A. Implementation of algorithms in R

R code implementing causal isotonic calibration with user-supplied (cross-fitted) nuisance estimates and predictions is provided in the Github package

B. Algorithm for causal isotonic calibration with cross-fitted nuisance estimates

Algorithm 4 Causal isotonic calibration (cross-fitted nuisances)

Require: predictor τ, dataset $\mathcal{D}_n$, # of cross-fitting splits k

1: partition $\mathcal{D}_n$ into datasets $\mathcal{T}^{(1)}, \mathcal{T}^{(2)}, \dots, \mathcal{T}^{(k)}$;
2: **for** $s = 1, 2, \dots, k$ **do**
3: let $j(i) = s$ for each $i \in \mathcal{T}^{(s)}$;
4: get estimate $\chi_{n,s}$ of χ_0 from $\mathcal{D}_n \backslash \mathcal{T}^{(s)}$;
5: **end for**
6: perform isotonic regression using pooled out-of-fold estimates to find

$$\theta_n^* = \operatorname*{argmin}_{\theta \in \mathcal{F}_{iso}} \frac{1}{n} \sum_{i=1}^{n} \left[\chi_{n,j(i)}(O_i) - (\theta \circ \tau)(W_i)\right]^2;$$

7: set $\tau_n^* := \theta_n^* \circ \tau$;
Ensure: τ_n^*

C. Technical proofs

Unless stated otherwise, the function τ_n^* denotes a calibrated predictor obtained using Algorithm 1 with a predictor τ, training dataset $\mathcal{E}_m$, and calibration dataset $\mathcal{C}_\ell = \mathcal{D}_n \backslash \mathcal{E}_m$ as described in Section 4.

C.1. Notation & definitions

Let $\mathcal{T} := \{\tau(w) : w \in \mathcal{W}\}$ denote the range of the predictor τ, which is a bounded subset of $\mathbb{R}$ by Condition 4.4. We redefine $\mathcal{F}_{iso} \subset \{\theta : \mathcal{T} \to \mathbb{R}; \theta \text{ is monotone nondecreasing}\}$ to denote the family of nondecreasing functions on $\mathcal{T}$ uniformly bounded by

$$B := \sup_{m \in \mathbb{N}} \sup_{\mathcal{E}_m} \sup_{o \in \mathcal{O}} \left[|\chi_0(o)| + |\chi_m(o)|\right],$$

where the second supremum is over all possible realizations of the training dataset $\mathcal{E}_m$. We necessarily have that B is nonrandom and finite by Lemma C.2. Redefining $\mathcal{F}_{iso}$ to be bounded allows us to directly apply certain maximal inequalities for empirical processes indexed by $\mathcal{F}_{iso}$. Since the isotonic regression estimator is obtained by locally averaging the pseudo-outcome χ_m (Barlow & Brunk, 1972), the unconstrained isotonic regression solution satisfies this bound and falls in the interior of this class almost surely. Moreover, $\mathcal{F}_{iso}$ is a convex subset of the space of monotone nondecreasing functions. Let $\mathcal{F}_{TV} \subset \{\theta : \mathbb{R} \to \mathbb{R}; \theta \text{ is of bounded variation}\}$ denote the space of functions with total variation uniformly bounded by three times the total variation of the function θ_0 where θ_0 is as in condition 4.5. Additionally, let $\mathcal{F}_{\tau,iso} := \{\theta \circ \tau : \mathcal{W} \to \mathbb{R}; \theta \in \mathcal{F}_{iso}\}$ be the family of functions obtained by composing nondecreasing functions in $\mathcal{F}_{iso}$ with τ, and let $\mathcal{F}_{\tau,TV} := \{\theta \circ \tau : \mathcal{W} \to \mathbb{R}; \theta \in \mathcal{F}_{TV}\}$ be the family of functions obtained by composing functions in $\mathcal{F}_{TV}$ with τ. Let $\mathcal{F}_{Lip,m} := \{o \mapsto [\tau_2(w) - \tau_1(w)][\chi_m(o) - \tau_2(w)] : \mathcal{O} \to \mathbb{R}; \tau_2 \in \mathcal{F}_{\tau,TV}, \tau_1 \in \mathcal{F}_{\tau,iso}\}$, where χ_m is the estimated pseudo-outcome function. Finally, for a function class $\mathcal{F}$, let $N(\epsilon, \mathcal{F}, L_2(P))$ denote the $\epsilon-$covering number (van der Vaart & Wellner, 1996) of $\mathcal{F}$ and define the uniform entropy integral of $\mathcal{F}$ by

$$\mathcal{J}(\delta, \mathcal{F}) := \int_0^\delta \sup_Q \sqrt{\log N(\epsilon, \mathcal{F}, L_2(Q))}\, d\epsilon\,,$$

where the supremum is taken over all discrete probability distributions Q. In contrast to the definition provided in van der Vaart & Wellner (1996), we do not define the uniform entropy integral relative to an envelope function for the function class $\mathcal{F}$. We can do this since all function classes we consider are uniformly bounded. Thus, any uniformly bounded envelope function will only change the uniform entropy integral as defined in van der Vaart & Wellner (1996) by a constant.

In the results below, we will use the following empirical process notation: for a $P-$measurable function f, we denote $\int f(o)dP(o)$ by Pf, and so, letting P_ℓ denote the empirical distribution of $\mathcal{C}_\ell$, $P_\ell f$ equals $\frac{1}{\ell}\sum_{i\in\mathcal{I}_\ell} f(O_i)$ with $\mathcal{I}_\ell$ indexing observations of $\mathcal{C}_\ell \subset \mathcal{D}_n$. We also let $\|f\|_P^2 := Pf^2$; to simplify notation, we omit the dependency in P and use $\|f\|^2$ instead of $\|f\|_P^2$. Finally, for two quantities x and y, we use the expression $x \lessapprox y$ to mean that x is upper bounded by y times a universal constant that may only depend on global constants that appear in conditions 4.1-4.5

C.2. Technical lemmas

The following lemma is a key component of our proof of Theorem 4.6.

Lemma C.1. *For a calibrated predictor τ_n^* obtained using Algorithm 1, and any real-valued function r, we have that*

$$\sum_{i\in\mathcal{I}_\ell}[r\circ\tau_n^*(W_i)]\,[\tau_n^*(W_i)-\chi_m(O_i)]=0\,. \tag{5}$$

Proof. Note that $\tau_n^*(w)$ can be expressed pointwise for any $w\in\mathcal{W}$ as $\theta_n^*\circ\tau(w)=a_0+\sum_{j=1}^J a_j 1(\tau(w)\geq u_j)$ for a piecewise constant function θ_n^* determined by coefficients $\{a_j\}_{j=0}^J$ and jump points $\{u_j\}_{j=1}^J$ (Barlow & Brunk, 1972). By monotonicity, we necessarily have $a_0\in\mathbb{R}$ and $\{a_j\}_{j=1}^J$ are positive coefficients.

Let $R_n(\theta):=\sum_{i\in\mathcal{I}_\ell}[\theta\circ\tau(W_i)-\chi_m(O_i)]^2$ denote the least-squares risk used in the isotonic regression step. Fix an arbitrary jump point $\bar{u}_j$, and let $\xi_n:\mathbb{R}^2\to\mathbb{R}$ denote the function $\xi_n(\varepsilon,h):=\theta_n^*(h)+\varepsilon 1(h\geq\bar{u}_j)$. Note that $\delta>0$ can be chosen to be sufficiently small that, for all $|\varepsilon|\leq\delta$, $h\mapsto\xi_n(\varepsilon,h)$ is nondecreasing — for instance, $\delta=\min\{a_j\}_{j=1}^J$ suffices. Thus, for sufficiently small $\delta>0$, $h\mapsto\xi_n(\varepsilon,h)$ lies in the space of monotone nondecreasing function for all $|\varepsilon|\leq\delta$. In a slight abuse of notation, we let $R_n(\xi_n(\varepsilon)):=\sum_{i\in\mathcal{I}_\ell}[\xi_n(\varepsilon,\tau(W_i))-\chi_m(O_i)]^2$ and $R_n(\xi_n(-\varepsilon)):=\sum_{i\in\mathcal{I}_\ell}[\xi_n(-\varepsilon,\tau(W_i))-\chi_m(O_i)]^2$.

Now, because θ_n^* minimizes $\theta\mapsto R_n(\theta)$ over the space of monotone nondecreasing functions, for all $\varepsilon\geq 0$, it holds that both $R_n(\xi_n(\varepsilon))-R_n(\tau_n^*)\geq 0$ and $R_n(\xi_n(-\varepsilon))-R_n(\tau_n^*)\geq 0$. Moreover, when $\varepsilon=0$, $R_n(\xi_n(0))-R_n(\tau_n^*)=0$. Therefore, if ε is sufficiently close to 0, the derivative with respect to ε of $R_n(\xi_n(\varepsilon))-R_n(\tau_n^*)$ must be non-negative, and $R_n(\xi_n(-\varepsilon))-R_n(\tau_n^*)$ must be non-positive. Hence, it must be true that

$$\frac{d}{d\varepsilon}[R_n(\xi_n(\varepsilon))-R_n(\theta_n^*)]\Big|_{\varepsilon=0}\geq 0\ \text{ and }\ \frac{d}{d\varepsilon}[R_n(\xi_n(-\varepsilon))-R_n(\theta_n^*)]\Big|_{\varepsilon=0}\leq 0\,.$$

This, in turn, implies that

$$2\sum_{i\in\mathcal{I}_\ell}1(\tau(W_i)\geq\bar{u}_j)\,[\tau_n^*(W_i)-\chi_m(O_i)]\geq 0\ \text{ and }\ 2\sum_{i\in\mathcal{I}_\ell}1(\tau(W_i)\geq\bar{u}_j)\,[\tau_n^*(W_i)-\chi_m(O_i)]\leq 0\,,$$

and so, it follows that $\sum_{i\in\mathcal{I}_\ell}1(\tau(W_i)\geq\bar{u}_j)\,[\tau_n^*(W_i)-\chi_m(O_i)]=0$. Because the jump point $\bar{u}_j$ was arbitrary, we have that for all functions of the form $s(w)=b_0+\sum_{j=1}^J b_j 1(\tau(w)\geq u_j)$ with coefficients $\{b_j\}_{j=0}^J$, we can show that

$$\sum_{i\in\mathcal{I}_\ell}s(W_i)\,[\tau_n^*(W_i)-\chi_m(O_i)]=0$$

by taking linear combinations of $1(\tau(w)\geq u_j)$ and noting that the score equations are linear in s. The main result of this lemma follows from the fact that, since both τ_n^* and $r\circ\tau_n^*$ can be expressed in this form, for any real-valued function r, we have that

$$\sum_{i\in\mathcal{I}_\ell}r\circ\tau_n^*(W_i)\,[\tau_n^*(W_i)-\chi_m(O_i)]=0\,.$$

□

Lemma C.2. *Conditions 4.1, 4.2 and 4.4 imply that the function classes $\mathcal{F}_{iso}$, $\mathcal{F}_{\tau,TV}$, $\mathcal{F}_{\tau,iso}$ and $\mathcal{F}_{Lip,m}$ are bounded.*

Proof. By Conditions 4.1, 4.2 and 4.4, we know that $\chi_m(o)$ is bounded uniformly over all observations $o\in\mathcal{O}$ and realizations of $\mathcal{E}_m$, that is, there exists a finite fixed constant B such that $\operatorname{ess\,sup}_{m\in\mathbb{N},o\in\mathcal{O}}\chi_m(o)\leq B/2$. Hence, as defined in the previous section, $\mathcal{F}_{iso}$ is uniformly bounded. Moreover, because $\mathcal{F}_{iso}$ is bounded, it directly implies that

$\mathcal{F}_{\tau,iso}$ is bounded. Noting that functions of finite variation are bounded, in view of Condition 4.5, we have that $\mathcal{F}_{TV}$ is uniformly bounded by some constant that depends neither on θ nor τ. This implies that $\mathcal{F}_{\tau,TV}$ is uniformly bounded. Finally, because $\mathcal{F}_{\tau,TV}$, $\mathcal{F}_{\tau,iso}$, χ_m and the potential outcomes are uniformly bounded, the function class $\mathcal{F}_{Lip,m}$ is also uniformly bounded. □

Lemma C.3. *Under conditions 4.5 and the conditions of Lemma C.2, the function $\tau' \mapsto E[Y_1 - Y_0 \mid \tau_n^*(W) = \tau']$ has total variation bounded above by three times the total variation of θ_0, where θ_0 is as in Condition 4.5.*

Proof. Since the function θ_n^* is nondecreasing and piecewise constant, we have

$$E[Y_1 - Y_0 \mid (\theta_n^* \circ \tau)(W) = \tau'] = E[Y_1 - Y_0 \mid \tau(W) \in B_{\tau'}]$$

for the set $B_{\tau'} := \{z \in \mathcal{T} : \theta_n^*(z) = \tau'\}$, where $B_{\tau'} = \{z \in \mathcal{T} : a(\tau') \leq z < b(\tau')\}$ for some endpoints $a(\tau'), b(\tau') \in \mathbb{R}$. The law of total expectation further implies that

$$E[Y_1 - Y_0 \mid \tau(W) \in B_{\tau'}] = E[\theta_0 \circ \tau(W) \mid \tau(W) \in B_{\tau'}] \ ,$$

where θ_0 is such that $\theta_0 \circ \tau(W) = \gamma_0(\tau, W)$ P-almost surely. By Condition 4.5, the function θ_0 is of bounded total variation. Heuristically, since $\tau' \mapsto E[\theta_0 \circ \tau(W) \mid \tau(W) \in B_{\tau'}]$ is obtained by locally averaging θ_0 within the bins $(B_{\tau'} : \tau')$, its total variation should also be bounded. We show this formally as follows. Note first that

$$E[\theta_0 \circ \tau(W) \mid \tau(W) \in B_{\tau'}] = E[\theta_0^+ \circ \tau(W) \mid \tau(W) \in B_{\tau'}] - E[\theta_0^- \circ \tau(W) \mid \tau(W) \in B_{\tau'}] \ ,$$

where θ_0^+ and θ_0^- are two bounded, nondecreasing functions satisfying the Jordan decomposition $\theta_0 = \theta_0^+ - \theta_0^-$ (Theorem 4, Section 5.2 of Royden, 1963). Moreover, we can choose θ_0^+ such that $\theta_0^+(\infty) - \theta_0^+(-\infty)$ is equal to the total variation of θ_0. Since $\|\theta_0^-\|_{TV} = \|\theta_0 - \theta_0^+\|_{TV} \leq \|\theta_0\|_{TV} + \|\theta_0^+\|_{TV}$, we have that $\|\theta_0^-\|_{TV}$ is bounded by $2\|\theta_0\|_{TV}$.

Since θ_n^* is nondecreasing, by definition, we have that $t_1 < t_2$ implies that $x_1 < x_2$ for any $x_1 \in B_{t_1}$ and $x_2 \in B_{t_2}$. It follows that both $\tau' \mapsto E[\theta_0^+ \circ \tau(W) \mid \tau(W) \in B_{\tau'}]$ and $\tau' \mapsto E[\theta_0^- \circ \tau(W) \mid \tau(W) \in B_{\tau'}]$ are nondecreasing; furthermore, they are also bounded. By Theorem 4 of Royden (1963), a function is of bounded variation if and only if it is the difference between two bounded nondecreasing functions. We conclude that $\tau' \mapsto E[Y_1 - Y_0 \mid \theta_n^* \circ \tau(W) = \tau'] = E[\theta_0^+ \circ \tau(W) \mid \tau(W) \in B_{\tau'}] - E[\theta_0^- \circ \tau(W) \mid \tau(W) \in B_{\tau'}]$ is of bounded variation. Moreover, its total variation norm is bounded above by the sum of the total variation norm of $E[\theta_0^+ \circ \tau(W) \mid \tau(W) \in B_{\tau'}]$ and that of $E[\theta_0^- \circ \tau(W) \mid \tau(W) \in B_{\tau'}]$. We recall that the total variation of monotone functions is simply the difference between the left and right endpoints of the monotone function, and that

$$\operatorname*{ess\,inf}_{w \in \mathcal{W}} (\theta_0^+ \circ \tau)(w) \leq E[\theta_0^+ \circ \tau(W) \mid \tau(W) \in B_{\tau'}] \leq \operatorname*{ess\,sup}_{w \in \mathcal{W}} (\theta_0^+ \circ \tau)(w),$$

and similarly for $\theta_0^- \circ \tau$. As a consequence, the total variation norms of $E[\theta_0^+ \circ \tau(W) \mid \tau(W) \in B_{\tau'}]$ and $E[\theta_0^- \circ \tau(W) \mid \tau(W) \in B_{\tau'}]$ are bounded by the total variation norm of θ_0^+ and that of θ_0^-, respectively. Using the sublinearity of the total variation norm, we conclude that $\tau' \mapsto E[Y_1 - Y_0 \mid \theta_n^* \circ \tau(W) = \tau']$ has total variation norm bounded above by $3\|\theta_0\|_{TV}$. □

C.3. Proofs of theorems

PROOF OF THEOREM 4.6

Proof. Conditioning on $\mathcal{D}_n$, we have that

$$\begin{aligned}
&E\left\{[\gamma_0(\tau_n^*, W) - \tau_n^*(W)]\left[\chi_0(O) - \tau_n^*(W)\right] \mid \mathcal{D}_n\right\} \\
&\quad = E\{E\left\{[\gamma_0(\tau_n^*, W) - \tau_n^*(W)]\left[\chi_0(O) - \tau_n^*(W)\right] |W\right\} \mid \mathcal{D}_n\} \\
&\quad = E\{[\gamma_0(\tau_n^*, W) - \tau_n^*(W)]\left[\tau_0(W) - \tau_n^*(W)\right] \mid \mathcal{D}_n\} \\
&\quad = E\{E\left\{[\gamma_0(\tau_n^*, W) - \tau_n^*(W)]\left[\tau_0(W) - \tau_n^*(W)\right] |\tau_n^*(W)\right\} \mid \mathcal{D}_n\} \\
&\quad = E\{[\gamma_0(\tau_n^*, W) - \tau_n^*(W)]\left[\gamma_0(\tau_n^*, W) - \tau_n^*(W)\right] \mid \mathcal{D}_n\} \\
&\quad = E\{[\gamma_0(\tau_n^*, W) - \tau_n^*(W)]^2 \mid \mathcal{D}_n\} \ .
\end{aligned}$$

The above equality implies that

$$\begin{aligned}\int \{\gamma_0(\tau_n^*, w) - \tau_n^*(w)\}^2 \, dP(w) &= \int \{\gamma_0(\tau_n^*, w) - \tau_n^*(w)\} \{\chi_0(o) - \tau_n^*(w)\} \, dP(o) \\ &= \int \{\gamma_0(\tau_n^*, w) - \tau_n^*(w)\} \{\chi_0(o) - \chi_m(o)\} \, dP(o) \\ &\quad + \int \{\gamma_0(\tau_n^*, w) - \tau_n^*(w)\} \{\chi_m(o) - \tau_n^*(w)\} \, dP(o) \, .\end{aligned} \tag{6}$$

Note that, by Lemma C.1, for each real-valued function r, τ_n^* satisfies the equation

$$\frac{1}{\ell} \sum_{i \in \mathcal{I}_\ell} r(\tau_n^*(W_i)) \, [\chi_m(O_i) - \tau_n^*(W_i)] = 0 \, .$$

Setting $r(\tau') := E[Y_1 - Y_0 \mid \tau_n^*(W) = \tau'] - \tau'$, we conclude that

$$\int \{\gamma_0(\tau_n^*, w) - \tau_n^*(w)\} \{\chi_m(o) - \tau_n^*(w)\} \, dP_\ell(o) = 0 \, .$$

Subtracting the above score equation from the second summand in (6), we obtain that

$$\begin{aligned}\int \{\gamma_0(\tau_n^*, w) - \tau_n^*(w)\}^2 \, dP(w) &= \int \{\gamma_0(\tau_n^*, w) - \tau_n^*(w)\} \{\chi_0(o) - \chi_m(o)\} \, dP(o) \\ &\quad + \int \{\gamma_0(\tau_n^*, w) - \tau_n^*(w)\} \{\chi_m(o) - \tau_n^*(w)\} \, d(P - P_\ell)(o) \, .\end{aligned} \tag{7}$$

This may be written in shorthand as $\|\gamma_0(\tau_n^*, \cdot) - \tau_n^*\|^2 = (I) + (II)$ with

$$\begin{aligned}(I) &:= P\{[\gamma_0(\tau_n^*, \cdot) - \tau_n^*](\chi_0 - \chi_m)\} \\ (II) &:= (P - P_\ell)\{[\gamma_0(\tau_n^*, \cdot) - \tau_n^*](\chi_m - \tau_n^*)\} \, .\end{aligned}$$

In order to show the desired result, we will bound both (I) and (II).

We can bound (I) using the law of iterated conditional expectations and the Cauchy-Schwarz inequality. First, conditioning on $\mathcal{E}_m$, we note that

$$\begin{aligned}&P\{[\gamma_0(\tau_n^*, \cdot) - \tau_n^*](\chi_0 - \chi_m)\} \\ &= \int \{\gamma_0(\tau_n^*, w) - \tau_n^*(w)\} \, E[\chi_0(O) - \chi_m(O) \mid W = w, \mathcal{E}_m] \, dP(w) \\ &\leq \|\gamma_0(\tau_n^*, \cdot) - \tau_n^*\| \, \|E[\chi_0(O) \mid W = \cdot\,] - E[\chi_m(O) \mid W = \cdot\,, \mathcal{E}_m]\| \, .\end{aligned} \tag{8}$$

Next, we express the second norm in (8) in terms of $\|\pi_m - \pi_0\|$ and $\|\mu_m - \mu_0\|$. Recalling that $E[\chi_0(O) \mid W = w] = \tau_0(w)$, we have that

$$\begin{aligned}&E[\chi_m(O) \mid W = w, \mathcal{E}_m] - E[\chi_0(O) \mid W = w] \\ &= \mu_m(1, w) - \mu_0(1, w) - [\mu_m(0, w) - \mu_0(0, w)] + \frac{\pi_0(w)}{\pi_m(w)} [\mu_0(1, w) - \mu_m(1, w)] \\ &\quad + \frac{1 - \pi_0(w)}{1 - \pi_m(w)} [\mu_0(0, w) - \mu_m(0, w)] \\ &= \left[\frac{\pi_0(w) - \pi_m(w)}{\pi_m(w)}\right] [\mu_0(1, w) - \mu_m(1, w)] + \left[\frac{\pi_m(w) - \pi_0(w)}{1 - \pi_m(w)}\right] [\mu_0(0, w) - \mu_m(0, w)] \, .\end{aligned}$$

By Condition 4.2, $P(1 - \eta > \pi_m(W) > \eta) = 1$ for some $\eta > 0$. The latter condition combined with the Cauchy-Schwarz inequality gives that $\|E[\chi_0(O) \mid W = \cdot\,] - E[\chi_m(O) \mid W = \cdot\,, \mathcal{E}_m]\|$ is bounded above by

$$\|[\pi_m(\cdot) - \pi_0(\cdot)][\mu_0(0, \cdot) - \mu_m(0, \cdot)]\| + \|[\pi_m(\cdot) - \pi_0(\cdot)][\mu_0(1, \cdot) - \mu_m(1, \cdot)]\| \, .$$

By Condition 4.2, we also have that for any P-measurable function $h : \mathcal{W} \to \mathbb{R}$

$$\begin{aligned}\int h(w)^2[\mu_0(1,w) - \mu_m(1,w)]^2 dP(w) &= \iint h(w)^2[\mu_0(a,w) - \mu_m(a,w)]^2 \frac{a}{\pi_0(w)} P(da,dw) \\ &\leq \frac{1}{\eta} \iint h(w)^2[\mu_0(a,w) - \mu_m(a,w)]^2 P(da,dw) \,.\end{aligned}$$

The same bound holds for $\int h(w)^2[\mu_0(0,w) - \mu_m(0,w)]^2 dP(w)$. Setting $h : w \mapsto \pi_m(w) - \pi_0(w)$, we conclude

$$\|E[\chi_m(O) \,|\, W = \cdot\,, \mathcal{E}_m] - E[\chi_0(O) \,|\, W = \cdot\,]\| \lessapprox \|(\pi_m - \pi_0)(\mu_0 - \mu_m)\| \,. \tag{9}$$

Together, (8) and (9) yield that (I) is bounded above by

$$P\{[\gamma_0(\tau_n^*, \cdot) - \tau_n^*](\chi_0 - \chi_m)\} \lessapprox \|\gamma_0(\tau_n^*, \cdot) - \tau_n^*\| \, \|(\pi_m - \pi_0)\,(\mu_0 - \mu_m)\| \,. \tag{10}$$

We now find an upper bound for (II). We claim that, conditionally on $\mathcal{E}_m$, the random functions appearing in this empirical process term are contained in fixed and uniformly bounded function classes. To see this, we note that $\tau_n^* = \theta_n^* \circ \tau$ for some $\theta_n^* \in \mathcal{F}_{iso}$ and, as a consequence, $\tau_n^* \in \mathcal{F}_{\tau,iso}$, a uniformly bounded function class by Lemma C.2, P_0-almost surely. By Lemma C.3, the function $w \mapsto \gamma_0(\tau_n^*, w)$ falls in $\mathcal{F}_{\tau,TV}$. This further implies that $o \mapsto \{E[Y_1 - Y_0 \,|\, \tau_n^*(W) = \tau_n^*(w)] - \tau_n^*(w)\}\{\chi_m(o) - \tau_n^*(w)\} \in \mathcal{F}_{Lip,m}$, which is a uniformly bounded function class by Lemma C.2.

Next, we let $C := \operatorname{ess\,sup}_{x \in \mathcal{T}} |\theta_0(x)|$ and define $K := B + C$, where we recall that $B := \sup_{m \in \mathbb{N}} \sup_{\mathcal{E}_m} \operatorname{ess\,sup}_{o \in \mathcal{O}} \{|\chi_0(o)| + |\chi_m(o)|\}$. Furthermore, we set $\delta_n := \|\gamma_0(\tau_n^*, \cdot) - \tau_n^*\|$, which is a random rate. For any given rate δ, we define

$$S_n(\delta) := \sup_{\tau_1 \in \mathcal{F}_{\tau,TV}, \tau_2 \in \mathcal{F}_{\tau,iso} : \|\tau_1 - \tau_2\| \leq \delta} (P - P_\ell)\{(\tau_1 - \tau_2)(\chi_m - \tau_2)\} = \sup_{f \in \mathcal{F}_{Lip,m} : \|f\| \leq \delta K} (P - P_\ell) f \,.$$

As a consequence of the above, we have that $(II) \leq S_n(\delta_n)$. Due to the randomness in δ_n, the above cannot be further upper-bounded immediately. To bound the term above, we will take a $\delta > 0$ that is deterministic conditional on $\mathcal{E}_m$, and upper-bound $\phi_n(\delta) := E\{S_n(\delta)\}$, where the expectation is also taken over $\mathcal{D}_n$. To bound the above term, we will use empirical process techniques with the function classes $\mathcal{F}_{iso}$, $\mathcal{F}_{\tau,TV}$, $\mathcal{F}_{\tau,iso}$ and $\mathcal{F}_{Lip,m}$. To do so, we must study the uniform entropy integral

$$\mathcal{J}(\delta, \mathcal{F}) := \int_0^\delta \sup_Q \sqrt{N(\varepsilon, \mathcal{F}, \|\cdot\|_Q)}\, d\varepsilon$$

for each of these function classes. By Lemma C.2, all these function classes are uniformly bounded. We note that, conditional on $\mathcal{E}_m$ so that χ_m is fixed, $\mathcal{F}_{Lip,m}$ is a multivariate Lipschitz transformation of $\mathcal{F}_{\tau,TV}$ and $\mathcal{F}_{\tau,iso}$, and therefore, by Theorem 2.10.20 of (van der Vaart & Wellner, 1996), we have that $\mathcal{J}(\delta, \mathcal{F}_{Lip,m}) \lessapprox \mathcal{J}(\delta, \mathcal{F}_{\tau,TV}) + \mathcal{J}(\delta, \mathcal{F}_{\tau,iso})$. Since functions of bounded total variation can be written as a difference of nondecreasing monotone functions, we have by the same theorem that $\mathcal{J}(\delta, \mathcal{F}_{TV}) \lessapprox \mathcal{J}(\delta, \mathcal{F}_{iso})$. We claim the same upper bound holds up to a constant for $\mathcal{F}_{\tau,TV}$ and $\mathcal{F}_{\tau,iso}$. We establish this explcitly for $\mathcal{F}_{\tau,iso}$ below; the result for $\mathcal{F}_{\tau,TV}$ follows from an identical argument. We note that

$$\mathcal{J}(\delta, \mathcal{F}_{\tau,iso}) = \int_0^\delta \sup_Q \sqrt{N(\varepsilon, \mathcal{F}_{\tau,iso}, \|\cdot\|_Q)}\, d\varepsilon = \int_0^\delta \sup_Q \sqrt{N(\varepsilon, \mathcal{F}_{iso}, \|\cdot\|_{Q \circ \tau^{-1}})}\, d\varepsilon = \mathcal{J}(\delta, \mathcal{F}_{iso}) \,,$$

where $Q \circ \tau^{-1}$ is the push-forward probability measure for the random variable $\tau(W)$. We now proceed with bounding $\phi_n(\delta)$. Applying Theorem 2.10.20 of (van der Vaart & Wellner, 1996), we obtain, for any $\delta > 0$ deterministic conditionally on $\mathcal{E}_m$, that

$$\begin{aligned} E\left[S_n(\delta) \,|\, \mathcal{E}_m\right] &\lessapprox \ell^{-1/2} \mathcal{J}(\delta, \mathcal{F}_{Lip,m}) \left(1 + \frac{\mathcal{J}(\delta, \mathcal{F}_{Lip,m})}{\sqrt{\ell}\delta^2}\right) \\ &\lessapprox \ell^{-1/2} \mathcal{J}(\delta, \mathcal{F}_{iso}) \left(1 + \frac{\mathcal{J}(\delta, \mathcal{F}_{iso})}{\sqrt{\ell}\delta^2}\right), \end{aligned} \tag{11}$$

where the right-hand side can only be random through δ.

We can now proceed with the main argument that gives a rate of convergence for δ_n. First, we note that combining Equations 7 and 10 yields that the event

$$\left\{\|\gamma_0(\tau_n^*,\cdot)-\tau_n^*\|^2 \le \|\gamma_0(\tau_n^*,\cdot)-\tau_n^*\|\, \|(\pi_m-\pi_0)(\mu_m-\mu_0)\| + S_n(\delta_n)\right\}$$

occurs with probability one. We then proceed with a peeling argument to account for the randomness of δ_n. Let ε_n be any given sequence that is deterministic conditional on $\mathcal{E}_m$, and define A_s as the event $\left\{2^{s+1}\varepsilon_n \ge \|\gamma_0(\tau_n^*,\cdot)-\tau_n^*\| \ge 2^s\varepsilon_n\right\}$ as well as the random quantity $\epsilon_m^{nuis} := \|(\pi_m-\pi_0)(\mu_m-\mu_0)\|$. Then, for any $S>0$, we have that

$$\begin{aligned}\left(\|\gamma_0(\tau_n^*,\cdot)-\tau_n^*\| \ge 2^S\varepsilon_n\right) &= \sum_{s=S}^{\infty} P\left(2^{s+1}\varepsilon_n \ge \|\gamma_0(\tau_n^*,\cdot)-\tau_n^*\| \ge 2^s\varepsilon_n\right) = \sum_{s=S}^{\infty} P(A_s)\\ &= \sum_{s=S}^{\infty} P\left(A_s, \delta_n^2 \le \delta_n\epsilon_m^{nuis} + S_n(\delta_n)\right). \qquad (12)\end{aligned}$$

In all the events in the above sum, we have that $S_n(\delta_n) \le S_n(2^{s+1}\varepsilon_n)$ since $\delta_n = \|\gamma_0(\tau_n^*,\cdot)-\tau_n^*\|$. Next, manipulating the inequalities in the above events, we have that

$$\begin{aligned}\left\{A_s, \delta_n^2 \le \delta_n\epsilon_m^{nuis} + S_n(\delta_n)\right\} &\subseteq \left\{A_s, \delta_n^2 \le 2^{s+1}\varepsilon_n\epsilon_m^{nuis} + S_n(2^{s+1}\varepsilon_n)\right\}\\ &\subseteq \left\{2^{2s}\varepsilon_n^2 \le \delta_n^2 \le 2^{s+1}\varepsilon_n\epsilon_m^{nuis} + S_n(2^{s+1}\varepsilon_n)\right\}\\ &\subseteq \left\{2^{2s}\varepsilon_n^2 \le 2^{s+1}\varepsilon_n\epsilon_m^{nuis} + S_n(2^{s+1}\varepsilon_n)\right\},\end{aligned}$$

which implies that the sum in (12) is upper bounded by

$$\sum_{s=S}^{\infty} P\left(2^{2s}\varepsilon_n^2 \le 2^{s+1}\varepsilon_n\epsilon_m^{nuis} + S_n(2^{s+1}\varepsilon_n)\right).$$

Using (11) and Markov's inequality, we find that

$$\begin{aligned}&\sum_{s=S}^{\infty} P\left(2^{2s}\varepsilon_n^2 \le 2^{s+1}\varepsilon_n\epsilon_m^{nuis} + S_n(2^{s+1}\varepsilon_n)\right)\\ &\le \sum_{s=S}^{\infty} E\left\{P\left(2^{2s}\varepsilon_n^2 \le 2^{s+1}\varepsilon_n\epsilon_m^{nuis} + S_n(2^{s+1}\varepsilon_n) \,|\, \mathcal{E}_m\right)\right\}\\ &\le \sum_{s=S}^{\infty} E\left\{\frac{2^{s+1}\varepsilon_n\epsilon_m^{nuis} + E[S_n(2^{s+1}\varepsilon_n) \,|\, \mathcal{E}_m]}{2^{2s}\varepsilon_n^2}\right\}\\ &\lessapprox \sum_{s=S}^{\infty} E\left[\frac{\epsilon_m^{nuis}}{2^{s-1}\varepsilon_n} + \frac{\mathcal{J}(2^{s+1}\varepsilon_n,\mathcal{F}_{iso})}{2^{2s}\sqrt{\ell}\varepsilon_n^2}\left(1+\frac{\mathcal{J}(2^{s+1}\varepsilon_n,\mathcal{F}_{iso})}{\sqrt{\ell}2^{2s+1}\varepsilon_n^2}\right)\right].\end{aligned}$$

As a consequence of Lemma C.2 and the covering number bound for bounded monotone functions given in Theorem 2.7.5 of van der Vaart & Wellner (1996), we have that $\mathcal{J}(2^{s+1}\varepsilon_n,\mathcal{F}_{iso}) = 2^{s/2+1/2}\sqrt{\varepsilon_n}$. Using this fact, we find that

$$\frac{\mathcal{J}(2^{s+1}\varepsilon_n,\mathcal{F}_{iso})}{2^{2s}\sqrt{\ell}\varepsilon_n^2} \lessapprox \frac{1}{2^s}\frac{\mathcal{J}(\varepsilon_n,\mathcal{F}_{iso})}{\sqrt{\ell}\varepsilon_n^2},$$

from which it follows that

$$\frac{\mathcal{J}(2^{s+1}\varepsilon_n,\mathcal{F}_{iso})\left(1+\frac{\mathcal{J}(2^{s+1}\varepsilon_n,\mathcal{F}_{iso})}{\sqrt{\ell}2^{2s+1}\varepsilon_n^2}\right)}{2^{2s}\sqrt{\ell}\varepsilon_n^2} \lessapprox 2^{-s}\frac{\mathcal{J}(\varepsilon_n,\mathcal{F}_{iso})\left(1+\frac{\mathcal{J}(\varepsilon_n,\mathcal{F}_{iso})}{\sqrt{\ell}\varepsilon_n^2}\right)}{\sqrt{\ell}\varepsilon_n^2}.$$

We now choose $\varepsilon_n := \max\{\ell^{-1/3}, \|(\pi_m-\pi_0)(\mu_m-\mu_0)\|\}$, which indeed is deterministic conditional on $\mathcal{E}_m$. This choice ensures that $\mathcal{J}(\varepsilon_n,\mathcal{F}_{iso}) \lessapprox \sqrt{\ell}\varepsilon_n^2$ and $\epsilon_m^{nuis} = \|(\pi_m-\pi_0)(\mu_m-\mu_0)\| \lessapprox \varepsilon_n$, so that

$$\frac{\epsilon_m^{nuis}}{2^{s-1}\varepsilon_n} + \frac{\mathcal{J}(2^{s+1}\varepsilon_n,\mathcal{F}_{iso})}{2^{2s}\sqrt{\ell}\varepsilon_n^2}\left(1+\frac{\mathcal{J}(2^{s+1}\varepsilon_n,\mathcal{F}_{iso})}{\sqrt{\ell}2^{2s+1}\varepsilon_n^2}\right) \lessapprox \frac{1}{2^s},$$

where the right-hand side is nonrandom. Thus, we have that

$$P\left(\|\gamma_0(\tau_n^*,\cdot)-\tau_n^*\|\geq 2^S\varepsilon_n\right) \lessapprox \sum_{s=S}^{\infty}\frac{1}{2^s}\xrightarrow[S\to\infty]{} 0\,.$$

As a consequence, for every $\varepsilon > 0$, we can find a constant 2^S sufficiently large such that $P\left(\|\gamma_0(\tau_n^*,\cdot)-\tau_n^*\|\geq 2^S\varepsilon_n\right) < \varepsilon$. In other words, we have shown that $\|\gamma_0(\tau_n^*,\cdot)-\tau_n^*\| = O_P(\varepsilon_n)$ for our choice of ε_n, and so, $CAL(\tau_n^*) = \|\gamma_0(\tau_n^*,\cdot)-\tau_n^*\|^2 = O_P(\varepsilon_n^2)$. The result follows from that the fact that $\varepsilon_n^2 \leq \ell^{-2/3} + \|(\pi_m-\pi_0)(\mu_m-\mu_0)\|^2$. □

PROOF OF THEOREM 4.7

Proof. By the definition of the pointwise median stated in Section 2.1, for each covariate value $w \in \mathcal{W}$, there exists some random index $j_n(w)$ such that $\tau_n^*(w) = \tau_{n,j_n(w)}^*(w)$. (We note here that this property may fail for other definitions of the median when k is even.) Thus, we have that $|\gamma_0(\tau_n^*,w)-\tau_n^*(w)| = |\gamma_0(\tau_{n,j_n(w)}^*,w)-\tau_{n,j_n(w)}^*(w)| \leq \sum_{s=1}^k |\gamma_0(\tau_{n,s}^*,w)-\tau_{n,s}^*(w)|$, and so,

$$\begin{aligned}\|\gamma_0(\tau_n^*,\cdot)-\tau_n^*\| \;&\leq\; \left\|\sum_{s=1}^k |\gamma_0(\tau_{n,s}^*,\cdot)-\tau_{n,s}^*|\right\| \;\leq\; \sum_{s=1}^k \left\|\gamma_0(\tau_{n,s}^*,\cdot)-\tau_{n,s}^*\right\| \\ &\leq\; \sqrt{k\sum_{s=1}^k \left\|\gamma_0(\tau_{n,s}^*,\cdot)-\tau_{n,s}^*\right\|^2}\,,\end{aligned}$$

where the final inequality follows from the Cauchy-Schwarz inequality. Squaring both sides gives that $\mathrm{CAL}(\tau_n^*) \leq k\sum_{s=1}^k \mathrm{CAL}(\tau_{n,s}^*)$, as desired. □

PROOF OF THEOREM 4.8

Proof. As before, we may write $\tau_n^* = \theta_n^* \circ \tau$ for some $\theta_n^* \in \mathcal{F}_{iso}$ that minimizes the empirical risk

$$R_n(\theta) : \theta \mapsto \sum_{i\in\mathcal{I}_\ell}\left[\chi_m(O_i)-\theta\circ\tau(W_i)\right]^2$$

over $\mathcal{F}_{iso}$. For any given $\theta \in \mathcal{F}_{iso}$, the one-sided path $\{\varepsilon \mapsto \theta_n^* + \varepsilon(\theta-\theta_n^*) : \varepsilon \in [0,1]\}$ through θ_n^* lies entirely in $\mathcal{F}_{iso}$ since $\mathcal{F}_{iso}$ is a convex space. Furthermore, we have that

$$-2\sum_{i\in\mathcal{I}_\ell}(\theta-\theta_n^*)\circ\tau(W_i)[\chi_m(O_i)-\theta_n^*\circ\tau(W_i)] = \lim_{\varepsilon\downarrow 0}\frac{R_n(\theta_n^*+\varepsilon(\theta-\theta_n^*))-R_n(\theta_n^*)}{\varepsilon} \geq 0 \tag{13}$$

for all $\theta \in \mathcal{F}_{iso}$. The oracle isotonic risk minimizer τ_0^* can be expressed as $\tau_0^* = \theta_0 \circ \tau$ where $\theta_0 := \mathrm{argmin}_{\theta\in\mathcal{F}_{iso}}\|\theta\circ\tau-\tau_0\|$. Taking $\theta = \theta_0$ in (13), we obtain the inequality

$$\sum_{i\in\mathcal{I}_\ell}[(\theta_0-\theta_n^*)\circ\tau(W_i)][\chi_m(O_i)-\theta_n^*\circ\tau(W_i)] \leq 0\,. \tag{14}$$

Rearranging terms and adding and subtracting $P_\ell\{[(\theta_0-\theta_n^*)\circ\tau](\chi_0)\}$ in the above inequality implies that $P_\ell\{[(\theta_0-\theta_n^*)\circ\tau](\chi_m-\chi_0)\} \leq P_\ell\{[(\theta_0-\theta_n^*)\circ\tau](\theta_n^*\circ\tau-\chi_0)\}$. Adding and subtracting $P\{[(\theta_0-\theta_n^*)\circ\tau](\theta_n^*\circ\tau-\chi_0)\}$ yields that

$$\begin{aligned}&P_\ell\{[(\theta_0-\theta_n^*)\circ\tau](\chi_m-\chi_0)\} - (P_\ell-P)\{[(\theta_0-\theta_n^*)\circ\tau](\theta_n^*\circ\tau-\chi_0)\}\\ &\leq\; P\{[(\theta_0-\theta_n^*)\circ\tau](\theta_n^*\circ\tau-\chi_0)\}\,.\end{aligned} \tag{15}$$

Next, adding and subtracting $P\{(\theta_0\circ\tau)[(\theta_0-\theta_n^*)\circ\tau]\}$, we have that

$$\begin{aligned}&P\{[(\theta_0-\theta_n^*)\circ\tau](\theta_n^*\circ\tau-\chi_0)\}\\ &=\; P\{[(\theta_0-\theta_n^*)\circ\tau][\theta_n^*\circ\tau-E[\chi_0(O)\,|\,W=\cdot]]\}\\ &=\; P\{[(\theta_0-\theta_n^*)\circ\tau](\theta_n^*\circ\tau-\tau_0)\}\\ &=\; P\{[(\theta_0-\theta_n^*)\circ\tau][(\theta_n^*-\theta_0)\circ\tau]\} + P\{[(\theta_0-\theta_n^*)\circ\tau](\theta_0\circ\tau-\tau_0)\}\\ &=\; P\{[(\theta_0-\theta_n^*)\circ\tau](\theta_0\circ\tau-\tau_0)\} - \|(\theta_0-\theta_n^*)\circ\tau\|^2\,,\end{aligned} \tag{16}$$

where we used the fact that $E[\chi_0(O) \,|\, W = w] = \tau_0(w)$. Next, we note that θ_0 minimizes the population risk function $\theta \mapsto E_P[\tau_0(W) - \theta \circ \tau(W)]^2$ over $\mathcal{F}_{iso}$. As a consequence, the same argument used to derive (14) can be used to obtain that $P\{[(\theta - \theta_0) \circ \tau](\tau_0 - \theta_0 \circ \tau)\} \leq 0$ for any $\theta \in \mathcal{F}_{iso}$. Taking $\theta = \theta_n^*$, we find that

$$P\{[(\theta_0 - \theta_n^*) \circ \tau](\theta_0 \circ \tau - \tau_0)\} \leq 0 \,. \tag{17}$$

Combining (16) and (17), we obtain that

$$P\{[(\theta_0 - \theta_n^*) \circ \tau](\theta_n^* \circ \tau - \chi_0)\} \leq - \left\|(\theta_0 - \theta_n^*) \circ \tau\right\|^2 . \tag{18}$$

Finally, combining (15) and (18), we obtain the following inequality

$$\left\|(\theta_0 - \theta_n^*) \circ \tau\right\|^2 \leq -P_\ell\{[(\theta_0 - \theta_n^*) \circ \tau](\chi_m - \chi_0)\} + (P_\ell - P)\{[(\theta_0 - \theta_n^*) \circ \tau](\theta_n^* \circ \tau - \chi_0)\} \,.$$

Adding and subtracting $P\{[(\theta_0 - \theta_n^*) \circ \tau](\chi_m - \chi_0)\}$ and noting that $\tau_0^* - \tau_n^* = (\theta_0 - \theta_n^*) \circ \tau$, we finally obtain the key inequality

$$\begin{aligned} \left\|\tau_0^* - \tau_n^*\right\|^2 \;\leq\; & P[(\tau_0^* - \tau_n^*)(\chi_0 - \chi_m)] + (P - P_\ell)[(\tau_0^* - \tau_n^*)(\chi_m - \chi_0)] \\ & + (P_\ell - P)[(\tau_0^* - \tau_n^*)(\tau_n^* - \chi_0)] \,. \end{aligned} \tag{19}$$

The above is similar to (7) in the proof of Theorem 4.6, and a similar proof technique is used to establish a convergence rate for τ_n^*. Specifically, we use the Cauchy-Schwarz inequality to bound the first term on the right-hand side of (19) in terms of $\|\tau_0^* - \tau_n^*\|$, and empirical process techniques to bound the remaining terms in terms of a function of $\|\tau_0^* - \tau_n^*\|$ with high probability. Using a similar approach as for the derivation of (10), we can upper-bound the first term of the right-hand side of (19) as $P[(\tau_0^* - \tau_n^*)(\chi_0 - \chi_m)] \leq \|\tau_0^* - \tau_n^*\| \|(\pi_m - \pi_0)(\mu_m - \mu_0)\|$. The second term in the right-hand side of (19) can be examined as follows. We let $\mathcal{F}_{4,m} := \{(\tau_1 - \tau_2)(\chi_m - \chi_0); \tau_1, \tau_2 \in \mathcal{F}_{\tau,iso}\}$, and define $Q := \sup_{o \in \mathcal{O}} \chi_0(o)$, which is finite in view of Conditions 4.1 and 4.2. Additionally, we let $R := Q + B$, and define for any fixed $\delta \in \mathbb{R}$

$$Z_{1,n}(\delta) := \sup_{\theta_1, \theta_2 \in \mathcal{F}_{iso} : \|(\theta_1 - \theta_2) \circ \tau\| \leq \delta R} (P - P_\ell)\{[(\theta_1 - \theta_2) \circ \tau](\chi_m - \chi_0)\} = \sup_{f \in \mathcal{F}_{4,m} : \|f\| \leq \delta R} (P - P_\ell) f.$$

Letting $\delta_{1,n} := \|\tau_0^* - \tau_n^*\|$, we have that $(P - P_\ell)[(\tau_0^* - \tau_n^*)(\chi_m - \chi_0)] \leq Z_{1,n}(\delta_{1,n})$. We note that $\mathcal{F}_{4,m}$ is a Lipschitz transformation of the function classes $\mathcal{F}_{\tau,iso}$ and $\mathcal{F}_{\tau,iso}$, and so, for every $\delta > 0$ that is deterministic conditional on $\mathcal{E}_m$, we have that

$$\psi_{1,n}(\delta \mid \mathcal{E}_m) := E[Z_{1,n}(\delta) \mid \mathcal{E}_m] \;\lessapprox\; \ell^{-1/2} \mathcal{J}(\delta, \mathcal{F}_{iso}) \left(1 + \frac{\mathcal{J}(\delta, \mathcal{F}_{iso})}{\sqrt{\ell}\delta^2}\right)$$

in view of Theorem 2.10.20 of (van der Vaart & Wellner, 1996) and the results outlined in Theorem 4.6, where the right-hand side can only be random through δ. Finally, the third term in (19) can be studied as follows. We let $\mathcal{F}_5 := \{(\tau_1 - \tau_2)(\tau_2 - \chi_0) : \tau_1, \tau_2 \in \mathcal{F}_{\tau,iso}\}$, and for any given $\delta > 0$, we define

$$Z_{2,n}(\delta) := \sup_{\theta_1, \theta_2 \in \mathcal{F}_{iso} : \|(\theta_1 - \theta_2) \circ \tau\| \leq \delta G} (P - P_\ell)\{[(\theta_1 - \theta_2) \circ \tau](\theta_2 - \chi_0)\} = \sup_{f \in \mathcal{F}_5 : \|f\| \leq \delta G} (P - P_\ell) f$$

with $G := Q + B$. We note that $\mathcal{F}_5$ is a Lipschitz transformation of $\mathcal{F}_{\tau,iso}$. Hence, similarly as above, for any $\delta > 0$ that is nonrandom conditional on $\mathcal{E}_m$, we have that

$$\psi_{2,n}(\delta \mid \mathcal{E}_m) := E[Z_{2,n}(\delta) \mid \mathcal{E}_m] \;\lessapprox\; \ell^{-1/2} \mathcal{J}(\delta, \mathcal{F}_{iso}) \left(1 + \frac{\mathcal{J}(\delta, \mathcal{F}_{iso})}{\sqrt{\ell}\delta^2}\right),$$

where the right-hand side can only berandom through δ. Defining $\epsilon_m^{nuis} := \|(\pi_m - \pi_0)(\mu_m - \mu_0)\|$, by a similar peeling argument as in Theorem 4.6, for any rate ε_n that is nonrandom conditional on $\mathcal{E}_m$, we can show that

$$\begin{aligned} P\left(\|\tau_0^* - \tau_n^*\| \geq 2^S \varepsilon_n\right) \;\leq\; & \sum_{s=S}^{\infty} E\left[\frac{2^{s+1}\varepsilon_n \epsilon_m^{nuis} + \psi_{1,n}(2^{s+1}\varepsilon_n \mid \mathcal{E}_m) + \psi_{2,n}(2^{s+1}\varepsilon_n \mid \mathcal{E}_m)}{2^{2s}\varepsilon_n^2}\right] \\ \lessapprox\; & \sum_{s=S}^{\infty} E\left[\frac{\epsilon_m^{nuis}}{2^{s-1}\varepsilon_n} + \frac{\mathcal{J}(2^{s+1}\varepsilon_n, \mathcal{F}_{iso})}{2^{2s}\sqrt{\ell}\varepsilon_n^2}\left(1 + \frac{\mathcal{J}(2^{s+1}\varepsilon_n, \mathcal{F}_{iso})}{\sqrt{\ell}2^{2s+1}\varepsilon_n^2}\right)\right] . \end{aligned}$$

Then, by the same arguments used in Theorem 4.6 and the same choice of $\mathcal{E}_m$-random ε_n, we can establish that $\|\tau_0^* - \tau_n^*\| = O_P(\ell^{-1/3}) + O_P(\|(\pi_m - \pi_0)(\mu_m - \mu_0)\|)$. By the triangle inequality and the fact that $\tau_0^* = \operatorname{argmin}_{\theta \circ \tau : \theta \in \mathcal{F}_{iso}} \|\tau_0 - \theta \circ \tau\|$ implies $\|\tau_0 - \tau_0^*\| \leq \|\tau_0 - \tau\|$, we find that $\|\tau_0 - \tau_n^*\| \leq \|\tau_0 - \tau_0^*\| + \|\tau_0^* - \tau_n^*\| \leq \|\tau_0 - \tau\| + \|\tau_0^* - \tau_n^*\|$. Combining these bounds, we find that $\|\tau_0 - \tau_n^*\| \leq \|\tau_0 - \tau\| + O_P(\ell^{-1/3}) + O_P(\|(\pi_m - \pi_0)(\mu_m - \mu_0)\|)$. □

C.4. Statement and proof of generalized Theorem 4.6 for random predictor

Here, we consider the same setup as Theorem 4.6 but allow τ_n^* to be obtained from a random predictor τ_m, as long as τ_m is built using only data in $\mathcal{E}_m$.

Condition C.4 (independence of predictor). The predictor $w \mapsto \tau_m(w)$ is independent of $\mathcal{C}_\ell$.

Theorem C.5 (Calibration with random predictors)**.** *Provided Conditions 4.1–C.4 hold, it holds that*

$$\mathrm{CAL}(\tau_n^*) = O_P\left(\ell^{-2/3} + \|(\pi_m - \pi_0)(\mu_m - \mu_0)\|^2\right).$$

Proof. Arguing exactly as in Theorem 4.6 with τ taken to be τ_m and conditioning on $\mathcal{E}_m$ as needed, we obtain the basic inequality stating that

$$\|\gamma_0(\tau_n^*, \cdot) - \tau_n^*\|^2 \leq P\{[\gamma_0(\tau_n^*, \cdot) - \tau_n^*](\chi_0 - \chi_m)\} + (P - P_\ell)\{[\gamma_0(\tau_n^*, \cdot) - \tau_n^*](\chi_m - \tau_n^*)\}$$

P-almost surely, where $\tau_n^* := \theta_n^* \circ \tau_m$. To establish the result of the theorem, we only need to make minor modifications to the proof of Theorem 4.6 to allow τ to be replaced by τ_m. We sketch those modifications here. A core component of the proof of Theorem 4.6 involved upper-bounding $E[S_n(\delta) \mid \mathcal{E}_m]$; this must now be done with $S_n(\delta)$ defined as

$$\sup_{\tau_1 \in \mathcal{F}_{\tau_m, TV}, \tau_2 \in \mathcal{F}_{\tau_m, iso} : \|\tau_1 - \tau_2\| \leq \delta} (P - P_\ell)[(\tau_1 - \tau_2)(\chi_m - \tau_2)] = \sup_{f \in \mathcal{F}_{Lip,m} : \|f\| \leq \delta K} (P - P_\ell) f$$

with τ_m now a random predictor. Previously, we showed that $E[S_n(\delta) \mid \mathcal{E}_m]$ can be bounded by a nonrandom constant depending on n, m and δ that is independent of $\mathcal{E}_m$. To do so, we showed that the random function class $\mathcal{F}_{Lip,m}$ is fixed conditional on $\mathcal{E}_m$, uniformly bounded, and has uniform entropy integral bounded by the uniform entropy integral of $\mathcal{F}_{iso}$. It suffices to show that this remains true when τ is replaced by τ_m. Since τ_m is obtained from $\mathcal{E}_m$, as with χ_m, the predictor τ_m is deterministic conditionally on $\mathcal{E}_m$. As a consequence, the function classes $\mathcal{F}_{\tau_m, TV}$ and $\mathcal{F}_{\tau_m, iso}$, which are now random through τ_m, are fixed conditional on $\mathcal{E}_m$. Since $\mathcal{F}_{Lip,m}$ is obtained from a Lipschitz transformation of elements of $\mathcal{F}_{\tau_m, TV}$ and $\mathcal{F}_{\tau_m, iso}$, we have that $\mathcal{F}_{Lip,m}$ is also fixed conditional on $\mathcal{E}_m$. Moreover, by the same argument as in the proof of Lemma C.2, which also holds for random τ, these function classes are uniformly bounded by a nonrandom constant almost surely. Finally, the preservation of the uniform entropy integral argument of the proof of Theorem 4.6 is valid with τ random. With these modifications to the proof of Theorem 4.6, the result follows. □

D. Simulation studies

D.1. Data-generating mechanisms

In simulation studies, data units were generated as follows for the two scenarios considered.

Scenario 1:

1. generate $W_1, W_2, \ldots, W_4$ independently from the uniform distribution on $(-1, +1)$;
2. given $(W_1, W_2, W_3, W_4) = (w_1, w_2, w_3, w_4)$, generate A as a Bernoulli random variable with success probability $\pi_0(w_1, w_2, w_3, w_4) := \mathrm{expit}\{-0.25 - w_1 + 0.5w_2 - w_3 + 0.5w_4\}$;
3. given $(W_1, W_2, W_3, W_4) = (w_1, w_2, w_3, w_4)$ and $A = a$, generate Y as a Bernoulli random variable with success probability $\mu_0(a, w_1, w_2, \ldots, w_4) := \mathrm{expit}\{1.5 + 1.5a + 2a|w_1||w_2| - 2.5(1-a)|w_2|w_3 + 2.5w_3 + 2.5(1-a)\sqrt{|w_4|} - 1.5aI(w_2 < 0.5) + 1.5(1-a)I(w_4 < 0)\}$.

Scenario 2:

- generate $W_1, W_2, \ldots, W_{20}$ independently from the uniform distribution on $(-1, +1)$;
- given $(W_1, W_2, \ldots, W_{20}) = (w_1, w_2, \ldots, w_{20})$, generate A as a Bernoulli random variable with success probability $\pi_0(w_1, w_2, \ldots, w_{20}) := \mathrm{expit}\{0.2 - 0.5w_1 - 0.5w_2 - 0.5w_3 + 0.5w_4 - 0.5w_5 + 0.5w_6 - 0.5w_7 - 0.5w_8 - 0.5w_9 - 0.2w_{10} + 0.5w_{11} - w_{12} + w_{13} - 1.5w_{14} + w_{15} - w_{16} + 2w_{17} - w_{18} + 1.5w_{19} - w_{20}\}$;

- given $(W_1, W_2, \ldots, W_{20}) = (w_1, w_2, \ldots, w_{20})$ and $A = a$, generate Y as a normal random variable with mean $\mu_0(a, w_1, w_2, \ldots, w_{20}) = -0.5 + 3.5a + 3aw_1 + 6.5(1-a)w_2 + 1.5aw_3 + 4(1-a)w_4 + 2.5aw_5 - 6(1-a)w_6 + 1aw_7 + 4.5(1-a)w_8 + aw_9 + 2.5(1-a)w_{10} + 1.5w_{11} - 2.5w_{12} + w_{13} - 1.5w_{14} + 3w_{15} - 2w_16 + 3w_{17} - w_{18} + 1.5w_{19} - 2w_{20}$ and unit variance.

Coefficients of the propensity score logistic regression models above were selected such that the probabilities of treatment were bounded between 0.05 and 0.95 in the low-dimensional case (Scenario 1), and between 0.01 and 0.99 in the high-dimensional setting (Scenario 2).

D.2. Implementation of the causal isotonic calibrator

In our simulation studies, we followed Algorithm 3 to fit the causal isotonic calibrator. In particular, we estimated the components of χ_0 (i.e., μ_0 and π_0) using the Super Learner (van der Laan et al., 2007) in Scenario 1, and penalized regression in Scenario 2. Super learner is an ensemble learning approach that uses cross-validation to select a convex combination of a library of candidate prediction methods. Table 1 shows the library of prediction models we used to estimate μ_0 and π_0. Note that all of our models for the outcome regression were misspecified in Scenario 1 because of the nonlinearities in the true outcome regression. However, in both scenarios, the propensity score estimator was a consistent estimator of the true propensity score. Additionally, for numerical stability, we imposed a threshold on the estimated propensity scores such that it took values between 0.01 and 0.99. We used the R package sl3 (Coyle et al., 2021) to implement the estimation procedure. Finally, we used the R function isoreg to performed the isotonic regression step.

Table 1. Information on the set of estimators used by the Super Learner to estimate the pseudo-outcome components. Abbreviations: generalized additive models (GAM), generalized linear model (GLM), generalized linear model with lasso regularization (GLMnet), gradient boosted trees (GBRT), random forests (RF), multivariate adaptive regression splines (MARS).

scenario	library for μ_0	library for π_0
1	logistic regression, GLMnet, GAM, GBRT with depth $\in \{2, 3, 5, 6, 8\}$, RF, MARS	logistic regression, GLMnet, GAM, GBRT with depth $\in \{2, 4, 6\}$
2	GLMnet	GLMnet

D.3. Performance metrics

We estimated the performance metrics as follows. With a slight abuse of notation, let $\hat\tau$ denote an arbitrary estimated treatment effect predictor or its calibrated version. For each fitted $\hat\tau$ in a given simulation, we computed its mean squared error by taking the empirical mean of the squared difference between the fitted values of the CATE estimator and τ_0,

$$\widehat{\text{MSE}}(\hat\tau) := \frac{1}{n_{\mathcal{V}}} \sum_{i: w_i \in \mathcal{V}} [\hat\tau(w_i) - \tau_0(w_i)]^2.$$

We obtained the estimated calibration measure in two steps. We recall that the calibration measure for a given predictor τ is

$$\int [\gamma_0(\tau, w) - \tau(w)]^2 \, dP_W(w) \, .$$

First, we estimated $\gamma_0(\hat\tau, w)$ using an independent dataset of 100,000 observations and fitted gradient boosted regression trees with the fitted values of the treatment effect predictors as covariates and the true CATE as outcome. For each simulation setting and CATE estimator, the depths of each of the regression trees were obtained using cross-validation in a separate simulation. Let $\hat\gamma_0(\hat\tau, w)$ denote the estimated function. In the second step, we used the sample $\mathcal{V}$ to estimate the calibration measure as

$$\widehat{\text{CAL}}(\tau) := \frac{1}{n_{\mathcal{V}}} \sum_{i: w_i \in \mathcal{V}} [\tau_0(w_i) - \hat\tau(w_i)] \, [\hat\gamma_0(\hat\tau, w_i) - \hat\tau(w_i)] \, .$$

The above measure has the advantage of having less bias with respect to $\text{CAL}(\hat\tau)$ than the plug-in estimator $n_{\mathcal{V}}^{-1} \sum_{i: w_i \in \mathcal{V}} [\hat\gamma_0(\hat\tau, w_i) - \hat\tau(w_i)]^2$.

E. Simulation results

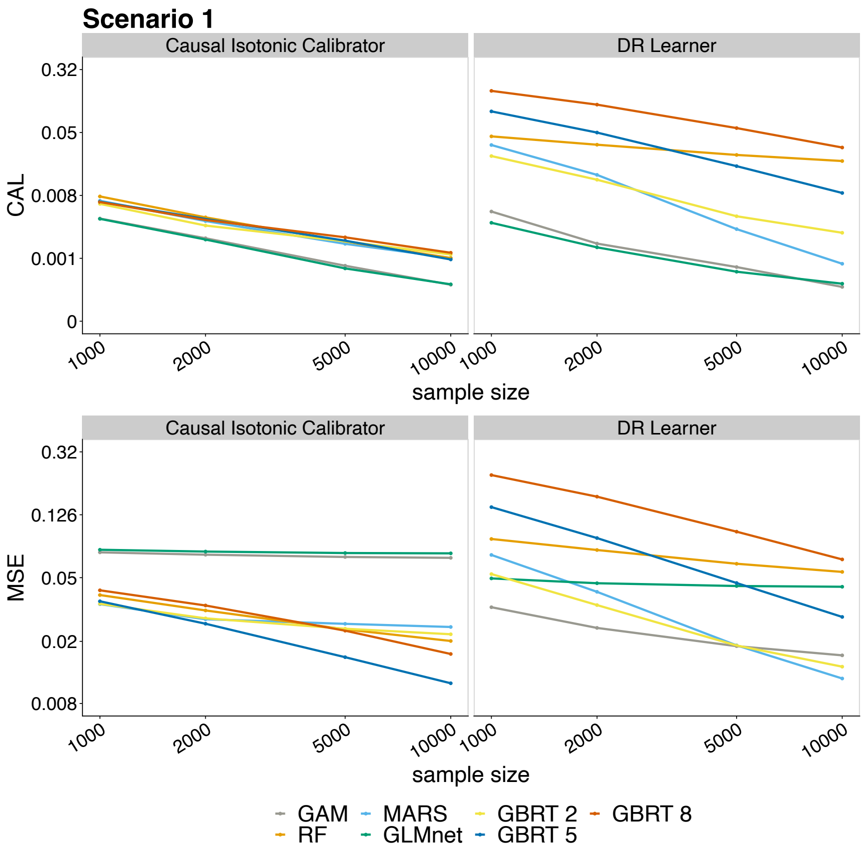
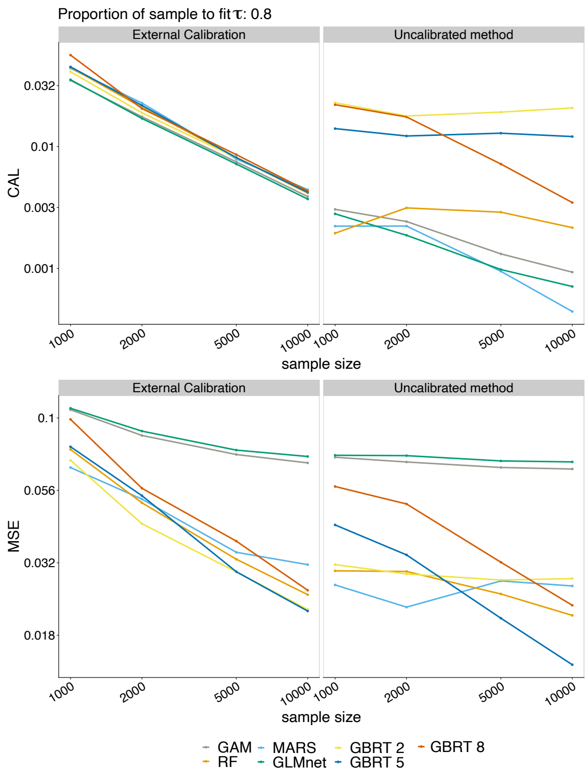
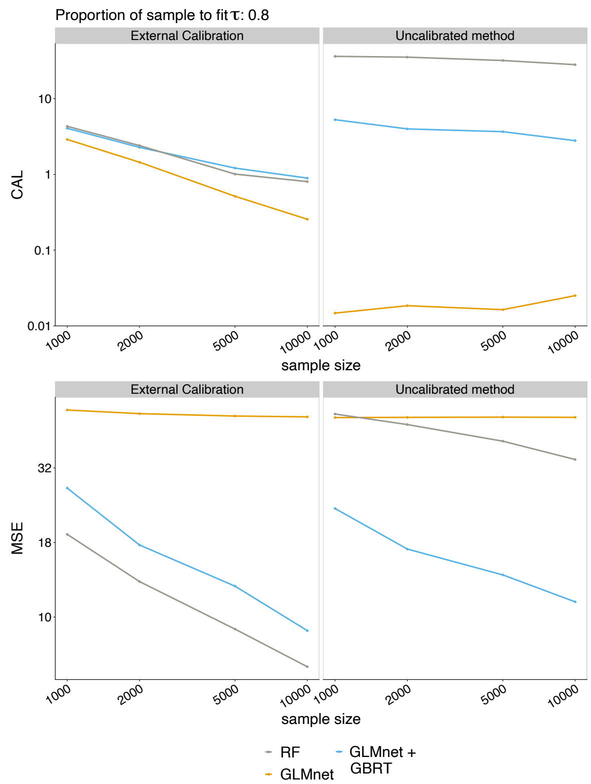


Figure 3. Calibration error and MSE in Scenario 1. The panels show the calibration error (top) and MSE (bottom) using the calibrated (left) and uncalibrated Double Robust Learner (right) predictors as a function of sample size. The y-axes are on a log scale.



(a) Scenario 1 calibration measure and MSE simulation results for causal calibration approach with an external hold-out dataset. The top left and right panels show the calibration measure and using conventional calibration and the uncalibrated estimator, respectively. Similarly, the bottom plots show MSE for the calibrated and uncalibrated estimators. Results for GLM and GBRT with depths of 3 and 6 are omitted because they were nearly identical to results shown for GLMnet and GBRT with other depths, respectively.



(b) Scenario 2 calibration measure and MSE simulation results for causal calibration approach with hold-out dataset. The top left and right panels show the calibration error using conventional calibration and the uncalibrated estimator, respectively. Similarly, the bottom plots show the MSE for the calibrated and uncalibrated estimators.

Figure 4. Causal isotonic calibration with a hold-out dataset external to the training dataset: Monte-Carlo estimates of calibration measure and MSE for calibrated vs uncalibrated predictors for Scenarios 1 and 2.

Table 2. Scenario 1 bias within bins of predictions for the calibrated and uncalibrated estimators. Each row shows the resulting bias for a given CATE estimator, and the Cal column indicates if it is calibrated or not. The columns are organized by sample size, and within each sample size, we show the results for the bias in the upper and lower deciles. Abbreviations: calibrated (cal), estimator (est), generalized additive models (GAM), generalized linear model (GLM), generalized linear model with lasso regularization (GLMnet), gradient boosted regression trees (GBRT), random forests (RF), multivariate adaptive regression splines (MARS).

Sample Size		1000		2000		5000		10000	
Cal	CATE estimator	Lower Decile	Upper Decile	Lower Decile	Upper Decile	Lower Decile	Upper Decile	Lower Decile	Upper Decile
yes	MARS	-0.01	-0.02	0	-0.01	0	-0.01	-0.02	0.02
no	MARS	-0.02	-0.03	-0.01	-0.02	-0.02	-0.01	-0.05	0.03
yes	GAM	-0.02	0.01	0	0.01	0	0.03	-0.01	0.05
no	GAM	0.02	-0.06	0.03	-0.07	0.02	-0.05	0.02	-0.04
yes	GLM	-0.01	0.02	-0.01	0.02	0	0.05	-0.01	0.06
no	GLM	0.02	-0.01	0.02	-0.02	0.02	-0.01	0.02	-0.01
yes	GLMnet	-0.01	0.02	-0.01	0.02	0	0.05	-0.01	0.06
no	GLMnet	0.02	-0.02	0.02	-0.02	0.03	-0.01	0.03	-0.01
yes	RF	-0.01	0	0	0	-0.03	0.03	-0.04	0.04
no	RF	-0.09	0.04	-0.08	0.05	-0.08	0.04	-0.06	0.03
yes	GBRT 2	-0.01	0	0	-0.01	-0.01	0.01	0	0.02
no	GBRT 2	0.1	-0.16	0.11	-0.16	0.12	-0.15	0.13	-0.14
yes	GBRT 3	-0.02	-0.02	0	-0.02	-0.01	0	-0.02	0.01
no	GBRT 3	0.02	-0.14	0.02	-0.14	0.03	-0.1	0.02	-0.08
yes	GBRT 5	-0.01	-0.02	0	-0.01	0	0	-0.01	0
no	GBRT 5	-0.04	-0.04	-0.01	-0.06	-0.07	0.01	-0.11	0.05
yes	GBRT 6	0	-0.03	0.01	-0.02	0	-0.01	-0.01	0
no	GBRT 6	-0.07	0	-0.04	-0.03	-0.11	0.06	-0.16	0.1
yes	GBRT 8	0.01	-0.04	0.02	-0.03	0	-0.01	-0.01	-0.01
no	GBRT 8	-0.14	0.08	-0.1	0.04	-0.19	0.14	-0.22	0.17

Table 3. Scenario 2 bias within bins of predictions for the calibrated and uncalibrated estimators. Each row shows the resulting bias for a given CATE estimator, and the Cal column indicates if it is calibrated or not. The columns are organized by sample size, and within each sample size, we show the results for the bias in the upper and lower deciles. Abbreviations: calibrated (cal), generalized linear model with lasso regularization (GLMnet), gradient boosted regression trees with GLMNet screening (GLMNet scr + GBRT).

Sample Size		1000		2000		5000		10000	
Cal	CATE estimator	Lower Decile	Upper Decile	Lower Decile	Upper Decile	Lower Decile	Upper Decile	Lower Decile	Upper Decile
yes	GLMnet	0	0	0	0	0.01	0.01	0.01	0.01
no	GLMnet	0.18	0.19	0.2	0.18	0.15	0.16	0.14	0.12
yes	GLMnet scr + GBRT	-0.23	-0.06	-0.18	-0.06	-0.33	-0.08	-0.37	-0.1
no	GLMnet scr + GBRT	-0.36	-0.16	-0.34	-0.15	-0.4	-0.2	-0.35	-0.22
yes	random forest	-0.09	-0.03	-0.06	-0.03	-0.14	-0.04	-0.21	-0.04
no	random forest	-0.9	-0.75	-0.86	-0.7	-0.95	-0.82	-0.98	-0.87