+\frac{\tilde{u}}{B_1^2}K_{\mathcal{G},n}\right)^{-1}\tilde{q},$$

where $P=\frac{1}{n^2}K_{\mathcal{G},n}R_n\Gamma_{\mathcal{H}'}K_{\mathcal{H}',n}R_nK_{\mathcal{G},n}+\lambda_{\mathcal{G}}K_{\mathcal{G},n}$, $\tilde{q}=\frac{1}{n^2}K_{\mathcal{G},n}R_n\Gamma_{\mathcal{H}'}\tilde{K}_{\mathcal{H}',n}^T\tilde{d}_n$, and $\tilde{u}$ is the largest solution to the nonlinear equation

$$\tilde{q}^TK_{\mathcal{G},n}^{-1/2}\left(K_{\mathcal{G},n}^{-1/2}PK_{\mathcal{G},n}^{-1/2}+\frac{u}{B_1^2}I_n\right)^{-2}K_{\mathcal{G},n}^{-1/2}\tilde{q}=B_1^2.$$

B.3 *Estimators Accounting for Two-Phase Sampling*

Consider a two-phase sampling procedure where each observation $i \in \{1, \ldots, n\}$ has an exposure measurement indicator Δ_i, which equals 1 if the measurement is available, and a weight S_i, representing the number of observations the i-th unit represents. To account for this sampling design, the estimator $\hat{h}$ for h_0 is defined as the solution to the following weighted optimization problem:

$$\hat{h} = \arg\min_{h \in \mathcal{H}} \max_{g \in \mathcal{G}'} \mathbb{E}_n \left[\Delta \cdot S \left\{ g(X, L, Z) \left[Y - h(X, L, W) \right] - g(X, L, Z)^2 \right\}\right] - \lambda_{\mathcal{G}'} \|g\|_{\mathcal{G}'} + \lambda_{\mathcal{H}} \|h\|_{\mathcal{H}}^2.$$

Let $m = \sum_{i=1}^{n} \Delta_i$ denote the number of observations with available exposure measurements, and let $\ell_1, \ldots, \ell_m$ be the indexes of these observations. Using an approach similar to the one described in the previous section, the closed-form solution for $\hat{h}$ is given by

$$\hat{h}(x, l, w) = \sum_{j=1}^{m} \gamma_j^* K_{\mathcal{H}}((x, l, w), (x_{\ell_j}, l_{\ell_j}, w_{\ell_j})),$$

where

$$\gamma^* = \left(K_{\mathcal{H},m} S_m \Gamma_{\mathcal{G}'} K_{\mathcal{G}',m} S_m K_{\mathcal{H},m} + n^2 \lambda_{\mathcal{H}} K_{\mathcal{H},m} \right)^{\dagger} K_{\mathcal{H},m} S_m \Gamma_{\mathcal{G}'} K_{\mathcal{G}',m} S_m \tilde{Y}_m.$$

In this expression, $K_{\mathcal{H},m}$ and $K_{\mathcal{G}',m}$ are the $m \times m$ empirical kernel matrices evaluated at the observations indexed by $\ell_1, \ldots, \ell_m$, S_m is an $m \times m$ diagonal matrix containing the weights $s_{n_1}, \ldots, s_{n_m}$ on its diagonal, $\Gamma_{\mathcal{G}'} = \frac{1}{4} \left[\frac{1}{n} S_m K_{\mathcal{G}',m} + \lambda_{\mathcal{G}'} I_m \right]^{-1}$, and $\tilde{Y}_m = (y_{\ell_1}, \ldots, y_{\ell_m})^T$. If $(\gamma^*)^T K_{\mathcal{H},n} \gamma^* > B_1$ for the pre-specified B_1, we obtain a constrained solution as outlined in the previous section.

Similarly, the closed-form expression for $\hat{\Gamma}$ is given by

$$\hat{\Gamma}(x, l, z) = \sum_{j=1}^{n} \theta_j^* K_{\mathcal{G}}((x, l, z), (x_{\ell_j}, l_{\ell_j}, z_{\ell_j})),$$

where the coefficients are

$$\theta^* = \left(K_{\mathcal{G},m} S_m R_m \Gamma_{\mathcal{H}'} K_{\mathcal{H}',m} R_m S_m K_{\mathcal{G},m} + n^2 \lambda_{\mathcal{G}} K_{\mathcal{G},m} \right)^{\dagger} K_{\mathcal{G},m} S_m R_m \Gamma_{\mathcal{H}'} \tilde{K}_{\mathcal{H}',m}^T S_m \tilde{d}_m.$$

Here, $K_{\mathcal{G},m}$ and $K_{\mathcal{H}',m}$ are the $m \times m$ empirical kernel matrices evaluated at the observations indexed by $\ell_1, \ldots, \ell_m$, $\tilde{K}_{\mathcal{H}',m}$ is given by $\tilde{K}_{\mathcal{H}',m} := \left[\; K_{\mathcal{H}'} \left[(q(x_{\ell_i}, l_{\ell_i}), l_{\ell_i}, w_{\ell_i}), (x_{\ell_j}, l_{\ell_j}, w_{\ell_j}) \right] \; \right]$,

R_m is an $m \times m$ diagonal matrix containing the entries $I\left\{X_{\ell_j} \in q(\mathcal{S}(L_{\ell_j}), L_{\ell_j})\right\}$ for $j = 1, \ldots, m$ on its diagonal, $\Gamma_{\mathcal{H}'} = \frac{1}{4}\left[\frac{1}{n} K_{\mathcal{H}',m} S_m R_m + \lambda_{\mathcal{H}'} I_m\right]^{-1}$, and $\tilde{d}_m$ is the m-dimensional column vector with entries $d_i = I\left\{(X_{\ell_i}, L_{\ell_i}) \in \mathcal{S}\right\}$ for $i = 1, \ldots, m$. If $(\theta^*)^T K_{\mathcal{G},n} \theta^* > B_1$, a constrained solution can be obtained as described previously. The estimator $\hat{g}$ for g_0 is $\hat{g}(x, l, z) = I\left\{x \in q(\mathcal{S}(l), l)\right\} \cdot \hat{\Gamma}(x, l, z)$.

Finally, for $k = 1, \ldots, K$, the estimator $\hat{\psi}_k$ of the target parameter within fold k is computed using a standard weighting approach. Specifically, for policies q whose image lies within the support of X, the estimator $\hat{\psi}_k$ is given by

$$\hat{\psi}_k = \frac{1}{|I_k|} \sum_{i: O_i \in I_k} \Delta_i S_i \left\{ h^{(-k)}\left[q\left(X_i, L_i\right), L_i, W_i\right] + g^{(-k)}(X_i, L_i, Z_i)\left[Y_i - h^{(-k)}(X_i, L_i, W_i)\right]\right\}.$$

Web Appendix C. Numerical Experiments and Implementation

Recall that $\mathcal{TN}_{[a,b]}(\mu, \sigma^2)$ denotes the probability distribution obtained from truncating a normally distributed variable $\mathcal{N}(\mu, \sigma^2)$ within the interval $[a, b]$, and that $\mathcal{TN}_c(\mu, \sigma^2)$ denotes a specific case where $a = \mu - 3\sigma$ and $b = \mu + 3\sigma$. Here, we present a more general data-generating process (DGP). For each $i = 1, \ldots, n$, we generated variables with bounded support as follows: the observed confounder $L_i \sim \mathcal{TN}_c(0, 1)$, the unobserved confounder $U_i \sim \mathcal{TN}_c(\beta_1 L_i, 1)$, the negative controls $Z_i \sim \mathcal{TN}_c(\beta_2 L_i + \beta_3 U_i, 1)$ and $W_i \sim \mathcal{TN}_c(\beta_4 L_i + \beta_5 U_i, 1)$, the exposure $X_i \sim \mathcal{TN}_{[c,d]}(\beta_6 L_i + \beta_7 U_i + \beta_8 Z_i, 1)$, and the outcome $Y_i \sim \text{Bernoulli}\,(p_i)$ where $p_i = \{1 + \exp\left(-\left[\mu + \beta_9 L_i + \beta_{10} U_i + \beta_{11} X_i + \beta_{12} W_i + \gamma X_i^2\right]\right)\}^{-1}$. The DGP described in the main manuscript used $(\beta_1, \ldots, \beta_{12}) = (0.5, 0.2, \beta_z, 0.5, \beta_w, 0.3, 1, 0, 0.5, -1, -1.5, 0)$, $c = -2$, $d = 2$, $\mu = -1$, and $\gamma = -0.75$. This DGP additionally includes the parameter β_8 that allows the negative control treatment to have an effect on the primary treatment, and the parameter β_{12} that allows the negative control outcome affect the outcome of interest. In the main manuscript, we set $\beta_8 = \beta_{12} = 0$. We also consider the counterfactual MTP mean

under the more general shift policy:

$$q^{\delta,\varepsilon,a}(x) = x + \delta \cdot I\{c \leqslant x \leqslant d - \varepsilon - \delta\} + \frac{a \cdot \delta}{\delta + \varepsilon}[d - x] \cdot I\{d - \varepsilon - \delta < x \leqslant d\}, \tag{13}$$

for parameters $\delta > 0$, $\varepsilon \geqslant 0$, and $a \in \{0, 1\}$, satisfying $c + \delta < d - \varepsilon$, with $\varepsilon > 0$ when $a = 1$. In the main manuscript, we studied $\mathbb{E}\left[Y\left\{q^{\delta=0.4,\varepsilon=1,a=1}(X)\right\}\right]$. Note that policies $q^{\delta,\varepsilon,a}$ with $a = 0$ can be seen as shift interventions defined only in the restricted population $[c, d-\delta-\varepsilon]$.

C.1 *Verifying Assumption 1-8*

From the DGP, it is clear that assumptions 1-5 are satisfied. As for assumption 6, from expression 13, it is easy to see that $q^{\delta,\varepsilon,a}$ is strictly monotone and almost everywhere differentiable over $[c, d - (1 - a)(\delta + \varepsilon)]$, that is, over $[c, d]$ when $a = 1$ and over $[c, d - \delta - \varepsilon]$ when $a = 0$. Its inverse is well-defined over $[c+\delta, d-(1-a)\varepsilon]$ and it is given by $(q^{\delta,\varepsilon,a})^{-1}(x) = x - \delta V(x)$, where $V(x) = I\{c + \delta \leqslant x \leqslant d - \varepsilon\} + a \cdot \frac{d-x}{\varepsilon} \cdot I\{d - \varepsilon < x \leqslant d\}$.

To verify assumptions 7 and 8, we provide solutions to the bridge equations (2) and (3). For our DGP, when β_{10}/β_5 is small and $\beta_{12} = 0$ (no effect of W on Y), a good approximation to an outcome bridge function is:

$$h(x, l, w) \approx \left\{1 + \exp\left(\mu - \log\left(1 + 0.9733\varphi_3^2\right) + \beta_{10}x + \left[\beta_9 - \frac{\beta_4\beta_{10}}{\beta_5}\right] l + \frac{\beta_{10}}{\beta_5}w + \gamma x^2\right)\right\}^{-1}.$$

In section G.1 we provide an approximation when β_{10}/β_5 is not small or $\beta_{12} \neq 0$. Even though this is only an approximation to a solution of the outcome bridge equation (2), in section C.3, we show that estimators that rely on these parametric models have excellent performance.

As for the treatment bridge equation (3), under our DGP, an exact solution when $\beta_8 = 0$ (no effect of Z on X) is given by

$$g(x, l, z) = A(x) \exp\left(\delta\left\{x - \left[\beta_6 - \frac{\beta_2\beta_7}{\beta_3}\right] l - \frac{\beta_7}{\beta_3}z - \frac{\delta}{2}\left[1 + \frac{\beta_7^2}{\beta_3^2}\right] V(x)\right\} V(x)\right),$$

where

$$A(x) = \frac{\Phi(3) - \Phi(-3)}{\Phi\left(3 + \frac{\delta\beta_7}{\beta_3}V(x)\right) - \Phi\left(-3 + \frac{\delta\beta_7}{\beta_3}V(x)\right)}\left[I\{c + \delta \leqslant x \leqslant d\} + \frac{\delta}{\varepsilon} \cdot I\{d - \varepsilon < x \leqslant d\}\right],$$

and $\Phi(\cdot)$ denotes the CDF of a standard normal distribution. In Web Appendix G, we show that the proposed functions solve (approximately and exactly) the bridge equations (2) and (3), respectively.

C.2 *Cross-Validation to Select the Bridge Functions*

Motivated by the regularized minimax approach for estimating the bridge functions and the result in Theorem 4, we implement the cross-validation procedure described in Section 5 of the main manuscript, using the following empirical risk functions, defined for any functions $h \in \mathcal{L}^2(X, L, W)$ and $g \in \mathcal{L}^2(X, L, Z)$ as:

$$\mathcal{R}_n^1(h) := \sup_{g \in \mathcal{G}'} \mathbb{E}_n \left[g(X, L, Z) \left\{ Y - h(X, L, W) \right\} - g(X, L, Z)^2 \right] - \lambda_{\mathcal{G}'} \|g\|_{\mathcal{G}'}^2,$$

and

$$\begin{aligned} \mathcal{R}_n^2(g) := \sup_{h \in \mathcal{H}'} \mathbb{E}_n \big[& h \left\{ q(X, L), L, W \right\} \cdot I \left\{ (X, L) \in \mathcal{S} \right\} \\ & - I \left\{ X \in q(\mathcal{S}(L), L) \right\} \cdot h(X, L, W) g(X, L, Z) \\ & - I \left\{ X \in q(\mathcal{S}(L), L) \right\} \cdot h(X, L, W)^2 \big] - \lambda_{\mathcal{H}'} \|h\|_{\mathcal{H}'}^2. \end{aligned}$$

Suppose that a set of candidate bridge functions are estimated using fold 2 (training set) and tested in fold 3 (testing set). Let n_3 denote the number of observations in folds 3, and let $\left\{ \ell_1^{(3)}, \ldots, \ell_{n_3}^{(3)} \right\}$ represent the indices of these observations. When the function classes correspond to reproducing kernel Hilbert spaces, the empirical risk functions take the following forms:

$$\mathcal{R}_n^1(h) := \frac{1}{4} \left\{ \xi_{n_3}(h) \right\}^T \left[\frac{1}{n_3} K_{\mathcal{G}', n_3}^0 + \lambda_{\mathcal{H}'} I_{n_3} \right]^{-1} K_{\mathcal{G}', n_3}^0 \left\{ \xi_{n_3}(h) \right\}$$

and

$$\mathcal{R}_n^2(g) := \frac{1}{4} \left\{ \zeta_{n_3}(g) \right\}^T \left[\frac{1}{n_3} K_{\mathcal{H}', n_3}^0 R_{n_3} K_{\mathcal{H}', n_3}^0 + \lambda_{\mathcal{H}'} K_{\mathcal{H}', n_3}^0 \right]^{-1} K_{\mathcal{H}', n_3}^0 \left\{ \zeta_{n_3}(g) \right\}.$$

Here, $K_{\mathcal{G}', n_3}^0$ and $K_{\mathcal{H}', n_3}^0$ are $n_3 \times n_3$ matrices whose (i, j)-th entries are obtained by evaluating the kernel functions $K_{\mathcal{G}'}^0$ and $K_{\mathcal{H}'}^0$ at observations $\ell_i^{(3)}$ and $\ell_j^{(3)}$, and R_{n_3} is the diagonal matrix

with values $I\left\{x_{\ell_i^{(3)}} \in q\big(\mathcal{S}(l_{\ell_i^{(3)}}), l_{\ell_i^{(3)}}\big)\right\}$ for $i = 1, \ldots, n_3$ in its diagonal. The terms $\xi_{n_3}(h)$ and $\zeta_{n_3}(g)$ are defined as

$$\xi_{n_3}(h) = \frac{1}{n_3}\left[\tilde{Y}_{n_3} - \tilde{h}_{n_3}\right] \quad \text{and} \quad \zeta_{n_3}(h) = \frac{1}{n_3}\left[\left(\tilde{K}^0_{\mathcal{H}',n_3}\right)^T \tilde{d}_{n_3} - K^0_{\mathcal{H}',n_3} R_{n_3} \tilde{g}_{n_3}\right],$$

where $\tilde{Y}_{n_3} = \left(y_{\ell_1^{(3)}}, \ldots, y_{\ell_{n_3}^{(3)}}\right)^T$ is the vector of observed outcomes in fold 3. The term $\tilde{h}_n = \left[h(x_{\ell_1^{(3)}}, l_{\ell_1^{(3)}}, w_{\ell_1^{(3)}}), \ldots, h(x_{\ell_{n_3}^{(3)}}, l_{\ell_{n_3}^{(3)}}, w_{\ell_{n_3}^{(3)}})\right]^T$ represents the evaluations of h at the observations of fold 3, while $\tilde{g}_n = \left[g(x_{\ell_1^{(3)}}, l_{\ell_1^{(3)}}, z_{\ell_1^{(3)}}), \ldots, g(x_{\ell_{n_3}^{(3)}}, l_{\ell_{n_3}^{(3)}}, z_{\ell_{n_3}^{(3)}})\right]^T$ represents the evaluations of g at the same observations. The matrix $\tilde{K}^0_{\mathcal{H},n_3}$ is the $n_3 \times n_3$ is an matrix whose (i,j)-entry is given by $K^0_{\mathcal{H}'}\left\{(q(x_{\ell_i^{(3)}}, l_{\ell_i^{(3)}}), l_{\ell_i^{(3)}}, w_{\ell_i^{(3)}}), (x_{\ell_j^{(3)}}, l_{\ell_j^{(3)}}, w_{\ell_j^{(3)}})\right\}$. The bandwidth for the kernel function $K^0_{\mathcal{G}',n_3}$ is set to 1/4 times the median of the pairwise euclidean distance between the observed vectors $(X_{\ell_i}, L_{\ell_i}, Z_{\ell_i})$. The bandwidth for $K^0_{\mathcal{H}',n_3}$ is chosen analogously. The parameters $\lambda'_{\mathcal{G}}$ and $\lambda_{\mathcal{H}'}$ are set both to $\frac{\log(n_3)}{n_3}$.

C.3 *Proximal Parametric Estimators*

The data-generating mechanism suggests the following parametric model for h:

$$h(x, l, w; \varphi) = \left(1 + \frac{\mathbb{E}[\widetilde{W}^2]}{2}\varphi_3^2\right)\left\{1 + \exp\left(-[\varphi_0 + \varphi_1 x + \varphi_2 l + \varphi_3 w + \varphi_4 x^2]\right)\right\}^{-1}$$

and the following model for g:

$$g(x, l, z; \eta) = A(x)\exp\left\{[\eta_0 x + \eta_1 l + \eta_2 z + \eta_3 V(x)]\, V(x)\right\},$$

where $V(x) = I\{c + \delta \leqslant x \leqslant d - \varepsilon\} + \frac{d-x}{\varepsilon} \cdot I\{d - \varepsilon < x \leqslant d\}$ and

$$A(x) = \frac{\Phi(3) - \Phi(-3)}{\Phi\left(3 - \eta_2 V(x)\right) - \Phi\left(-3 - \eta_2 V(x)\right)}\left[I\left\{c + \delta \leqslant x \leqslant d\right\} + \frac{\delta}{\varepsilon} \cdot I\left\{d - \varepsilon \leqslant x \leqslant d\right\}\right].$$

Additionally, the integral equations (4) and (5) motivate parametric estimators $\hat{\varphi}$ for φ and $\hat{\eta}$ for η , obtained by solving the following estimating equations:

$$\mathbb{E}_n\left[\{Y - h(X, L, W; \varphi)\}\,(1, X, L, Z, X^2)^T\right] = (0,0,0,0,0)^T,$$

$$\mathbb{E}_n\left[I\left\{(X, L) \in \mathcal{S}\right\} \cdot \left(1, q^{\delta,\varepsilon,a}(X)\, L, W\right)^T - g(X, L, Z; \eta)(1, X, L, W)^T\right] = (0,0,0,0)^T.$$

Once estimators $\hat{\varphi}$ and $\hat{\eta}$ have been obtained, we compute the following three parametric

estimators:

$$\hat{\psi}_{p,OR} = \mathbb{E}_n \left[h\{q^{\delta,\varepsilon,a}(X, L), W, L; \hat{\varphi}\} \right],$$

$$\hat{\psi}_{p,IPW} = \mathbb{E}_n \left[Y \cdot g(X, Z, L; \hat{\eta}) \right],$$

$$\hat{\psi}_{p,DR} = \mathbb{E}_n \left[h\{q^{\delta,\varepsilon,a}(X, L), W, L; \hat{\varphi}\} + g(X, Z, L; \hat{\eta}) \cdot \{Y - h(X, W, L; \hat{\eta})\} \right].$$

We performed the numerical experiments for the DGPs described in the main manuscript. Given that the parametric approach is less computationally demanding, we considered sample sizes $n \in \{750, 3000, 12000, 48000\}$ across 16 scenarios formed by combining $\beta_3 \in \{2, 1, 0.5, 0.25\}$ and $\beta_5 \in \{-2, -1, -0.5, -0.25\}$. The values $\beta_3 = 0.25$ and $\beta_5 = -0.25$ correspond to conditional correlations $\text{Cor}(Z, U|L) = 0.235$ and $\text{Cor}(W, U|L) = -0.235$. For each configuration, we performed 500 repetitions of the data-generating and estimation procedures. Figure 1 presents boxplots of the three parametric estimators based on the parametric models used for estimating the bridge functions. The model for the treatment bridge function is correctly specified, while the model for the outcome bridge function is nearly correctly specified. For the data-generating mechanisms considered, we observed that the challenge of estimating the bridge functions under weak conditional correlation between the negative controls and the unmeasured confounder diminishes as the sample size increases.

[Figure 1 about here.]

C.4 *Performance of the Proximal Estimator Under No Unmeasured Confounding*

To assess the performance of our estimator in the absence of unmeasured confounding, we generated data under a data-generating mechanism similar to that described in Section 5 of the manuscript, but with β_7 set to zero. We focused solely on the scenario where $\beta_z = 2$ and $\beta_w = -2$, in which both proxies are highly correlated with the variable U. Under this configuration, Z and W act as precision variables. The target parameter for the same policy under all the data generating processes was 0.2081. Data were generated for sample

sizes $n \in \{750, 1500, 3000\}$. For each configuration, we performed 500 repetitions of the data-generating and estimation procedures. Figure 2 displays boxplots, variance ratios, and coverage probabilities for the proximal DR estimator and its competitors. In this setting, both proximal and non-proximal estimators show low bias, estimated asymptotic variances closely aligned with empirical variances, and coverage probabilities at the nominal level.

[Figure 2 about here.]

C.5 *Proximal Estimation of the Counterfactual MTP Mean in a Restricted Population*

To illustrate the performance of our proximal estimator for a target parameter as described in Strategy I—outlined in Example 1 of the main manuscript—we conducted numerical experiments using the data-generating processes described there, but only for $\beta_z = -\beta_w \in \{2, 1, 0.5\}$. Here, however, we focus on estimating $\mathbb{E}\{Y(X + 0.4)|X \in [-2, 1.6]\}$, the counterfactual MTP mean under a shift intervention applied to the restricted population $\{X : X \in [-2, 1.6]\}$, which under all data-generating mechanisms is 0.2728. This corresponds to the policy $q^{\delta=0.4,\varepsilon=0,a=1}(X)$ described previously, for which assumptions 1-8 have already been shown to hold. For each setting, we performed 500 repetitions of the data-generating and estimation procedures. Figure **??** presents boxplots, variance ratios, and coverage probabilities of the proximal DR estimator. Our proximal cross-fitted estimator demonstrates performance comparable to that observed for the estimand under Strategy II.

C.6 *Performance of the Proximal Estimator Under DGPs with arrows $Z \to X$ and $W \to Y$*

We also conducted numerical experiments to evaluate the performance of out estimator under data-generating processes where Z influences X and W affects Y. Specifically, we simulated data with bounded support using the DGP described at the beginning of Web Appendix C with $(\beta_1, \ldots, \beta_{12}) = (0.5, 0.2, \beta_z, 0.5, \beta_w, 0.3, 1, 0.3, 0.5, -1, -1.5, -0.3)$, $c = -2$, $d = 2$, $\mu = -1$, and $\gamma = -0.75$. We considered three scenarios with $\beta_z = -\beta_w \in \{2, 1, 0.5\}$. The target

estimand remained the same as in the main manuscript, namely $\mathbb{E}[Y\left\{q^{\delta=0.4,\varepsilon=1,a=1}(X)\right\}]$ which equals 0.2185 when $\beta_w = -2$, 0.2400 when $\beta_w = -1$, and 0.2488 when $\beta_w = -0.5$. For each setting, we performed 500 repetitions of the data-generating and estimation procedures. Figure **??** displays boxplots, variance ratios, and coverage probabilities for the proximal DR estimator and its competitors. As in the numerical experiments reported in the main manuscript, our proximal estimator exhibits low bias, with the average estimated asymptotic variance closely matching the empirical variance, and confidence interval coverage near the nominal level. The proximal estimator also outperforms the non-proximal alternatives in terms of bias and coverage, with this advantage diminishing as the conditional correlation between the negative controls and the unmeasured confounder weakens—a pattern consistent with the main manuscript results for the sample sizes explored, though only the proximal estimator is theoretically guaranteed to converge to the true parameter as sample size increases.

C.7 *Vignette Application to a Single Data Set*

We provide R code and a mock dataset simulating data from the ENSEMBLE trial in the GitHub repository to demonstrate the use of our `pmtp` R function for implementing the proximal one-step cross-fitted estimator.

The mock dataset `sim_trial_data.csv` include the following variables:

- Outcome: `Y` (taking values 0 or 1).
- Covariates: `L1, L2, L3`.
- Biomarker: univariate biomarker `X`.
- Negative control treatment: `Z`.
- Negative control outcome: `W`.
- Sampling weights: `wt`.

The function `pmtp()` estimates the counterfactual MTP mean $\mathbb{E}\left[Y\left\{q(X)\right\}|X \in \mathcal{S}\right]$ of a

policy $q(x)$ over a population of interest $\mathcal{S} \subset \text{supp}\,(X)$, using the proximal one-step cross-fitted estimator described in Section A.5. It estimates the bridge functions using reproducing kernel Hilbert spaces generated by Gaussian kernels as outlined in sections B.2, B.3, and C.2. It supports both binary and continuous outcomes, but does not handle time-to-event (censored) outcomes.

The `pmtp()` function requires the following inputs:

- `data`: A `data.frame` with the dataset to analyze.
- `trt`: String for the biomarker (treatment) variable name.
- `outcome`: String for outcome variable name.
- `covariates`: String or vector of strings for covariate(s) name(s).
- `nct`: String or vector of strings for negative control treatment variable(s) name(s).
- `nco`: String or vector of strings for negative control outcome variable(s) name(s).
- `policy`: A list of policy functions. Functions should depend only on the treatment variable.

The following inputs are optional. If not provided, default values will be used:

- `weights`: String for weights variable name.
- `ind_S`: A numeric vector of 0s and 1s indicating membership in the target population $\mathcal{S}$. Default is a vector of 1s (i.e., all units are in $\mathcal{S}$).
- `theta`: Scaling factor for the regularization parameter in the risk function used during cross-validation to select the optimal bridge functions. Default is `1`.
- `K_folds`: Integer indicating the number of folds used for cross-fitting. Must be at least 3. Default is `3`.
- `lm_H_list`, `lm_G_list`: Lists of scaling factors c_1 for the regularization parameters $c_1 \left(\frac{\log(n)}{n}\right)^{1/2}$ in the outer minimization step of the outcome and treatment bridge function estimation, respectively. Defaults are `10^seq(-5, -1, by = 1)`.
- `lm_Gh_list`, `lm_Hg_list`: Lists of scaling factors c_2 for the regularization parameters $c_2 \frac{\log(n)}{n}$

in the inner maximization step of the outcome and treatment bridge function estimation, respectively. Defaults are `10^seq(-1, 2, by = 1)`.

- `bw_int_scale_list`, `bw_ext_scale_list`: Lists of scaling factors for the bandwidths of the Gaussian kernels used in the internal and external steps of the bridge function estimation procedure. Defaults are `1/4` and `c(1/4,1/2,1,2,4)`, respectively.
- `bw_int_fixed_scale`: Scale factor used for the internal bandwidth of the Gaussian kernel used in the risk function for cross-validation. Default is `1/4`.
- `bw0_H`, `bw0_G`: Optional initial bandwidths for the function classes $\mathcal{H}$ and $\mathcal{H}'$, and $\mathcal{G}$ and $\mathcal{G}'$, respectively. Default is `NULL`. If not provided, the initial bandwidths are set to the median of the pairwise Euclidean distances between the corresponding observed vectors.
- `inf.fun`: Logical flag indicating whether to return the influence function evaluated at the sample observations. Default is `FALSE`.
- `show.prog`: Logical flag indicating whether to display progress messages. Default is `FALSE`.
- `control.folds`: Logical flag indicating whether fold assignment in cross-fitting should be stratified based on the distribution of the treatment variable. Default is `FALSE`.

The function outputs a `data.frame` with the one-step estimator for each specified policy and also returns a `proximal.mtp` object with detailed about the estimation procedure. In addition, the `summary.proximal.mtp()` function reports, for each specified policy, the proportion of observations with treatment value within the image of the policy, the proportion of observations whose treatment values fall within the image of the policy, along with standard errors and 95% confidence intervals for the estimators.

Input to conduct analysis on the mock dataset

We illustrate our estimator for the following two policies:

```
policy_q1 <- function(x) {
  (x+0.4)*(x+0.4<=3.5-1)+(0.4*3.5+1*x)/(0.4+1)*(x+0.4>3.5-1)
```

```
}
policy_q2 <- function(x) {
  (x+0.8)*(x+0.8<=3.5-1)+(0.8*3.5+1*x)/(0.8+1)*(x+0.8>3.5-1)
}

obj.pmtp = pmtp(data = sim_trial_data,
                trt = "X",
                outcome = "Y",
                covariates = c("L1", "L2", "L3"),
                nct = "Z",
                nco = "W",
                weights = "wt",
                policy = list(policy_q1, policy_q2),
                control.folds = TRUE)
```

Output using the `summary.proximal.mtp()` *function*

```
summary(obj.pmtp)


Summary of Proximal MTP Estimation
-------------------------------------------------
Number of observations: 1000
Number in target population: 1000 (100%)
-Proportion in the image of Policy_q^1: 38%
-Proportion in the image of Policy_q^2: 27%


-----------------------------------------------------------------
Proximal Doubly Robust Estimators and 95% Confidence Intervals:
```

```
------------------------------------------------------------------

Estimator: Proximal One-Step

           Estimate Std.Error CI.Lower CI.Upper

Policy_q^1   0.0146    0.0025   0.0098   0.0195

Policy_q^2   0.0066    0.0018   0.0030   0.0101
```

Web Appendix D. Supporting Lemmas

In this section, we present Lemmas that are used in the proofs of the results provided in Web Appendix A and Web Appendix B.

LEMMA 5 (Picard's Theorem): *Let $K : H_1 \to H_2$ be a compact operator with singular system $(\sigma_j, \varphi_j, \phi_j)_{j=1}^{\infty}$ and r_0 be a given function in H_2. Then, the equation of the first kind $Kh = r_0$ has solutions if and only if*

i) $r_0 \in \mathcal{N}(K^*)^{\perp}$, *where* $\mathcal{N}(K^*) = \{q : K^*q = 0\}$ *is the null space of the adjoint operator* K^*.

ii) $\sum_{j=1}^{\infty} \sigma_j^{-2} |\langle r, \phi_j \rangle|^2 < +\infty$.

LEMMA 6: *Under assumption 6, for any $h \in \mathcal{L}^2(X, L, W)$, it holds that*

$$\mathbb{E}\left[h\left\{q\left(X, L\right), L, W\right\} \cdot I\left\{(X, L) \in \mathcal{S}\right]\right\} = E\left\{h(X, L, W)\alpha_0(X, L, W)\right\}.$$

Lemma 6 implies that $\alpha_0(X, L, W)$ is the Riesz representer of the linear functional given by $h \mapsto \mathbb{E}\left[h\left\{q\left(X, L\right), L, W\right\} \cdot I\left\{(X, L) \in \mathcal{S}\right\}\right]$.

LEMMA 7: *Let $h^{\dagger}$ and $g^{\dagger}$ be any solutions of the observed equations (4) and (5), respectively. Then, under assumption 6, for any $h \in \mathcal{L}^2(X, L, W)$ and $g \in \mathcal{L}^2(X, L, Z)$, it holds*

that

$$\mathbb{E}\left\{\phi(O;h,g)-\phi(O;h^\dagger,g^\dagger)\right\}=\mathbb{E}\left[\left\{(h-h^\dagger)(X,L,W)\right\}\left\{(g^\dagger-g)(X,L,Z)\right\}\right].$$

In particular, it holds that

$$\mathbb{E}\left[h^\dagger\left\{q\left(X,L\right),L,W\right\}\cdot I\left\{(X,L)\in\mathcal{S}\right\}\right]=\mathbb{E}\left\{\phi(O;h^\dagger,g^\dagger)\right\}=\mathbb{E}\left\{Yg^\dagger\left(X,L,Z\right)\right\}$$

LEMMA 8: *Under assumptions 1 - 3, for any* $(x,l,u)\in supp(X,L,U)$ *such that* $x\in\mathcal{S}(l)$, *it holds that*

$$\mathbb{E}\left[Y\left\{q\left(X,L\right)\right\}|X=x,L=l,U=u\right]=\mathbb{E}\left\{Y|X=q\left(x,l\right),L=l,U=u\right\}.$$

Moreover, for any h_0 *solving equation (2), it holds that*

$$\mathbb{E}\left[Y\left\{q\left(X,L\right)\right\}|X=x,L=l,U=u\right]=\mathbb{E}\left[h_0\left\{q(X,L),L,W\right\}|X=x,L=l,U=u\right].$$

LEMMA 9 (Mean-squared-continuity):

i) Suppose $|\alpha_0(X,L,W)|\leqslant B$ *for some* B *and that assumption 6 holds. Then, for any* $h\in\mathcal{L}^2(X,L,W)$, *it holds that*

$$\left\|h\left\{q(X,L),L,W\right\}\cdot I\left\{(X,L)\in\mathcal{S}\right\}\right\|_2^2=O\left(\|h\|_2^2\right).$$

ii) Suppose $|Y|\leqslant B$ *for some* B. *Then, for any* $g\in\mathcal{L}^2(X,L,Z)$, *it holds that*

$$\left\|Yg(X,L,Z)\right\|_2^2=O\left(\|g\|_2^2\right).$$

LEMMA 10: *Let* h_0 *and* g_0 *be the minimum-norm solutions of equations (4) and (5), respectively, and assume that* $\hat{h}$ *and* $\hat{g}$ *are norm consistent estimators for* h_0 *and* g_0 *in the sense that* $\left\|\hat{h}-h_0\right\|_2=o_p(1)$ *and* $\|\hat{g}-g_0\|_2=o_p(1)$. *Suppose* $|\alpha_0(X,L,W)|\leqslant B$, $|Y|\leqslant B$, *either* $\|h_0\|_\infty+\|\hat{g}\|_\infty\leqslant B$ *or* $\|\hat{h}\|_\infty+\|g_0\|_\infty\leqslant B$ *for some* B, *and that assumption 6 holds. Then* $\phi(O;\hat{h},\hat{g})$ *is a norm consistent estimator of* $\phi(O;h_0,g_0)$, *i.e.,* $\left\|\phi(O;\hat{h},\hat{g})-\phi(O;h_0,g_0)\right\|_2=o_p(1)$. *Furthermore, if* $\hat{h}$ *and* $\hat{g}$ *are independent of* $O_1,\ldots,O_n$, $\mathbb{E}_n\left[\phi(O;\hat{h},\hat{g})\right]$ *converges in probability to* $\mathbb{E}\left\{\phi(O;h_0,g_0)\right\}=\phi_0$.

LEMMA 11 (Talagrand concentration for empirical process, Theorem 3.27 of Wainwright (2019)): *Consider a countable class of functions $\mathcal{F}$, uniformly bounded by b, and define*

$$Z := \sup_{f \in \mathcal{F}} \left\{ \frac{1}{n} \sum_{i=1}^{n} f(X_i) \right\}.$$

Then, for all $u > 0$, the random variable Z satisfies the upper tail bound

$$P\{Z \geqslant \mathbb{E}(Z) + u\} \leqslant 2 \exp\left(-\frac{nu^2}{8ev^2 + 4bu} \right),$$

where $v^2 = 2b\mathbb{E}(Z) + \sup_{f \in \mathcal{F}} \mathbb{E}\{f(X)^2\}$.

While Lemma 11 assumes a countable class of functions, our methodology extends to uncountable classes. To account for this, we adapt a result inspired by the ideas and techniques from Steinwart and Christmann (2008), which allows the application of Lemma 11 to certain uncountable uniformly bounded function classes.

LEMMA 12: *Let $\mathcal{F}$ be a separable space of measurable functions $f : \mathcal{X} \to \mathbb{R}$ equipped with a metric d. For each $f \in \mathcal{F}$, define the function $g(\cdot, f) : \mathcal{X}^n \to \mathbb{R}$ by*

$$g(x, f) := \frac{1}{n} \sum_{i=1}^{n} f(x_i) \quad \textit{for all } x = (x_1, \ldots, x_n) \in \mathcal{X}^n.$$

Assume that the mapping $f \mapsto g(x, f)$ is continuous for all $x \in \mathcal{X}^n$. Then, for any dense subset $\mathcal{S} \subset \mathcal{F}$ and any $x \in \mathcal{X}^n$,

$$\sup_{f \in \mathcal{F}} g(x, f) = \sup_{f \in \mathcal{S}} g(x, f).$$

Furthermore, the random variable

$$Z := \sup_{f \in \mathcal{F}} g(X_1, \ldots, X_n, f)$$

is well defined.

We now introduce notation that is required for the next three Lemmas. Let $\ell : \mathbb{R} \times \mathcal{Z} \to \mathbb{R}$ be a loss function. We say that ℓ is L-Lipschitz in its first argument if

$$|\ell(y, z) - \ell(\tilde{y}, z)| \leqslant L|y - \tilde{y}|,$$

for all $y, \tilde{y} \in \mathbb{R}$ and $z \in \mathcal{Z}$. Additionally, for any $f \in \mathcal{F}$, where $\mathcal{F}$ is a class of real-valued functions, we let $\mathcal{L}_f$ denote the random variable $\ell\left(f(x), z\right)$.

LEMMA 13 (Lemma 14.21 of Wainwright (2019)): *Consider a separable metric space $\mathcal{F}$ of functions $f : \mathcal{X} \to \mathbb{R}$, uniformly bounded by 1, such that the mapping $f \mapsto g(x, f)$, defined in Lemma 12, is continuous for all $x \in \mathcal{X}^n$. Let $f^* \in \mathcal{F}$ be any fixed function. Assume that the loss function ℓ is L-Lipschitz in its first argument, and define*

$$Z_n(r) := \sup_{f \in \mathcal{F}: \|f - f^*\|_2 \leqslant r} \left| \mathbb{E}_n\left(\mathcal{L}_f - \mathcal{L}_{f^*}\right) - \mathbb{E}\left(\mathcal{L}_f - \mathcal{L}_{f^*}\right) \right|.$$

Then, the following holds:

$$P\left\{Z_n(r) \geqslant 8L \cdot \mathcal{R}\left(\mathcal{F} - f^*, r\right) + u\right\} \leqslant 2\exp\left(-c_1 \frac{nu^2}{L^2 r^2 + 64L\mathcal{R}_n\left(\mathcal{F} - f^*, r\right) + Lu}\right),$$

where $c_1 = \frac{1}{8e}$.

Moreover, if δ_n is any solution to the inequality

$$\mathcal{R}\left(star\left(\mathcal{F} - f^*\right), \delta\right) \leqslant \delta^2,$$

then for each $r \geqslant \delta_n$,

$$P\left\{Z_n(r) \geqslant 8Lr\delta_n + u\right\} \leqslant 2\exp\left(-c_1 \frac{nu^2}{L^2 r^2 + 64Lr\delta_n + Lu}\right).$$

LEMMA 14 (Lemma 12 of Foster et al. (2023)): *Consider a separable metric space $\mathcal{F}$ of functions $f : \mathcal{X} \to \mathbb{R}$, uniformly bounded by 1, such that the mapping $f \mapsto g(x, f)$, defined in Lemma 12, is continuous for all $x \in \mathcal{X}^n$. Let $f^* \in \mathcal{F}$ be any fixed function. In addition, consider a loss function ℓ that is L-Lipschitz in its first argument. Let $n \geqslant 3$, and let $\delta_n^2 \geqslant c_1 \frac{\log(\log(n))}{n}$ be any solution to the inequality*

$$\mathcal{R}\left(star\left(\mathcal{F} - f^*\right), \delta\right) \leqslant \delta^2,$$

where $c_1 = 32e + \frac{1024e}{L}$. Moreover, consider the following event:

$$\mathcal{E}_1 = \left\{\exists f \in \mathcal{F} : \|f - f^*\|_2 \geqslant \delta_n \text{ and } \left|\mathbb{E}_n\left(\mathcal{L}_f - \mathcal{L}_{f^*}\right) - \mathbb{E}\left(\mathcal{L}_f - \mathcal{L}_{f^*}\right)\right| \geqslant 10L\delta_n \|f - f^*\|_2\right\}.$$

Then $P(\mathcal{E}_1) \leqslant 40 \exp\left(-c_2 n \delta_n^2\right)$ *where* $c_2 = \frac{L}{32e(L+32)}$.

LEMMA 15 (Lemma 14 of Foster and Syrgkanis (2023)): *Consider a separable metric space* $\mathcal{F}$ *of functions* $f : \mathcal{X} \to \mathbb{R}$*, uniformly bounded by 1, such that the mapping* $f \mapsto g(x, f)$*, defined in Lemma 12, is continuous for all* $x \in \mathcal{X}^n$*. Let* $f^* \in \mathcal{F}$ *be any fixed function. In addition, consider a loss function* ℓ *that is L-Lipschitz in its first argument. Let* $n \geqslant 3$*, and let* $\delta_n^2 \geqslant c_1 \frac{\log(\log(n))}{n}$ *be any solution to the inequality*

$$\mathcal{R}\left(star\left(\mathcal{F} - f^*\right), \delta\right) \leqslant \delta^2,$$

where $c_1 = 32e + \frac{1024e}{L}$*. Then, for some universal constants* $c_2, c_3 > 0$*, with probability at least* $1 - c_2 \exp\left(-c_3 n \delta_n^2\right)$*,*

$$\left|\mathbb{E}_n\left(\mathcal{L}_f - \mathcal{L}_{f^*}\right) - \mathbb{E}\left(\mathcal{L}_f - \mathcal{L}_{f^*}\right)\right| \leqslant 10 L \delta_n \left\{\|f - f^*\|_2 + \delta_n\right\}, \quad \forall f \in \mathcal{F}. \tag{14}$$

LEMMA 16 (Lemma 10 of Bennett et al. (2023)): *If* $\hat{h}$ *optimizes the regularized objective:*

$$\hat{h} = \arg \min_{h \in \mathcal{H}_{B_1}} \max_{g \in \mathcal{G}'_{B_2}} \left\{\mathbb{E}_n\left[g(X, L, Z)\left\{Y - h(X, L, W)\right\} - g(X, L, Z)^2\right] - \lambda_{\mathcal{G}'}\|g\|_{\mathcal{G}'}^2 + \lambda_{\mathcal{H}}\|h\|_{\mathcal{H}}^2\right\},$$

and the assumptions of Theorem 5 are satisfied, then, for any $h_* \in T_{\mathcal{H}}^{-1/2}\left(\mathcal{H}_{B_1}\right)$*, with probability* $1 - \varsigma$*, it holds that*

$$\frac{1}{4}\left\{\left\|\mathcal{T}\left(h_0 - \hat{h}\right)\right\|_2^2 - \left\|\mathcal{T}\left(h_0 - T_{\mathcal{H}}^{1/2} h_*\right)\right\|_2^2\right\} \leqslant \frac{1}{2}\left\|\mathcal{T}\left(h_0 - T_{\mathcal{H}}^{1/2} h_*\right)\right\|_2^2 + \lambda_{\mathcal{H}}\left(\left\|T_{\mathcal{H}}^{1/2} h_*\right\|_{\mathcal{H}}^2 - \left\|\hat{h}\right\|_{\mathcal{H}}^2\right) + \lambda_{\mathcal{G}'} B_2^2 + O\left(\delta_n \left\|\mathcal{T}\left(\hat{h} - T_{\mathcal{H}}^{1/2} h_*\right)\right\|_2 + \delta_n^2\right).$$

LEMMA 17: *If* $\hat{\Gamma}$ *optimizes the regularized objective:*

$$\hat{\Gamma} = \arg\min_{g\in\mathcal{G}_{B_1}} \max_{h\in\mathcal{H}'_{B_2}} \{\mathbb{E}_n[h\{q(X,L),L,W\}\cdot I\{(X,L)\in\mathcal{S}\} - I\{X\in q(\mathcal{S}(L),L)\}h(X,L,W)g(X,L,Z) - I\{X\in q(\mathcal{S}(L),L)\}h(X,L,W)^2] - \lambda_{\mathcal{H}'}\|h\|^2_{\mathcal{H}'} + \lambda_{\mathcal{G}}\|g\|^2_{\mathcal{G}}\},$$

and the assumptions of Theorem 7 are satisfied, then, for any $g_* \in T_{\mathcal{G}}^{-1/2}(\mathcal{G}_{B_1})$*, with probability* $1-\varsigma$*, it holds that*

$$\frac{1}{4}\left\{\left\|\tilde{I}_q\circ\mathcal{T}^*\left(\tilde{g}_0-\hat{\Gamma}\right)\right\|_2^2 - \left\|\tilde{I}_q\circ\mathcal{T}^*\left(\tilde{g}_0-T_{\mathcal{G}}^{1/2}g_*\right)\right\|_2^2\right\} \leqslant \frac{1}{2}\left\|\tilde{I}_q\circ\mathcal{T}^*\left(\tilde{g}_0-T_{\mathcal{G}}^{1/2}g_*\right)\right\|_2^2 + \lambda_{\mathcal{G}}\left(\left\|T_{\mathcal{G}}^{1/2}g_*\right\|_{\mathcal{G}}^2 - \left\|\hat{\Gamma}\right\|_{\mathcal{G}}^2\right) + \lambda_{\mathcal{H}'}B_2^2 + O\left(\delta_n\left\|\tilde{I}_q\circ\mathcal{T}^*\left(\hat{\Gamma}-T_{\mathcal{G}}^{1/2}g_*\right)\right\|_2+\delta_n^2\right).$$

Web Appendix E. Proofs

E.1 *Proofs of main results*

Proof. [Theorem 1]

Proof of part i). Take $(x,l,z)\in\text{supp}(X,L,Z)$ such that $x\in\mathcal{S}(l)\cup q(\mathcal{S}(l),l)$. Hence, if $h_0(x,l,w)$ solves the integral equation (2), then

$$\begin{aligned}
\int h_0(x,l,w)p_{W|X,L,Z}(w|x,l,z)\,dw &= \int h_0(x,l,w)\left\{\int p_{(W,U)|X,L,Z}(w,u|x,l,z)\,du\right\}dw \\
&= \int h_0(x,l,w)\int p_{W|X,L,U}(w|x,l,u)\,p_{U|X,L,Z}(u|x,l,z)\,dudw \quad \text{by assumption 5} \\
&= \int\left\{\int h_0(x,l,w)\,p_{W|X,L,U}(w|x,l,u)\,dw\right\}p_{U|X,L,Z}(u|x,l,z)\,du \\
&= \int \mathbb{E}(Y|X=x,L=l,U=u)\,p_{U|X,L,Z}(u|x,l,z)\,du \qquad h_0 \text{ solves (2)} \\
&= \int \mathbb{E}(Y|X=x,L=l,U=u,Z=z)\,p_{U|X,L,Z}(u|x,l,z)\,du \qquad \text{by assumption 4} \\
&= \mathbb{E}(Y|X=x,L=l,Z=z).
\end{aligned}$$

Thus, $h_0(x, l, w)$ solves the integral equation (4).

Proof of part ii). Take $(x, l, w) \in \text{supp}(X, L, W)$ such that $x \in \mathcal{S}(l) \cup q(\mathcal{S}(l), l)$ and the map $x' \mapsto q(x', l)$ is differentiable at $x' = x$. Hence, if $g_0(x, l, z)$ solves the integral equation (3), then

$$
\begin{aligned}
\int g_0(x,l,z) p_{Z|X,L,W}(z|x,l,w)\,dz &= \int g_0(x,l,z) \left\{ \int p_{(Z,U)|X,L,W}(z,u|x,l,w)\,du \right\} dz \\
&= \int g_0(x,l,z) \int p_{Z|X,L,U}(z|x,l,u)\, p_{U|X,L,W}(u|x,l,w)\,dudz \quad \text{by assumption 5} \\
&= \int \left\{ \int g_0(x,l,z)\, p_{Z|X,L,U}(z|x,l,u)\,dz \right\} p_{U|X,L,W}(u|x,l,w)\,du \\
&= \int \alpha_0(x,l,u) p_{U|X,L,W}(u|x,l,w)\,du \qquad g_0 \text{ solves (3)} \\
&= \int I\{x \in q(\mathcal{S}(l),l)\} \frac{dq^{-1}(x,l)}{dx} \cdot \frac{p_{X|L,U}\{q^{-1}(x,l)|l,u\}}{p_{X|L,U}(x|l,u)} p_{U|X,L,W}(u|x,l,w)\,du \\
&\overset{\text{(i)}}{=} I\{x \in q(\mathcal{S}(l),l)\} \frac{dq^{-1}(x,l)}{dx} \int \frac{p_{U|X,L,W}(u|x,l,w)}{p_{X|L,U,W}(x|l,u,w)} p_{X|L,U,W}\{q^{-1}(x,l)|l,u,w\}\,du \\
&= I\{x \in q(\mathcal{S}(l),l)\} \frac{dq^{-1}(x,l)}{dx} \int \frac{p_{U|L,W}(u|l,w)}{p_{X|L,W}(x|l,w)} p_{X|L,U,W}\{q^{-1}(x,l)|l,u,w\}\,du \\
&= I\{x \in q(\mathcal{S}(l),l)\} \frac{dq^{-1}(x,l)}{dx} \frac{1}{p_{X|L,W}(x|l,w)} \int p_{(X,U)|L,W}\{q^{-1}(x,l),u|l,w\}\,du \\
&= I\{x \in q(\mathcal{S}(l),l)\} \frac{dq^{-1}(x,l)}{dx} \frac{p_{X|L,W}\{q^{-1}(x,l)|l,w\}}{p_{X|L,W}(x|l,w)} \\
&= \alpha_0(x,l,w),
\end{aligned}
$$

where the equality in (i) follows from assumption 5. Thus, $g_0(x, l, z)$ solves equation (5).

Proof. [Theorem 2]

Case 1. Assumption 7 holds and equation (5) has at least one solution. Let h_0 and $g^\dagger$ be

solutions to equations (2) and (5), respectively. Then,

$$\begin{aligned}
\psi_0 &= \mathbb{E}\big[Y\left\{q(X,L)\right\}|(X,L)\in\mathcal{S}\big] \\
&= \frac{\mathbb{E}\big[Y\left\{q(X,L)\right\}\cdot I\left\{(X,L)\in\mathcal{S}\right\}\big]}{P\left\{(X,L)\in\mathcal{S}\right\}} \\
&= \frac{\mathbb{E}\left\{\mathbb{E}\big[Y\left\{q(X,L)\right\}|X,L,U\big]\cdot I\left\{(X,L)\in\mathcal{S}\right\}\right\}}{P\left\{(X,L)\in\mathcal{S}\right\}} \\
&= \frac{\mathbb{E}\left\{\mathbb{E}\big[h_0\left\{q(X,L),L,W\right\}|X,L,U\big]\cdot I\left\{(X,L)\in\mathcal{S}\right\}\right\}}{P\left\{(X,L)\in\mathcal{S}\right\}} \qquad \text{by Lemma 8} \\
&= \frac{\mathbb{E}\big[h_0\left\{q(X,L),L,W\right\}\cdot I\left\{(X,L)\in\mathcal{S}\right\}\big]}{P\left\{(X,L)\in\mathcal{S}\right\}}.
\end{aligned}$$

By Theorem 1, h_0 also solves equation (4), then, by Lemma 7,

$$\frac{\mathbb{E}\left\{Yg^{\dagger}\left(X,L,Z\right)\right\}}{P\left\{(X,L)\in\mathcal{S}\right\}} = \frac{\mathbb{E}\left[h_0\left\{q\left(X,L\right),L,W\right\}\cdot I\left\{(X,L)\in\mathcal{S}\right\}\right]}{P\left\{(X,L)\in\mathcal{S}\right\}} = \psi_0,$$

which is the treatment bridge representation.

Now, let $h^{\dagger}$ be any solution to equation (4). By Lemma 7,

$$\frac{\mathbb{E}\left[h^{\dagger}\left\{q\left(X,L\right),L,W\right\}\cdot I\left\{(X,L)\in\mathcal{S}\right\}\right]}{P\left\{(X,L)\in\mathcal{S}\right\}} = \frac{\mathbb{E}\left[Yg^{\dagger}\left(X,L,Z\right)\right]}{P\left\{(X,L)\in\mathcal{S}\right\}} = \psi_0,$$

which is the outcome bridge representation.

Lastly, for any $h\in\mathcal{L}^2(X,L,W)$ and any $g\in\mathcal{L}^2(X,L,Z)$, by Lemma 7,

$$\mathbb{E}\left\{\phi(O;h,g^{\dagger})\right\} = \mathbb{E}\left\{\phi(O;h^{\dagger},g^{\dagger})\right\} = \mathbb{E}\left[h^{\dagger}\left\{q\left(X,L\right),L,W\right\}\cdot I\left\{(X,L)\in\mathcal{S}\right\}\right],$$

and

$$\mathbb{E}\left\{\phi(O;h^{\dagger},g)\right\} = \mathbb{E}\left\{\phi(O;h^{\dagger},g^{\dagger})\right\} = \mathbb{E}\left\{Yg^{\dagger}\left(X,L,Z\right)\right\}.$$

Hence

$$\frac{\mathbb{E}\left\{\phi(O;h,g^{\dagger})\right\}}{P\left\{(X,L)\in\mathcal{S}\right\}} = \psi_0 = \frac{\mathbb{E}\left\{\phi(O;h^{\dagger},g)\right\}}{P\left\{(X,L)\in\mathcal{S}\right\}},$$

which is the double robust representation.

Case 2. Assumption 8 holds and equation (4) has at least one solution. Let g_0 and $h^{\dagger}$ be

solutions to equations (3) and (4), respectively. Then,

$$
\begin{aligned}
\psi_0 &= \mathbb{E}\big[Y\left\{q(X,L)\right\}|(X,L)\in\mathcal{S}\big] \\
&= \frac{\mathbb{E}\big\{\mathbb{E}\left[Y\left\{q(X,L)\right\}|X,L,U\right]\cdot I\left\{(X,L)\in\mathcal{S}\right\}\big\}}{P\left\{(X,L)\in\mathcal{S}\right\}} \\
&= \iiint \frac{I\left\{(x,l)\in\mathcal{S}\right\}}{P\left\{(X,L)\in\mathcal{S}\right\}}\mathbb{E}\left\{Y|X=q(x,l),L=l,U=u\right\}p_{X,L,U}(x,l,u)dxdldu \\
&= \iint\left[\int \frac{I\left\{(x,l)\in\mathcal{S}\right\}}{P\left\{(X,L)\in\mathcal{S}\right\}}\mathbb{E}\left\{Y|X=q(x,l),L=l,U=u\right\}p_{X|L,U}(x|l,u)dx\right]p_{L,U}(l,u)dldu \\
&\overset{\text{(i)}}{=} \iiint \frac{I\left\{\tilde{x}\in q(\mathcal{S}(l),l)\right\}}{P\left\{(X,L)\in\mathcal{S}\right\}}\mathbb{E}\big(Y|X=\tilde{x},L=l,U=u\big)\frac{dq^{-1}(\tilde{x},l)}{d\tilde{x}}p_{X|L,U}\{q^{-1}(\tilde{x},l)|l,u\}p_{L,U}(l,u)d\tilde{x}dldu \\
&= \frac{1}{P\left\{(X,L)\in\mathcal{S}\right\}}\iiint \mathbb{E}\big[Y|X=\tilde{x},L=l,U=u\big]\cdot\alpha_0(\tilde{x},l,u)\cdot p_{X,L,U}(\tilde{x},l,u)d\tilde{x}dldu \\
&\overset{\text{(ii)}}{=} \frac{1}{P\left\{(X,L)\in\mathcal{S}\right\}}\mathbb{E}\left[\mathbb{E}\big(Y|X,L,U\big)\cdot\mathbb{E}\big\{g_0(X,L,Z)|X,L,U\big\}\right] \qquad g_0 \text{ solves (3)} \\
&= \frac{\mathbb{E}\left\{Yg_0(X,L,Z)\right\}}{P\left\{(X,L)\in\mathcal{S}\right\}},
\end{aligned}
$$

where the equality in (i) follows from assumption 6, which justifies applying the change of variable $\tilde{x} = q(x,l)$ in x for each l, and from the fact that $(x,l)\in\mathcal{S}$ if and only if $q(x,l)\in q(\mathcal{S}(l),l)$. For the equality in (ii), note that $\tilde{x}\in q(\mathcal{S}(l),l)$ implies $\tilde{x} = q(x,l)$ for some $x\in\mathcal{S}(l)$, and by the definitions of $\mathcal{S}$ and $\mathcal{S}(l)$, $\tilde{x}\in\text{supp}\,(X|L=l)$. Then, for any $u\in\text{supp}\,(U|L)$, assumption 3 implies $\tilde{x}\in\text{supp}\,(X|L=l,U=u)$, and consequently $(\tilde{x},l,u)\in\text{supp}\,(X,L,U)$. Therefore, $(\tilde{x},l,u)$ lies in the domain where g_0 solves equation (3). By Theorem 1, g_0 also solves equation (5). Thus, by Lemma 7,

$$
\frac{\mathbb{E}\left[h^\dagger\left\{q\left(X,L\right),L,W\right\}\cdot I\left\{(X,L)\in\mathcal{S}\right\}\right]}{P\left\{(X,L)\in\mathcal{S}\right\}} = \frac{\mathbb{E}\left\{Yg_0\left(X,L,Z\right)\right\}}{P\left\{(X,L)\in\mathcal{S}\right\}} = \psi_0,
$$

which is the outcome bridge representation. From here, the other two representations are derived analogously to Case 1.

Proof. [Proof of Theorem 3] We use empirical process notation, defining for any function $f(O)$, $\mathbb{G}_m^k[f] := \sqrt{m}\left\{\mathbb{E}_m^k[f] - \mathbb{E}[f]\right\}$, where $m = n/K$ and $\mathbb{E}_m^k$ denotes the empirical expectation over data in the fold I_k. The target parameter can be written as $\psi_0 = \frac{\phi_0}{p_0}$, where, recall, $p_0 = P\left\{(X,L)\in\mathcal{S}\right\} = \mathbb{E}\left[I\left\{(X,L)\in\mathcal{S}\right\}\right]$ and $\phi_0 = \mathbb{E}\left[\phi(O;h_0,g_0)\right]$. Using this notation, the

expression for $\sqrt{n}\left(\frac{1}{\hat{p}}\cdot\frac{1}{K}\sum_{k=1}^{K}\hat{\phi}_k-\psi_0\right)$ can be decompose as

$$\begin{aligned}\sqrt{n}\left(\frac{1}{\hat{p}}\cdot\frac{1}{K}\sum_{k=1}^{K}\hat{\phi}_k-\psi_0\right)&=\sqrt{n}\left\{\frac{1}{\hat{p}}\cdot\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}_m^k\left[\phi\left\{O;\hat{h}^{(-k)},\hat{g}^{(-k)}\right\}\right]-\psi_0\right\}\\&=\underbrace{\sqrt{n}\frac{1}{K}\left(\frac{1}{\hat{p}}-\frac{1}{p_0}\right)\sum_{k=1}^{K}\mathbb{E}_m^k\left[\phi\left\{O;\hat{h}^{(-k)},\hat{g}^{(-k)}\right\}-\phi_0\right]}_{T_1}\\&\quad+\underbrace{\sqrt{n}\frac{1}{K}\left(\frac{1}{\hat{p}}-\frac{1}{p_0}\right)\sum_{k=1}^{K}\mathbb{E}_m^k\left[\phi_0\right]}_{T_2}\\&\quad+\underbrace{\sqrt{n}\left\{\frac{1}{K}\cdot\frac{1}{p_0}\sum_{k=1}^{K}\mathbb{E}_m^k\left[\phi\left\{O;\hat{h}^{(-k)},\hat{g}^{(-k)}\right\}\right]-\psi_0\right\}}_{T_3}\end{aligned}$$

We will show that $T_1=o_p(1)$, that $T_2=-\sqrt{n}\mathbb{E}_n\left[\frac{\psi_0}{p_0}I\left\{(X,L)\in\mathcal{S}\right\}-\psi_0\right]+o_p(1)$, and that $T_3=\sqrt{n}\mathbb{E}_n\left[\frac{\phi(O;h_0,g_o)}{p_0}-\psi_0\right]+\sqrt{n}R_n+o_p(1)$.

Since $\hat{p}=\frac{1}{n}\sum_{i=1}^{n}I\left\{(X_i,L_i)\in\mathcal{S}\right\}$, by the central limit theorem (CLT), it follows that $\sqrt{n}\left(\hat{p}-p_0\right)=O_p(1)$, and consequently $\hat{p}-p_0=O_p\left(n^{-1/2}\right)$. Moreover, using the first order Taylor expansion of the function $f(v)=1/v$ around $v=p_0$, we have

$$\frac{1}{\hat{p}}=\frac{1}{p_0}-\frac{\hat{p}-p_0}{p_0^2}+\frac{1}{4}\frac{(\tilde{p}-p_0)^2}{p_0^3},$$

with $\tilde{p}\in\left(\min\left\{\hat{p},p_0\right\},\max\left\{\hat{p},p_0\right\}\right)$. Since $(\tilde{p}-p_0)^2\leqslant(\hat{p}-p_0)^2=O_p\left(n^{-1/2}\right)O_p\left(n^{-1/2}\right)=O_p\left(n^{-1/2}\right)o_p(1)=o_p\left(n^{-1/2}\right)$, it follows that

$$\sqrt{n}\left(\frac{1}{\hat{p}}-\frac{1}{p_0}\right)=-\frac{\sqrt{n}\left(\hat{p}-p_0\right)}{p_0^2}+o_p(1)=-\sqrt{n}\mathbb{E}_n\left[\frac{I\left\{(X,L)\in\mathcal{S}\right\}-p_0}{p_0^2}\right]+o_p(1),$$

and $\sqrt{n}\left(\frac{1}{\hat{p}}-\frac{1}{p_0}\right)=O_p(1)$.

By Lemma 10, for each $k=1,\ldots,K$, we have $\mathbb{E}_m^k\left[\phi\left\{O;\hat{h}^{(-k)},\hat{g}^{(-k)}\right\}\right]-\phi_0=o_p(1)$. Then,

$$T_1=\frac{1}{K}\sqrt{n}\left(\frac{1}{\hat{p}_k}-\frac{1}{p_0}\right)\sum_{k=1}^{K}\mathbb{E}_m^k\left[\phi\left\{O;\hat{h}^{(-k)},\hat{g}^{(-k)}\right\}-\phi_0\right]=\frac{1}{K}O_p(1)\sum_{k=1}^{K}o_p(1)=o_p(1),$$

The term T_2 is handled as follows:

$$T_2 = \sqrt{n}\left(\frac{1}{\hat{p}} - \frac{1}{p_0}\right)\phi_0$$
$$= -\frac{\phi_0}{p_0^2}\sqrt{m}\mathbb{E}_n\left[I\left\{(X,L)\in\mathcal{S}\right\} - p_0\right] + o_p(1) = -\sqrt{n}\mathbb{E}_n\left[\frac{\psi_0}{p_0}I\left\{(X,L)\in\mathcal{S}\right\} - \psi_0\right] + o_p(1).$$

The term T_3 can be decomposed as

$$T_3 = = \underbrace{\frac{1}{p_0}\cdot\frac{1}{\sqrt{K}}\sum_{k=1}^{K}\left\{\mathbb{G}_m^k\left[\phi\left\{O;\hat{h}^{(-k)},\hat{g}^{(-k)}\right\}\right] - \mathbb{G}_m^k\left[\phi\left(O;h_0,g_0\right)\right]\right\}}_{T_4}$$
$$+ \underbrace{\frac{1}{\sqrt{K}}\sum_{k=1}^{K}\mathbb{G}_m^k\left[\frac{\phi\left(O;h_0,g_0\right)}{p_0}\right]}_{T_5}$$
$$+ \underbrace{\frac{1}{p_0}\cdot\frac{1}{\sqrt{K}}\sum_{k=1}^{K}\sqrt{m}\left\{\mathbb{E}\left[\phi\left\{O;\hat{h}^{(-k)},\hat{g}^{(-k)}\right\}\right] - \phi_0\right\}}_{T_6}.$$

We will show that $T_4 = o_p(1)$, $T_5 = \sqrt{n}\mathbb{E}_n\left[\frac{\phi(O;h_0,g_o)}{p_0} - \psi_0\right]$, and $T_6 = \sqrt{n}R_n$.

To analyze T_1, for each $k = 1,\ldots,K$, we define

$$A_m^k := \mathbb{G}_m^k\left[\phi\left\{O;\hat{h}^{(-k)},\hat{g}^{(-k)}\right\}\right] - \mathbb{G}_m^k\left[\phi\left(O;h_0,g_0\right)\right].$$

We will demonstrate that $\text{var}\left(A_m^k|I_k^c\right) = o_p(1)$.

Since $\mathbb{E}_m^k$ only operates over data in the fold I_k, we have

$$\begin{aligned}\text{var}\left(A_m^k|I_k^c\right) &= m\text{var}\left(\mathbb{E}_m^k\left[\phi\left\{O;\hat{h}^{(-k)},\hat{g}^{(-k)}\right\}\right] - \mathbb{E}_m^k\left[\phi\left(O;h_0,g_0\right)\right]\Big|I_k^c\right)\\ &= \text{var}\left[\phi\left\{O;\hat{h}^{(-k)},\hat{g}^{(-k)}\right\} - \phi\left(O;h_0,g_0\right)\Big|I_k^c\right]\\ &\leqslant \mathbb{E}\left(\left[\phi\left\{O;\hat{h}^{(-k)},\hat{g}^{(-k)}\right\} - \phi\left(O;h_0,g_0\right)\right]^2\Bigg|I_k^c\right)\\ &= \left\|\phi\left\{O;\hat{h}^{(-k)},\hat{g}^{(-k)}\right\} - \phi\left(O;h_0,g_0\right)\right\|_2^2.\end{aligned}$$

By Lemma 10, it follows that $\text{var}\left(A_m^k|I_k^c\right) = o_p(1)$.

Since $\mathbb{E}\left[A_m^k|I_k^c\right] = 0$, it follows that $\mathbb{E}\left[\left(A_m^k\right)^2|I_k^c\right] = \text{var}\left(A_m^k|I_k^c\right) = o_p(1)$. By Chebyshev inequality, for any $\varepsilon > 0$, we have

$$\Pr\left[\left|A_m^k\right| > \varepsilon|I_k^c\right] \leqslant \frac{\text{var}\left(A_m^k|I_k^c\right)}{\varepsilon^2} = o_p(1).$$

Let Q^k denote the random variable $Q^k := \Pr\left[\left|A_m^k\right| > \varepsilon | I_k^c\right]$, with support contained in $[0,1]$. For any $\varepsilon' > 0$, applying the law of iterated expectation gives

$$\begin{aligned}
\Pr\left[\left|A_m^k\right| > \varepsilon\right] &= \mathbb{E}\left[I\left(\left|A_m^k\right| > \varepsilon\right)\right] \\
&= \mathbb{E}\left\{\mathbb{E}\left[I\left(\left|A_m^k\right| > \varepsilon\right) \middle| I_k^c\right]\right\} \\
&= \mathbb{E}\left\{\Pr\left[\left|A_m^k\right| > \varepsilon \middle| I_k^c\right]\right\} \\
&= \int_0^{\varepsilon'/2} \Pr\left[\left|A_m^k\right| > \varepsilon \middle| I_k^c\right] dP(Q^k) + \int_{\varepsilon'/2}^1 \Pr\left[\left|A_m^k\right| > \varepsilon \middle| I_k^c\right] dP(Q^k) \\
&\leqslant \frac{\varepsilon'}{2} + \Pr\left[\left|Q^k\right| > \varepsilon'/2\right].
\end{aligned}$$

As $Q^k = o_p(1)$, for m large enough, we have

$$\Pr\left[\left|A_m^k\right| > \varepsilon\right] \leqslant \frac{\varepsilon'}{2} + \frac{\varepsilon'}{2} = \varepsilon'.$$

Hence, $A_m^k = o_p(1)$. Since k was arbitrary, we conclude that $T_4 = o_p(1)$.

The term T_5 is handled as follows:

$$\begin{aligned}
T_2 &= \frac{1}{\sqrt{K}} \sum_{k=1}^K \mathbb{G}_m^k \left[\frac{\phi\left(O; h_0, g_0\right)}{p_0}\right] \\
&= \frac{\sqrt{m}}{\sqrt{K}} \sum_{k=1}^K \left\{\mathbb{E}_m^k \left[\frac{\phi\left(O; h_0, g_0\right)}{p_0}\right] - \psi_0\right\} \\
&= \frac{\sqrt{m}}{\sqrt{K}} \frac{1}{m} \sum_{k=1}^K \sum_{i: O_i \in I_k} \left[\frac{\phi\left(O_i; h_0, g_0\right)}{p_0} - \psi_0\right] \\
&= \sqrt{n} \frac{1}{n} \sum_{i=1}^n \left[\frac{\phi\left(O_i; h_0, g_0\right)}{p_0} - \psi_0\right].
\end{aligned}$$

In the term T_6, for each $k = 1, \ldots, K$, using Lemma 7 with $h = \hat{h}^{(-k)}$ and $g = \hat{g}^{(-k)}$ yields

$$\sqrt{m}\left\{\mathbb{E}\left[\phi\left\{O; \hat{h}^{(-k)}, \hat{g}^{(-k)}\right\}\right] - \phi_0\right\} = \sqrt{m}\mathbb{E}\left[\left\{\hat{h}^{(-k)} - h_0\right\} \cdot \left\{g_0 - \hat{g}^{(-k)}\right\}\right].$$

Hence,

$$T_6 = \frac{1}{p_0} \frac{\sqrt{m}}{\sqrt{K}} \sum_{k=1}^K \mathbb{E}\left[\left\{\hat{h}^{(-k)} - h_0\right\} \cdot \left\{g_0 - \hat{g}^{(-k)}\right\}\right] = \sqrt{n} R_n.$$

Finally, if $R_n = o_p\left(n^{-1/2}\right)$, we have

$$\sqrt{n}\left(\hat{\psi}_{CF}^{DR} - \psi_0\right) = \sqrt{n}\mathbb{E}_n\left[\frac{\phi(O; h_0, g_0)}{p_0} - \psi_0 \frac{I\left\{(X, L) \in \mathcal{S}\right\}}{p_0}\right] + o_p(1).$$

Since $\mathbb{E}\left[\frac{\phi(O;h_0,g_0)}{p_0} - \psi_0\frac{I\{(X,L)\in\mathcal{S}\}}{p_0}\right] = 0$ and

$$\mathbb{E}\left\{\left[\frac{\phi(O;h_0,g_0)}{p_0} - \psi_0\frac{I\left\{(X,L)\in\mathcal{S}\right\}}{p_0}\right]^2\right\} \leqslant 2\left\|\frac{\phi(O;h_0,g_0)}{p_0}\right\|_2^2 + 2\left\|\psi_0\frac{I\left\{(X,L)\in\mathcal{S}\right\}}{p_0}\right\|_2^2$$
$$= \frac{2\mathbb{E}\left[\phi(O;h_0,g_0)^2\right]}{p_0^2} + \frac{2\psi_0^2}{p_0} < \infty,$$

by the CLT, $\sqrt{n}\left(\hat{\psi}_{CF} - \psi_0\right)$ converges in distribution to a normal distribution with mean zero and variance $\tau^2 = \mathbb{E}\left\{\left[\frac{\phi(O;h_0,g_0)}{p_0} - \psi_0\frac{I\{(X,L)\in\mathcal{S}\}}{p_0}\right]^2\right\}$. This completes the proof.

Proof. [Theorem 4] Recall that

$$h_* = \arg\min_{h\in\mathscr{L}^2(X,L,W)}\left\{\left\|\widetilde{\mathcal{T}}(q_0 - h)\right\|_2^2 + \lambda_{\mathcal{H}}\|h\|_2^2\right\},$$

where $\widetilde{\mathcal{T}} = \frac{1}{2}\mathcal{T}\circ T_{\mathcal{H}}^{1/2}$ and $q_0 = T_{\mathcal{H}}^{-1/2}h_0$. Since $\widetilde{\mathcal{T}}$ is a bounded linear operator, the regularized solution h_* (see Theorem 16.4 of Kress (2010)) is given by

$$h_* = \left(\widetilde{\mathcal{T}}^*\circ\widetilde{\mathcal{T}} + \lambda_{\mathcal{H}}I\right)^{-1}\widetilde{\mathcal{T}}^*\circ\widetilde{\mathcal{T}}q_0.$$

Hence,

$$\left\|h_* - T_{\mathcal{H}}^{-1/2}h_0\right\|_2 = \left\|\left\{I - \left(\widetilde{\mathcal{T}}^*\circ\widetilde{\mathcal{T}} + \lambda_{\mathcal{H}}I\right)^{-1}\widetilde{\mathcal{T}}^*\circ\widetilde{\mathcal{T}}\right\}T_{\mathcal{H}}^{-1/2}h_0\right\|_2$$
$$= \left\|r_{\lambda_{\mathcal{H}}}\left[\widetilde{\mathcal{T}}^*\circ\widetilde{\mathcal{T}}\right]T_{\mathcal{H}}^{-1/2}h_0\right\|_2,$$

where for any bounded linear self-adjoint operator $B : \mathscr{L}^2(X,L,W)\to\mathscr{L}^2(X,L,W)$, the operator $r_\lambda(B)$ is defined as $r_\lambda(B) := I - (B+\lambda I)^{-1}B$.

Incorporating the source condition (assumption 13), we have

$$\left\|r_{\lambda_{\mathcal{H}}}\left[\widetilde{\mathcal{T}}^*\circ\widetilde{\mathcal{T}}\right]T_{\mathcal{H}}^{-1/2}h_0\right\|_2 = \left\|r_{\lambda_{\mathcal{H}}}\left[\widetilde{\mathcal{T}}^*\circ\widetilde{\mathcal{T}}\right]\left(\widetilde{\mathcal{T}}^*\circ\widetilde{\mathcal{T}}\right)^{\beta/2}w_0\right\|_2$$
$$\leqslant \left\|r_{\lambda_{\mathcal{H}}}\left[\widetilde{\mathcal{T}}^*\circ\widetilde{\mathcal{T}}\right]\left(\widetilde{\mathcal{T}}^*\circ\widetilde{\mathcal{T}}\right)^{\beta/2}\right\|_2\cdot\|w_0\|_2$$
$$\leqslant \sup_{\theta\in\left[0,\left\|\widetilde{\mathcal{T}}^*\circ\widetilde{\mathcal{T}}\right\|\right]}\left|\theta^{\beta/2}r_{\lambda_{\mathcal{H}}}(\theta)\right|\cdot\|w_0\|_2$$
$$\leqslant \|w_0\|_2\left(\frac{\tilde{\eta}}{4}\right)^{(\beta/2-1)\cdot I\{\beta\geqslant 2\}}\lambda_{\mathcal{H}}^{\min\{\beta/2,1\}},$$

where the second inequality follows from formula (4.30c) in Theorem 4.17 of Hohage (2002) applied to the functional calculus mapping $r_{\lambda_{\mathcal{H}}}$ at B, and the third inequality follows from

the fact that

$$\left|\theta^{\beta/2} r_{\lambda_{\mathcal{H}}}(\theta)\right| = \frac{\lambda_{\mathcal{H}}\theta^{\beta/2}}{\lambda_{\mathcal{H}}+\theta} \leqslant \left\|\widetilde{\mathcal{T}}^* \circ \widetilde{\mathcal{T}}\right\|^{(\beta/2-1)\cdot I\{\beta\geqslant 2\}} \lambda_{\mathcal{H}}^{\min\{\beta/2,1\}} \leqslant \left(\frac{\tilde{\eta}}{4}\right)^{(\beta/2-1)\cdot I\{\beta\geqslant 2\}} \lambda_{\mathcal{H}}^{\min\{\beta/2,1\}}$$

for all $\theta \in \left[0, \left\|\widetilde{\mathcal{T}}^* \circ \widetilde{\mathcal{T}}\right\|\right]$. To see this, we consider two cases.

- **Case 1**: $\beta \geqslant 2$. In this case, for all $\theta \in \left[0, \left\|\widetilde{\mathcal{T}}^* \circ \widetilde{\mathcal{T}}\right\|\right]$, we have

$$\left|\theta^{\beta/2} r_{\lambda_{\mathcal{H}}}(\theta)\right| = \frac{\lambda_{\mathcal{H}}\theta^{\beta/2}}{\lambda_{\mathcal{H}}+\theta} \leqslant \frac{\theta}{\lambda_{\mathcal{H}}+\theta} \cdot \theta^{\beta/2-1} \cdot \lambda_{\mathcal{H}} \leqslant \left\|\widetilde{\mathcal{T}}^* \circ \widetilde{\mathcal{T}}\right\|^{\beta/2-1} \cdot \lambda_{\mathcal{H}} \leqslant \left(\frac{\tilde{\eta}}{4}\right)^{\beta/2-1} \cdot \lambda_{\mathcal{H}},$$

  where the last inequality holds because $\left\|\widetilde{\mathcal{T}}^* \circ \widetilde{\mathcal{T}}\right\| \leqslant \tilde{\eta}/4$.

- **Case 2**. $\beta \in (0,2)$. In this case, maximizing $f(\theta) := \theta^{\beta/2} r_{\lambda_{\mathcal{H}}}(\theta) = \frac{\lambda_{\mathcal{H}}\theta^{\beta/2}}{\lambda_{\mathcal{H}}+\theta}$ over $\theta \in \left[0, \left\|\widetilde{\mathcal{T}}^* \circ \widetilde{\mathcal{T}}\right\|\right]$ is equivalent to maximizing $\log f(\theta)$. To maximize $\log f(\theta)$, we take the first derivative and equate to 0:

$$\frac{d\log f(\theta)}{d\theta} = \frac{\beta}{2}\cdot\frac{1}{\theta} - \frac{1}{\lambda_{\mathcal{H}}+\theta} = 0 \quad \Rightarrow \quad \theta^* = \frac{\beta\lambda_{\mathcal{H}}}{2-\beta}.$$

  Since $\frac{d\log f(\theta)}{d\theta} > 0$ for $\theta < \theta^*$ and $\frac{d\log f(\theta)}{d\theta} < 0$ for $\theta > \theta^*$, $\log f(\theta)$ attains its maximum at $\theta = \theta^*$. Thus,

$$\begin{aligned}
\sup_{\theta\in\left[0,\left\|\widetilde{\mathcal{T}}^*\circ\widetilde{\mathcal{T}}\right\|\right]} \left|\theta^{\beta/2} r_{\lambda_{\mathcal{H}}}(\theta)\right| \leqslant \sup_{\theta\in[0,\infty)} \left|\theta^{\beta/2} r_{\lambda_{\mathcal{H}}}(\theta)\right| &\leqslant \frac{\lambda_{\mathcal{H}}\left(\frac{\beta\lambda_{\mathcal{H}}}{2-\beta}\right)^{\beta/2}}{\lambda_{\mathcal{H}}+\frac{\beta\lambda_{\mathcal{H}}}{2-\beta}} \\
&= \frac{\lambda_{\mathcal{H}}\cdot\lambda_{\mathcal{H}}^{\beta/2}\left(\frac{\beta}{2-\beta}\right)^{\beta/2}}{\lambda_{\mathcal{H}}\cdot\frac{2}{2-\beta}} \\
&= \lambda_{\mathcal{H}}^{\beta/2}\left(\frac{2-\beta}{2}\right)^{1-\beta/2}\left(\frac{\beta}{2}\right)^{\beta/2} \\
&\leqslant \lambda_{\mathcal{H}}^{\beta/2}.
\end{aligned}$$

Therefore, we have shown that

$$\left\|h_* - T_{\mathcal{H}}^{-1/2} h_0\right\|_2^2 \leqslant \|w_0\|_2^2 \left(\frac{\tilde{\eta}}{4}\right)^{\max\{0,\beta-2\}} \lambda_{\mathcal{H}}^{\min\{\beta,2\}}.$$

To derive the other bound, note that

$$\begin{aligned}
\frac{1}{4}\left\|\mathcal{T}\left(T_{\mathcal{H}}^{1/2}h_* - h_0\right)\right\|_2^2 &= \left\|\widetilde{\mathcal{T}} \circ \left(h_* - T_{\mathcal{H}}^{-1/2}h_0\right)\right\|_2^2 \\
&= \left\langle \widetilde{\mathcal{T}} \circ \left(h_* - T_{\mathcal{H}}^{-1/2}h_0\right), \widetilde{\mathcal{T}} \circ \left(h_* - T_{\mathcal{H}}^{-1/2}h_0\right)\right\rangle_2 \\
&= \left\langle \widetilde{\mathcal{T}}^* \circ \widetilde{\mathcal{T}} \circ \left(h_* - T_{\mathcal{H}}^{-1/2}h_0\right), \left(h_* - T_{\mathcal{H}}^{-1/2}h_0\right)\right\rangle_2 \\
&\overset{(i)}{=} \left\langle \left(\widetilde{\mathcal{T}}^* \circ \widetilde{\mathcal{T}}\right)^{1/2} \circ \left(h_* - T_{\mathcal{H}}^{-1/2}h_0\right), \left(\widetilde{\mathcal{T}}^* \circ \widetilde{\mathcal{T}}\right)^{1/2} \left(h_* - T_{\mathcal{H}}^{-1/2}h_0\right)\right\rangle_2 \\
&= \left\|\left(\widetilde{\mathcal{T}}^* \circ \widetilde{\mathcal{T}}\right)^{1/2} \circ \left(h_* - T_{\mathcal{H}}^{-1/2}h_0\right)\right\|_2^2 \\
&= \left\|r_{\lambda_{\mathcal{H}}}\left[\widetilde{\mathcal{T}}^* \circ \widetilde{\mathcal{T}}\right]\left(\widetilde{\mathcal{T}}^* \circ \widetilde{\mathcal{T}}\right)^{(\beta+1)/2} w_0\right\|_2^2,
\end{aligned}$$

where the equality in (i) holds because $\left(\widetilde{\mathcal{T}}^*\widetilde{\mathcal{T}}\right)^{1/2}$ is self adjoint (see Theorem 4.15 of Hohage (2002)) and the last equality from applying Lemma 4.16 of Hohage (2002) to the self-adjoint operator $\left(\widetilde{\mathcal{T}}^* \circ \widetilde{\mathcal{T}}\right)^{1/2}$.

Repeating the arguments for the previous result, but replacing β by $\beta + 1$, we get

$$\left\|\mathcal{T}\left(T_{\mathcal{H}}^{1/2}h_* - h_0\right)\right\|_2^2 \leqslant 4\|w_0\|_2^2 \left(\frac{\tilde{\eta}}{4}\right)^{\max\{0,\beta-1\}} \lambda_{\mathcal{H}}^{\min\{\beta+1,2\}}.$$

This completes the proof.

Proof. [Proof of Theorem 5] We adapted the proof of Theorem 4 of Bennett et al. (2023) to our estimator. For given functions h_0, h_*, and $\hat{h}$, we define the following function:

$$L(\tau) := \frac{1}{4}\left\|\mathcal{T}\left[h_0 - T_{\mathcal{H}}^{1/2}h_* - \tau\left(\hat{h} - T_{\mathcal{H}}^{1/2}h_*\right)\right]\right\|_2^2 + \lambda_{\mathcal{H}}\left\|h_* + \tau\left(T_{\mathcal{H}}^{-1/2}\hat{h} - h_*\right)\right\|_2^2.$$

This function is strongly convex with constant and positive second derivative:

$$\frac{\partial^2 L(\tau)}{\partial \tau^2} \equiv 2\left\{\frac{1}{4}\left\|\mathcal{T}\left(\hat{h} - T_{\mathcal{H}}^{1/2}h_*\right)\right\|_2^2 + \lambda_{\mathcal{H}}\left\|T_{\mathcal{H}}^{-1/2}\hat{h} - h_*\right\|_2^2\right\}.$$

The function achieves its minimum at $\tau = 0$. Evaluating $L(\tau)$ at $\tau = 0$ yields:

$$L(0) = \frac{1}{4}\left\|\mathcal{T}\left(h_0 - T_{\mathcal{H}}^{1/2}h_*\right)\right\|_2^2 + \lambda_{\mathcal{H}}\|h_*\|_2^2 = \frac{1}{4}\left\|\mathcal{T}\left(h_0 - T_{\mathcal{H}}^{1/2}h_*\right)\right\|_2^2 + \lambda_{\mathcal{H}}\left\|T_{\mathcal{H}}^{1/2}h_*\right\|_{\mathcal{H}}^2.$$

and at $\tau = 1$,

$$L(1) = \frac{1}{4}\left\|\mathcal{T}\left(h_0 - \hat{h}\right)\right\|_2^2 + \lambda_{\mathcal{H}}\left\|T_{\mathcal{H}}^{-1/2}\hat{h}\right\|_2^2 = \frac{1}{4}\left\|\mathcal{T}\left(h_0 - \hat{h}\right)\right\|_2^2 + \lambda_{\mathcal{H}}\left\|\hat{h}\right\|_{\mathcal{H}}^2.$$

Using the fact that $\frac{1}{2}\frac{\partial^2 L(\tau)}{\partial \tau^2} = L(1) - L(0)$, it follows that:

$$\frac{1}{4}\left\|\mathcal{T}\left(\hat{h} - T_{\mathcal{H}}^{1/2}h_*\right)\right\|_2^2 + \lambda_{\mathcal{H}}\left\|T_{\mathcal{H}}^{-1/2}\hat{h} - h_*\right\|_2^2 = \underbrace{\frac{1}{4}\left\{\left\|\mathcal{T}\left(h_0 - \hat{h}\right)\right\|_2^2 - \left\|\mathcal{T}\left(h_0 - T_{\mathcal{H}}^{1/2}h_*\right)\right\|_2^2\right\}}_{(\star)} + \lambda_{\mathcal{H}}\left(\left\|\hat{h}\right\|_{\mathcal{H}}^2 - \left\|T_{\mathcal{H}}^{1/2}h_*\right\|_{\mathcal{H}}^2\right).$$

Next, observe that $\left\|T_{\mathcal{H}}^{1/2}h_*\right\|_{\mathcal{H}} \leqslant \|h_0\|_{\mathcal{H}}$. Otherwise, the following contradiction arises:

$$\frac{1}{4}\|\mathcal{T}(h_0 - h_0)\|_2^2 + \lambda_{\mathcal{H}}\|h_0\|_{\mathcal{H}}^2 < \frac{1}{4}\left\|\mathcal{T}(h_0 - T_{\mathcal{H}}^{1/2}h_*)\right\|_2^2 + \lambda_{\mathcal{H}}\left\|T_{\mathcal{H}}^{1/2}h_*\right\|_{\mathcal{H}}^2$$

contradicting that $T_{\mathcal{H}}^{1/2}h_* = \arg\min_{h\in\mathcal{H}}\left\{\frac{1}{4}\|\mathcal{T}(h_0 - h)\|_2^2 + \lambda_{\mathcal{H}}\|h\|_{\mathcal{H}}^2\right\}$. Hence, $T_{\mathcal{H}}^{1/2}h_* \in \mathcal{H}_{B_1}$. Consequently, we can apply Lemma 16 to bound the term $(\star)$ above with high probability as follows:

$$\frac{1}{4}\left\{\left\|\mathcal{T}\left(h_0 - \hat{h}\right)\right\|_2^2 - \left\|\mathcal{T}\left(h_0 - T_{\mathcal{H}}^{1/2}h_*\right)\right\|_2^2\right\} \leqslant \frac{1}{2}\left\|\mathcal{T}\left(h_0 - T_{\mathcal{H}}^{1/2}h_*\right)\right\|_2^2 + \lambda_{\mathcal{H}}\left(\left\|T_{\mathcal{H}}^{1/2}h_*\right\|_{\mathcal{H}}^2 - \left\|\hat{h}\right\|_{\mathcal{H}}^2\right) + \lambda_{\mathcal{G}'}B_2^2 + O\left(\delta_n\left\|\mathcal{T}\left(\hat{h} - T_{\mathcal{H}}^{1/2}h_*\right)\right\|_2 + \delta_n^2\right).$$

Combining this with the expression for $\frac{1}{4}\left\|\mathcal{T}\left(\hat{h} - T_{\mathcal{H}}^{1/2}h_*\right)\right\|_2^2 + \lambda_{\mathcal{H}}\left\|T_{\mathcal{H}}^{-1/2}\hat{h} - h_*\right\|_2^2$ gives:

$$\frac{1}{4}\left\|\mathcal{T}\left(\hat{h} - T_{\mathcal{H}}^{1/2}h_*\right)\right\|_2^2 + \lambda_{\mathcal{H}}\left\|T_{\mathcal{H}}^{-1/2}\hat{h} - h_*\right\|_2^2 \leqslant \frac{1}{2}\left\|\mathcal{T}\left(h_0 - T_{\mathcal{H}}^{1/2}h_*\right)\right\|_2^2 + \lambda_{\mathcal{G}'}B_2^2 + O\left(\delta_n\left\|\mathcal{T}\left(\hat{h} - T_{\mathcal{H}}^{1/2}h_*\right)\right\|_2 + \delta_n^2\right).$$

Applying the inequality between the arithmetic and geometric means (AM-GM) to the term $\delta_n\left\|\mathcal{T}\left(\hat{h} - T_{\mathcal{H}}^{1/2}h_*\right)\right\|_2$, and re-arranging terms, we obtain

$$\lambda_{\mathcal{H}}\left\|T_{\mathcal{H}}^{-1/2}\hat{h} - h_*\right\|_2^2 + \frac{1}{8}\left\|\mathcal{T}\left(\hat{h} - T_{\mathcal{H}}^{1/2}h_*\right)\right\|_2^2 \leqslant \frac{1}{2}\left\|\mathcal{T}\left(h_0 - T_{\mathcal{H}}^{1/2}h_*\right)\right\|_2^2 + \lambda_{\mathcal{G}'}B_2^2 + O\left(\delta_n^2\right). \tag{15}$$

Noting that $\left\|\mathcal{T}\left(\hat{h} - T_{\mathcal{H}}^{1/2}h_*\right)\right\|_2^2 \geqslant 0$, we have

$$\lambda_{\mathcal{H}}\left\|T_{\mathcal{H}}^{-1/2}\hat{h} - h_*\right\|_2^2 \leqslant \frac{1}{2}\left\|\mathcal{T}\left(h_0 - T_{\mathcal{H}}^{1/2}h_*\right)\right\|_2^2 + \lambda_{\mathcal{G}'}B_2^2 + O\left(\delta_n^2\right).$$

Using the inequality $\|A + B\|_2^2 \leqslant 2\|A\|_2^2 + 2\|B\|_2^2$ and the last expression, we get

$$\begin{aligned}\left\|T_{\mathcal{H}}^{-1/2}\left(\hat{h} - h_0\right)\right\|_2^2 &\leqslant 2\left\|T_{\mathcal{H}}^{-1/2}\hat{h} - h_*\right\|_2^2 + 2\left\|h_* - T_{\mathcal{H}}^{-1/2}h_0\right\|_2^2 \\ &\leqslant \frac{1}{\lambda_{\mathcal{H}}}\left\|\mathcal{T}\left(h_0 - T_{\mathcal{H}}^{1/2}h_*\right)\right\|_2^2 + \frac{2\lambda_{\mathcal{G}'}B_2^2}{\lambda_{\mathcal{H}}} + O\left(\frac{\delta_n^2}{\lambda_{\mathcal{H}}}\right) + 2\left\|h_* - T_{\mathcal{H}}^{-1/2}h_0\right\|_2^2.\end{aligned}$$

Using that

$$\left\|\hat{h}-h_0\right\|_2^2=\left\|T_{\mathcal{H}}^{1/2}\circ T_{\mathcal{H}}^{-1/2}\left(\hat{h}-h_0\right)\right\|_2^2\leqslant\left\|T_{\mathcal{H}}^{1/2}\right\|\cdot\left\|T_{\mathcal{H}}^{-1/2}\left(\hat{h}-h_0\right)\right\|_2^2\leqslant\tilde{\eta}\left\|T_{\mathcal{H}}^{-1/2}\left(\hat{h}-h_0\right)\right\|_2^2$$

and the bounds from Theorem 4, with high probability, we have that

$$\left\|\hat{h}-h_0\right\|_2^2=O\left(\frac{\delta_n^2}{\lambda_{\mathcal{H}}}+\frac{\lambda_{\mathcal{G}'}}{\lambda_{\mathcal{H}}}+\|w_0\|_2^2\lambda_{\mathcal{H}}^{\min\{\beta,1\}}\right).$$

To derive the other bound, we use again the inequality $\|A+B\|_2^2\leqslant 2\|A\|_2^2+2\|B\|_2^2$ along with expression (15), to obtain

$$\begin{aligned}\left\|\mathcal{T}\left(\hat{h}-h_0\right)\right\|_2^2\leqslant&2\left\|\mathcal{T}\left(\hat{h}-T_{\mathcal{H}}^{1/2}h_*\right)\right\|_2^2+2\left\|\mathcal{T}\left(T_{\mathcal{H}}^{1/2}h_*-h_0\right)\right\|_2^2\\ \leqslant&10\left\|\mathcal{T}\left(h_0-T_{\mathcal{H}}^{1/2}h_*\right)\right\|_2^2+16\lambda_{\mathcal{G}'}B_2^2+O\left(\delta_n^2\right).\end{aligned}$$

Using the bounds from Theorem 4, with high probability, we have

$$\left\|\mathcal{T}\left(\hat{h}-h_0\right)\right\|_2^2=O\left(\delta_n^2+\lambda_{\mathcal{G}'}+\|w_0\|_2^2\lambda_{\mathcal{H}}^{\min\{\beta+1,2\}}\right).$$

This completes the proof.

Proof. [Theorem 6] Recall that g^* is defined as

$$g_*=\arg\min_{g\in\mathscr{L}^2(X,L,Z)}\left\{\left\|\widetilde{\mathcal{T}}_G(\pi_0-g)\right\|_2^2+\lambda_{\mathcal{G}}\|g\|_2^2\right\},$$

where $\widetilde{\mathcal{T}}_G=\frac{1}{2}\mathcal{T}^*\circ I_q\circ T_{\mathcal{G}}^{1/2}$ and $\pi_0=T_{\mathcal{G}}^{-1/2}\tilde{g}_0$. Since $\widetilde{\mathcal{T}}_G$ is a bounded linear operator, the regularized solution g_* (see Theorem 16.4 of Kress (2010)) is given by

$$g_*=\left(\widetilde{\mathcal{T}}_G^*\circ\widetilde{\mathcal{T}}_G+\lambda_{\mathcal{G}}I\right)^{-1}\widetilde{\mathcal{T}}_G^*\circ\widetilde{\mathcal{T}}_G\pi_0.$$

Consequently,

$$\left\|g_*-T_{\mathcal{G}}^{-1/2}\tilde{g}_0\right\|_2=\left\|\left\{I-\left(\widetilde{\mathcal{T}}_G^*\circ\widetilde{\mathcal{T}}_G+\lambda_{\mathcal{G}}I\right)^{-1}\widetilde{\mathcal{T}}_G^*\circ\widetilde{\mathcal{T}}_G\right\}T_{\mathcal{G}}^{-1/2}\tilde{g}_0\right\|_2.$$

Since $\pi_0=T_{\mathcal{G}}^{-1/2}\tilde{g}_0=\left(\widetilde{\mathcal{T}}_G^*\circ\widetilde{\mathcal{T}}_G\right)^{\tilde{\beta}/2}z_0$, applying an approach analogous to the proof of 4 yields the desired bounds.

Proof. [Theorem 7] For given functions g_0, g_* and $\hat{g}$, we define the function:

$$L(\tau):=\frac{1}{4}\left\|\mathcal{T}^*\left[g_0-I_q\circ T_{\mathcal{H}}^{1/2}h_*-\tau I_q\left(\hat{\Gamma}-T_{\mathcal{H}}^{1/2}g_*\right)\right]\right\|_2^2+\lambda_{\mathcal{G}}\left\|g_*+\tau\left(T_{\mathcal{G}}^{-1/2}\hat{\Gamma}-g_*\right)\right\|_2^2,$$

which is strongly convex function with constant positive second derivative:

$$\frac{\partial^2 L(\tau)}{\partial \tau^2} \equiv 2\left\{\frac{1}{4}\left\|\tilde{I}_q \circ \mathcal{T}^*\left(\hat{\Gamma} - T_{\mathcal{G}}^{1/2} g_*\right)\right\|_2^2 + \lambda_{\mathcal{G}}\left\|T_{\mathcal{G}}^{-1/2}\hat{\Gamma} - g_*\right\|_2^2\right\}.$$

Clearly, $L(\tau)$ achieves its minimum at $\tau = 0$. Evaluating $L(\tau)$ at $\tau = 0$, we obtain:

$$L(0) = \frac{1}{4}\left\|\mathcal{T}^*\left(g_0 - I_q \circ T_{\mathcal{G}}^{1/2} g_*\right)\right\|_2^2 + \lambda_{\mathcal{G}}\left\|g_*\right\|_2^2 = \frac{1}{4}\left\|\tilde{I}_q \circ \mathcal{T}^*\left(\tilde{g}_0 - T_{\mathcal{G}}^{1/2} g_*\right)\right\|_2^2 + \lambda_{\mathcal{G}}\left\|T_{\mathcal{G}}^{1/2} g_*\right\|_{\mathcal{G}}^2.$$

and at $\tau = 1$,

$$L(1) = \frac{1}{4}\left\|\mathcal{T}^*\left(g_0 - I_q\hat{\Gamma}\right)\right\|_2^2 + \lambda_{\mathcal{G}}\left\|T_{\mathcal{G}}^{-1/2}\hat{\Gamma}\right\|_2^2 = \frac{1}{4}\left\|\tilde{I}_q \circ \mathcal{T}^*\left(\tilde{g}_0 - \hat{\Gamma}\right)\right\|_2^2 + \lambda_{\mathcal{G}}\left\|\hat{\Gamma}\right\|_{\mathcal{G}}^2.$$

Noting that $\frac{1}{2}\frac{\partial^2 L(\tau)}{\partial \tau^2} = L(1) - L(0)$, it follows that:

$$\begin{aligned}
&\frac{1}{4}\left\|\tilde{I}_q \circ \mathcal{T}^*\left(\hat{\Gamma} - T_{\mathcal{G}}^{1/2} g_*\right)\right\|_2^2 + \lambda_{\mathcal{G}}\left\|T_{\mathcal{G}}^{-1/2}\hat{\Gamma} - g_*\right\|_2^2 \\
&\quad = \frac{1}{4}\left\{\left\|\tilde{I}_q \circ \mathcal{T}^*\left(\tilde{g}_0 - \hat{\Gamma}\right)\right\|_2^2 - \left\|\tilde{I}_q \circ \mathcal{T}^*\left(\tilde{g}_0 - T_{\mathcal{G}}^{1/2} g_*\right)\right\|_2^2\right\} + \lambda_{\mathcal{G}}\left(\left\|\hat{\Gamma}\right\|_{\mathcal{G}}^2 - \left\|T_{\mathcal{G}}^{1/2} g_*\right\|_{\mathcal{G}}^2\right).
\end{aligned}$$

From this point, we can repeat the steps analogously to those in the proof of Theorem 5, replacing Lemma 16 with Lemma 17.

E.2 *Proofs of Lemmas*

Proof. [Lemma 1] Let $d_0 \in \mathcal{L}^2(V_d)$ satisfy $r_2(V_f)\mathbb{E}\left\{d_0(V_d)r_1(V)|V_f\right\} = \beta(V_f)$. For any $d \in \mathcal{L}^2(V_d)$ and any $f \in \mathcal{L}^2(V_f)$, we have

$$\begin{aligned}
&\mathbb{E}\big\{m(f;V) - d(V_d)r_1(V)r_2(V_f)f(V_f) - c \cdot r_2(V_f)f^2(V_f)\big\} \\
&\quad = \mathbb{E}\big\{\beta(V_f)f(V_f) - d(V_d)r_1(V)r_2(V_f)f(V_f) - c \cdot r_2(V_f)f^2(V_f)\big\} \\
&\quad = \mathbb{E}\left[r_2(V_f)\left\{\mathbb{E}\big[\left\{d_0(V_d) - d(V_d)\right\}r_1(V)|V_f\big]f(V_f) - cf^2(V_f)\right\}\right] \\
&\quad = -c\mathbb{E}\left\{r_2(V_f)\left(\frac{1}{2c}\mathbb{E}\big[\left\{d_0(V_d) - d(V_d)\right\}r_1(V)|V_f\big] - f(V_f)\right)^2\right\} \\
&\qquad + \frac{1}{4c}\mathbb{E}\left\{r_2(V_f)\mathbb{E}\big[\left\{d_0(V_d) - d(V_d)\right\}r_1(V)|V_f\big]^2\right\}.
\end{aligned}$$

Using that $r_2(V_f)^2 \equiv r_2(V_f)$, this expression is maximized when

$$r_2(V_f)f(V_f) = \frac{1}{2c}r_2(V_f)\mathbb{E}\left\{d_0(V_d) - d(V_d)|V_f\right\}$$

yielding:

$$\max_{f\in\mathcal{L}^2(V_f)} \mathbb{E}\big\{m(f;V) - d(V_d)r_1(V)r_2(V_f)f(V_f) - c\cdot r_2(V_f)f^2(V_f)\big\}$$
$$= \frac{1}{4c}\mathbb{E}\left\{r_2(V_f)\mathbb{E}\big[\left\{d_0(V_d) - d(V_d)\right\}r_1(V)|V_f\big]^2\right\}.$$

Given that the right-hand side is non-negative, we obtain:

$$\min_{d\in\mathcal{L}^2(V_d)}\max_{f\in\mathcal{L}^2(V_f)} \mathbb{E}\big\{m(f;V) - d(V_d)r_1(V)r_2(V_f)f(V_f) - c\cdot r_2(V_f)f^2(V_f)\big\} \geqslant 0.$$

Furthermore, choosing $d = d_0$ makes the last expression equal to zero, so d_0 belongs to:

$$\arg\min_{d\in\mathcal{L}^2(V_d)}\max_{f\in\mathcal{L}^2(V_f)} \mathbb{E}\big\{m(f;V) - d(V_d)r_1(V)r_2(V_f)f(V_f) - c\cdot r_2(V_f)f^2(V_f)\big\}.$$

Finally, suppose $d_1 \in \mathcal{L}^2(V_d)$ is another function that minimizes the same objective. Then, $r_2(V_f)\mathbb{E}\left\{\mathbb{E}\left[\left\{d_0(V_d) - d_1(V_d)\right\}r(V)|V_f\right]^2\right\} = 0$, implying $r_2(V_f)\mathbb{E}\left[\left\{d_0(V_d) - d_1(V_d)\right\}r(V)|V_f\right] = 0$ almost everywhere. Hence, $r_2(V_f)\mathbb{E}\left\{d_1(V_d)r(V)|V_f\right\} = r_2(V_f)\mathbb{E}\left\{d_0(V_d)r(V)|V_f\right\} = \beta(V_f)$ almost everywhere. This completes the proof.

Proof. [Lemma 2] The proof follows from a direct application of Picard's Theorem (Lemma 5 in Web Appendix D). Observe that, for each $(x,l) \in \text{supp}\,(X,L)$, the integral equation (2) can be rewritten as $\mathcal{T}^W_{x,l}h = r_0$ with $r_0 = \mathbb{E}(Y|X=x,L=l,U)$. Also, the operator $\left(\mathcal{T}^W_{x,l}\right)^*$: $\mathcal{L}^2\left(U|X=x,L=l\right) \to \mathcal{L}^2\left(W|X=x,L=l\right)$ defined for any $q \in \mathcal{L}^2\left(U|X=x,L=l\right)$ by

$$\left[\left(\mathcal{T}^W_{x,l}\right)^* q\right](x,l,W) := \mathbb{E}\left\{q(X,L,U)|X=x,L=l,W\right\},$$

is the adjoint of $\mathcal{T}^W_{x,l}$. Then, under assumption 9a, the kernel of $\left(\mathcal{T}^W_{x,l}\right)^*$ is trivial, i.e., $\mathcal{N}\left[\left(\mathcal{T}^W_{x,l}\right)^*\right] = \{0\}$. Consequently, by assumption 11a, $r_0 \in \mathcal{L}^2(U|X=x,L=l) = \mathcal{N}\left[\left(\mathcal{T}^W_{x,l}\right)^*\right]^\perp$, which is the condition (i) of Lemma 5. Furthermore, assumption 10 guarantees that the operator $\mathcal{T}^W_{x,l}$ is compact, while assumption 11b ensures that condition (ii) of Lemma 5 is met. Therefore, by Lemma 5, there exists a solution to the integral equation (2).

Proof. [Lemma 3]

Proof or part i). Suppose $h_0(x, l, w)$ solves equation (4). Then

$$\begin{aligned}
\mathbb{E}\{\mathbb{E}(Y|X, L, U)|X = x, L = l, Z = z\} &= \mathbb{E}\left\{\mathbb{E}(Y|X, L, Z, U)|X = x, L = l, Z = z\right\} \text{ by assumption 4} \\
&= \mathbb{E}(Y|X = x, L = l, Z = z) \\
&= \int h_0(x, l, w) p_{W|X,L,Z}(w|x, l, z) dw \qquad h_0 \text{ solves (4)} \\
&= \int h_0(x, l, w) \left\{ \int p_{(W,U)|X,L,Z}(w, u|x, l, z) du \right\} dw \\
&= \iint h_0(x, l, w) p_{W|X,L,U}(w|x, l, u) p_{U|X,L,Z}(u|x, l, z) du dw \qquad \text{by assumption 5} \\
&= \int \left\{ \int h_0(x, l, w) p_{W|X,L,U}(w|x, l, u) dw \right\} p_{U|X,L,Z}(u|x, l, z) du \\
&= \mathbb{E}\left\{ \int h_0(x, l, w) p_{W|X,L,U}(w|x, l, u) dw \Big| X = x, L = l, Z = z \right\}.
\end{aligned}$$

Consequently, we have

$$\mathbb{E}\left[\left\{ \mathbb{E}(Y|X, L, U) - \int h_0(x, l, w) p_{W|X,L,U}(w|x, l, u) dw \right\} \Big| X = x, L = l, Z = z \right] = 0.$$

Then, by assumption 9b, for almost every u such that $(x, l, u) \in \text{supp}\,(X, L, U)$,

$$\mathbb{E}(Y|X = x, L = l, U = u) - \int h_0(x, l, w)\, p_{W|X,L,U}(w|x, l, u)\, dw = 0,$$

which shows that h_0 also solves the integral equation (2).

Proof or part ii). Suppose $g_0(x, l, z)$ solves equation (5). Then

$$
\begin{aligned}
&\mathbb{E}\{\alpha_0(X, L, U)|X = x, L = l, W = w\} \\
&\quad = \int I\big\{x \in q\left(\mathcal{S}(l), l\right)\big\} \frac{dq^{-1}(x, l)}{dx} \frac{p_{X|L,U}\left\{q^{-1}\left(x, l\right)|l, u\right\}}{p_{X|L,U}\left(x|l, u\right)} p_{U|X,L,W}(u|x, l, w)du \\
&\quad \overset{\text{(i)}}{=} I\big\{x \in q\left(\mathcal{S}(l), l\right)\big\} \frac{dq^{-1}(x, l)}{dx} \int \frac{p_{U|X,L,W}(u|x, l, w)}{p_{X|L,U,W}\left(x|l, u, w\right)} p_{X|L,U,W}\left\{q^{-1}\left(x, l\right)|l, u, w\right\} du \\
&\quad = I\big\{x \in q\left(\mathcal{S}(l), l\right)\big\} \frac{dq^{-1}(x, l)}{dx} \int \frac{p_{U|L,W}(u|l, w)}{p_{X|L,W}\left(x|l, w\right)} p_{X|L,U,W}\left\{q^{-1}\left(x, l\right)|l, u, w\right\} du \\
&\quad = I\big\{x \in q\left(\mathcal{S}(l), l\right)\big\} \frac{dq^{-1}(x, l)}{dx} \frac{1}{p_{X|L,W}\left(x|l, w\right)} \int p_{(X,U)|L,W}\left\{q^{-1}\left(x, l\right), u|l, w\right\} du \\
&\quad = I\big\{(x \in q\left(\mathcal{S}(l), l\right)\big\} \frac{dq^{-1}(x, l)}{dx} \frac{p_{X|L,W}\left\{q^{-1}\left(x, l\right)|l, w\right\}}{p_{X|L,W}\left(x|l, w\right)} \\
&\quad = \alpha_0(x, l, w) \\
&\quad = \int g_0(x, l, z) p_{Z|X,L,W}(z|x, l, w)dz \qquad\qquad h_0 \text{ solves (5)} \\
&\quad = \int g_0(x, l, z) \left\{\int p_{(Z,U)|X,L,W}(z, u|x, l, w)du\right\} dz \\
&\quad = \iint g_0(x, l, z) p_{Z|X,L,U}(z|x, l, u) p_{U|X,L,W}(u|x, l, w)dudz \qquad \text{by assumption 5} \\
&\quad = \int \left\{\int g_0(x, l, z) p_{Z|X,L,U}(z|x, l, u)dz\right\} p_{U|X,L,W}(u|x, l, w)du \\
&\quad = \mathbb{E}\left\{\int g_0(x, l, z) p_{Z|X,L,U}(z|x, l, u)dw|X = x, L = l, W = w\right\},
\end{aligned}
$$

where the equality in (i) follows from assumption (5). Using assumption assumption 9a, we conclude as in part i).

Proof. [Lemma 4] Let $\psi(P_0)$ be the target parameter with $P_0 \in \mathcal{M}$ where $\mathcal{M}$ is the collection of all distributions satisfying the assumptions of Theorem 2. Under such assumptions, $\psi(P_0)$ can be expressed as follows:

$$\psi(P_0) = \frac{\mathbb{E}_{P_0}\left\{Y g_0(X, L, Z)\right\}}{\mathbb{E}_{P_0}\left[I\left\{(X, L) \in \mathcal{S}\right\}\right]},$$

where g_0 denotes the minimum-norm solution to the integral equation (5) under P_0.

Let P_t denote a parametric submodel with score at $t = 0$ denoted by S and such that

$P_t|_{t=0} = P_0$. Let

$$\alpha_{P_t}(x,l,w) := \{x \in q(\mathcal{S}(l),l)\} \frac{dq^{-1}(x,l)}{dx} \frac{p_{t,X,L,W}\{q^{-1}(x,l),l,w\}}{p_{t,X,L,W}(x,l,w)},$$

where $P_{t,X|L,W}|_{t=0} = P_{X|L,W}$, and let g_{P_t} be a curve such that $g_{P_t}|_{t=0} = g_0$, $\frac{dg_{P_t}}{dt}|_{t=0}$ exist, and

$$\alpha_{P_t}(X,L,W) = \mathbb{E}_{P_t}\left[g_{P_t}(X,L,Z)|X,L,W\right].$$

The previous condition implies that

$$\frac{d}{dt}\alpha_{P_t}(X,L,W)\Big|_{t=0} = \mathbb{E}_{P_0}\left\{\frac{d}{dt}g_{P_t}(X,L,Z)\big|_{t=0}\Big|X,L,W\right\} + \mathbb{E}_{P_0}\big\{g_0(X,L,Z)S(Z|X,L,W)|X,L,W\big\}.$$

From the definition of $\alpha_{P_t}(x,l,w)$, we have

$$\frac{d}{dt}\alpha_{P_t}(X,L,W)\Big|_{t=0} = \alpha_0(X,L,W)\left\{S(q^{-1}(X,L),L,W) - S(X,L,W)\right\},$$

hence,

$$\begin{aligned}\mathbb{E}_{P_0}\left\{\frac{d}{dt}g_{P_t}(X,L,Z)\big|_{t=0}\Big|X,L,W\right\} =& \alpha_0(X,L,W)\left\{S(q^{-1}(X,L),L,W) - S(X,L,W)\right\} \\ &- \mathbb{E}_{P_0}\big\{g_0(X,L,Z)S(Z|X,L,W)|X,L,W\big\}.\end{aligned}$$

The perturbed parameter is given by

$$\psi(P_t) = \frac{\mathbb{E}_{P_t}\{Y g_{P_t}(X,L,Z)\}}{\mathbb{E}_{P_t}\left[I\{(X,L)\in\mathcal{S}\}\right]}.$$

Then

$$\begin{aligned}\frac{d\psi(P_t)}{dt}\Big|_{t=0} =& \underbrace{\frac{\mathbb{E}_{P_0}\{Y g_0(X,L,Z)S(O)\}}{\mathbb{E}_{P_0}\left[I\{(X,L)\in\mathcal{S}\}\right]}}_{T_1} + \underbrace{\frac{\mathbb{E}_{P_0}\left\{Y\frac{d}{dt}g_0(X,L,Z)|_{t=0}\right\}}{\mathbb{E}_{P_0}\left[I\{(X,L)\in\mathcal{S}\}\right]}}_{T_2} \\ &\underbrace{- \frac{\mathbb{E}_{P_0}\left[I\{(X,L)\in\mathcal{S}\}S(X,L)\right]}{\left\{\mathbb{E}_{P_0}\left[I\{(X,L)\in\mathcal{S}\}\right]\right\}^2}\mathbb{E}_{P_0}\{Y g_0(X,L,Z)\}}_{T_3}.\end{aligned}$$

The term T_1 is already in the desired form. For the numerator of T_2, let h_0 be the minimum-

norm solution to the integral equation (4), then

$$\begin{aligned}\mathbb{E}_{P_0}\left\{Y\frac{d}{dt}g_0(X,L,Z)|_{t=0}\right\} &= \mathbb{E}_{P_0}\left\{\mathbb{E}_{P_0}(Y|X,L,Z)\frac{d}{dt}g_0(X,L,Z)|_{t=0}\right\}\\ &= \mathbb{E}_{P_0}\left[\mathbb{E}_{P_0}\{h_0(X,L,W)|X,L,Z\}\frac{d}{dt}g_0(X,L,Z)|_{t=0}\right]\\ &= \mathbb{E}_{P_0}\left\{h_0(X,L,W)\frac{d}{dt}g_0(X,L,Z)|_{t=0}\right\}\\ &= \mathbb{E}_{P_0}\left[h_0(X,L,W)\mathbb{E}_{P_0}\left\{\frac{d}{dt}g_0(X,L,Z)|_{t=0}\big|X,L,W\right\}\right].\end{aligned}$$

Using the expression for $\mathbb{E}_{P_0}\left\{\frac{d}{dt}g_0(X,L,Z)|_{t=0}\big|X,L,W\right\}$ yields

$$\begin{aligned}\mathbb{E}_{P_0}\left\{Y\frac{d}{dt}g_0(X,L,Z)|_{t=0}\right\} =& \mathbb{E}_{P_0}\left\{h_0(X,L,W)\alpha_0(X,L,W)S(q^{-1}(X,L),L,W)\right\}\\ &- \mathbb{E}_{P_0}\left\{h_0(X,L,W)\alpha_0(X,L,W)S(X,L,W)\right\} \qquad (16)\\ &- \mathbb{E}_{P_0}\left[h_0(X,L,W)\mathbb{E}_{P_0}\{g_0(X,L,Z)S(Z|X,L,W)|X,L,W\}\right].\end{aligned}$$

For the first term in the right hand side, we have

$$\begin{aligned}&\mathbb{E}_{P_0}\left\{h(X,L,W)\alpha_0(X,L,W)S(q^{-1}(X,L),L,W)\right\}\\ &\quad= \iiint h(x,l,w)\alpha_0(x,l,w)S(q^{-1}(x,l),l,w)p_{X|L,W}(x|l,w)p_{L,W}(l,w)dxdldw\\ &\quad= \iiint h(x,l,w)I\{x\in q\left(\mathcal{S}(l),l\right)\}\frac{dq^{-1}(x,l)}{dx}\frac{p_{X|L,W}\left\{q^{-1}\left(x,l\right)|l,w\right\}}{p_{X|L,W}\left(x|l,w\right)}\\ &\qquad\qquad\times S(q^{-1}(x,l),l,w)p_{X|L,W}(x|l,w)p_{L,W}(l,w)dxdldw\\ &\quad\overset{\text{(i)}}{=} \iiint h_0\left\{q(\tilde{x},l),l,w\right\}I\left\{(\tilde{x},l)\in\mathcal{S}\right\}\frac{dq^{-1}\left\{q(\tilde{x},l),l\right\}}{dx}\cdot\frac{dq\left(\tilde{x},l\right)}{d\tilde{x}}p_{X|L,W}(\tilde{x}|l,w)\\ &\qquad\qquad\times S(\tilde{x},l,w)p_{L,W}(l,w)d\tilde{x}dldw\\ &\quad\overset{\text{(ii)}}{=} \iint h\left\{q(\tilde{x},l),l,w\right\}I\left\{(\tilde{x},l)\in\mathcal{S}\right\}S(\tilde{x},l,w)p_{X,L,W}(\tilde{x},l,w)d\tilde{x}dldw\\ &\quad= \mathbb{E}_{P_0}\left[h\left\{q(X,L),L,W\right\}\cdot I\left\{(X,L)\in\mathcal{S}\right\}S(X,L,W)\right],\end{aligned}$$

where the equality in (i) follows from assumption 6, which allows us to apply the change of variable $\tilde{x}=q^{-1}(x,l)$ in x for each $l\in\text{supp}\,(L)$, and from the fact that $x\in q(\mathcal{S}(l),l)$ if and only if $(q^{-1}(x,l),l)\in\mathcal{S}$. For the equality in (ii), we used that for each $l\in\text{supp}\,(L)$,

$$1=\frac{d\tilde{x}}{d\tilde{x}}=\frac{d\left[q^{-1}\left\{q(\tilde{x},l),l\right\}\right]}{d\tilde{x}}=\frac{dq^{-1}\left\{q(\tilde{x},l),l\right\}}{dx}\cdot\frac{dq\left(\tilde{x},l\right)}{d\tilde{x}}.$$

Using that $\mathbb{E}_{P_0}\{S(Y,Z|X,L,W|X,L,W)\}=0$ and $S(O)=S(Y,Z|X,L,W)+S(X,L,W)$, we obtain

$$\begin{aligned}
&\mathbb{E}_{P_0}\left[h\left\{q(X,L),L,W\right\}\cdot I\left\{(X,L)\in\mathcal{S}\right\}S(X,L,W)\right]\\
&\qquad=\mathbb{E}_{P_0}\left[h\left\{q(X,L),L,W\right\}\cdot I\left\{(X,L)\in\mathcal{S}\right\}S(X,L,W)\right]\\
&\qquad\quad+\mathbb{E}_{P_0}\Big[h\left\{q(X,L),L,W\right\}\cdot I\left\{(X,L)\in\mathcal{S}\right\}\underbrace{\mathbb{E}_{P_0}\left\{S(Y,Z|X,L,W)|X,L,W\right\}}_{=0}\Big]\\
&\qquad=\mathbb{E}_{P_0}\left[h\left\{q(X,L),L,W\right\}\cdot I\left\{(X,L)\in\mathcal{S}\right\}S(O)\right].
\end{aligned}$$

In addition, using that $S(O)=S(Y|X,L,Z,W)+(Z|X,L,W)+S(X,L,W))$ and that $\mathbb{E}_{P_0}\{S(Y|X,L,Z,W)|X,L,Z,W\}=0$, the negative terms in the right hand side of (16) are handled as follows:

$$\begin{aligned}
&\mathbb{E}_{P_0}\left\{h_0(X,L,W)\alpha_0(X,L,W)S(X,L,W)\right\}\\
&\qquad\qquad+\mathbb{E}_{P_0}\left[h_0(X,L,W)\mathbb{E}_{P_0}\big\{g_0(X,L,Z)S(Z|X,L,W)|X,L,W\big\}\right]\\
&\qquad=\mathbb{E}_{P_0}\Big[h_0(X,L,W)\underbrace{\mathbb{E}_{P_0}\left\{g_0(X,L,Z)|X,L,W\right\}}_{=\alpha_0(X,L,W)}S(X,L,W)\Big]\\
&\qquad\qquad+\mathbb{E}_{P_0}\left[h_0(X,L,W)g_0(X,L,Z)S(Z|X,L,W)\right]\\
&\qquad=\mathbb{E}_{P_0}\left[h_0(X,L,W)g_0(X,L,W)\left\{S(X,L,W)+S(Z|X,L,W)\right\}\right]\\
&\qquad\qquad+\mathbb{E}_{P_0}\Big[h_0(X,L,W)g_0(X,L,Z)\underbrace{\mathbb{E}_{P_0}\left\{S(Y|Z,X,L,W)|X,L,Z,W\right\}}_{=0}\Big]\\
&\qquad=\mathbb{E}_{P_0}\left\{h_0(X,L,W)g_0(X,L,Z)S(O)\right\}.
\end{aligned}$$

Hence, the term T_2 is equal to

$$\mathbb{E}_{P_0}\left\{\frac{h\left\{q(X,L),L,W\right\}\cdot I\left\{(X,L)\in\mathcal{S}\right\}-h_0(X,L,W)g_0(X,L,W)}{\mathbb{E}_{P_0}\left[I\left\{(X,L)\in\mathcal{S}\right\}\right]}S(O)\right\}.$$

For the term $\mathbb{E}_{P_0}\left[I\left\{(X,L)\in\mathcal{S}\right\}S(X,L)\right]$ of T_3, using that $S(O)=S(Y,Z,W|X,L)+$

$S(X,L)$ and $\mathbb{E}_{P_0}\{S(Y,Z,W|X,L)|X,L\}=0$, we have

$$\begin{aligned}\mathbb{E}_{P_0}\left[I\left\{(X,L)\in\mathcal{S}\right\}S(X,L)\right] =&\mathbb{E}_{P_0}\left[I\left\{(X,L)\in\mathcal{S}\right\}S(X,L)\right]\\ &+\mathbb{E}_{P_0}\left[I\left\{(X,L)\in\mathcal{S}\right\}\mathbb{E}_{P_0}\left\{S(Y,Z,W|X,L)|X,L\right\}\right]\\ =&\mathbb{E}_{P_0}\left[I\left\{(X,L)\in\mathcal{S}\right\}S(O)\right].\end{aligned}$$

Therefore, we have shown that

$$\frac{d\psi(P_t)}{dt}\Big|_{t=0}=\mathbb{E}_{P_0}\left[\left\{\frac{\phi(O:h_0,g_0)}{\mathbb{E}_{P_0}\left[I\left\{(X,L)\in\mathcal{S}\right\}\right]}-\psi_0\frac{I\left\{(X,L)\in\mathcal{S}\right\}}{\mathbb{E}_{P_0}\left[I\left\{(X,L)\in\mathcal{S}\right\}\right]}\right\}S(O)\right].$$

Since

$$\mathbb{E}_{P_0}\left\{\frac{\phi(O:h_0,g_0)}{\mathbb{E}_{P_0}\left[I\left\{(X,L)\in\mathcal{S}\right\}\right]}-\psi_0\frac{I\left\{(X,L)\in\mathcal{S}\right\}}{\mathbb{E}_{P_0}\left[I\left\{(X,L)\in\mathcal{S}\right\}\right]}\right\}=0,$$

an influence function for ψ_0 is given by

$$\psi(P)^1=\frac{\phi(O:h_0,g_0)}{\mathbb{E}_{P_0}\left[I\left\{(X,L)\in\mathcal{S}\right\}\right]}-\psi_0\frac{I\left\{(X,L)\in\mathcal{S}\right\}}{\mathbb{E}_{P_0}\left[I\left\{(X,L)\in\mathcal{S}\right\}\right]}.$$

Proof. [Lemma 5] See proof in pages 311-312 of Kress (2010).

Proof. [Lemma 6] For any $h\in\mathcal{L}^2(X,L,W)$, we have

$$\begin{aligned}&\iiint h_0(x,l,w)\alpha_0(x,l,w)p_{X,L,W}(x,l,w)dxdldw\\ &\quad=\iiint h_0(x,l,w)I\left\{x\in q\left(\mathcal{S}(l),l\right)\right\}\frac{dq^{-1}\left(x,l\right)}{dx}\frac{p_{X|L,W}\left\{q^{-1}\left(x,l\right)|l,w\right\}}{p_{X|L,W}\left(x|l,w\right)}p_{X,L,W}(x,l,w)dxdldw\\ &\quad=\iint\left[h_0(x,l,w)I\left\{x\in q\left(\mathcal{S}(l),l\right)\right\}\frac{dq^{-1}\left(x,l\right)}{dx}p_{X|L,W}\left\{q^{-1}\left(x,l\right)|l,w\right\}dx\right]p_{L,W}(l,w)dldw\\ &\quad\overset{\text{(i)}}{=}\iiint h_0\left\{q(\tilde{x},l),l,w\right\}I\left\{(\tilde{x},l)\in\mathcal{S}\right\}\frac{dq^{-1}\left\{q(\tilde{x},l),l\right\}}{dx}\cdot\frac{dq\left(\tilde{x},l\right)}{d\tilde{x}}p_{X|L,W}(\tilde{x}|l,w)p_{L,W}(l,w)d\tilde{x}dldw\\ &\quad\overset{\text{(ii)}}{=}\iiint h_0\left\{q(\tilde{x},l),l,w\right\}I\left\{(\tilde{x},l)\in\mathcal{S}\right\}p_{X,L,W}(\tilde{x},l,w)d\tilde{x}dldw,\end{aligned}$$

where the equality in (i) follows from assumption 6, which allows us to apply the change of variable $\tilde{x}=q^{-1}(x,l)$ in x for each $l\in\text{supp}\left(L\right)$, and from the fact that $x\in q(\mathcal{S}(l),l)$ if and only if $(q^{-1}(x,l),l)\in\mathcal{S}$. For the equality in (ii), we used that for each $l\in\text{supp}\left(L\right)$,

$$1=\frac{d\tilde{x}}{d\tilde{x}}=\frac{d\left[q^{-1}\left\{q(\tilde{x},l),l\right\}\right]}{d\tilde{x}}=\frac{dq^{-1}\left\{q(\tilde{x},l),l\right\}}{dx}\cdot\frac{dq\left(\tilde{x},l\right)}{d\tilde{x}}.$$

This concludes the proof.

Proof. [Lemma 7] Let $h^\dagger$ and $g^\dagger$ be any solutions to the observed equations (4) and (5), respectively; and take any $h \in \mathcal{L}^2(X, L, W)$ and $g \in \mathcal{L}^2(X, L, Z)$. Then,

$$
\begin{aligned}
&\mathbb{E}\big\{\phi(O; h, g) - \phi(O; h^\dagger, g^\dagger)\big\} \\
&= \mathbb{E}\left[(h - h^\dagger)\left\{q(X, L), L, W\right\} \cdot I\left\{(X, L) \in \mathcal{S}\right\}\right] + \mathbb{E}\left[g(X, L, Z)\left\{Y - h(X, L, W)\right\}\right] \\
&\quad - \mathbb{E}\Big[g^\dagger(X, L, Z) \underbrace{\mathbb{E}\left\{Y - h^\dagger(X, L, W) | X, L, Z\right\}}_{=0 \text{ since } h^\dagger \text{ solves (4)}}\Big] \\
&= \mathbb{E}\left\{(h - h^\dagger)(X, L, W)\alpha_0(X, L, W)\right\} + \mathbb{E}\big\{g(X, L, Z) \cdot \mathbb{E}\left(Y | X, L, Z\right)\big\} \qquad \text{by Lemma 6} \\
&\quad - \mathbb{E}\big\{g(X, L, Z) \cdot h(X, L, W)\big\} \\
&= \mathbb{E}\Big[(h - h^\dagger)(X, L, W) \underbrace{\mathbb{E}\left\{g^\dagger(X, L, Z) | X, L, W\right\}}_{g^\dagger \text{ solves (5)}}\Big] + \mathbb{E}\Big[g(X, L, Z) \underbrace{\mathbb{E}\left\{h^\dagger(X, L, W) | X, L, Z\right\}}_{h^\dagger \text{ solves (4)}}\Big] \\
&\quad - \mathbb{E}\left[g(X, L, Z) \cdot h(X, L, W)\right] \\
&= \mathbb{E}\left[\left\{(h - h^\dagger)(X, L, W)\right\}\left\{(g^\dagger - g)(X, L, Z)\right\}\right].
\end{aligned}
$$

Taking $h \equiv h^\dagger$ and $g \equiv 0$ gives

$$
\mathbb{E}\left[h^\dagger\left\{q\left(X, L\right), L, W\right\} \cdot I\left\{(X, L) \in \mathcal{S}\right\}\right] = \mathbb{E}\left\{\phi(O; h^\dagger, 0)\right\} = \mathbb{E}\left\{\phi(O; h^\dagger, g^\dagger)\right\},
$$

and taking $h \equiv 0$ and $g \equiv g^\dagger$ gives

$$
\mathbb{E}\left\{Y g^\dagger(X, L, Z)\right\} = \mathbb{E}\left\{\phi(O; 0, g^\dagger)\right\} = \mathbb{E}\left\{\phi(O; h^\dagger, g^\dagger)\right\}.
$$

This concludes the proof.

Proof. [Lemma 8]

Take $(x, l, u) \in \operatorname{supp}(X, L, U)$ such that $x \in \mathcal{S}(l)$. By the definitions of $\mathcal{S}$ and $\mathcal{S}(l)$, $q(x, l) \in \operatorname{supp}(X | L = l)$. Then, by assumption 3, $q(x, l) \in \operatorname{supp}(X | L = l, U = u)$, which

implies that $(q(x,l),l,u) \in \text{supp}\,(X,L,U)$. Then,

$$\begin{aligned}\mathbb{E}\left[Y\left\{q\left(X,L\right)\right\}|X=x,L=l,U=u\right] &= \mathbb{E}\left[Y\left\{q\left(x,l\right)\right\}|X=x,L=l,U=u\right] \\ &= \mathbb{E}\left[Y\left\{q\left(x,l\right)\right\}|X=q(x,l),L=l,U=u\right] \\ &= \mathbb{E}\left\{Y|X=q(x,l),L=l,U=u\right\},\end{aligned}$$

where the second equality follows from assumption 2 and the last one from assumption 1.

Now, let h_0 be a solution to equation (2). Since $(q(x,l),l,u) \in \text{supp}\,(X,L,U)$ and $q(x,l) \in q(\mathcal{S}(l),l)$, $(q(x,l),l,u)$ lies in the domain where h_0 solves equation (2). Then,

$$\begin{aligned}\mathbb{E}\left[Y|X=q(x,l),L=l,U=u\right] &= \int h_0\{q(x,l),l,w\}p_{W|X,L,U}\{w|q(x,l),l,u\}dw \\ &= \int h_0\{q(x,l),l,w\}p_{W|X,L,U}(w|x,l,u)dw \qquad \text{by assumption 5} \\ &= \mathbb{E}\big[h_0\{q(X,L),L,W\}|X=x,L=l,U=u\big].\end{aligned}$$

This concludes the proof.

Proof. [Lemma 9]

Proof of part i). For any $h \in \mathcal{L}^2(X,L,W)$, we have

$$\begin{aligned}&\left\|h\left\{q(X,L),L,W\right\}\cdot I\left\{(X,L)\in\mathcal{S}\right\}\right\|_2^2 \\ &= \iint\left[\int h^2\left\{q(x,l),l,w\right\}I\left\{(x,l)\in\mathcal{S}\right\}p_{X|L,W}(x,l,w)dx\right]p_{L,W}(l,w)dldw \\ &\overset{\text{(i)}}{=} \iiint h^2\left(\tilde{x},l,w\right)I\left\{\tilde{x}\in q(\mathcal{S}(l),l)\right\}\frac{dq^{-1}\left(\tilde{x},l\right)}{d\tilde{x}}p_{X|L,W}(q^{-1}(\tilde{x},l)|l,w)p_{L,W}(l,w)d\tilde{x}dldw \\ &= \iiint h^2\left(\tilde{x},l,w\right)I\left\{\tilde{x}\in q(\mathcal{S}(l),l)\right\}\frac{dq^{-1}\left(\tilde{x},l\right)}{d\tilde{x}}\frac{p_{X|L,W}\left\{q^{-1}\left(\tilde{x},l\right)|l,w\right\}}{p_{X|L,W}\left(\tilde{x}|l,w\right)}p_{X,L,W}(\tilde{x},l,w)d\tilde{x}dldw \\ &= \iiint h^2(\tilde{x},l,w)\alpha_0(\tilde{x},l,w)p_{X,L,W}(\tilde{x},l,w)d\tilde{x}dldw \\ &\leqslant B\iiint h^2\left(\tilde{x},l,w\right)p_{X,L,W}(\tilde{x},l,w)d\tilde{x}dldw \qquad\qquad \alpha_0 \text{ is bounded} \\ &= O\left(\|h\|_2^2\right),\end{aligned}$$

where the equality in (i) follows from the change of variable $\tilde{x} = q(x,l)$ in x for each $l \in \text{supp}\,(L)$, which is justified by assumption 6.

Proof of part ii). For any $g \in \mathcal{L}^2(X, L, W)$, we have

$$\left\|Yg(X,L,Z)\right\|_2^2 = \mathbb{E}\left[\{Yg(X,L,Z)\}^2\right] \leqslant B^2\mathbb{E}\left\{g(X,L,Z)^2\right\} = O\left(\|g\|_2^2\right).$$

Proof. [Lemma 10] From the definition of ϕ, we have

$$\begin{aligned}\phi(O;\hat{h},\hat{g}) - \phi(O;h_0,g_0) = &\left(\hat{h} - h_0\right)\{q(X,L),L,W\} \cdot I\{(X,L) \in \mathcal{S}\} \\ &+ Y\left(\hat{g} - g_0\right)(X,L,Z) \\ &- \hat{h}(X,L,W)\hat{g}(X,L,Z) + h_0(X,L,W)g_0(X,L,Z),\end{aligned}$$

and by the triangle inequality

$$\begin{aligned}\left\|\phi(O;\hat{h},\hat{g}) - \phi(O;h_0,g_0)\right\|_2 \leqslant &\left\|\left(\hat{h} - h_0\right)\{q(X,L),L,W\} \cdot I\{(X,L) \in \mathcal{S}\}\right\|_2 \\ &+ \left\|Y\left(\hat{g} - g_0\right)(X,L,Z)\right\|_2 + \left\|\hat{h}\cdot\hat{g} - h_0\cdot g_0\right\|_2.\end{aligned}$$

By Lemma 9, Lemma 2.12 in Van der Vaart (2010), and the fact that $\hat{h}$ and g are norm-consistent estimators of h_0 and g_0, we get

$$\left\|\left(\hat{h} - h_0\right)\{q(X,L),L,W\} \cdot I\{(X,L) \in \mathcal{S}\}\right\|_2 = O_p\left(\left\|\hat{h} - h_0\right\|_2\right) = O_p(o_p(1)) = o_p(1),$$

and

$$\left\|Y\left(\hat{g} - g_0\right)(X,L,Z)\right\|_2 = O_p\left(\left\|\hat{g} - g_0\right\|_2\right) = O_p(o_p(1)) = o_p(1).$$

The term $\left\|\hat{h}\cdot\hat{g} - h_0\cdot g_0\right\|_2$ is upper bounded by

$$\begin{aligned}&\min\left\{\left\|\hat{h}(\hat{g} - g_0)\right\|_2^2 + \left\|g_0(\hat{h} - h_0)\right\|_2^2, \left\|\hat{g}(\hat{h} - h_0)\right\|_2^2 + \left\|h_0(\hat{g} - g_0)\right\|_2^2\right\} \\ &\quad\leqslant \min\left\{\|\hat{h}\|_\infty\left\|\hat{g} - g_0\right\|_2^2 + \|g_0\|_\infty\left\|\hat{h} - h_0\right\|_2^2, \|\hat{g}\|_\infty\left\|\hat{h} - h_0\right\|_2^2 + \|h_0\|_\infty\left\|\hat{g} - g_0\right\|_2^2\right\}\end{aligned}$$

which is $o_p(1)$ since $\left\|\hat{h} - h_0\right\|_2^2 = o_p(1)$, $\left\|\hat{g} - g_0\right\|_2^2 = o_p(1)$ and either $\|\hat{h}\|_\infty + \|g_0\|_\infty \leqslant B$ or $\|\hat{g}\|_\infty + \|h_0\|_\infty \leqslant B$.

To get the other result, note that

$$\mathbb{E}_n\left[\phi\left\{O;\hat{h},\hat{g}\right\}\right]-\mathbb{E}\left\{\phi(O;h_0,g_0\right\}=\underbrace{\mathbb{E}_n\left[\phi\left\{O;\hat{h},\hat{g}\right\}\right]-\mathbb{E}_n\left\{\phi(O;h_0,g_0)\right\}}_{T_1}$$
$$+\underbrace{\mathbb{E}_n\left\{\phi(O;h_0,g_0)\right\}-\mathbb{E}\left\{\phi(O;h_0,g_0)\right\}}_{T_2}.$$

For any $\epsilon > 0$, using that we have independent observations $O_1,\ldots,O_n$ which are also independent of $\hat{h}$ and $\hat{g}$ and Markov inequality, we have

$$\begin{aligned}P\left[\left|\mathbb{E}_n\left[\phi\left\{O;\hat{h},\hat{g}\right\}\right]-\mathbb{E}_n\left\{\phi(O;h_0,g_0\right\}\right|\geqslant\varepsilon\right]&=P\left[\left|\phi\left\{O;\hat{h},\hat{g}\right\}-\phi(O;h_0,g_0)\right|\geqslant\varepsilon\right]\\&=P\left[\left(\phi\left\{O;\hat{h},\hat{g}\right\}-\phi(O;h_0,g_0)\right)^2\geqslant\varepsilon^2\right]\\&\leqslant\frac{\left\|\phi\left\{O;\hat{h},\hat{g}\right\}-\phi(O;h_0,g_0)\right\|_2^2}{\varepsilon^2}\to 0\text{ as }n\to\infty.\end{aligned}$$

Hence, $T_1 = o_p(1)$. Next, by the weak law of large numbers, it follows that $T_2 = o_p(1)$. Therefore, $\mathbb{E}_n\left[\phi\left\{O;\hat{h},\hat{g}\right\}\right]-\mathbb{E}\left\{\phi(O;h_0,g_0\right\}=o_p(1)$.

Proof. [Proof of Lemma 11] See proof in pages 89-91 of Wainwright (2019).

Proof. [Proof of Lemma 12] Fix $x=(x_1,\ldots,x_n)\in\mathcal{X}^n$ and let $\mathcal{S}$ be a dense subset of $\mathcal{F}$. Define

$$A_1(x):=\sup_{f\in\mathcal{F}}g(x,f)\quad\text{and}\quad A_2(x):=\sup_{f\in\mathcal{S}}g(x,f).$$

We will show that $A_1(x)=A_2(x)$.

Since $\mathcal{S}\subset\mathcal{F}$, it follows that $A_2(x)\leqslant A_1(x)$. Suppose that $A_2(x)<A_1(x)$ and set $\varepsilon:=A_1(x)-A_2(x)>0$. By the definition of the supremum, there exists $f_\varepsilon\in\mathcal{F}$ such that

$$A_1(x)-\frac{\varepsilon}{2}<g(x,f_\varepsilon).$$

By the continuity of the mapping $f\mapsto g(x,f)$, there exists $\delta>0$ such that $d(f,f_\varepsilon)<\delta$ implies $|g(x,f)-g(x,f_\varepsilon)|<\frac{\varepsilon}{2}$. Since $\mathcal{S}$ is dense in $\mathcal{F}$, there exits $f_s\in\mathcal{S}$ such that $d(f_s,f_\varepsilon)<$

δ. Then,

$$g(x, f_s) > g(x, f_\varepsilon) - \frac{\varepsilon}{2} > A_1(x) - \varepsilon = A_2(x),$$

which contradicts the definition of $A_2(x)$. Thus, $A_1(x) = A_2(x)$, and we conclude that

$$\sup_{f \in \mathcal{F}} g(x, f) = \sup_{f \in \mathcal{S}} g(x, f).$$

Next, since $g(\cdot, f) : \mathcal{X}^n \to \mathbb{R}$ is measurable for all $f \in \mathcal{S}$, and the mapping $f \mapsto \tilde{g}(x, f)$ is continuous for all $x \in \mathcal{X}^n$, Lemma A.3.17 of Steinwart and Christmann (2008) ensures that the mapping $(x, f) \mapsto g(x, f)$ is measurable on $\mathcal{X}^n \times \mathcal{S}$. Because $\mathcal{S}$ is countable, the function $\sup_{f \in \mathcal{S}} g(x, f)$ is also measurable. Hence,

$$Z := \sup_{f \in \mathcal{S}} g(X_1, \ldots, X_n, f) = \sup_{f \in \mathcal{F}} g(X_1, \ldots, X_n, f)$$

is a well-defined random variable. This completes the proof.

Proof. [Proof of Lemma 13] The proof relies on an application of Talagrand's concentration inequality for empirical processes, which is stated in Lemma 11.

Define the classes $\widetilde{\mathcal{F}} := \{\mathcal{F} - f^*\} \cup \{-\mathcal{F} + f^*\}$ and $\mathcal{G}_{\mathcal{L}}(r) := \left\{ \mathcal{L}_{\tilde{f}} - \mathbb{E}[\mathcal{L}_{\tilde{f}}] \mid \tilde{f} \in \widetilde{\mathcal{F}}, \|\tilde{f}\|_2 \leqslant r \right\}$, and the random variable

$$Z_n^g(r) := \sup_{g \in \mathcal{G}_{\mathcal{L}}(r)} \frac{1}{n} \sum_{i=1}^{n} g(X_i, Z_i).$$

Using this notation, observe that

$$Z_n(r) = Z_n^g(r).$$

Since $\mathcal{F}$ is uniformly bounded by 1, and the loss function ℓ is L-Lipschitz, the class $\mathcal{G}_{\mathcal{L}}(r)$ is uniformly bounded by $b = 4L$. Furthermore, for any $g \in \mathcal{G}_{\mathcal{L}(r)}$, there exists a corresponding function $f_g \in \widetilde{\mathcal{F}}$ such that $g = \mathcal{L}_{f_g} - \mathbb{E}[\mathcal{L}_{f_g}]$, and $\|f_g\|_2 \leqslant r$. By the Lipschitz condition

$$\mathbb{E}[g(X, Z)^2] = \operatorname{var}\left(\mathcal{L}_{f_g}\right) \leqslant \mathbb{E}\left[\left(\mathcal{L}_{f_g}\right)^2\right] \leqslant L^2 \mathbb{E}\left[f_g^2\right] \leqslant L^2 r^2.$$

Therefore, $\sup_{g \in \mathcal{G}_{\mathcal{L}}(r)} \mathbb{E}\left[g(X, Z)^2\right] \leqslant L^2 r^2$.

Next, applying Lemmas 11 and 12 to control the upper tail of $Z_n^g(r)$, we get for any $u > 0$,

$$\Pr\left[Z_n^g(r) \geqslant \mathbb{E}\left[Z_n^g(r)\right] + u\right] \leqslant 2\exp\left(-c_1 \frac{nu^2}{L^2 r^2 + 8L\mathbb{E}\left[Z_n^g(r)\right] + Lu}\right),$$

where $c_1 = \frac{1}{8e}$. Since $Z_n^g(r) = Z_n(r)$ and $\mathbb{E}\left[Z_n^g(r)\right] \leqslant 2\mathbb{E}\left[Z_n(r)\right]$, it follows that

$$\Pr\left[Z_n(r) \geqslant 2\mathbb{E}\left[Z_n(r)\right] + u\right] \leqslant 2\exp\left(-c_1 \frac{nu^2}{L^2 r^2 + 16L\mathbb{E}\left[Z_n(r)\right] + Lu}\right).$$

We now turn to bounding $\mathbb{E}\left[Z_n(r)\right]$. We proceed as follows:

$$\begin{aligned}
\mathbb{E}\left[Z_n(r)\right] &= \mathbb{E}_{X,Z}\left[\sup_{f\in\mathcal{F}:\|f-f^*\|_2\leqslant r}\left|\mathbb{E}_n\left(\mathcal{L}_f - \mathcal{L}_{f^*}\right) - \mathbb{E}\left(\mathcal{L}_f - \mathcal{L}_{f^*}\right)\right|\right] \\
&= \mathbb{E}_{X,Z}\left[\sup_{\|f-f^*\|_2\leqslant r}\left|\frac{1}{n}\sum_{i=1}^n\left\{\left(\mathcal{L}_f - \mathcal{L}_{f^*}\right) - \mathbb{E}\left(\mathcal{L}_f - \mathcal{L}_{f^*}\right)\right\}\right|\right] \\
&= \mathbb{E}_{X,Z}\left[\sup_{\|f-f^*\|_2\leqslant r}\left|\frac{1}{n}\sum_{i=1}^n\left\{\ell\left(f(X_i), Z_i\right) - \ell\left(f^*(X_i), Z_i\right) - \mathbb{E}\left[\ell\left(f(X_i), Z_i\right) - \ell\left(f^*(X_i), Z_i\right)\right]\right\}\right|\right] \\
&\overset{(i)}{=} \mathbb{E}_{X,Z}\left\{\sup_{\|f-f^*\|_2\leqslant r}\left|\mathbb{E}_{X',Z'}\left[\frac{1}{n}\sum_{i=1}^n\left\{\ell\left(f(X_i), Z_i\right) - \ell\left(f^*(X_i), Z_i\right) - \left[\ell\left(f(X_i'), Z_i'\right) - \ell\left(f^*(X_i'), Z_i'\right)\right]\right\}\right]\right|\right\} \\
&\overset{(ii)}{\leqslant} \mathbb{E}_{X,Z,X',Z'}\left[\sup_{\|f-f^*\|_2\leqslant r}\left|\frac{1}{n}\sum_{i=1}^n\left\{\ell\left(f(X_i), Z_i\right) - \ell\left(f^*(X_i), Z_i\right) - \left[\ell\left(f(X_i'), Z_i'\right) - \ell\left(f^*(X_i'), Z_i'\right)\right]\right\}\right|\right],
\end{aligned}$$

where in (i) we used a symmetrization argument by introducing independent copies (X_i', Z_i') for $i = 1, \ldots, n$, and in (ii) we used the inequality $\sup_{h\in\mathcal{H}}\mathbb{E}\left\{|h(Y)|\right\} \leqslant \mathbb{E}\left\{\sup_{h\in\mathcal{H}}|h(Y)|\right\}$, which is valid for any class $\mathcal{H}$ of real-valued functions.

Now, let $\varepsilon_1, \ldots, \varepsilon_n$ denote n i.i.d. Rademacher random variables, independent of (X_i, Z_i) and (X_i', Z_i') for all $i = 1, \ldots, n$. Given this independence assumption, for any function $f \in \mathcal{F}$, the random vector with components $\varepsilon_i\left\{\ell\left(f(X_i), Z_i\right) - \ell\left(f^*(X_i), Z_i\right) - \left[\ell\left(f(X_i'), Z_i'\right) - \ell\left(f^*(X_i'), Z_i'\right)\right]\right\}$ has the same joint distribution as the random vector with components $\ell\left(f(X_i), Z_i\right) - \ell\left(f^*(X_i), Z_i\right) -$

$[\ell(f(X_i'), Z_i') - \ell(f^*(X_i'), Z_i')]$. Hence,

$$\begin{aligned}
&\mathbb{E}_{X,Z,X',Z'}\left[\sup_{\|f-f^*\|_2\leqslant r}\left|\frac{1}{n}\sum_{i=1}^{n}\left\{\ell(f(X_i),Z_i)-\ell(f^*(X_i),Z_i)-[\ell(f(X_i'),Z_i')-\ell(f^*(X_i'),Z_i')]\right\}\right|\right]\\
&=\mathbb{E}_{X,Z,X',Z',\varepsilon}\left[\sup_{\|f-f^*\|_2\leqslant r}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\left\{\ell(f(X_i),Z_i)-\ell(f^*(X_i),Z_i)-[\ell(f(X_i'),Z_i')-\ell(f^*(X_i'),Z_i')]\right\}\right|\right]\\
&\leqslant 2\mathbb{E}_{X,Z,\varepsilon}\left[\sup_{\|f-f\|_2\leqslant r}\left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\left\{\ell(f(X_i),Z_i)-\ell(f^*(X_i),Z_i)\right\}\right|\right]\\
&\overset{(iii)}{\leqslant} 4L\mathbb{E}_{X,\varepsilon}\left[\sup_{\|f-f^*\|_2\leqslant r}\left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\left\{f(X_i)-f^*(X_i)\right\}\right|\right]\\
&=4L\mathcal{R}_n(\mathcal{F}-f^*,r),
\end{aligned}$$

where in (iii) we used the Ledoux-Talagrand contraction inequality (see inequality (5.61) of Wainwright (2019)). Thus, we have shown that

$$\begin{aligned}
\Pr[Z_n(r)\geqslant 8L\mathcal{R}_n(\mathcal{F}-f^*,r)+u] &\leqslant \Pr[Z_n(r)\geqslant 2\mathbb{E}[Z_n(r)]+u]\\
&\leqslant 2\exp\left(-c_1\frac{nu^2}{L^2r^2+64L\mathcal{R}_n(\mathcal{F}-f^*,r)+Lu}\right).
\end{aligned}$$

Finally, since the class $star(\mathcal{F}-f^*)$ is star-shaped, by the arguments in the proof of Lemma 13.6 of Wainwright (2019), it follows that the function $\delta\mapsto\frac{\mathcal{R}(star(\mathcal{F}-f^*),\delta)}{\delta}$ is non-decreasing on the interval $(0,\infty)$. Hence, noting that $\mathcal{F}-f^*\subset star(\mathcal{F}-f^*)$, for any $r\geqslant\delta_n$,

$$\frac{\mathcal{R}(\mathcal{F}-f^*,r)}{r}\leqslant\frac{\mathcal{R}(star(\mathcal{F}-f^*),r)}{r}\leqslant\frac{\mathcal{R}(star(\mathcal{F}-f^*),\delta_n)}{\delta_n}\leqslant\delta_n,$$

and consequently,

$$\begin{aligned}
\Pr[Z_n(r)\geqslant 8Lr\delta_n+u] &\leqslant \Pr[Z_n(r)\geqslant 8L\mathcal{R}_n(\mathcal{F}-f^*,r)+u]\\
&\leqslant c_1\exp\left(-c_2\frac{nu^2}{L^2r^2+64Lr\delta_n+Lu}\right).
\end{aligned}$$

Proof. [Proof of Lemma 14] This proof uses a peeling argument For $m=1,2,\ldots,$ define the events

$$\mathcal{S}_m:=\left\{f\in\mathcal{F}\mid\theta^{m-1}\delta_n\leqslant\|f-f^*\|_2\leqslant\theta^m\delta_n\right\},$$

for $\theta=10/9$. Note that for any function $f\in\mathcal{F}$, since $\|f\|_2\leqslant\sup_{x\in\mathcal{X}}|f(x)|\leqslant 1$, it follows

that $\|f - f^*\|_2 \leqslant 2$. Hence, any function $f \in \mathcal{F} \cap \{\|f - f^*\|_2 \geqslant \delta_n\}$ belongs to some $\mathcal{S}_m$ for $m \in \{1, \ldots, M\}$, where $M \leqslant \max\{20\log(2/\delta_n), 1\}$. This upper bound for M follows from noting that M must satisfy $\theta^{M-1}\delta_n \leqslant 2 \leqslant \theta^M \delta_n$. If $\delta_n > 2/\theta$, $M = 1$ is enough. If $\delta_n \leqslant 2/\theta$, $M \leqslant \frac{\log(2/\delta_n)}{\log\theta} + 1 \leqslant 2\frac{\log(2/\delta_n)}{\log\theta} \leqslant 20\log(2/\delta_n)$. Hence, $\mathcal{E}_1 \subset \bigcup_{m=1}^{M} [\mathcal{E}_1 \cap \mathcal{S}_m]$.

Now, if the event $\mathcal{E}_1 \cap \mathcal{S}_m$ occurs for some $m \in \{1, \ldots, M\}$, then there is a function f with $\theta^{m-1}\delta_n \leqslant \|f - f^*\|_2 \leqslant r_m := \theta^m \delta_n$ such that

$$\left|\mathbb{E}_n\left(\mathcal{L}_f - \mathcal{L}_{f^*}\right) - \mathbb{E}\left(\mathcal{L}_f - \mathcal{L}_{f^*}\right)\right| \geqslant 10L\delta_n \|f - f^*\|_2 \geqslant 10L\delta_n \frac{r_m}{\theta} = 9L\delta_n r_m.$$

Consequently, we have $\Pr[\mathcal{E}_1 \cap \mathcal{S}_m] \leqslant \Pr[Z(r_m) \geqslant 9L\delta_n r_m]$, where $Z(r)$ was defined in Lemma 13. Applying Lemma 13 with $r = r_m$ and $u = Lr_m\delta_n$, we obtain that $\Pr[\mathcal{E}_1 \cap \mathcal{S}_m]$ happens with probability at most $2\exp\left(-\frac{1}{8e} \cdot \frac{nL^2 r_m^2 \delta_n^2}{L^2 r_m^2 + 64 L r_m \delta_n + L^2 r_m \delta_n}\right) \leqslant 2\exp\left(-c_3 n\delta_n^2\right)$ with $c_3 = \frac{L}{16e(L+32)}$ where the last bound was obtained by using $\delta_n \leqslant r_m$ in the denominator. If $M = 1$, we have concluded the proof. Otherwise, $M \leqslant 20\log(2/\delta_n)$, giving that

$$\Pr[\mathcal{E}_1] \leqslant \sum_{m=1}^{M} \Pr[\mathcal{E}_1 \cap \mathcal{S}_m] \leqslant 2M\exp\left(-c_3 n\delta_n^2\right) \leqslant 40\exp\left(-c_3 n\delta_n^2 + \log(\log(2/\delta_n))\right).$$

Finally, the condition $\delta_n^2 \geqslant c_1 \frac{\log(\log(n))}{n}$ implies that $\log(\log(2/\delta_n)) \leqslant \frac{c_3}{2} n\delta_n^2$ for $c_1 = 32e + \frac{1024e}{L}$, and consequently that

$$\Pr[\mathcal{E}_1] \leqslant 40\exp\left(-c_2 n\delta_n^2\right),$$

with $c_2 = \frac{c_3}{2} = \frac{L}{32e(L+32)}$, as desired.

To see that $\delta_n^2 \geqslant c_1 \frac{\log(\log(n))}{n}$ implies $\log(\log(2/\delta_n)) \leqslant \frac{c_3}{2} n\delta_n^2$ for $c_1 = 32e + \frac{1024e}{L}$, first note that we are only interested in $\delta_n \in (0, 2]$, otherwise, the set $\{f \in \mathcal{F} : \|f - f^*\|_2 \geqslant \delta_n\}$ would be empty. Then

$$\frac{\delta_n}{2} \geqslant \left(\frac{\delta_n}{2}\right)^2 \geqslant \frac{c_1}{4}\frac{\log(\log(n))}{n}.$$

Now, for $n \geqslant 3$ and $c_1 \geqslant \frac{4}{\log(\log(3))}$, it follows that $n \geqslant 2/\delta_n$. Hence, taking $c_1 \geqslant \max\left\{\frac{4}{\log(\log(3))}, \frac{2}{c_3}\right\}$, it follows that

$$\log(\log(2/\delta_n)) \leqslant \log(\log(n)) \leqslant \frac{n\delta_n^2}{c_1} \leqslant \frac{c_3}{2} n\delta_n^2.$$

Hence, the desired condition is always satisfied for $c_1 = \max\left\{\frac{4}{\log(\log(3))}, 32e + \frac{1024e}{L}\right\} = 32e + \frac{1024e}{L}$.

Proof. [Proof of Lemma 15] Consider the events:

$$\mathcal{E}_0 = \left\{Z_n(\delta_n) \geqslant 9L\delta_n^2\right\},$$

$$\mathcal{E}_1 = \left\{\exists f \in \mathcal{F} : \|f - f^*\|_2 \geqslant \delta_n \text{ and } \left|\mathbb{E}_n\left(\mathcal{L}_f - \mathcal{L}_{f^*}\right) - \mathbb{E}\left(\mathcal{L}_f - \mathcal{L}_{f^*}\right)\right| \geqslant 10L\delta_n \|f - f^*\|_2\right\},$$

where Z_n was defined in Lemma 13. Note that if (14) is violated, then either $\mathcal{E}_0$ or $\mathcal{E}_1$ must occur. Applying Lemma 13 with $r = \delta_n$ and $u = L\delta_n^2$ implies that $\mathcal{E}_0$ happens with probability at most $2\exp\left(-c_4 n\delta_n^2\right)$ where $c_4 = \frac{L}{16e(L+32)}$. In addition, by Lemma 14 we obtain that $\mathcal{E}_1$ happens with probability at most $40\exp\left(-c_5 n\delta_n^2\right)$ where $c_5 = \frac{L}{32e(L+32)}$. Thus, given that $-c_4 \leqslant -c_5$, (14) is violated with probability at most $80\exp\left(-c_5 n\delta_n^2\right)$. Therefore, (14) occurs with probability at least $1 - c_2\exp(-c_3 n\delta_n^2)$ where $c_2 = 80$ and $c_3 = \frac{L}{32e(L+32)}$.

Proof. [Proof of Lemma 16] For any $h \in \mathcal{H}_{B_1}$, define $g_h := \frac{1}{2}\mathcal{T}(h_0 - h)$. By the triangle inequality, we have $\|h_0 - h\|_{\mathcal{H}} \leqslant \|h_0\|_{\mathcal{H}} + \|h\|_{\mathcal{H}} \leqslant B + B_1 = 2B + 1$. Thus, assumptions 14 and 15 ensure that $g_h \in \mathcal{G}'_{B_2}$.

Now, consider the squared norm of $\frac{1}{2}\mathcal{T}(h_0 - \hat{h})$, which can be written as:

$$\frac{1}{4}\left\|\mathcal{T}\left(h_0 - \hat{h}\right)\right\|_2^2 = \mathbb{E}\left[Y g_{\hat{h}}(X, L, Z) - \hat{h}(X, L, W) g_{\hat{h}}(X, L, Z) - g_{\hat{h}}(X, L, Z)^2\right].$$

Since $\mathcal{H}$ and $\mathcal{G}'$ are reproducing kernel Hilbert spaces generated by a Gaussian kernel, they satisfy the conditions required in Lemma 12 (see details in Section F.3). Moreover, the loss function $\ell_1(v) = v$ is 1-Lipschitz and $\ell_2(v) = v^2$ is 2-Lipschitz over $v \in [-1, 1]$. Then, applying Lemma 15, we can upper bound the above by its empirical counterpart, yielding, with high probability,

$$\begin{aligned}\frac{1}{4}\left\|\mathcal{T}\left(h_0 - \hat{h}\right)\right\|_2^2 \leqslant\ & \mathbb{E}_n\left[Y g_{\hat{h}}(X, L, Z) - \hat{h}(X, L, W) g_{\hat{h}}(X, L, Z) - g_{\hat{h}}(X, L, Z)^2\right] \\ & + O\left(\delta_n\left[\sqrt{\mathbb{E}\left[Y^2 g_{\hat{h}}(X, L, Z)^2\right]} + \sqrt{\mathbb{E}\left[\hat{h}(X, L, W)^2 g_{\hat{h}}(X, L, Z)^2\right]} + \|g_{\hat{h}}\|_2\right] + \delta_n^2\right),\end{aligned}$$

Next, applying Lemma 9b to $\mathbb{E}\left[Y^2 g_{\hat{g}}^2(X, L, Z)\right]$, and using the fact that $\hat{h}$ is uniformly

bounded, we upper bound the previous expression by

$$\mathbb{E}_n\left[Yg_{\hat{h}}(X,L,Z) - \hat{h}(X,L,W)g_{\hat{h}}(X,L,Z) - g_{\hat{h}}(X,L,Z)^2\right] + O\left(\delta_n\|g_{\hat{h}}\|_2 + \delta_n^2\right).$$

Using the result of Yinxiang Wu et al. (PENDING FINAL CITATION), the last expression is equivalent to

$$\mathbb{E}_n\left[Yg_{\hat{h}}(X,L,Z) - \hat{h}(X,L,W)g_{\hat{h}}(X,L,Z) - g_{\hat{h}}(X,L,Z)^2\right] - \lambda_{\mathcal{G}'}\|g_{\hat{h}}\|_{\mathcal{G}'}^2 + \lambda_{\mathcal{G}'}\|g_{\hat{h}}\|_{\mathcal{G}'}^2 + O\left(\delta_n\|g_{\hat{h}}\|_2 + \delta_n^2\right)$$

which, can be further upper bounded by

$$\sup_{g\in\mathcal{G}'_{B_2}}\left\{\mathbb{E}_n\left[Yg(X,L,Z) - \hat{h}(X,L,W)g(X,L,Z) - g(X,L,Z)^2\right] - \lambda_{\mathcal{G}'}\|g\|_{\mathcal{G}'}^2\right\}$$
$$+ \lambda_{\mathcal{G}'}B_2^2 + O\left(\delta_n\|g_{\hat{h}}\|_2 + \delta_n^2\right).$$

We now make use of the fact that $\hat{h}$ is the empirical risk minimizer to obtain:

$$\sup_{g\in\mathcal{G}'_{B_2}}\left\{\mathbb{E}_n\left[Yg(X,L,Z) - \hat{h}(X,L,W)g(X,L,Z) - g(X,L,Z)^2\right] - \lambda_{\mathcal{G}'}\|g\|_{\mathcal{G}'}^2\right\}$$
$$+ \lambda_{\mathcal{G}'}B_2^2 + O\left(\delta_n\|g_{\hat{h}}\|_2 + \delta_n^2\right)$$
$$\leqslant \sup_{g\in\mathcal{G}'_{B_2}}\left\{\mathbb{E}_n\left[Yg(X,L,Z) - T_{\mathcal{H}}^{1/2}h_*(X,L,W)g(X,L,Z) - g(X,L,Z)^2\right] - \lambda_{\mathcal{G}'}\|g\|_{\mathcal{G}'}^2\right\}$$
$$+ \lambda_{\mathcal{G}'}B_2^2 + \lambda_{\mathcal{H}}\left(\left\|T_{\mathcal{H}}^{1/2}h_*\right\|_{\mathcal{H}}^2 - \left\|\hat{h}\right\|_{\mathcal{H}}^2\right) + O\left(\delta_n\|g_{\hat{h}}\|_2 + \delta_n^2\right).$$

Applying Lemma 15 to the terms inside the empirical expectation and supreme as before, allow us to upper bound the last expression by

$$\sup_{g\in\mathcal{G}'_{B_2}}\left\{\mathbb{E}\left[Yg(X,L,Z) - T_{\mathcal{H}}^{1/2}h_*(X,L,W)g(X,L,Z) - g(X,L,Z)^2\right] + O\left(\delta_n\|g\|_2\right) - \lambda_{\mathcal{G}'}\|g\|_{\mathcal{G}'}^2\right\}$$
$$+ \lambda_{\mathcal{G}'}B_2^2 + \lambda_{\mathcal{H}}\left(\left\|T_{\mathcal{H}}^{1/2}h_*\right\|_{\mathcal{H}}^2 - \left\|\hat{h}\right\|_{\mathcal{H}}^2\right) + O\left(\delta_n\|g_{\hat{h}}\|_2 + \delta_n^2\right)$$

Next, we apply the AM-GM inequality to the term $O\left(\delta_n\|g\|_2\right)$ inside the supremum and use the fact that $-\lambda_{\mathcal{G}'}\|g\|_{\mathcal{G}'}^2 \leqslant -\frac{1}{2}\lambda_{\mathcal{G}'}\|g\|_{\mathcal{G}'}^2$, to upper bound the last expression by

$$\sup_{g\in\mathcal{G}'_{B_2}}\left\{\mathbb{E}\left[Yg(X,L,Z) - T_{\mathcal{H}}^{1/2}h_*(X,L,W)g(X,L,Z) - \frac{1}{2}g(X,L,Z)^2\right] - \frac{1}{2}\lambda_{\mathcal{G}'}^h\|g\|_{\mathcal{G}'}^2\right\}$$
$$+ \lambda_{\mathcal{G}'}B_2^2 + \lambda_{\mathcal{H}}\left(\left\|T_{\mathcal{H}}^{1/2}h_*\right\|_{\mathcal{H}}^2 - \left\|\hat{h}\right\|_{\mathcal{H}}^2\right) + O\left(\delta_n\|g_{\hat{h}}\|_2 + \delta_n^2\right),$$

which is equivalent to

$$\frac{1}{2}\left\|\mathcal{T}\left(h_0 - T_{\mathcal{H}}^{1/2}h_*\right)\right\|_2^2 + \lambda_{\mathcal{H}}\left(\left\|T_{\mathcal{H}}^{1/2}h_*\right\|_{\mathcal{H}}^2 - \left\|\hat{h}\right\|_{\mathcal{H}}^2\right) + \lambda_{\mathcal{G}'}B_2^2 + O\left(\delta_n\|g_{\hat{h}}\|_2 + \delta_n^2\right).$$

Hence, we have established that

$$\frac{1}{4}\left\|\mathcal{T}\left(h_0 - \hat{h}\right)\right\|_2^2 \leqslant \frac{1}{2}\left\|\mathcal{T}\left(h_0 - T_{\mathcal{H}}^{1/2}h_*\right)\right\|_2^2 + \lambda_{\mathcal{H}}\left(\left\|T_{\mathcal{H}}^{1/2}h_*\right\|_{\mathcal{H}}^2 - \left\|\hat{h}\right\|_{\mathcal{H}}^2\right) + \lambda_{\mathcal{G}'}B_2^2 + O\left(\delta_n\|g_{\hat{h}}\|_2 + \delta_n^2\right).$$

Furthermore, since $\|g_{\hat{h}}\|_2 = \left\|\mathcal{T}\left(h_0 - \hat{h}\right)\right\|_2$, using the triangle inequality

$$\left\|\mathcal{T}\left(h_0 - \hat{h}\right)\right\|_2 \leqslant \left\|\mathcal{T}\left(h_0 - T_{\mathcal{H}}^{1/2}h_*\right)\right\|_2 + \left\|\mathcal{T}\left(\hat{h} - T_{\mathcal{H}}^{1/2}h_*\right)\right\|_2$$

yields

$$\frac{1}{4}\left\|\mathcal{T}\left(h_0 - \hat{h}\right)\right\|_2^2 \leqslant \frac{1}{2}\left\|\mathcal{T}\left(h_0 - T_{\mathcal{H}}^{1/2}h_*\right)\right\|_2^2 + \lambda_{\mathcal{H}}\left(\left\|T_{\mathcal{H}}^{1/2}h_*\right\|_{\mathcal{H}}^2 - \left\|\hat{h}\right\|_{\mathcal{H}}^2\right) + \lambda_{\mathcal{G}'}B_2^2 + O\left(\delta_n\left\|\mathcal{T}\left(h_0 - T_{\mathcal{H}}^{1/2}h_*\right)\right\|_2 + \delta_n\left\|\mathcal{T}\left(\hat{h} - T_{\mathcal{H}}^{1/2}h_*\right)\right\|_2 + \delta_n^2\right).$$

Finally, applying the AM-GM to the term $O\left(\delta_n\left\|\mathcal{T}\left(h_0 - T_{\mathcal{H}}^{1/2}h_*\right)\right\|_2\right)$, with high probability, we obtain

$$\frac{1}{4}\left\{\left\|\mathcal{T}\left(h_0 - \hat{h}\right)\right\|_2^2 - \left\|\mathcal{T}\left(h_0 - T_{\mathcal{H}}^{1/2}h_*\right)\right\|_2^2\right\} \leqslant \frac{1}{2}\left\|\mathcal{T}\left(h_0 - T_{\mathcal{H}}^{1/2}h_*\right)\right\|_2^2 + \lambda_{\mathcal{H}}\left(\left\|T_{\mathcal{H}}^{1/2}h_*\right\|_{\mathcal{H}}^2 - \left\|\hat{h}\right\|_{\mathcal{H}}^2\right) + \lambda_{\mathcal{G}'}B_2^2 + O\left(\delta_n\left\|\mathcal{T}\left(\hat{h} - T_{\mathcal{H}}^{1/2}h_*\right)\right\|_2 + \delta_n^2\right),$$

which completes the proof.

Proof. [Proof of Lemma 17] For any $g \in \mathcal{G}_{B_1}$, define $h_g := \frac{1}{2}\mathcal{T}^*(\tilde{g}_0 - g)$. By the triangle inequality, we have $\|\tilde{g}_0 - g\|_{\mathcal{G}} \leqslant \|\tilde{g}_0\|_{\mathcal{G}} + \|g\|_{\mathcal{G}} \leqslant B + B_1 = 2B + 1$. Thus, assumptions 17 and 18 ensure that $h_g \in \mathcal{H}'_{B_2}$.

Now, since $\tilde{g}_0$ is a solution to the integral equation (5), by the proof of Lemma 1, the squared norm of $\frac{1}{2}\tilde{I}_q \circ \mathcal{T}^*(\tilde{g}_0 - \hat{\Gamma})$, can be written as:

$$\begin{aligned}\frac{1}{4}\left\|\tilde{I}_q \circ \mathcal{T}^*\left(\tilde{g}_0 - \hat{\Gamma}\right)\right\|_2^2 = \mathbb{E}\big[&h\left\{q(X,L),L,W\right\}\cdot I\left\{(X,L)\in\mathcal{S}\right\}\\ &- I\left\{X \in q(\mathcal{S}(L),L)\right\}h_{\hat{\Gamma}}(X,L,W)g(X,L,Z)\\ &- I\left\{X \in q(\mathcal{S}(L),L)\right\}h_{\hat{\Gamma}}(X,L,W)^2\big]\end{aligned}$$

From here, the proof proceeds analogously to that of Lemma 16, but with Lemma 9b replaced by Lemma 9a.

Web Appendix F. Feasibility of the Assumptions on the Function Classes for Gaussian Reproducing Kernel Hilbert Spaces generated

F.1 *Injectivity of the Integral Operator Associated with a Gaussian kernel*

Let $\mathcal{V} = \mathcal{X} \times \mathcal{L} \times \mathcal{W}$, and suppose that the probability measure $P(v)$ on $\mathcal{V}$ is dominated by the Lebesgue measure. The integral operator $T_{\mathcal{H}}$ associated with the Gaussian kernel $K(v, v') = \exp\left(-\frac{\|v-v'\|_2^2}{2\sigma^2}\right)$ is defined as

$$(T_{K_{\mathcal{H}}}h)(v) = \int_{\mathcal{V}} K(v, v')h(v')dP(v') = \int_{\mathbb{R}^3} \exp\left(-\frac{\|v - v'\|_2^2}{2\sigma^2}\right) h(v')I\{v' \in \mathcal{V}\}p(v')dv'.$$

To establish injectivity, assume $T_{K_{\mathcal{H}}}h = 0$ for some function $h \in \mathscr{L}^2(X, L, W)$. Taking the Fourier transform on both sides and leveraging the convolution property of the integral operator, we have

$$\widehat{K}(\xi)\widehat{\tilde{h}}(\xi) = 0,$$

where $\widehat{K}(\xi) = (2\pi\sigma^2)^{3/2} \exp\left(-\frac{\sigma^2\|\xi\|_2^2}{2}\right)$ is the Fourier transform of the Gaussian kernel, and $\tilde{h}(v) = h(v)p(v)I\{v \in \mathcal{V}\}$.

Since $\widehat{K}(\xi) > 0$ for all $\xi \in \mathbb{R}^3$, it follows that $\widehat{\tilde{h}}(\xi) = 0$ for all $\xi \in \mathbb{R}^3$. By the injectivity of the Fourier transform, this implies $\tilde{h}(v)(v)I\{v \in \mathcal{V}\} = 0$ for all $v \in \mathbb{R}^3$. Thus, $h(v) = 0$ for all $v \in \mathcal{V}$, proving that $T_{K_{\mathcal{H}}}$ is injective.

A more general proof of injectivity for Gaussian kernels over any finite measure on $\mathbb{R}^d$ can be found in Theorem 4.47 of Steinwart and Christmann (2008).

F.2 *Closedness and Lipschitz Assumptions*

The following result from Dikkala et al. (2020) establishes sufficient conditions on the data distributions to ensure that assumptions 14 and 15 are satisfied.

LEMMA 18 (Lemma 7 of Dikkala et al. (2020)): *Suppose that: (i)* $p(w|x,l,z)$ *belongs to* $\mathcal{G}'$ *for all* $w \in \mathcal{W}$*, (ii) there exists a functional* $\kappa : \mathcal{X} \times \mathcal{L} \times \mathcal{W} \to \mathbb{R}$ *such that* $|h(x,l,w)| \leqslant \kappa(x,l,w)\|h\|_{\mathcal{H}}$ *for all* $h \in \mathcal{H}$*, and (iii)* $L := \int \kappa(x,l,w)\|p(w|x,l,z)\|_{\mathcal{G}'}dw < \infty$*. Then,* $\mathcal{T}h \in \mathcal{G}'$ *with* $\|\mathcal{T}h\|_{\mathcal{G}'} \leqslant L\|h\|_{\mathcal{H}}$*.*

Given that reproducing kernel Hilbert spaces $\mathcal{G}'$ generated by Gaussian kernels $K_{\mathcal{G}'}$ are dense in $\mathscr{L}^2(X,L,Z)$ (Theorems 4.63 of Steinwart and Christmann (2008)), condition (i) in Lemma is feasible.

For condition (ii), using the reproducing kernel property and the Cauchy-Schwarz inequality, we have:

$$|h(x,l,w)| = |\langle h, K_{\mathcal{H}}[\cdot;(x,l,w)]\rangle_{\mathcal{H}}| \leqslant \|h\|_{\mathcal{H}} \cdot \|K_{\mathcal{H}}[\cdot;(x,l,w)]\|_{\mathcal{H}} = \|h\|_{\mathcal{H}},$$

where, for a Gaussian kernel,

$$\|K_{\mathcal{H}}[\cdot;(x,l,w)]\|^2_{\mathcal{H}} = \langle K_{\mathcal{H}}[\cdot;(x,l,w)], K_{\mathcal{H}}[\cdot;(x,l,w)]\rangle_{\mathcal{H}} = K_{\mathcal{H}}[(x,l,w);(x,l,w)] = 1.$$

Thus, the functional $\kappa(x,l,w) \equiv 1$ satisfies condition (ii). This derivation supports Remark 5, which states that functions in a class $\mathcal{F}_B$ are uniformly bounded by B if $\mathcal{F}$ is a reproducing kernel Hilbert space generated by a Gaussian kernel.

To assess the feasibility of condition (iii) with $\kappa(x,l,w) \equiv 1$, observe:

$$\int \kappa(x,l,w)\|p(w|x,l,z)\|_{\mathcal{G}'}dw = \int \|p(w|x,l,z)\|_{\mathcal{G}'}dw.$$

Let $\mathcal{T}_{K_{\mathcal{G}'}} : \mathscr{L}^2(X,L,W) \to \mathscr{L}^2(X,L,W)$ denote the integral operator associated with $K_{\mathcal{G}'}$:

$$\mathcal{T}_{K_{\mathcal{G}'}}g(x,l,z) := \int K_{\mathcal{G}'}(x,l,z;x',l',z')\, g(x',l',z')p(x',l',z')dx'dl'dz',$$

Under an assumption analogous to 12, the operator $\mathcal{T}_{K_{\mathcal{G}'}}$ is Hilbert-Schmidt and admits a singular value decomposition. Specifically, it has a sequence of eigenfunctions $(\phi_j)_{j=1}$ forming an orthonormal basis of $\mathscr{L}^2(X,L,Z)$ and a sequence of non-negative eigenvalues $(\mu_j)_{j=1}^{\infty}$ converging to zero. Consequently, $\mathcal{G}'$ admits the representation:

$$\mathcal{G}' \equiv \left\{ g = \sum_{j=1}^{\infty} \gamma_j \varphi_j \Big| \text{ for some } (\gamma_j)_{j=1}^{\infty} \text{ with } \sum_{j=1}^{\infty} \frac{\gamma_j^2}{\mu_j} < \infty \right\}, \tag{17}$$

with the inner product:

$$\langle g, g' \rangle_{\mathcal{G}'} := \sum_{j=1}^{\infty} \frac{\langle g, \phi_j \rangle_2 \langle g', \phi_j \rangle_2}{\mu_j}.$$

Condition (iii) holds if the conditional densities $p(w|x, l, z)$ can be expressed in terms of the orthonormal basis $(\phi_j)_{j=1}^{\infty}$ with significant support on a finite subset of eigenfunctions. This corresponds to assuming a certain level of smoothness for $p(w|x, l, z)$.

F.3 *Verifying assumptions on Lemma 12*

Let $\mathcal{F}$ be a reproducing kernel Hilbert space of functions $f : \mathcal{X} \to \mathbb{R}$ generated by a Gaussian kernel $K_{\mathcal{F}}$ and equipped with the norm $\|\cdot\|_{\mathcal{F}}$. For each $f \in \mathcal{F}$, define the function $g(\cdot, f) : \mathcal{X}^n \to \mathbb{R}$ as follows:

$$g(x, f) := \frac{1}{n} \sum_{i=1}^{n} f(x_i), \quad \text{for all } x = (x_1, \ldots, x_n) \in \mathcal{X}^n.$$

Fix $x = (x_1, \ldots, x_n) \in \mathcal{X}^n$ and consider any $f, f' \in \mathcal{F}$. By the reproducing kernel property, the Cauchy-Schwarz inequality, and the fact that $\|K_{\mathcal{F}}(\cdot, x_i)\|_{\mathcal{F}} \equiv 1$ for all $i = 1, \ldots, n$, we obtain:

$$|g(x, f) - g(x, f')| = \left| \frac{1}{n} \sum_{i=1}^{n} (f - f')(x_i) \right| = \left| \frac{1}{n} \sum_{i=1}^{n} \langle f - f', K_{\mathcal{F}}(\cdot, x_i) \rangle \right| \leqslant \|f - f'\|_{\mathcal{F}}.$$

This establishes that the mapping $f \mapsto g(x, f)$ is continuous. Since $x \in \mathcal{X}^n$ was arbitrary, the continuity holds for all $x \in \mathcal{X}^n$.

Moreover, assumption 12 guarantees that $\mathcal{F}$ is separable. Let $\mathcal{S}$ be a dense subset of $\mathcal{F}$. For any $B > 0$, it is straightforward to check that $\mathcal{S} \cap \mathcal{F}_B$ is a dense subset of $\mathcal{F}_B$. Finally, if $\tilde{F}$ is another reproducing kernel Hilbert space generated by a Gaussian kernel, it is also straightforward to show that the product space $\mathcal{F} \cdot \tilde{\mathcal{F}}$ is separable. Consequently, Lemma 12 can be applied to the function classes defined in Theorem 5.

Web Appendix G. Existence of Bridge Functions for our Data Generating Mechanism

G.1 *Derivation of an Outcome Bridge Function*

Recall that, under our data-generating mechanism, the outcome satisfies

$$\mathbb{E}\left[Y|X,L,U\right] = \left\{1+\exp\left(-[\mu+\beta_9 L+\beta_{10}U+\beta_{11}X+\beta_{12}W+\gamma X^2]\right)\right\}^{-1}.$$

This suggests considering the following parametric model for h:

$$h(X,L,W;\varphi) = A(\varphi)\left\{1+\exp\left(-[\varphi_0+\varphi_1 X+\varphi_2 L+\varphi_3 W+\varphi_4 X^2]\right)\right\}^{-1}.$$

First, consider the expression

$$\varphi_0+\varphi_1 X+\varphi_2 L+\varphi_3 W+\varphi_4 X^2 = \varphi_0+\varphi_3\widetilde{W}+\varphi_1 X+(\varphi_2+\varphi_3\beta_4)L+\varphi_3\beta_5 U+\varphi_4 X^2,$$

where $\widetilde{W} = W-\beta_4 L-\beta_5 U$ is a random variable with distribution $\mathcal{TN}_c(0,1)$.

Now, consider the function $f(v,t) = v\left\{1+\exp(-t)\right\}^{-1}$. Its second-order Taylor approximation at (a,b) is:

$$\begin{aligned}
f(v,t) &\approx f(a,b)+(v-a)\frac{\partial f(a,b)}{\partial v}+(t-b)\frac{\partial f(a,b)}{\partial t}+\frac{(t-b)^2}{2}\frac{\partial^2 f(a,b)}{\partial t^2}+(v-a)(t-b)\frac{\partial^2 f(a,b)}{\partial v\partial t}\\
&= f(a,b)+(v-a)\frac{f(a,b)}{a}+(t-b)f(a,b)\left[1-\frac{f(a,b)}{a}\right]\\
&\quad+\frac{(t-b)^2}{2}f(a,b)\left[1-\frac{f(a,b)}{a}\right]\left[1-\frac{2f(a,b)}{a}\right]+(v-a)(t-b)\frac{f(a,b)}{a}\left[1-\frac{f(a,b)}{a}\right]
\end{aligned}$$

Using this approximation with $a=1$, $v=A(\varphi)$, $B=\varphi_0+\varphi_1 X+(\varphi_2+\varphi_3\beta_4)L+\varphi_3\beta_5 U+\varphi_4 X^2$, and $T=B+\varphi_3\widetilde{W}$, and the fact that $\mathbb{E}[\widetilde{W}]=0$, gives

$$\begin{aligned}
&\mathbb{E}\left[h(X,L,W;\varphi)|X,L,U\right]\\
&\quad= \mathbb{E}\left[f(1,T)|X,L,U\right]\\
&\quad\approx f(1,B)+\left[A(\varphi)-1\right]f(1,B)+\frac{\varphi_3^2}{2}f(1,B)[1-f(1,B)][1-2f(1,B)]\mathbb{E}[\widetilde{W}^2]\\
&\quad\approx \left(A(\varphi)+\frac{\varphi_3^2}{2}\mathbb{E}[\widetilde{W}^2]\right)f(1,B)\\
&\quad= \left(A(\varphi)+\frac{\varphi_3^2}{2}\mathbb{E}[\widetilde{W}^2]\right)\left\{1+\exp\left(-[\varphi_0+\varphi_1 X+(\varphi_2+\varphi_3\beta_4)L+\varphi_3\beta_5 U+\varphi_4 X^2]\right)\right\}^{-1}.
\end{aligned}$$

Using a similar analysis, an approximation of $\mathbb{E}\left[Y|X,L,U\right]$ is given by

$$\left(1+\frac{\beta_{12}^2}{2}\mathbb{E}[\widetilde{W}^2]\right)\left\{1+\exp\left(-[\mu+\beta_{11}X+(\beta_9+\beta_{12}\beta_4)L+(\beta_{12}\beta_5+\beta_{10})U+\gamma_4X^2]\right)\right\}^{-1}$$

Hence, in order to approximately solve the integral equation (2), we must have

$$\begin{aligned}&\left(1+\frac{\beta_{12}^2}{2}\mathbb{E}[\widetilde{W}^2]\right)\left\{1+\exp\left(-[\mu+\beta_{11}X+(\beta_9+\beta_{12}\beta_4)L+(\beta_{12}\beta_5+\beta_{10})U+\gamma_4X^2]\right)\right\}^{-1}\\ &\qquad\approx\left(A(\varphi)+\frac{\varphi_3^2}{2}\mathbb{E}[\widetilde{W}^2]\right)\left\{1+\exp\left(-[\varphi_0+\varphi_1X+(\varphi_2+\varphi_3\beta_4)L+\varphi_3\beta_5U+\varphi_4X^2]\right)\right\}^{-1},\end{aligned}$$

where $\mathbb{E}[\widetilde{W}^2]\approx 0.9733369$ is a constant.

This approximation implies that $(\varphi_1,\varphi_2,\varphi_3,\varphi_4)\approx\left(\beta_{11},\beta_9-\frac{\beta_4\beta_{10}}{\beta_5},\beta_{12}+\frac{\beta_{10}}{\beta_5},\gamma\right)$, and suggests taking $A(\varphi)\approx 1+(\beta_{12}^2-\varphi_3^2)\mathbb{E}[\widetilde{W}^2]/2$, assuming β_{12} is known. This assumption holds when W does not affect Y, i.e., when $\beta_{12}=0$. However, this choice can be problematic because $A(\varphi)$ may become negative when β_5 is small—that is, when W is a weak proxy for U. To address this issue, we instead chose $A(\varphi)=1+\frac{\varphi_3^2-\beta_{12}^2}{2}\mathbb{E}[\widetilde{W}^2]$, which yields stable solutions. Under this selection, when β_{10}/β_5 is small, setting

$$\varphi_0\approx\mu+\log\left(1+\beta_{12}^2\mathbb{E}\left[\widetilde{W}^2\right]\right)-\log\left(1+\left[\beta_{12}+\frac{\beta_{10}}{\beta_5}\right]^2\mathbb{E}\left[\widetilde{W}^2\right]\right)$$

ensures that the proposed function h approximately solves the integral equation (2). In the general case, an approximation of φ can be found by solving a non-linear system of equations. We used such equations for estimating φ in our implementation. Our numerical experiments demonstrates that this selection of $A(\varphi)$ leads to stable solutions.

G.2 *Derivation of a Treatment Bridge Function*

From the data-generating mechanism and policy, when $\beta_8=0$, we have

$$\begin{aligned}\alpha_0(x,l,w)&=I\left\{x\in q\left[\mathcal{S}(l)\right]\right\}\frac{dq^{-1}(x,l)}{dx}\frac{p\left(q^{-1}(x|l,u)\right)}{p(x|l,u)}\\ &=A_0(x)\exp\left\{-\frac{1}{2}\left[q^{-1}(x)-x\right]\left[q^{-1}(x)+x-2\left(\beta_6l+\beta_7u\right)\right]\right\}\\ &=A_0(x)\exp\left\{\frac{\delta V(x)}{2}\left[2x-\delta V(x)-2\left(\beta_6l+\beta_7u\right)\right]\right\}\\ &=A_0(x)\exp\left\{\delta V(x)\left[x-\beta_6l-\beta_7u-\frac{\delta}{2}V(x)\right]\right\}\end{aligned}$$

where $A_0(x) = \left(I\left\{c+\delta \leqslant x \leqslant d\right\} + \frac{\delta}{\varepsilon} I\left\{d-\varepsilon \leqslant x \leqslant d\right\}\right)$ and, recall, $V(x) = I\{c+\delta \leqslant x < c-\varepsilon\} + \frac{d-x}{\varepsilon} \cdot I\{d-\varepsilon \leqslant x \leqslant d\}$. This suggests the following parametric model for g:

$$g(x,l,z;\eta) = A(x)\exp\left\{\left[\eta_0 x + \eta_1 l + \eta_2 z + \eta_3 V(x)\right] V(x)\right\}.$$

Noting that the moment generating function of $Z|X,L,U$ is

$$\mathbb{E}\left[\exp(tZ)|X,L,U\right] = \frac{\Phi(3-t)-\Phi(-3-t)}{\Phi(3)-\Phi(-3)} \exp\left(\left[\beta_2 L + \beta_3 U\right] t + \frac{t^2}{2}\right),$$

where t could be a function of (x,l,u), it follows that

$$\begin{aligned}\mathbb{E}\left[g(X,L,Z;\eta)|X,L,U\right] &= A(X)\exp\left\{\left[\eta_0 X + \eta_1 L + \eta_3 V(X)\right] V(X)\right\} \mathbb{E}\left[\eta_2 V(X) Z|X,L,U\right] \\ &= A_1(X)\exp\left\{\left[\eta_0 X + \left(\eta_1 + \eta_2\beta_2\right) L + \eta_2\beta_3 U + \left(\eta_3 + \frac{\eta_2^2}{2}\right) V(X)\right] V(X)\right\},\end{aligned}$$

where $A_1(x) = \frac{\Phi(3-\eta_2 V(x))-\Phi(-3-\eta_2 V(x))}{\Phi(3)-\Phi(-3)} A(x)$. Hence, taking

$$A(x) = \frac{\Phi(3)-\Phi(-3)}{\Phi\left(3+\frac{\delta\beta_7}{\beta_3}V(x)\right) - \Phi\left(-3+\frac{\delta\beta_7}{\beta_3}V(x)\right)} A_0(x)$$

and

$$\eta = \left(\delta, -\delta\left[\beta_6 - \frac{\beta_2\beta_7}{\beta_3}\right], -\delta\frac{\beta_7}{\beta_3}, -\frac{\delta^2}{2}\left[1+\frac{\beta_7^2}{\beta_3^2}\right]\right)^T,$$

ensures that the proposed function g solves the integral equation (3).

When $\beta_8 \neq 0$, an approximate solution can be obtained using a Taylor expansion of the conditional density $p_{X|L,U}(x|l,u) = \int p_{X|L,U,Z}(x|l,u,z) p_{Z|L,U}(z|l,u) dz$. For simplicity, we omit the details of this approximation. Nevertheless, in the simulation setting described in Section C.6, the parametric model proposed for the treatment bridge function exhibits very good performance.

References

Robins, J. (1986). A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Mathematical modelling* **7**, 1393–1512.

Diaz Munoz, I. and van der Laan, M. (2012). Population intervention causal effects based on stochastic interventions. *Biometrics* **68**, 541–549.

Haneuse, S. and Rotnitzky, A. (2013). Estimation of the effect of interventions that modify the received treatment. *Statistics in medicine* **32**, 5260–5277.

Miao, W., Geng, Z., and Tchetgen Tchetgen, E. J. (2018). Identifying causal effects with proxy variables of an unmeasured confounder. *Biometrika* **105**, 987–993.

Tchetgen Tchetgen, E., Ying, A., Cui, Y., Shi, X., and Miao, W. (2020). An Introduction to Proximal Causal Learning. Preprint 2020, arXiv:2009.10982v1

Cui, Y., Pu, H., Shi, X., Miao, W., and Tchetgen Tchetgen, E. J. (2023). Semiparametric proximal causal inference. *Journal of the American Statistical Association*, 1-12.

Fong, Y., McDermott, A. B., Benkeser, D., Roels, S., Stieh, D. J., Vandebosch, A., et al. (2022). Immune correlates analysis of the ENSEMBLE single Ad26. COV2. S dose vaccine efficacy clinical trial. *Nature microbiology* **7**, 1996–2010.

Sadoff, J., Gray, G., Vandebosch, A., Cárdenas, V., Shukarev, G., Grinsztejn, B., et al. (2022).Final analysis of efficacy and safety of single-dose Ad26.COV2.S. *New England Journal of Medicine* **386**, 847–860.

Rubin, D.B. (1980). Randomization analysis of experimental data: The Fisher randomization test comment. *Journal of the American statistical association* **75**, 591–593.

Kallus, N., Mao, X., and Uehara, M. (2022). Causal inference under unmeasured confounding with negative controls: A minimax learning approach. Preprint 2022, arXiv:2103.14029v4.

Ghassami, A. E., Ying, A., Shpitser, I., and Tchetgen Tchetgen, E.J. (2022). Minimax Kernel Machine Learning for a Class of Doubly Robust Functionals with Application to Proximal Causal Inference. Preprint 2022, arXiv:2104.02929v3.

Díaz, I., Williams, N., Hoffman, K. L., and Schenck, E. J. (2023). Nonparametric causal

effects based on longitudinal modified treatment policies. *Journal of the American Statistical Association* **118**, 846–857.

Dikkala, N., Lewis, G., Mackey, L., and Syrgkanis, V. (2020). Minimax estimation of conditional moment models. *Advances in Neural Information Processing Systems* **33**, 12248-12262.

Wainwright, M.J. (2019). *High-dimensional statistics: A non-asymptotic viewpoint.* Cambridge university press

Bennett, A., Kallus, N., Mao, X., Newey, W., Syrgkanis, V. and Uehara, M. (2023). Inference on Strongly Identified Functionals of Weakly Identified Functions. Preprint 2023, arXiv:2208.08291v3.

Bennett, A., Kallus, N., Mao, X., Newey, W., Syrgkanis, V. and Uehara, M. (2023). Source condition double robust inference on functionals of inverse problems. Preprint 2023, arXiv:2307.13793.

Hohage, T (2002). Lecture notes on inverse problems. Lectures given at the University of Göttingen.

Ying, A., Miao, W., Shi, X., and Tchetgen Tchetgen, E. J. (2023). Proximal causal inference for complex longitudinal studies. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **85**, 684–704

Shi, X., Miao, W., Nelson, J. C., and Tchetgen Tchetgen, E. J. (2023). Multiply robust causal inference with double-negative control adjustment for categorical unmeasured confounding. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **82**, 521–540.

Chernozhukov, V., Newey, W., Singh, R., and Syrgkanis, V. (2020). Adversarial Estimation of Riesz Representers. Preprint, arXiv:2101.00009v1

Carrasco, M., Florens, J.P., Renault, E. (2007). Linear inverse problems in structural

econometrics estimation based on spectral decomposition and regularization. *Handbook of econometrics* **6**, 5633–5751.

Cavalier, L. (2011). Inverse problems in statistics. In *Inverse problems and high-dimensional estimation. Stats in the Château Summer School, August 31-September 4, 2009,*. p. 3-96. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011.

Kress, R. (2010). *Linear Integral Equations.* 3rd edition. New York: Springer.

Van der Vaart, A. W. (2000). *Asymptotic statistics.* Cambridge: Cambridge university press.

Paulsen, V. I., and Raghupathi, M. (2016). *An introduction to the theory of reproducing kernel Hilbert spaces* Cambridge: Cambridge university press.

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

Schölkopf, B., Herbrich, R., Smola, A. J. (2001). A generalized representer theorem. In *International conference on computational learning theory.* Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 416–426.

Boyd, S. (2004). *Convex optimization.* Cambridge UP.

Steinwart, I. and Christmann, A. (2008). *Support Vector Machines.* New York: Springer.

Figure 1. Boxplots of the parametric estimators for numerical experiments with data generated under varying conditional correlations of the proxies Z and W with the unmeasured confounder U and sample sizes $n = 750, 1500, 3000, 15000, 50000, 100000$. The dashed line indicates the true parameter value.

Figure 2. (Panel A) Boxplots of the estimators and (Panel B) coverage probabilities for numerical experiments with data generated under no unmeasured confounding and sample sizes $n = 750, 1500, 3000$. In panel A, the dashed line indicates the true parameter value, while in panel B, it represents the nominal coverage level of 0.95. Non-proximal estimators are based on the assumption of no unmeasured confounding.