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 $E^p_X$ and $E^p_Y$ 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. $1 - \delta$:

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

$$\alpha_1 = \left( \frac{8\kappa(Q_1 + \kappa\|E^p_X\|_{\mathcal{H}_\delta}\ln(2/\delta))}{n_1\xi}\right)^{1/2}.$$

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

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

$$\alpha_2 = \left( \frac{8\kappa(Q_2 + \kappa\|E^p_Y\|_{\mathcal{H}_\delta}\ln(2/\epsilon))}{n_2\xi}\right)^{1/2}.$$

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

$$\mathcal{E}_{\gamma,m}^r(H) = \frac{1}{m} \sum_{i=1}^m \|y_{\gamma,i} - H\psi_{\gamma,i}\|_{\mathcal{H}_\delta}^2 + \xi \|H\|_{\mathcal{H}_\delta}^2,$$

$$H_{\xi,m}^r = \arg\min_{H} \mathcal{E}_{\gamma,m}^r(H),$$

where the true transformed inputs and outputs are given by

$$\psi_{\gamma,i} = \psi(x_i) + (\sqrt{\gamma} - 1)E^p_X\phi(z_i) \in \mathcal{H}_\delta, \quad y_{\gamma,i} = y_i + (\sqrt{\gamma} - 1)E^p_Y\phi(z_i) \in \mathcal{Y}.$$

The closed form solution of $H_{\xi,m}^r$ 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 $H_{\xi,m}^r$ to $\mathcal{E}_{\gamma,m}^r(H)$ exists, is unique, and

$$T = \frac{1}{m} \sum_{i=1}^m T_{\psi_{\gamma,i}}, \quad g = \frac{1}{m} \sum_{i=1}^m \Omega_{\psi_{\gamma,i}}y_{\gamma,i}, \quad H_{\xi,m}^r = (T + \xi)^{-1} \circ g.$$

We then define the following terms.
Definition 1  Fix $\eta \in (0, 1)$ and define the following constants

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

For the excess error of KAR estimator $\hat{H}_{\xi}^{\gamma, m}$, 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

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

where

$$S_{-1} = \|\sqrt{T} \circ (\hat{T} + \xi)^{-1}(\hat{g} - g)\|_{\mathcal{H}_\alpha}^2,$$

$$S_0 = \|\sqrt{T} \circ (\hat{T} + \xi)^{-1}(T - \hat{T})H_{\xi}^{\gamma, m}\|_{\mathcal{H}_\alpha}^2,$$

$$S_1 = \|\sqrt{T} \circ (\hat{T} + \xi)^{-1}(g - TH^\gamma)\|_{\mathcal{H}_\alpha}^2,$$

$$S_2 = \|\sqrt{T} \circ (\hat{T} + \xi)^{-1}(T - T)(H_{\xi}^\gamma - H^\gamma)\|_{\mathcal{H}_\alpha}^2,$$

$$A(\xi) = \|\sqrt{T}(H_{\xi}^\gamma - H^\gamma)\|_{\mathcal{H}_\alpha}.$$  

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

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

$$S_0 \leq \frac{4}{\xi} 4BL^2r_x^{2\eta} H_{\xi}^{\gamma, m} H_{\mathcal{H}_\alpha},$$

$$S_1 \leq 32 \ln^2(6\eta) \left[ \frac{BM^2}{m^2 \xi} + \frac{\Sigma^2}{m} \frac{b_\gamma}{\beta_1 / b_\gamma} + \frac{\pi}{\sin(\pi) \xi} \right],$$

$$S_2 \leq 8 \ln^2(6\eta) \left[ \frac{A^2 \xi_{\gamma, \gamma}^{-1}}{m^2 \xi} + \frac{B \xi \gamma_{\gamma}}{m \xi} \right],$$

$$A(\xi) \leq \zeta \xi_{\gamma, \gamma}.$$  

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 $\sqrt{T} \circ (\hat{T} + \xi)^{-1}$ in $S_{-1}$ is given by Proposition 39 in Singh et al. [2019].

Lemma A6  If $\|\hat{\psi}_{\gamma} - \psi_{\gamma}\|_{\mathcal{H}_\alpha} \leq r_x$ w.p. $1 - \delta, \xi \leq \|T\|_{L(\mathcal{H}_\alpha)}$, $m$ is sufficiently large and Condition [3] holds, then w.p. $1 - \eta/3 - \delta$

$$\|\sqrt{T} \circ (\hat{T} + \xi)^{-1}\|_{L(\mathcal{H}_\alpha)} \leq \frac{2}{\sqrt{\xi}}.$$  

With the the error propagated from the estimators in the projection stage, we can bound $\hat{\psi}_{\gamma} - \psi_{\gamma}$ and $\hat{y}_{\gamma} - y_{\gamma}$ as shown in Lemma A7A8.

Lemma A7  Under Condition [4] $\forall \delta \in (0, 1)$, the following statement holds w.p. $1 - \delta:\forall z \in \mathcal{Z}, x \in \mathcal{X}$,

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

Proof 1  By definition, we have

$$\|\hat{\psi}_{\gamma} - \psi_{\gamma}\|_{\mathcal{H}_\alpha} = \|\sqrt{\gamma} - 1\| \left( E_{\alpha_1, X}^{n_1} - E_{X}^{p} \right) \phi(z) \|_{\mathcal{H}_\alpha} \phi(z)\|_{\mathcal{H}_\alpha} \leq |\sqrt{\gamma} - 1|\|E_{\alpha_1, X}^{n_1} - E_{X}^{p}\|_{\mathcal{H}_\alpha} \phi(z)\|_{\mathcal{H}_\alpha}. $$
This, together with Lemma A7 and Condition 1 ensures that w.p. 1 − \( \delta \)
\[
\| \hat{\psi}_r - \psi_r \|_{H_x} \leq r_x(\gamma, \delta, n_1, c_1) := \sqrt{\gamma} - 1|\kappa r E_x(\delta, n_1, c_1)|.
\]

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

Lemma A8 Under Condition 1 \( \forall \epsilon \in (0, 1) \), the following statement holds w.p. 1 − \( \epsilon \); \( \forall z \in Z, y \in Y \),
\[
\| \hat{y}_r - y_r \|_{H_y} \leq r_y(\gamma, \epsilon, n_2, c_2) := \sqrt{\gamma} - 1|\kappa r E_x(\epsilon, n_2, c_2)|.
\]

Proof 2 Lemma A8 is analogous to Lemma A7 by replacing \( \hat{\psi}_r \) with \( y_r \). The proof is thus omitted.

Combining Lemma A6 - A8, we can derive the bound of \( \hat{g} - g \) and then the bound of \( S_{-1} \).

Lemma A9 If \( \| \hat{\psi}_r - \psi_r \|_{H_x} \leq r_x \) w.p. \( 1 - \delta \) and \( \| \hat{y}_r - y_r \|_{Y} \leq r_y \) w.p. \( 1 - \epsilon \), then w.p. \( 1 - \delta - \epsilon \)
\[
\| \hat{g} - g \|_{H_\Omega}^2 \leq 3(L^2r_x^2r_y^2 + B^2r_y^2 + L^2r_x^2C^2).
\]

Proof 3 By definition, we have
\[
\hat{g} - g = \sum_{l=1}^{m} \left( \Omega_{\psi_r, l} \hat{\psi}_{r,l} - \Omega_{\psi_r, l}(x) y_{r,l} \right)
\]
\[
= \sum_{l=1}^{m} \left( \Omega_{\hat{\psi}_r, l} - \Omega_{\psi_r, l} \right) \{ \hat{y}_{r,l} - y_{r,l} \} + \Omega_{\psi_r, l} \{ \hat{y}_{r,l} - y_{r,l} \} + \{ \Omega_{\hat{\psi}_r, l} - \Omega_{\psi_r, l} \} y_{r,l}.
\]

We then have
\[
\| \hat{g} - g \|_{H_\Omega}^2 \leq \sum_{l=1}^{m} \left( \Omega_{\hat{\psi}_r, l} - \Omega_{\psi_r, l} \right) \| \{ \hat{y}_{r,l} - y_{r,l} \} \|_{H_\Omega}^2 + \| \Omega_{\psi_r, l} \|_{H_\Omega} \| \{ \hat{y}_{r,l} - y_{r,l} \} \|_{Y}^2
\]
\[
\leq \sum_{l=1}^{m} \left( \Omega_{\hat{\psi}_r, l} - \Omega_{\psi_r, l} \right) \| \{ \hat{y}_{r,l} - y_{r,l} \} \|_{Y}^2 + \| \Omega_{\psi_r, l} \|_{H_\Omega} \| \{ \hat{y}_{r,l} - y_{r,l} \} \|_{Y}^2
\]
\[
\leq \sum_{l=1}^{m} \| \{ \Omega_{\hat{\psi}_r, l} - \Omega_{\psi_r, l} \} \|_{Y}^2 + \| \Omega_{\psi_r, l} \|_{H_\Omega} \| \{ \hat{y}_{r,l} - y_{r,l} \} \|_{Y}^2
\]
\[
\leq \sum_{l=1}^{m} \| \{ \Omega_{\hat{\psi}_r, l} - \Omega_{\psi_r, l} \} \|_{H_\Omega} \| \{ \hat{y}_{r,l} - y_{r,l} \} \|_{Y}^2 + \| \Omega_{\psi_r, l} \|_{H_\Omega} \| \{ \hat{y}_{r,l} - y_{r,l} \} \|_{Y}^2
\]
\[
\leq 3(L^2r_x^2r_y^2 + B^2r_y^2 + L^2r_x^2C^2).
\]

Lemma A10 Under Condition 1 \( \| \hat{g} - g \|_{H_\Omega}^2 \leq 3(L^2r_x^2r_y^2 + B^2r_y^2 + L^2r_x^2C^2).
\]

Proof 4 We can derive from the definition of \( S_{-1} \) that
\[
S_{-1} \leq \| \sqrt{T} \circ (T + \xi)^{-1} \|_{L_2(H_\Omega)}^2 \| \hat{g} - g \|_{H_\Omega}^2.
\]

This, together with Lemma A6 and Lemma A9 ensures
\[
S_{-1} \leq 4\xi^2(L^2r_x^2r_y^2 + B^2r_y^2 + L^2r_x^2C^2).
\]
We then show the order of the sum $S_0 + S_1 + S_2 + A(\xi)$, which is adapted from Theorem 4 in [Singh et al. 2019].

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

$$f(m) = \frac{1}{m^{2+d_1}} + \frac{1}{m^{1+d_1}} + \frac{1}{m^{d_1}} + \frac{1}{m^{2+d_1}},$$

we then have

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

(i) If $d_1 \leq \frac{b_1(c_1+1)}{b_1c_1+1}$ then $O(f(m)) = O(m^{-\frac{d_1}{c_1+1}})$ with $\xi = m^{-\frac{d_1}{c_1+1}}$.

(ii) If $d_1 > \frac{b_1(c_1+1)}{b_1c_1+1}$ then $O(f(m)) = O(m^{-\frac{b_1c_1}{b_1c_1+1}})$ with $\xi = m^{-\frac{b_1c_1}{b_1c_1+1}}$.

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

$$E^p_X = (B_{XZ} + B_{XC} B_{CZ}),$$

and

$$E^p_Y = [B_{YZ} + B_{YCB_{CZ}} + B_{YX}(B_{XZ} + B_{XC} B_{CZ})].$$

As such, the transformed input and output has the form

$$\psi(\gamma) = B_{XC}\epsilon + \epsilon_X + \sqrt{\gamma}(B_{XZ} + B_{XC} B_{CZ})\phi(Z),$$

and

$$y(\gamma) = (B_{YC} + B_{YX} B_{XC})\epsilon + B_{YX}\epsilon_X + \epsilon_Y + \gamma[B_{YZ} + B_{YCB_{CZ}} + B_{YX}(B_{XZ} + B_{XC} B_{CZ})]\phi(Z).$$
Define relevant covariance matrix/operators as $\Sigma_C = \mathbb{E}[^c_c^T]$, $\Sigma_X = \mathbb{E}[^x_x \otimes ^x_x]$ and $\Sigma_Z = \mathbb{E}[^z_z \otimes ^z_z]$, 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

$$\mathbb{E}[\psi_\gamma(X) \otimes \psi_\gamma(X)] = \mathbb{E}[^c_c^T]B^X \mathbb{E}[^x_x \otimes ^x_x] + \gamma(B^Z \otimes \gamma(Z))B^X \mathbb{E}[^z_z \otimes ^z_z]B^C = B^X \Sigma_C B^C + \Sigma_X \gamma(B^Z \otimes \gamma(Z))B^C \Sigma_B(B^X + BZB^C).$$

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

$$(B^C + B^X B^C \Sigma_C B^C + B^Y X \Sigma_X) \gamma(B^Z + B^C B^Z + B^X X (B^Z + B^C B^Z)) \Sigma_B(B^X + BZB^C)$$

as $^c_c$, $^x_x$ and $^y_y$ are independent variables, which are also independent of $Z$. In overall, we have

$$H^\gamma = [(B^C + B^X B^C \Sigma_C B^C + B^Y X \Sigma_X) + \gamma(B^Z + B^Y B^Z + B^X X (B^Z + B^C B^Z)) \Sigma_B(B^X + BZB^C)]^{-1}$$

The bias of the target KAR estimator is then given by

$$H^\gamma - B^Y X = [(B^C + B^X B^C \Sigma_C B^C + B^Y X \Sigma_X + \gamma(B^Z + B^Y B^Z + B^X X (B^Z + B^C B^Z)) \Sigma_B(B^X + BZB^C)]^{-1}$$

Collecting all the common terms we get

$$H^\gamma - B^Y X = \left[\frac{B^C \Sigma_C B^C + \Sigma_X + \Sigma_B(B^X + BZB^C)}{\Sigma_B(B^X + BZB^C)}\right]^{-1}$$

Thus, $\forall x \in \mathcal{X}, y \in \mathcal{Y}$, consider the inner product $y^T (H^\gamma - B^Y X) \psi(x) = 0$ when the following holds: (i) $B^C = 0$ and $\gamma = 0$, or (ii) $B^Z + B^Y B^Z = 0$ and $\gamma = \infty$, or (iii) $B^C = 0$, $B^Z + B^Y B^Z = 0$ and $\gamma \geq 0$, or (iv) $\Sigma_B(B^X + BZB^C) = a\Sigma_X^y$ for some $a > 0$, and $\gamma = \infty$, or (v) $\Sigma_B(B^X + BZB^C) = -a\Sigma_X^y$ for some $a > 0$, and $\gamma = 1/c$. As such, we conclude $H^\gamma = B^Y X$.

### 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 $E^p_X$ and $E^p_Y$ by $E^0_{\alpha,X}$ and $E^0_{\alpha,Y}$, respectively. In the two-stage KAR procedure, instead, we approximate the two operators by $E^n_{\alpha,X}$ and $E^n_{\alpha,Y}$, respectively. Note that the estimated operators $E^n_{\alpha,X}$ and $E^n_{\alpha,Y}$ use the same $\alpha$. The shared $\alpha$ may fail to ensure the optimal approximation error for $E^n_{\alpha,X}$ and $E^n_{\alpha,Y}$ at the same time.

**Lemma A12** Under Condition 1 and Condition 2, where $F$ denotes 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 procedure, instead, we approximate the two operators by $E^0_{\alpha,X}$ and $E^0_{\alpha,Y}$, respectively. Note that the estimated operators $E^n_{\alpha,X}$ and $E^n_{\alpha,Y}$ use the same $\alpha$. The shared $\alpha$ may fail to ensure the optimal approximation error for $E^n_{\alpha,X}$ and $E^n_{\alpha,Y}$ at the same time.

**Lemma A12** Under Condition 1 and Condition 2, the following holds w.p. $1 - \delta$:

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

Under Condition 1 and Condition 2, the following holds w.p. $1 - \epsilon$:

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

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

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

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

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

**Lemma A12** above provides the upper bounds of the approximation errors for $E^n_{\alpha,X}$ and $E^n_{\alpha,Y}$, 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 $E^n_{\alpha,X}$ and $E^n_{\alpha,Y}$ at the same time, which disjoint sample sets projection estimators can guarantee by setting different $\alpha_1$ and $\alpha_2$ 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 $E^n_{\alpha,X}$ and $E^n_{\alpha,Y}$, can have larger order than using $E^n_{\alpha_1,X}$ and $E^n_{\alpha_2,Y}$ 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 [Chen and Christensen, 2018] studied in [Singh et al., 2019]. The structural model is set as follows,

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

The explanatory variables are generated from

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

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 $\gamma$s 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 $\gamma$ 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 $\zeta = 1m^{-0.5}$, where $c_\alpha > 0$ is a constant.
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 $E(Y | do(x))$, 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 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 $c_\alpha$ are chosen to be 1, 0.1, 3, 0.8, 3, 3, 1, 0.1, 3, 1, 3, 3, 3 and 2 for KAR with $\gamma$ being 0, 0.5, 1, 2, 5, 10, 100, KAR.2 with same $\gamma$ 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.

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

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,

$$Y = 0.75C - 0.25Z + 0.5X + 0.75,$$
where the data-generating process for $X$, $Z$ and $C$ remains the same as Section 5.1 in the main text,

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

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.

**References**

