Learning Nonlinear Causal Effect via Kernel Anchor Regression (Supplementary Material)

A PROOFS AND DERIVATIONS

A.1 PROOF OF THEOREM 1

Before proving Theorem 1, we introduce the exact bounds of the approximation errors for estimating E_X^p and E_Y^p in the disjoint sample sets projection stage. Lemma A1 and A2 below are adapted from Theorem 2 in Singh et al. [2019].

Lemma A1 *Under Condition 1, $\forall \delta \in (0,1)$, the following holds w.p. $1-\delta$:*

$$\|E^{n_1}_{\alpha_1,X} - E^p_X\|_{\mathcal{H}_\Gamma} \leq r_{E_1}(\delta, n_1, c_1) := \frac{\sqrt{\zeta_1}(c_1+1)}{4^{\frac{1}{c_1+1}}} \left(\frac{4\kappa(Q_1 + \kappa\|E^p_X\|_{\mathcal{H}_\Gamma}\ln(2/\delta)}{\sqrt{n_1\zeta_1}(c_1-1)}\right)^{\frac{c_1-1}{c_1+1}},$$
$$\alpha_1 = \left(\frac{8\kappa(Q_1 + \kappa\|E^p_X\|_{\mathcal{H}_\Gamma}\ln(2/\delta)}{\sqrt{n_1\zeta_1}(c_1-1)}\right)^{\frac{2}{c_1+1}}.$$

Lemma A2 *Under Condition 1 and Condition 2, $\forall \epsilon \in (0,1)$, the following holds w.p. $1-\epsilon$:*

$$\|E^{n_2}_{\alpha_2,Y} - E^p_Y\|_{\mathcal{H}_\Theta} \leq r_{E_2}(\epsilon, n_2, c_2) := \frac{\sqrt{\zeta_2}(c_2+1)}{4^{\frac{1}{c_2+1}}} \left(\frac{4\kappa(Q_2 + \kappa\|E^p_Y\|_{\mathcal{H}_\Theta}\ln(2/\epsilon)}{\sqrt{n_2\zeta_2}(c_2-1)}\right)^{\frac{c_2-1}{c_2+1}},$$
$$\alpha_2 = \left(\frac{8\kappa(Q_2 + \kappa\|E^p_Y\|_{\mathcal{H}_\Theta}\ln(2/\epsilon)}{\sqrt{n_2\zeta_2}(c_2-1)}\right)^{\frac{2}{c_2+1}}.$$

Recall that we define the population-level risk for the regression stage $\mathcal{E}^\gamma(H)$, population-level risk with regularization $\mathcal{E}^\gamma_\xi(H)$, and the empirical risk $\widehat{\mathcal{E}}^{\gamma,m}_\xi(H)$ with E^p_X and E^p_Y being replaced by $E^{n_1}_{\alpha_1,X}$ and $E^{n_2}_{\alpha_2,Y}$, respectively. Denote the optimal operator to $\mathcal{E}^\gamma_\xi(H)$ as $H^\gamma_\xi = \arg\min_H \mathcal{E}^\gamma_\xi(H)$. We now define the empirical risk $\mathcal{E}^\xi_{\gamma,m}(H)$ with true E^p_X and E^p_Y, and the corresponding optimal operator.

$$\mathcal{E}^{\gamma,m}_\xi(H) = \frac{1}{m}\sum_{l=1}^m \|y_{\gamma,l} - H\psi_{\gamma,l}\|^2_{\mathcal{Y}} + \xi\|H\|^2_{\mathcal{H}_\Omega}, \quad H^{\gamma,m}_\xi = \arg\min_H \mathcal{E}^{\gamma,m}_\xi(H),$$

where the true transformed inputs and outputs are given by

$$\psi_{\gamma,l} = \psi(x_l) + (\sqrt{\gamma}-1)E^p_X\phi(z_l) \in \mathcal{H}_\mathcal{X}, \quad y_{\gamma,l} = y_l + (\sqrt{\gamma}-1)E^p_Y\phi(z_l) \in \mathcal{Y}.$$

The closed form solution of $H^{\gamma,m}_\xi$ is given by Lemma A3 below, and it's adapted from Theorem 3 in Singh et al. [2019]

Lemma A3 *$\forall \xi > 0$, the solution $H^{\gamma,m}_\xi$ to $\mathcal{E}^{\gamma,m}_\xi$ exists, is unique, and*

$$\mathbf{T} = \tfrac{1}{m}\textstyle\sum_{l=1}^m T_{\psi_{\gamma,l}}, \quad \mathbf{g} = \tfrac{1}{m}\textstyle\sum_{l=1}^m \Omega_{\psi_{\gamma,l}} y_{\gamma,l}, \quad H^{\gamma,m}_\xi = (\mathbf{T}+\xi)^{-1}\circ\mathbf{g}.$$

We then define the following terms.

Definition 1 *Fix $\eta \in (0,1)$ and define the following constants*

$$C_\eta = 96\ln^2(6/\eta), \quad M = 2(C + \|H^\gamma\|_{\mathcal{H}_\Omega}\sqrt{B}), \quad \Sigma = \frac{M}{2}.$$

For the excess error of **KAR** estimator $\widehat{H}_\xi^{\gamma,m}$, we can bound it by five terms according to Proposition 32 in Singh et al. [2019].

Lemma A4 *The excess error can be bounded as follows*

$$\mathcal{E}^\gamma(\widehat{H}_\xi^{\gamma,m}) - \mathcal{E}^\gamma(H^\gamma) \le 5[S_{-1} + S_0 + \mathcal{A}(\xi) + S_1 + S_2],$$

where

$$\begin{aligned}
S_{-1} &= \|\sqrt{T}\circ(\widehat{\mathbf{T}}+\xi)^{-1}(\widehat{\mathbf{g}}-\mathbf{g})\|_{\mathcal{H}_\Omega}^2,\\
S_0 &= \|\sqrt{T}\circ(\widehat{\mathbf{T}}+\xi)^{-1}(\mathbf{T}-\widehat{\mathbf{T}})H_\xi^{\gamma,m}\|_{\mathcal{H}_\Omega}^2,\\
S_1 &= \|\sqrt{T}\circ(\widehat{\mathbf{T}}+\xi)^{-1}(\mathbf{g}-\mathbf{T}H^\gamma)\|_{\mathcal{H}_\Omega}^2,\\
S_2 &= \|\sqrt{T}\circ(\widehat{\mathbf{T}}+\xi)^{-1}(T-\mathbf{T})(H_\xi^\gamma - H^\gamma)\|_{\mathcal{H}_\Omega}^2,\\
\mathcal{A}(\xi) &= \|\sqrt{T}(H_\xi^\gamma - H^\gamma)\|_{\mathcal{H}_\Omega}^2.
\end{aligned}$$

For all five terms above, only $\widehat{\mathbf{g}}-\mathbf{g}$ in S_{-1} depends on the approximation error of E_Y^p. The bounds for other four terms are same to the **KIV** case. Below we introduce without proof the bond of S_0, S_1, S_2 and $\mathcal{A}(\xi)$ according to Theorem 7 in Singh et al. [2019].

Lemma A5 *Under Condition 1–3, if m is large enough and $\xi \le \|T\|_{L(\mathcal{H}_\Omega)}$ then $\forall\delta,\eta\in(0,1)$, the following holds up w.p. $1-\eta-\delta$:*

$$\begin{aligned}
S_0 &\le \frac{4}{\xi}4BL^2r_x^{2\iota}\|H_\xi^{\gamma,m}\|_{\mathcal{H}_\Omega}^2,\\
S_1 &\le 32\ln^2(6\eta)\left[\frac{BM^2}{m^2\xi} + \frac{\Sigma^2}{m}\beta^{1/b_\gamma}\frac{\pi/b_\gamma}{\sin(\pi/b_\gamma)\xi^{-1/b_\gamma}}\right],\\
S_2 &\le 8\ln^2(6/\eta)\left[\frac{4B^2\zeta\xi^{c_\gamma-1}}{m^2\xi} + \frac{B\zeta\xi^{c_\gamma}}{m\xi}\right],\\
\mathcal{A}(\xi) &\le \zeta\xi^{c_\gamma}.
\end{aligned}$$

To extend the convergence rate of **KIV** estimator to **KAR** estimator. We then illustrate the bound for S_{-1}. To begin with, the bound of term $\sqrt{T}\circ(\widehat{\mathbf{T}}+\xi)^{-1}$ in S_{-1} is given by Proposition 39 in Singh et al. [2019].

Lemma A6 *If $\|\widehat{\psi}_\gamma - \psi_\gamma\|_{\mathcal{H}_\mathcal{X}} \le r_x$ w.p. $1-\delta$, $\xi \le \|T\|_{L(\mathcal{H}_\Omega)}$, m is sufficiently large and Condition 3 holds, then w.p. $1-\eta/3-\delta$*

$$\|\sqrt{T}\circ(\widehat{\mathbf{T}}+\xi)^{-1}\|_{L(\mathcal{H}_\Omega)} \le \frac{2}{\sqrt{\xi}}.$$

With the the error propagated from the estimators in the projection stage, we can bound $\widehat{\psi}_\gamma - \psi_\gamma$ and $\widehat{y}_\gamma - y_\gamma$ as shown in Lemma A7–A8.

Lemma A7 *Under Condition 1, $\forall\delta\in(0,1)$, the following statement holds w.p. $1-\delta$: $\forall z\in\mathcal{Z}, x\in\mathcal{X}$,*

$$\|\widehat{\psi}_\gamma - \psi_\gamma\|_{\mathcal{H}_\mathcal{X}} \le r_x(\gamma,\delta,n_1,c_1) := |\sqrt{\gamma}-1|\kappa r_{E_1}(\delta,n_1,c_1).$$

Proof 1 *By definition, we have*

$$\begin{aligned}
\|\widehat{\psi}_\gamma - \psi_\gamma\|_{\mathcal{H}_\mathcal{X}} &= \left\|(\sqrt{\gamma}-1)\left(E_{\alpha_1,X}^{n_1} - E_X^p\right)\phi(z)\right\|_{\mathcal{H}_\mathcal{X}}\\
&\le |\sqrt{\gamma}-1|\|E_{\alpha_1,X}^{n_1} - E_X^p\|_{\mathcal{H}_\Gamma}\|\phi(z)\|_{\mathcal{H}_\mathcal{Z}}.
\end{aligned}$$

This, together with Lemma A1 and Condition 1, ensures that w.p. $1-\delta$

$$\|\widehat{\psi}_\gamma - \psi_\gamma\|_{\mathcal{H}_\mathcal{X}} \leq r_x(\gamma,\delta,n_1,c_1) := |\sqrt{\gamma}-1|\kappa r_{E_1}(\delta,n_1,c_1).$$

Remark A1 *Corollary 1 in Singh et al. [2019] is a special case of Lemma A7 with* $\gamma = 0$.

Lemma A8 *Under Condition 1– 2,* $\forall \epsilon \in (0,1)$*, the following statement holds w.p.* $1-\epsilon$*:* $\forall z \in \mathcal{Z}, y \in \mathcal{Y}$*,*

$$\|\widehat{y}_\gamma - y_\gamma\|_{\mathcal{H}_\mathcal{Y}} \leq r_y(\gamma,\epsilon,n_2,c_2) := |\sqrt{\gamma}-1|\kappa r_{E_2}(\epsilon,n_2,c_2).$$

Proof 2 *Lemma A8 is analogous to Lemma A7 by replacing* ψ_γ *with* y_γ*. The proof is thus omitted.*

Combining Lemma A6- A8, we can derive the bound of $\widehat{\mathbf{g}} - \mathbf{g}$ and then the bound of S_{-1}.

Lemma A9 *If* $\|\widehat{\psi}_\gamma - \psi_\gamma\|_{\mathcal{H}_\mathcal{X}} \leq r_x$ *w.p.* $1-\delta$ *and* $\|\widehat{y}_\gamma - y_\gamma\|_\mathcal{Y} \leq r_y$ *w.p.* $1-\epsilon$*, then w.p.* $1-\delta-\epsilon$

$$\|\widehat{\mathbf{g}} - \mathbf{g}\|^2_{\mathcal{H}_\Omega} \leq 3(L^2 r_x^{2\iota} r_y^2 + B^2 r_y^2 + L^2 r_x^{2\iota} C^2).$$

Proof 3 *By definition, we have*

$$\begin{aligned}
\widehat{\mathbf{g}} - \mathbf{g} &= \frac{1}{m}\sum_{l=1}^m \left(\Omega_{\widehat{\psi}_{\gamma,l}}\widehat{y}_{\gamma,l} - \Omega_{\psi_{\gamma,l}(x)} y_{\gamma,l}\right) \\
&= \frac{1}{m}\sum_{l=1}^m \left\{\Omega_{\widehat{\psi}_{\gamma,l}} - \Omega_{\psi_{\gamma,l}}\right\}\{\widehat{y}_{\gamma,l} - y_{\gamma,l}\} + \Omega_{\widehat{\psi}_{\gamma,l}}\{\widehat{y}_{\gamma,l} - y_{\gamma,l}\} + \left\{\Omega_{\widehat{\psi}_{\gamma,l}} - \Omega_{\psi_{\gamma,l}}\right\} y_{\gamma,l}.
\end{aligned}$$

We then have

$$\begin{aligned}
\|\widehat{\mathbf{g}} - \mathbf{g}\|^2_{\mathcal{H}_\Omega} &\leq \frac{3m}{m^2}\sum_{l=1}^m \| \left\{\Omega_{\widehat{\psi}_{\gamma,l}} - \Omega_{\psi_{\gamma,l}}\right\}\{\widehat{y}_{\gamma,l} - y_{\gamma,l}\} \|^2_{\mathcal{H}_\Omega} + \|\Omega_{\widehat{\psi}_{\gamma,l}}\{\widehat{y}_{\gamma,l} - y_{\gamma,l}\}\|^2_{\mathcal{H}_\Omega} \\
&\quad + \| \left\{\Omega_{\widehat{\psi}_{\gamma,l}} - \Omega_{\psi_{\gamma,l}}\right\} y_{\gamma,l}\|^2_{\mathcal{H}_\Omega} \\
&\leq \frac{3}{m}\sum_{l=1}^m \|\Omega_{\widehat{\psi}_{\gamma,l}} - \Omega_{\psi_{\gamma,l}}\|^2_{L(\mathcal{Y},\mathcal{H}_\Omega)}\|\widehat{y}_{\gamma,l} - y_{\gamma,l}\|^2_\mathcal{Y} + \|\Omega_{\psi_{\gamma,l}}\|^2_{L(\mathcal{Y},\mathcal{H}_\Omega)}\|\widehat{y}_{\gamma,l} - y_{\gamma,l}\|^2_\mathcal{Y} \\
&\quad + \|\Omega_{\widehat{\psi}_{\gamma,l}} - \Omega_{\psi_{\gamma,l}}\|^2_{L(\mathcal{Y},\mathcal{H}_\Omega)}\|y_{\gamma,l}\|^2_\mathcal{Y}.
\end{aligned}$$

By the boundedness and the Hölder property in Condition 3, we obtain that w.p. $1-\delta-\epsilon$*,*

$$\begin{aligned}
\|\widehat{\mathbf{g}} - \mathbf{g}\|^2_{\mathcal{H}_\Omega} &\leq \frac{3}{m}\sum_{l=1}^m L^2\|\widehat{\psi}_{\gamma,l} - \psi_{\gamma,l}\|^{2\iota}_{\mathcal{H}_\mathcal{X}}\|\widehat{y}_{\gamma,l} - y_{\gamma,l}\|^2_\mathcal{Y} + \|\Omega_{\psi_{\gamma,l}}\|^2_{L(\mathcal{Y},\mathcal{H}_\Omega)}\|\widehat{y}_{\gamma,l} - y_{\gamma,l}\|^2_\mathcal{Y} \\
&\quad + L^2\|\widehat{\psi}_{\gamma,l} - \psi_{\gamma,l}\|^{2\iota}_{\mathcal{H}_\mathcal{X}}\|y_{\gamma,l}\|^2_\mathcal{Y} \\
&\leq 3(L^2 r_x^{2\iota} r_y^2 + B^2 r_y^2 + L^2 r_x^{2\iota} C^2).
\end{aligned}$$

Lemma A10 *Under Condition 1–3, then w.p.* $1-\delta-\epsilon$

$$S_{-1} \leq \frac{4}{\xi} 3(L^2 r_x^{2\iota} r_y^2 + B^2 r_y^2 + L^2 r_x^{2\iota} C^2).$$

Proof 4 *We can derive from the definition of* S_{-1} *that*

$$S_{-1} \leq \|\sqrt{T} \circ (\widehat{\mathbf{T}} + \xi)^{-1}\|^2_{L(\mathcal{H}_\Omega)}\|\widehat{\mathbf{g}} - \mathbf{g}\|^2_{\mathcal{H}_\Omega}.$$

This, together with Lemma A6 and Lemma A9, ensures

$$S_{-1} \leq \frac{4}{\xi} 3(L^2 r_x^{2\iota} r_y^2 + B^2 r_y^2 + L^2 r_x^{2\iota} C^2).$$

We then show the order of the sum $S_0 + S_1 + S_2 + \mathcal{A}(\xi)$, which is adapted from Theorem 4 in Singh et al. [2019].

Lemma A11 *Under Condition 1– 3, choose* $\alpha_1 = n_1^{-\frac{1}{c_1+1}}$, $n_1 = m^{\frac{d_1(c_1+1)}{\iota(c_1-1)}}$, *where* $d_1 > 0$. *Let*

$$f(m) = \frac{1}{m^{2+d_1}\xi^3} + \frac{1}{m^{1+d_1}\xi^{2+1/b_\gamma}} + \frac{1}{m^{d_1}\xi} + \xi^{c_\gamma} + \frac{1}{m^2\xi} + \frac{1}{m\xi^{1/b_\gamma}},$$

we then have

$$O_p(S_0 + \mathcal{A}(\xi) + S_1 + S_2) = O(f(m)).$$

(i) If $d_1 \leq \frac{b_\gamma(c_\gamma+1)}{b_\gamma c_\gamma+1}$ *then* $O(f(m)) = O(m^{-\frac{d_1 c_\gamma}{c_\gamma+1}})$ *with* $\xi = m^{-\frac{d_1}{c_\gamma+1}}$;

(ii) If $d_1 > \frac{b_\gamma(c_\gamma+1)}{b_\gamma c_\gamma+1}$ *then* $O(f(m)) = O(m^{-\frac{b_\gamma c_\gamma}{b_\gamma c_\gamma+1}})$ *with* $\xi = m^{-\frac{b_\gamma}{b_\gamma c_\gamma+1}}$.

Proof 5 (Proof of Theorem 1) *The choices of* α_1, α_2 *and* n_1, n_2 *in the statement of Theorem 1 ensure that*

$$r_x^2 = O([(n_1^{-\frac{1}{2}})^{\frac{2}{c_1+1}}]^{2\iota}) = O(m^{-d_1}), \quad r_y^2 = O([(n_2^{-\frac{1}{2}})^{\frac{2}{c_2+1}}]^2) = O(m^{-d_2}).$$

Thus, by Lemma A10, we have $O_p(S_{-1}) = O_p(1/\xi(r_x^{2\iota} r_y^2 + r_y^2 + r_x^{2\iota})) = O_p(1/\xi\left\{m^{-d_1} + m^{-d_2} + m^{-d_1-d_2}\right\})$. *Since* $d_1, d_2 > 0$, *and* $d_1 \leq d_2$ *by Condition 4,* m^{-d_1}/ξ *then dominates two other terms in* S_{-1}.

Note that $f(m)$ *in Lemma A11 also includes* m^{-d_1}/ξ. *Therefore, given Condition 4, the sum of four terms* $S_0 + \mathcal{A}(\xi) + S_1 + S_2$ *dominates* S_{-1}, *which suggests that the approximation error of* E_Y^p *is dominated by that of* E_X^p. *We can then derive the result from Lemma A11.*

A.2 PROOF OF THEOREM 2

Proof 6 (Proof of Theorem 2) *Under the kernel structural equation model, simple calculation gives*

$$C = B_{CZ}\Phi(Z) + \epsilon_C, \tag{1}$$

$$\Psi(X) = (B_{XZ} + B_{XC}B_{CZ})\Phi(Z) + B_{XC}\epsilon_C + \epsilon_X, \tag{2}$$

$$Y = [B_{YZ} + B_{YC}B_{CZ} + B_{YX}(B_{XZ} + B_{XC}B_{CZ})]\Phi(Z) + (B_{YC} + B_{YX}B_{XC})\epsilon_C + B_{YX}\epsilon_X + \epsilon_Y. \tag{3}$$

We denote $B_{\square\triangle}$ *as the adjoint operator of* $B_{\triangle\square}$, $B_{\square\triangle} = B^*_{\triangle\square}$. *When no ambiguity arise, we use the transpose matrix notation* $B_{\square\triangle} = B^\top_{\triangle\square}$. *For instance,* $B_{XZ} = B^\top_{ZX}$, $B_{YC} = B^\top_{CY}$. *Recall that the transformed input and output in Equation* (16) *and Equation* (17) *has the form*

$$\psi_\gamma(X) = \psi(X) - E_X^p\phi(Z) + \sqrt{\gamma}E_X^p\phi(Z),$$

and

$$Y_\gamma = Y - E_Y^p\phi(Z) + \sqrt{\gamma}E_Y^p\phi(Z).$$

In the SEM case, the projections E_X^p *and* E_Y^p *into* $\phi(Z)$ *are noted by the (composition of) operators in Equation* (2) *and Equation* (3), *where*

$$E_X^p = (B_{XZ} + B_{XC}B_{CZ}),$$

and

$$E_Y^p = [B_{YZ} + B_{YC}B_{CZ} + B_{YX}(B_{XZ} + B_{XC}B_{CZ})].$$

As such, the transformed input and output has the form

$$\psi_\gamma(x) = B_{XC}\epsilon_C + \epsilon_X + \sqrt{\gamma}(B_{XZ} + B_{XC}B_{CZ})\phi(Z), \tag{4}$$

and

$$y_\gamma = (B_{YC} + B_{YX}B_{XC})\epsilon_C + B_{YX}\epsilon_X + \epsilon_Y + \gamma[B_{YZ} + B_{YC}B_{CZ} + B_{YX}(B_{XZ} + B_{XC}B_{CZ})]\phi(Z). \tag{5}$$

Define relevant covariance matrix/operators as $\Sigma_C = \mathbb{E}[\epsilon_C \epsilon_C^\top]$, $\Sigma_X = \mathbb{E}[\epsilon_X \otimes \epsilon_X]$ *and* $\Sigma_Z = \mathbb{E}[\phi(Z) \otimes \phi(Z)]$, *where* $\otimes$ *denotes the tensor outer product. Then the solution for the least square objective on the transformed input output can be written as*

$$H^\gamma = \mathbb{E}[Y_\gamma \psi_\gamma(X)](\mathbb{E}[\psi_\gamma(X) \otimes \psi_\gamma(X)])^{-1}.$$

Plug in the transformed terms in the form of Equation (4) *and Equation* (5)*, we have*

$$\begin{aligned}
&\mathbb{E}[\psi_\gamma(X) \otimes \psi_\gamma(X)] \\
&= \mathbb{E}[(B_{XC}\epsilon_C + \epsilon_X + \sqrt{\gamma}(B_{XZ} + B_{XC}B_{CZ})\phi(Z))(B_{XC}\epsilon_C + \epsilon_X + \sqrt{\gamma}(B_{XZ} + B_{XC}B_{CZ})\phi(Z))^\top] \\
&= B_{XC}\mathbb{E}[\epsilon_C\epsilon_C^\top]B_{CX} + \mathbb{E}[\epsilon_X \otimes \epsilon_X] + \gamma(B_{XZ} + B_{XC}B_{CZ})\mathbb{E}[\phi(Z) \otimes \phi(Z)](B_{ZX} + B_{ZC}B_{CX}) \\
&= B_{XC}\Sigma_C B_{CX} + \Sigma_X + \gamma(B_{XZ} + B_{XC}B_{CZ})\Sigma_Z(B_{ZX} + B_{ZC}B_{CX}).
\end{aligned}$$

Moreover, $\mathbb{E}[Y_\gamma \psi_\gamma(X)]$ *has the form*

$$\begin{aligned}
&(B_{YC} + B_{YX}B_{XC})\mathbb{E}[\epsilon_C\epsilon_C^\top]B_{CX} + B_{YX}\mathbb{E}[\epsilon_X \otimes \epsilon_X] + \\
&\gamma[B_{YZ} + B_{YC}B_{CZ} + B_{YX}(B_{XZ} + B_{XC}B_{CZ})]\mathbb{E}[\phi(Z) \otimes \phi(Z)](B_{ZX} + B_{ZC}B_{CX}) \\
=&(B_{YC} + B_{YX}B_{XC})\Sigma_C B_{CX} + B_{YX}\Sigma_X + \\
&\gamma[B_{YZ} + B_{YC}B_{CZ} + B_{YX}(B_{XZ} + B_{XC}B_{CZ})]\Sigma_Z(B_{ZX} + B_{ZC}B_{CX})
\end{aligned}$$

as ϵ_C, ϵ_X *and* ϵ_Y *are independent variables, which are also independent of* Z*. In overall, we have*

$$\begin{aligned}
H^\gamma =&[(B_{YC} + B_{YX}B_{XC})\Sigma_C B_{CX} + B_{YX}\Sigma_X \\
&+ \gamma[B_{YZ} + B_{YC}B_{CZ} + B_{YX}(B_{XZ} + B_{XC}B_{CZ})]\Sigma_Z(B_{ZX} + B_{ZC}B_{CX})] \\
&[B_{XC}\Sigma_C B_{CX} + \Sigma_X + \gamma(B_{XZ} + B_{XC}B_{CZ})\Sigma_Z(B_{ZX} + B_{ZC}B_{CX})]^{-1}
\end{aligned}$$

The bias of the target KAR estimator is then given by

$$\begin{aligned}
&H^\gamma - B_{YX} = \\
&\Big[(B_{YC} + B_{YX}B_{XC})\Sigma_C B_{CX} + B_{YX}\Sigma_X + \gamma[B_{YZ} + B_{YC}B_{CZ} + B_{YX}(B_{XZ} + B_{XC}B_{CZ})]\Sigma_Z(B_{ZX} + B_{ZC}B_{CX})\Big] \\
&\Big[B_{XC}\Sigma_C B_{CX} + \Sigma_X + \gamma(B_{XZ} + B_{XC}B_{CZ})\Sigma_Z(B_{ZX} + B_{ZC}B_{CX})\Big]^{-1} - B_{YX} \\
&= \Big[(B_{YC} + B_{YX}B_{XC})\Sigma_C B_{CX} + B_{YX}\Sigma_X + \gamma[B_{YZ} + B_{YC}B_{CZ} + B_{YX}(B_{XZ} + B_{XC}B_{CZ})]\Sigma_Z(B_{ZX} + B_{ZC}B_{CX}) \\
&\quad - B_{YX}(B_{XC}\Sigma_C B_{CX} + \Sigma_X + \gamma(B_{XZ} + B_{XC}B_{CZ})\Sigma_Z(B_{ZX} + B_{ZC}B_{CX}))\Big] \\
&\Big[B_{XC}\Sigma_C B_{CX} + \Sigma_X + \gamma(B_{XZ} + B_{XC}B_{CZ})\Sigma_Z(B_{ZX} + B_{ZC}B_{CX})\Big]^{-1}
\end{aligned}$$

Collecting all the common terms we get

$$\begin{aligned}
H^\gamma - B_{YX} =&\Big[\underbrace{B_{YC}\Sigma_C B_{CX}}_{\Sigma^\perp_{YX}} + \gamma\underbrace{(B_{YZ} + B_{YC}B_{CZ})\Sigma_Z(B_{ZX} + B_{ZC}B_{CX})}_{\Sigma^{||}_{YX}}\Big] \\
&\Big[B_{XC}\Sigma_C B_{CX} + \Sigma_X + \gamma(B_{XZ} + B_{XC}B_{CZ})\Sigma_Z(B_{ZX} + B_{ZC}B_{CX})\Big]^{-1}
\end{aligned}$$

Thus, $\forall x \in \mathcal{X}, y \in \mathcal{Y}$*, consider the inner product* $y^\top(H^\gamma - B_{YX})\psi(x) = 0$ *when the following holds: (i)* $B_{YC} = 0$ *and* $\gamma = 0$*, or (ii)* $B_{YZ} + B_{YC}B_{CZ} = 0$ *and* $\gamma = \infty$*, or (iii)* $B_{YC} = 0$*,* $B_{YZ} + B_{YC}B_{CZ} = 0$ *and* $\gamma \geq 0$*, or (iv)* $\Sigma^{||}_{YX} = a\Sigma^\perp_{YX}$ *for some* $a > 0$*, and* $\gamma = \infty$*, or (v)* $\Sigma^{||}_{XY} = -a\Sigma^\perp_{XY}$ *for some* $a > 0$*, and* $\gamma = 1/c$*. As such, we conclude* $H^\gamma = B_{XY}$.

A.3 CONVERGENCE RATE FOR KAR.2 ESTIMATOR

In this section, we will further discuss the convergence rate of **KAR.2** estimator, and show that the rate does not improve upon the convergence rate of **KAR** estimator.

In the three-stage KAR procedure, we approximate E_X^p and E_Y^p by $E_{\alpha_1,X}^{n_1}$ and $E_{\alpha_2,Y}^{n_2}$, respectively. In the two-stage KAR procedure, instead, we approximate the two operators by $E_{\alpha,X}^n$ and $E_{\alpha,Y}^n$, respectively. Note that the estimated operators $E_{\alpha,X}^n$ and $E_{\alpha,Y}^n$ use the same α. The shared α may fail to ensure the optimal approximation error for $E_{\alpha,X}^n$ and $E_{\alpha,Y}^n$ at the same time.

Lemma A12 *Under Condition 1, $\forall\delta \in (0,1)$, the following holds w.p. $1-\delta$:*

$$\|E_{\alpha,X}^n - E_X^p\|_{\mathcal{H}_\Gamma} \leq r_1(\alpha) := \frac{4\kappa(Q_1 + \kappa\|E_X^p\|_{\mathcal{H}_\Gamma})\ln(2/\delta)}{\sqrt{n}\alpha} + \alpha^{\frac{c_1-1}{2}}\sqrt{\zeta_1}.$$

Under Condition 1 and Condition 2, $\forall\epsilon \in (0,1)$, the following holds w.p. $1-\epsilon$:

$$\|E_{\alpha,Y}^n - E_Y^p\|_{\mathcal{H}_\Theta} \leq r_2(\alpha) := \frac{4\kappa(Q_2 + \kappa\|E_Y^p\|_{\mathcal{H}_\Theta})\ln(2/\epsilon)}{\sqrt{n}\alpha} + \alpha^{\frac{c_2-1}{2}}\sqrt{\zeta_2}.$$

Approximation error bound $r_1(\alpha)$ for $E_{\alpha,X}^n$ achieves its minimum at rate $O(n^{-\frac{c_1-1}{2(c_1+1)}})$ when

$$\alpha = \left(\frac{8\kappa(Q_1 + \kappa\|E_X^p\|_{\mathcal{H}_\Gamma})\ln(2/\delta)}{\sqrt{n\zeta_1}(c_1-1)}\right)^{\frac{2}{c_1+1}} = O(n^{\frac{-1}{c_1+1}});$$

and approximation error bound $r_2(\alpha)$ for $E_{\alpha,Y}^n$ achieves its minimum at rate $O(n^{-\frac{c_2-1}{2(c_2+1)}})$ when

$$\alpha = \left(\frac{8\kappa(Q_2 + \kappa\|E_Y^p\|_{\mathcal{H}_\Theta})\ln(2/\epsilon)}{\sqrt{n\zeta_2}(c_2-1)}\right)^{\frac{2}{c_2+1}} = O(n^{\frac{-1}{c_2+1}}).$$

Lemma A12 above provides the upper bounds of the approximation errors for $E_{\alpha,X}^n$ and $E_{\alpha,Y}^n$, and it's adapted from Theorem 2 in Singh et al. [2019]. We can see that if $c_1 \neq c_2$, we cannot claim the optimal convergence rate for $E_{\alpha,X}^n$ and $E_{\alpha,Y}^n$ at the same time, which disjoint sample sets projection estimators can guarantee by setting different α_1 and α_2 as shown in Lemma 1 and 2. In other words, in **KAR.2** procedure, the error propagated to the final stage, which are caused by using $E_{\alpha,X}^n$ and $E_{\alpha,Y}^n$, can have larger order than using $E_{\alpha_1,X}^{n_1}$ and $E_{\alpha_2,Y}^{n_2}$ separately in the **KAR** procedure. Therefore, we cannot ensure a same or improved convergence rate for **KAR.2** estimator compared to **KAR** estimator.

B ADDITIONAL SIMULATION DETAILS AND RESULTS

B.1 SYNTHETIC EXAMPLE IN KIV SETTING

In this section, we show the data generating process and implementation details for the example that follows the simulation case of learning counterfactual functions **?** studied in Singh et al. [2019]. The structural model is set as follows,

$$Y = C + \ln(|16X - 8| + 1)sgn(X - 0.5).$$

The explanatory variables are generated from

$$\begin{aligned}
\begin{pmatrix} C \\ V \\ W \end{pmatrix} &\sim N\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1, 0.5, 0 \\ 0.5, 1, 0 \\ 0, 0, 1 \end{pmatrix}\right), \\
X &= F\left(\frac{W+V}{\sqrt{2}}\right), \\
Z &= F(W),
\end{aligned}$$

where F denote the c.d.f of standard normal distribution. This structural model ensures that anchor Z is a valid instrumental variable, so that KIV is supposed to perform well in this case. We conduct kernel anchor regression with three-stage algorithm (KAR), kernel anchor regression with two-stage algorithm (KAR.2) and multiple γs and kernel instrument variable regression (KIV). Set $n_1 = 200$, $n_2 = 200$, $m = 600$, $n = n_1 + n_2 = 400$. For KAR and KAR.2, we set γ to be 0, 0.5, 1, 2, 5, 10, and 100. We set $\alpha_1 = c_\alpha n_1^{-0.5}$, $\alpha_2 = c_\alpha n_2^{-0.5}$, $\alpha = c_\alpha n^{-0.5}$, and $\xi = 1m^{-0.5}$, where $c_\alpha > 0$ is a constant

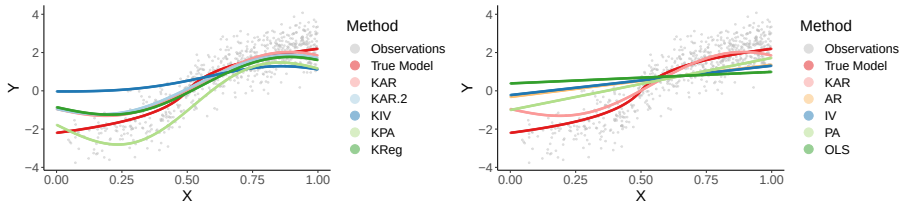
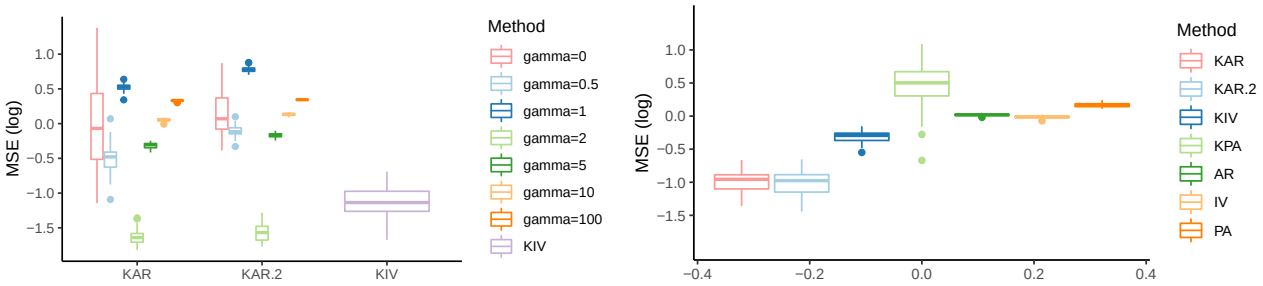


Figure B1: Variant synthetic example: fitted nonlinear (left) and linear (right) methods.



(a) MSE results in the KIV setting

(b) MSE results of all estimators in the variant case.

Figure B2: Experimental results for additional experiments.

chosen from $\{0.01, 0.05, 0.1, 0.5, 0.8, 1, 2, 3\}$ for each estimator separately to minimise the corresponding MSE. We use Gaussian kernel for all kernel methods, where the lengthscales are set according to median heuristic Gretton et al. [2012].

For each algorithm, we then implement 50 simulations and calculate MSE with respect to the true causal model $\mathbb{E}(Y|do(x))$, which can be computed from the structural model. As shown in Figure 2(a), though KIV performs better than most KAR and KAR.2 estimators, KAR and KAR.2 with $\gamma = 2$ defeat KIV in the KIV setting. This, together with the the fact that KIV defeat other non kernel-based approaches as shown in Singh et al. [2019], indicates that KAR also outperforms DeepIV and SmoothIV in this setting. The parameters c_αs are chosen to be 1, 0.1, 3, 0.8, 3, 3, 3, 1, 0.1, 3, 1, 3, 3, 3 and 2 for KAR with γ being 0, 0.5, 1, 2, 5, 10, 100, KAR.2 with same γ series and KIV, respectively.

B.2 ADDITIONAL SYNTHETIC DATA EXAMPLES

We also consider a variant case where the structural equation is same to the case in Section 5.1 in the main text

$$Y = 0.75C - 0.25Z + \ln(|16X - 8| + 1)sgn(X - 0.5),$$

and the explanatory variables are generated as

$$\begin{pmatrix} C \\ V \\ W \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1, 0.3, 0.2 \\ 0.3, 1, 0 \\ 0.2, 0, 1 \end{pmatrix}\right).$$

Instead, X and Z are set via the following transformation.

$$X = F\left(\frac{|W| + V}{\sqrt{2}}\right), \quad Z = F(|W|) - 0.5. \tag{6}$$

The fitted result of nonlinear and linear methods is shown in Figure B1. The MSE averaged over 50 simulations is shown in Figure 2(b). From the result, we can also see that the proposed kernel anchor regression estimators still performs the best among others under the variant case.

Moreover, we consider a case where the is structural equation is linear,

$$Y = 0.75C - 0.25Z + 0.5X + 0.75,$$

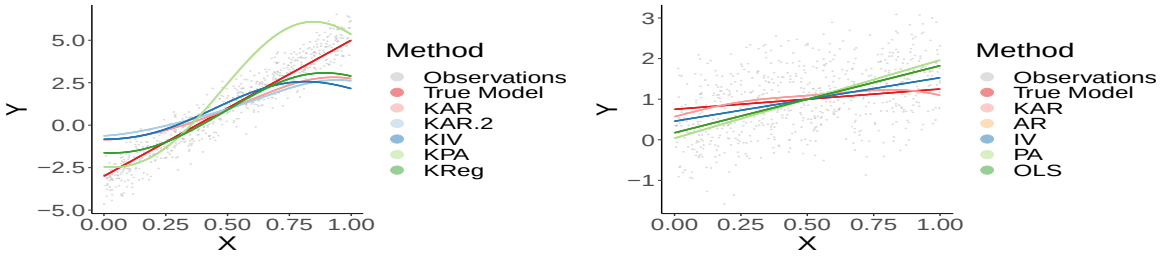


Figure B3: Linear SEM example: fitted nonlinear (left) and linear (right) methods.

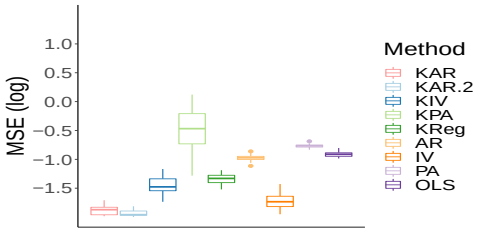


Figure B4: MSE for the linear SEM example.

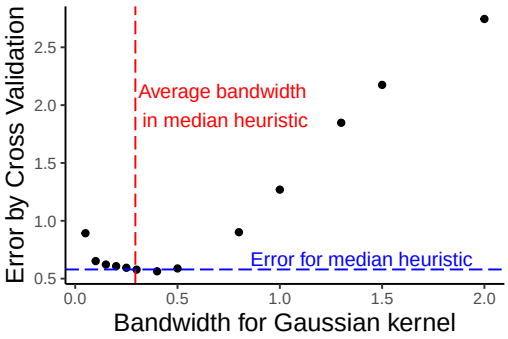
where the data-generating process for X, Z and C remains the same as Section 5.1 in the main text,

$$\begin{pmatrix} C \\ V \\ W \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1, 0.3, 0.2 \\ 0.3, 1, 0 \\ 0.2, 0, 1 \end{pmatrix} \right),$$

and

$$X = F\left(\frac{W+V}{\sqrt{2}}\right), \quad Z = F(W) - 0.5.$$

We compare KAR with the linear models to show the robustness and usefulness of the non-linear anchor regression. By cross-validation, we choose $\gamma = 3$ for KAR estimators. As shown in the Figure B3, KAR and KAR.2 are able to learn the linear relationship well and both methods achieve the lower MSE among others, outperforming the linear methods, as shown in Figure B4.



B.3 BANDWIDTH CHOICE FOR GAUSSIAN KERNEL

We conduct the experiment using different bandwidths for Gaussian kernels on the setting in Section 5; and plot the cross-validation error on the right. The median bandwidth, averaged over 50 trials, are plotted in red vertical line; and the average cross-validation error are plotted in blue horizontal line. The result Bandwidth for Gaussian kernel shows that the median heuristic bandwidth choice achieves close-to-optimal cross-validation error, which reassures the good results presented in the main text.