Appendix A: Proofs for asymptotic properties for the estimator

In this section, we will prove Theorem 1, to show the asymptotically linear expansion of $\hat{\theta}_Q$. Recalling the definition of θ_Q (20), denote by D_Q and N_Q the denominator and the numerator

$$D_Q = \mathbb{E}\left[\frac{T}{\pi^\star(X)}\lambda(\pi^\star(X),\alpha^\star)\right] \quad \text{and} \quad N_Q = \mathbb{E}\left[L\frac{T}{\pi^\star(X)}\lambda(\pi^\star(X),\alpha^\star)\right].$$

Observe that the numerator and denominator of θ_Q (and also $\hat{\theta}_Q$) are nearly identical: we can think of the denominator D_Q (and its empirical plug-in $\widehat{D}_Q$) as a special case of the numerator N_Q (and also $\hat{N}_Q$) but with $L := 1$. Below, we focus on asymptotic expansions for N_Q without loss of generality.

The following lemma shows it suffices to find the asymptotically linear expansion of the numerator and denominator of $\hat{\theta}_Q$ separately. We defer its proof to Appendix B.1.

LEMMA EC.1. *Assume that* $A_n \xrightarrow{p} A$, $B_n \xrightarrow{p} B$, $\sqrt{n}(A_n - A) = \frac{1}{\sqrt{n}}\sum_{i=1}^n a_i + o_P(1)$ *and* $\sqrt{n}(B_n - B) = \frac{1}{\sqrt{n}}\sum_{i=1}^n b_i + o_P(1)$. *Additionally, assume that* $B > 0$. *Then,*

$$\sqrt{n}\left(\frac{A_n}{B_n} - \frac{A}{B}\right) = \frac{1}{\sqrt{n}}\sum_{i=1}^n\left(\frac{1}{B}a_i - \frac{A}{B^2}b_i\right) + o_P(1).$$

Because of Lemma EC.1, it suffices to prove the following proposition; we give its proof in the rest of this section. Let $\widehat{N}_Q$ be the empirical plug-in for N_Q

$$\widehat{N}_Q := \sum_{i=1}^n l_i \frac{t_i}{\hat{\pi}(x_i)}\lambda(\hat{\pi}(x_i),\hat{\alpha}). \tag{EC.1}$$

PROPOSITION EC.1. $\sqrt{n}(\hat{N}_Q - N_Q) = \frac{1}{\sqrt{n}}\sum_{i=1}^n \psi_{N_Q}(w_i) + o_P(1)$, *where* $\psi_{N_Q}(w) = g(w) + h(w) + \delta(w)$ *and*

$$\begin{aligned}
g(w) &= l\frac{t}{\pi^*(x)}\lambda(\pi^*(x),\alpha^*) - N_Q \\
h(w) &= \mathbb{E}\left[L\frac{T}{\pi^*(X)}\frac{\partial}{\partial\alpha}\lambda(\pi^*(X),\alpha)\Big|_{\alpha=\alpha^*}\right](t-\alpha^*) \\
\delta(w) &= \mu_Q(x)\pi^*(x)(t-\pi^*(x))\frac{\partial}{\partial\pi}\left[\frac{\lambda(\pi,\alpha^*)}{\pi}\right]\Big|_{\pi=\pi^*(x)}.
\end{aligned}$$

To begin, recall that $\pi(x) = \gamma_2(x)/\gamma_1(x)$ and define

$$g(w,\gamma,\alpha) = l\frac{t}{\pi(x)}\lambda(\pi(x),\alpha) - N_Q \tag{EC.2}$$

so that

$$\widehat{N}_Q - N_Q = n^{-1} \sum_{i=1}^{n} g(w_i, \hat{\gamma}, \hat{\alpha}).$$

Since g is continuously differentiable with respect to α, expand $\hat{\alpha}$ around α^* as

$$g(w_i, \hat{\gamma}, \hat{\alpha}) = \nabla_\alpha g(w_i, \hat{\gamma}, \overline{\alpha})(\hat{\alpha} - \alpha^*) + g(w_i, \hat{\gamma}, \alpha^*)$$

where $\overline{\alpha}$ is a mean value. By Lemma EC.2 below, $n^{-1} \sum_{i=1}^{n} \nabla_\alpha g(w, \hat{\gamma}, \alpha) \xrightarrow{p} \mathbb{E}[\nabla_\alpha g(w, \gamma^*, \alpha)]$.

LEMMA EC.2. *Under the assumptions and definitions of Theorem 1,*

$$n^{-1} \sum_{i=1}^{n} \nabla_\alpha g(w, \hat{\gamma}, \overline{\alpha}) \xrightarrow{p} \mathbb{E}[\nabla_\alpha g(w, \gamma^*, \alpha^*)] =: G_\alpha$$

Proof of this lemma is deferred to Appendix B.2. Noting that $\hat{\alpha} = n^{-1} \sum_{i=1}^{n} t_i$ and letting

$$h(w) := \mathbb{E}[\nabla_\alpha g(w, \gamma^*, \alpha^*)](t - \alpha^*) = \mathbb{E}\left[L \frac{T}{\pi^*(X)} \nabla_\alpha \lambda(\pi^*(X), \alpha^*) \right] (t - \alpha^*),$$

we arrive at

$$\sum_{i=1}^{n} g(w_i, \hat{\gamma}, \hat{\alpha}) / \sqrt{n} = \sum_{i=1}^{n} [h(w_i) + g(w_i, \hat{\gamma}, \alpha^*)] / \sqrt{n} + o_P(1). \tag{EC.3}$$

To find an asymptotically linear representation of $\sum_{i=1}^{n} g(w_i, \hat{\gamma}, \alpha^*)/\sqrt{n}$, we will show that under the assumptions made in Theorem 1, $\sum_{i=1}^{n} g(w_i, \hat{\gamma}, \alpha^*)/\sqrt{n}$ satisfies the assumptions of Newey and McFadden (1994, Theorem 8.11), which we restate now for convenience. Note that while the theorem as stated in Newey and McFadden (1994) only shows asymptotic normality, in the proof they show asymptotic linearity, so we have modified the result to state the asymptotic linearity result, instead. We have also modified the result for the case where g is linear in the estimated parameter, in which case consistency of the estimated parameter does not need to be shown (Newey and McFadden 1994).

ASSUMPTION EC.1 **(N+M Assumptions 8.1-8.3 and Assumption from Theorem 8.11)**. *Let d be the dimension of X.*

1. *$K(u)$ is differentiable of order p, the derivatives of order p are bounded, $K(u)$ is zero outside of a bounded set, $\int K(u)du = 1$, there is a positive integer m such that for all $j < m$, $\int K(u)[\bigotimes_{l=1}^{j} u]du = 0$.*
2. *There is a version of $\gamma^*(x)$ that is continuously differentiable to order p with bounded derivatives on an open set containing $\mathcal{X}$.*

3. *There is* $r \geq 4$ *such that* $\mathbb{E}[\|A\|^r] < \infty$ *and* $\mathbb{E}[\|A\|^r | X = x] m_X(x)$ *is bounded.*
4. *The bandwidth* $\sigma = \sigma(n)$ *satisfies* $n\sigma^{2d+4p}/(\ln n)^2 \to \infty$ *and* $n\sigma^{2m} \to 0$.

ASSUMPTION EC.2 **(N+M Assumptions from Theorem 8.11)**. *Let* β *be the parameter of interest, and* $\hat{\beta}$ *its estimator where* $\hat{\beta} - \beta = \frac{1}{n}\sum_{i=1}^n g(W_i, \hat{\gamma})$.

1. $\mathbb{E}[g(W, \gamma^*)] = 0$
2. $\mathbb{E}[\|g(W, \gamma^*)\|^2] < \infty$
3. $\mathcal{X}$ *is a compact set.*

ASSUMPTION EC.3 **(N+M enumerated Assumptions from Theorem 8.11)**. *There is a vector of functionals* $G(w, \gamma)$ *that is linear in* γ *such that*

(i) for $\|\gamma - \gamma^*\|$ *small where the norm is the Sobolev norm (*$\|\gamma\| := \max_{\ell \leq p} \sup_{x \in \mathcal{X}} \|\partial^\ell \gamma(x)/\partial^\ell x\|$*)* , $\|g(w, \gamma) - g(w, \gamma^*) - G(w, \gamma - \gamma^*)\| \leq b(w)\|\gamma - \gamma^*\|^2$, *and* $\mathbb{E}[b(W)] < \infty$*;*

(ii) $\|G(w, \gamma)\| \leq c(w)\|\gamma\|$ *and* $\mathbb{E}[c(W)^2] < \infty$*;*

(iii) there is $v(x)$ *with* $\int G(w, \gamma) dM(w) = \int v(x)\gamma(x)dx$ *for all* $\|\gamma\| < \infty$*;*

(iv) $v(x)$ *is continuous almost everywhere,* $\int \|v(x)\| dx < \infty$*, and there is* $\epsilon > 0$ *such that* $\mathbb{E}[\sup_{\|\nu\| \leq \epsilon} \|v(X + \nu)\|^4] < \infty$.

THEOREM EC.1 **(N+M Theorem 8.11)**. *Let* γ^* *be the nuisance parameter and* $\hat{\gamma}$ *its kernel density estimate satisfying Assumption EC.1. Let* β *be the parameter of interest, and* $\hat{\beta}$ *its estimator satisfying Assumption EC.2. Assume there is a vector of functionals* $G(w, \gamma)$ *satisfying Assumption EC.3. Then for* $\delta(w) = v(x)a - \mathbb{E}[v(x)a]$,

$$\sum_{i=1}^n g(w_i, \hat{\gamma})/\sqrt{n} = \sum_{i=1}^n [g(w_i, \gamma^*) + \delta(w_i)]/\sqrt{n} + o_P(1).$$

We now proceed to use this result to prove the asymptotically linear representation

$$\sum_{i=1}^n g(w_i, \hat{\gamma}, \alpha^*)/\sqrt{n} = \sum_{i=1}^n [g(w_i, \gamma^*, \alpha^*) + \delta(w_i)]/\sqrt{n} + o_P(1) \tag{EC.4}$$

for our choice of $g(w, \gamma, \alpha)$. Then, the desired result follows from combining Equations (EC.4) and (EC.3), and then applying Lemma EC.1. What remains is to check the conditions of Theorem EC.1.

Verifying Assumption EC.1:

1. Assumed in Assumption 1
2. Assumed in Assumption 2

3. There is $r \geq 4$ such that $\mathbb{E}[\|A\|^r] < \infty$ and $\mathbb{E}[\|A\|^r | X = x] m_X(x)$ is bounded: recall that $A = [1, T]$ and T takes values in $\{0, 1\}$. Then let $r = 4$, $\|A\|^r \leq 2$, so that $\mathbb{E}[\|A\|^r] \leq 2 < \infty$, and and $\mathbb{E}[\|A\|^r \mid X = x] m_X(x) \leq 2B_{mU} < \infty$ by Assumption 2.
4. Assumed in Assumption 1

Verifying Assumption EC.2:

1. $\mathbb{E}[g(W, \gamma^*, \alpha^*)] = 0$: this holds by definition of g (Equation (EC.2)) and N_Q (Equation (EC.1)).
2. $\mathbb{E}[|g(W, \gamma^*, \alpha^*)|^2] < \infty$:

$$\mathbb{E}[|g(W, \gamma^*, \alpha^*)|^2] = \mathbb{E}\left[\left| L \frac{T}{\pi^*(X)} \lambda(\pi^*(X), \alpha^*) - N_Q \right|^2\right].$$

   This is finite since $N_Q < \infty$, and then by Cauchy-Schwarz since $\mathbb{E}[L^4] < B_{L^4}$ by Assumption 3, and $\left|\frac{T}{\pi^*(X)} \lambda(\pi^*(X), \alpha^*)\right|^4$ is bounded, since $T \in \{0, 1\}$ and

$$\frac{\lambda(\pi^*(X), \alpha^*)}{\pi^*(X)} = \frac{1 - \pi^*(X)}{(1 - \alpha^*)\pi^*(X) + \alpha^*(1 - \pi^*(X))} \leq \frac{1 - \delta_\pi}{\delta_\pi}$$

   with δ_π from Assumption 2.
3. Assumed in Assumption 2.

Verifying Assumption EC.3:

(i) For $\|\gamma - \gamma^*\|$ small (where the norm on $\|\gamma - \gamma^*\|$ is the Sobolev norm), $\left|g(w, \gamma, \alpha^*) - g(w, \gamma^*, \alpha^*) - G(w, \gamma - \gamma^*)\right| \leq b(w)\|\gamma - \gamma^*\|^2$, and $\mathbb{E}[b(W)] < \infty$: By Taylor-expanding g around $\gamma^*(x)$ (where in the following equations we abuse notation and also use g to mean $g(w, \gamma(x), \alpha)$, where the second argument is the value of the function γ evaluated at x, rather than the function γ. We also write $\nabla_\gamma g(\cdot)$ to denote the derivative of this new g with respect to its second argument, the value $\gamma(x)$, rather than to the function $\gamma(\cdot)$, and similarly for $\nabla^2_\gamma g(\cdot)$),

$$\begin{aligned} g(w, \gamma, \alpha^*) = g(w, \gamma^*(x), \alpha^*) + \nabla_\gamma g(w, \gamma^*(x), \alpha^*)^\top (\gamma(x) - \gamma^*(x)) \\ + \frac{1}{2}(\gamma(x) - \gamma^*(x))^\top \nabla^2_\gamma g(w, \overline{m}, \alpha^*)(\gamma(x) - \gamma^*(x)) \end{aligned}$$

where $\overline{m}$ is a mean value on the line between $\gamma(x)$ and $\gamma^*(x)$. Note that g is twice differentiable with respect to $\gamma(x)$ in an open set containing this line (since $\|\gamma - \gamma^*\|$ small, so that π is bounded away from 0 and 1) so that the expansion holds. Thus let $G(w, \gamma - \gamma^*)$ be the first-order term in the expansion above:

$$G(w, \gamma) := \nabla_\gamma g(w, \gamma(x), \alpha^*)^\top \gamma(x)$$

$$= lt \frac{\partial}{\partial \pi} \left(\frac{\lambda(\pi, \alpha^*)}{\pi} \right) \Big|_{\pi = \pi^*(x)} \nabla_\gamma \pi(\gamma^*(x))^\top \gamma(x)$$
$$= lt \frac{\partial}{\partial \pi} \left(\frac{\lambda(\pi, \alpha^*)}{\pi} \right) \Big|_{\pi = \pi^*(x)} m_X(x)^{-1} [-\pi^*(x), 1] \gamma(x)$$

where we used $\pi(\gamma(x)) = \gamma_2(x)/\gamma_1(x)$, $\gamma_1(x) = m_X(x)$ is the marginal density of X, and

$$\frac{\partial}{\partial \pi} \left(\frac{\lambda(\pi, \alpha^*)}{\pi} \right) \Big|_{\pi = \pi^*(x)} = -\frac{1 - \alpha^*}{((1-\alpha^*)\pi^*(x) + \alpha^*(1 - \pi^*(x)))^2}.$$

Then to verify the assumption,

$$\begin{aligned}
&\left| g(w, \gamma, \alpha^*) - g(w, \gamma^*, \alpha^*) - G(w, \gamma - \gamma^*) \right| \\
&= \frac{1}{2} \left| (\gamma(x) - \gamma^*(x))^\top \nabla_\gamma^2 g(w, \overline{m}, \alpha^*) (\gamma(x) - \gamma^*(x)) \right| \\
&\leq \frac{1}{2} \left\| \nabla_\gamma^2 g(w, \overline{m}, \alpha^*) \right\|_F \|\gamma - \gamma^*\|^2.
\end{aligned}$$

Thus to verify the assumption, let $b(w) = \frac{1}{2} \left\| \nabla_\gamma^2 g(w, \overline{m}, \alpha^*) \right\|_F$. Some simple calculus shows that $\nabla_\gamma^2 g(w, \overline{m}, \alpha^*) = lt \phi(\overline{m})$ where

$$\phi(\overline{m}) = \frac{\partial^2}{\partial \pi^2} \left(\frac{\lambda(\pi, \alpha^*)}{\pi} \right) \Big|_{\pi = \overline{\pi}} \nabla_\gamma \pi(\overline{m}) \nabla_\gamma \pi(\overline{m})^\top + \frac{\partial}{\partial \pi} \left(\frac{\lambda(\pi, \alpha^*)}{\pi} \right) \Big|_{\pi = \overline{\pi}} \nabla_\gamma^2 \pi(\overline{m}).$$

Then $\mathbb{E}[\|\phi(\overline{m})\|_F^2]$ is bounded: since $\|\gamma - \gamma^*\|$ is small, $\overline{m}$ is close to $\gamma^*(x)$, so that $\overline{m}_1$ is bounded away from 0, and $\overline{\pi}$ is bounded away from both 0 and 1, so that each term in $\phi(\overline{m})$ is bounded, so that $\mathbb{E}[\|\phi(\overline{m})\|_F^2]$ is bounded. Then, applying Cauchy-Schwarz, $\mathbb{E}[b(W)] \leq \frac{1}{2} \sqrt{\mathbb{E}[(LT)^2] \mathbb{E}[\|\phi(\overline{m})\|_F^2]} < \infty$ as $\mathbb{E}[(LT)^2] \leq \mathbb{E}[L^2] \leq \mathbb{E}[L^4] < B_{L^4}$ by Assumption 3.

(ii) $|G(w, \gamma)| \leq c(w) \|\gamma\|$ and $\mathbb{E}[c(w)^2] < \infty$:

$$\begin{aligned}
|G(w, \gamma)| &= lt \left| \frac{\partial}{\partial \pi} \left(\frac{\lambda(\pi, \alpha^*)}{\pi} \right) \Big|_{\pi = \pi^*(x)} m_X(x)^{-1} [-\pi^*(x), 1] \gamma(x) \right| \\
&\leq lt \left| \frac{\partial}{\partial \pi} \left(\frac{\lambda(\pi, \alpha^*)}{\pi} \right) \Big|_{\pi = \pi^*(x)} \right| m_X(x)^{-1} \|[-\pi^*(x), 1]\| \|\gamma(x)\| \\
&\leq lt \left| \frac{\partial}{\partial \pi} \left(\frac{\lambda(\pi, \alpha^*)}{\pi} \right) \Big|_{\pi = \pi^*(x)} \right| m_X(x)^{-1} \sup_{x \in \mathcal{X}} \left\{ \|[-\pi^*(x), 1]\| \right\} \|\gamma\| \\
&\leq ltC \|\gamma\|
\end{aligned}$$

for some constant C since $\left| \frac{\partial}{\partial \pi} \left(\frac{\lambda(\pi, \alpha^*)}{\pi} \right) \Big|_{\pi = \pi^*(x)} \right|$ is bounded as in the previous part, and $m_X(x)^{-1}, \pi^*(x)$ are bounded by Assumption 2. Here, the norm $\|\gamma\|$ is the Sobolev norm while $\|\gamma(x)\|$ is the Euclidean norm. Thus let $c(w) = ltC$, and $\mathbb{E}[c(W)^2] = C^2 \mathbb{E}[L^2 T^2] \leq C^2 \sqrt{\mathbb{E}[(LT)^4]} \leq C^2 \sqrt{\mathbb{E}[L^4]} \leq C^2 \sqrt{B_{L^4}} < \infty$ by Assumption 3.

(iii) There is $v(x)$ with $\int G(w,\gamma)dM(w) = \int v(x)\gamma(x)dx$:

Define $v(x)$ by rewriting $\int G(w,\gamma)dM(w)$:

$$\begin{aligned}\int G(w,\gamma)dM(w) &= \int lt\frac{\partial}{\partial\pi}\left[\frac{\lambda(\pi,\alpha^*)}{\pi}\right]\Big|_{\pi=\pi^*(x)} m_X(x)^{-1}[-\pi^*(x),1]\gamma(x)dM(w)\\ &= \int \mu_Q(x)\pi^*(x)\frac{\partial}{\partial\pi}\left[\frac{\lambda(\pi,\alpha^*)}{\pi}\right]\Big|_{\pi=\pi^*(x)}[-\pi^*(x),1]\gamma(x)dx\end{aligned}$$

where the last equality is essentially obtained by using iterated expectations to rewrite

$$\mathbb{E}[LT\xi(X)] = \mathbb{E}[\mathbb{E}[L \mid T=1,X]\mathbb{E}[T \mid X]\xi(X)]$$

since $T \in \{0,1\}$, for $\xi(x) = \frac{\partial}{\partial\pi}\left[\frac{\lambda(\pi,\alpha^*)}{\pi}\right]\Big|_{\pi=\pi^*(x)}[-\pi^*(x),1]\gamma(x)$ so that

$$v(x) = \mu_Q(x)\pi^*(x)\frac{\partial}{\partial\pi}\left[\frac{\lambda(\pi,\alpha^*)}{\pi}\right]\Big|_{\pi=\pi^*(x)}[-\pi^*(x),1].$$

(iv) $v(x)$ is continuous almost everywhere, $\int \|v(x)\|dx < \infty$, and there is $\epsilon > 0$ such that $\mathbb{E}[\sup_{\|\nu\|\le\epsilon}\|v(X+\nu)\|^4] < \infty$:

$v(x)$ is continuous almost everywhere since it is the product of functions that are continuous almost everywhere: $\pi^*(x)$ and $\mu_Q(x)$ are continuous almost everywhere by Assumptions 2 and 3, and $\frac{\partial}{\partial\pi}\left[\frac{\lambda(\pi,\alpha^*)}{\pi}\right]\Big|_{\pi=\pi^*(x)}$ is also continuous in x.

$\int \|v(x)\|dx < \infty$: $\|v(x)\|$ is bounded on $\mathcal{X}$ since it is the product of several terms that are each bounded on $\mathcal{X}$: $\mu_Q(x)$ is bounded by Assumption 3, $\pi^*(x)$ is bounded by Assumption 2, and $\frac{\partial}{\partial\pi}\left[\frac{\lambda(\pi,\alpha^*)}{\pi}\right]\Big|_{\pi=\pi^*(x)}$ is bounded since $\delta_\pi < \pi^*(x) < 1-\delta_\pi$ for $\delta_\pi > 0$ from in Assumption 2. The integral is finite since $\mathcal{X}$ is compact.

The sup condition is satisfied since $\|v(x)\|$ is bounded in $\mathcal{X}$.

Now that we have satisfied the conditions of Theorem EC.1 and we have $v(x)$, let $a = [1,t]^\top$ and define

$$\begin{aligned}\delta(w) &= v(x)a - \mathbb{E}[v(X)A]\\ &= \mu_Q(x)(t-\pi^*(x))\pi^*(x)\frac{\partial}{\partial\pi}\left[\frac{\lambda(\pi,\alpha^*)}{\pi}\right]\Big|_{\pi=\pi^*(x)}.\end{aligned}$$

Then by Theorem EC.1,

$$\sum_{i=1}^n g(w_i,\hat{\gamma},\alpha^*)/\sqrt{n} = \sum_{i=1}^n [g(w_i,\gamma^*,\alpha^*)+\delta(w_i)]/\sqrt{n} + o_P(1)$$

as desired.

Appendix B: More proofs for asymptotic properties of the estimator

B.1. Proof of Lemma EC.1

Define $r(a,b) = a/b$. It is continuously differentiable, so we can apply the mean value theorem to its derivative, so that for some choices of values $\overline{A}, \overline{B}$ between A_n and A, and B_n and B, respectively,

$$\begin{aligned}\sqrt{n}\left(\frac{A_n}{B_n} - \frac{A}{B}\right) &= \sqrt{n}(r(A_n,B_n) - r(A,B)) \\ &= \sqrt{n}\nabla r(\overline{A},\overline{B})^\top \begin{bmatrix} A_n - A \\ B_n - B \end{bmatrix} \\ &= [1/\overline{B}, -\overline{A}/\overline{B}^2] \begin{bmatrix} \sum_{i=1}^n a_i/\sqrt{n} + o_P(1) \\ \sum_{i=1}^n b_i/\sqrt{n} + o_P(1) \end{bmatrix}.\end{aligned}$$

Since $A_n \xrightarrow{p} A$ and $B_n \xrightarrow{p} B$, we have that $\overline{A} \xrightarrow{p} A$ and $\overline{B} \xrightarrow{p} B$, so

$$\sqrt{n}\left(\frac{A_n}{B_n} - \frac{A}{B}\right) = \frac{1}{\sqrt{n}}\sum_{i=1}^n \left(\frac{1}{B}a_i - \frac{A}{B^2}b_i\right) + o_P(1).$$

B.2. Proof of Lemma EC.2

We will use the following from Newey and McFadden (1994):

LEMMA EC.3 (**Direct consequence of N+M Lemma 8.10**). *If Assumption EC.1 is satisfied, then* $\sqrt{n}\|\hat{\gamma} - \gamma^*\|^2 \xrightarrow{p} 0$.

Since we are assuming Assumption EC.1 for showing Theorem 1, the conclusion holds.

Now, since $\frac{1}{n}\sum_{i=1}^n \nabla_\alpha g(w,\gamma^*,\alpha^*) \xrightarrow{p} \mathbb{E}[\nabla_\alpha g(w,\gamma^*,\alpha^*)]$ by law of large numbers, it suffices to show

$$\frac{1}{n}\sum_{i=1}^n \left[\nabla_\alpha g(w_i,\hat{\gamma},\overline{\alpha}) - \nabla_\alpha g(w_i,\gamma^*,\alpha^*)\right] \xrightarrow{p} 0.$$

Using the Markov inequality, it suffices to show that for $\|\hat{\gamma} - \gamma^*\|$ small enough,

$$\left|\nabla_\alpha g(w,\hat{\gamma},\overline{\alpha}) - \nabla_\alpha g(w,\gamma^*,\alpha^*)\right| \le b(w)\left[\|\hat{\gamma} - \gamma^*\| + \|\hat{\alpha} - \alpha^*\|\right]$$

for some $b(w)$ such that $\mathbb{E}[b(W)] < \infty$, since $\|\hat{\gamma} - \gamma^*\| \xrightarrow{p} 0$ and $\|\hat{\alpha} - \alpha^*\| \xrightarrow{p} 0$. Similar to verifying Assumption EC.3 (where again we abuse notation to write g on the RHS to take $\gamma(x)$ as the second argument, rather than $\gamma(\cdot)$, and we write ∇_γ to denote derivatives with respect to $\gamma(x)$ instead of $\gamma(\cdot)$),

$$\nabla_\alpha g(w,\hat{\gamma},\overline{\alpha}) - \nabla_\alpha g(w,\gamma^*,\alpha^*) = \nabla_{\gamma,\alpha}\nabla_\alpha g(w,\overline{m},\overline{\overline{\alpha}})^\top \begin{bmatrix} \hat{\gamma}(x) - \gamma^*(x) \\ \overline{\alpha} - \alpha^* \end{bmatrix}$$

where $(\overline{m}, \overline{\overline{\alpha}})$ is a mean value on the line between $(\hat{\gamma}(x), \overline{\alpha})$ and $(\gamma^*(x), \alpha^*)$. Then, note that $\hat{\gamma}(x) - \gamma^*(x) \leq \|\hat{\gamma} - \gamma^*\|$. Let

$$b(w) := \sup_{\overline{m}, \overline{\overline{\alpha}}} \|\nabla_{\gamma,\alpha} \nabla_\alpha g(w, \overline{m}, \overline{\overline{\alpha}})\|_\infty,$$

so that

$$\left|\nabla_\alpha g(w, \hat{\gamma}, \overline{\alpha}) - \nabla_\alpha g(w, \gamma^*, \alpha^*)\right| \leq b(w) \left[\|\hat{\gamma} - \gamma^*\| + \|\hat{\alpha} - \alpha^*\|\right].$$

Finally, $\mathbb{E}[b(W)] < \infty$ by an argument similar to the verification of Assumption EC.3.

Appendix C: Proofs for efficiency

We verify that our estimator in Theorem 1 attains the nonparametric efficiency bound, i.e. that our estimator has the best possible asymptotic variance. Instead of rigorously deriving the nonparametric efficiency bound, for brevity, we use heuristics from (Kennedy 2022) to obtain the efficient influence function of the estimand. To show the desired result, we then show the efficient influence function of the estimand is the same as the influence function of the estimator in Theorem 1.

C.1. Efficient influence function

We calculate the efficient influence function for the estimand using heuristics from (Kennedy 2022). To simplify notation, we first deal with only the numerator of θ_Q, which we denoted as

$$N_Q = \mathbb{E}_M \left[\ell(f(X), Y) \frac{T}{\pi^*(X)} \lambda(\pi^*(X), \alpha^*)\right] = \mathbb{E}_M \left[\mu_Q(X) \lambda(\pi^*(X), \alpha^*)\right].$$

To calculate the efficient influence function, as in (Kennedy 2022), we treat $\mathcal{X}$ as discrete, and use derivative rules with simple influence functions as building blocks. For notational simplicity, we omit *'s for π^*, α^* for the calculation of efficient influence functions.

$$\begin{aligned}
\mathbb{IF}\{\alpha\} &= T - \alpha \\
\mathbb{IF}\{p(x)\} &= \mathbf{1}\{X = x\} - p(x) \\
\mathbb{IF}\{p(t)\} &= \mathbf{1}\{T = t\} - p(t) \\
\mathbb{IF}\{\pi(x)\} &= \frac{\mathbf{1}\{X = x\}}{p(x)}(T - \pi(x)) \\
\mathbb{IF}\{\mu_Q(x)\} &= \frac{\mathbf{1}\{X = x, T = 1\}}{\mathbb{P}(X = x, T = 1)}(L - \mu_Q(x)) = \frac{T\mathbf{1}\{X = x\}}{p(x)\pi(x)}(L - \mu_Q(x)) \\
\mathbb{IF}\{\lambda(\pi(x), \alpha)\} &= \nabla_\pi \lambda(\pi(x), \alpha) \mathbb{IF}\{\pi(x)\} + \nabla_\alpha \lambda(\pi(x), \alpha) \mathbb{IF}\{\alpha\} \\
&= \nabla_\pi \lambda(\pi(x), \alpha) \frac{\mathbf{1}\{X = x\}}{p(x)}(T - \pi(x)) + \nabla_\alpha \lambda(\pi(x), \alpha)(T - \alpha)
\end{aligned}$$

Then

$$\begin{aligned}
\mathbb{IF}\{N_Q\} &= \sum_{x\in\mathcal{X}} \mathbb{IF}\{\mu_Q(x)\lambda(\pi(x),\alpha)p(x)\} \\
&= \sum_{x\in\mathcal{X}} \left(\mathbb{IF}\{\mu_Q(x)\}\lambda(\pi(x),\alpha)p(x) + \mu_Q(x)\mathbb{IF}\{\lambda(\pi(x),\alpha)\}p(x) + \mu_Q(x)\lambda(\pi(x),\alpha)\mathbb{IF}\{p(x)\}\right) \\
&= \sum_{x\in\mathcal{X}} \frac{T\mathbf{1}\{X=x\}}{p(x)\pi(x)}(L-\mu_Q(x))\lambda(\pi(x),\alpha)p(x) \\
&\quad + \sum_{x\in\mathcal{X}} \mu_Q(x)\nabla_\pi\lambda(\pi,\alpha)\frac{\mathbf{1}\{X=x\}}{p(x)}(T-\pi(x))p(x) \\
&\quad + \sum_{x\in\mathcal{X}} \mu_Q(x)\nabla_\alpha\lambda(\pi(x),\alpha)(T-\alpha)p(x) \\
&\quad + \sum_{x\in\mathcal{X}} \mu_Q(x)\lambda(\pi(x),\alpha)(\mathbf{1}\{X=x\}-p(x)) \\
&= (L-\mu_Q(X))\frac{T}{\pi(X)}\lambda(\pi(X),\alpha) + \mu_Q(X)(T-\pi(X))\nabla_\pi\lambda(\pi,\alpha) \\
&\quad + (T-\alpha)\mathbb{E}\left[\mu_Q(x)\nabla_\alpha\lambda(\pi(x),\alpha)\right] + \mu_Q(X)\lambda(\pi(X),\alpha) - N_Q.
\end{aligned}$$

C.2. Comparison

Now we show the efficient influence function from the previous section is the same as the influence function of the estimator in Theorem 1. We do so by first comparing the efficient influence function of the numerator of θ_Q, $\mathbb{IF}(N_Q)$, with the influence function of the estimator for the numerator of θ_Q, $\psi_{N1}(w)$. Then we do the same for the denominator, then θ_Q, then θ_P, then the terms in the decomposition (1). We start with the numerator of θ_Q, N_Q. Recall that from Proposition EC.1,

$$\sqrt{n}(\widehat{N}_Q - N_Q) = \sum_{i=1}^{n} \psi_{N1}(w_i)/\sqrt{n} + o_P(1)$$

where $\psi_{N1}(w) = g(w) + h(w) + \delta(w)$ with

$$\begin{aligned}
g(w) &= l\frac{t}{\pi^*(x)}\lambda(\pi^*(x),\alpha^*) - N_Q \\
h(w) &= \mathbb{E}\left[L\frac{T}{\pi^*(X)}\nabla_\alpha\lambda(\pi^*(X),\alpha^*)\right](t-\alpha^*) \\
\delta(w) &= \mu_Q(x)\pi^*(x)(t-\pi^*(x))\frac{\partial}{\partial\pi}\left[\frac{\lambda(\pi,\alpha^*)}{\pi}\right]\Big|_{\pi=\pi^*(x)}.
\end{aligned}$$

Note that we can also express $\delta(w)$ as

$$\delta(w) = \mu_Q(x)(t-\pi^*(x))\left(\nabla_\pi\lambda(\pi^*(x),\alpha^*) - \frac{\lambda(\pi^*(x),\alpha^*)}{\pi^*(x)}\right)$$

$$= \mu_Q(x)\left(\lambda(\pi^*(x),\alpha^*) - \frac{t}{\pi^*(x)}\lambda(\pi^*(x),\alpha^*) + (t-\pi^*(x))\nabla_\pi\lambda(\pi^*(x),\alpha^*)\right).$$

It is clear that $\psi_{N1}(W)$ is the same as $\mathbb{IF}\{N_Q\}$ from the previous section. Then

$$\sqrt{n}(\widehat{N}_Q - N_Q) = \frac{1}{\sqrt{n}}\sum_{i=1}^n \psi_{N1}(w_i) + o_P(1) \overset{d}{\rightsquigarrow} N(0, \mathrm{Var}(\mathbb{IF}\{N_Q\}))$$

where $\psi_{N1}(w) := g(w) + h(w) + \delta(w)$ as before. An analogous argument applies to the denominator D_Q, as D_Q is the same as N_Q but with L replaced by 1. Then similar to before,

$$\sqrt{n}(\widehat{D}_Q - D_Q) = \frac{1}{\sqrt{n}}\sum_{i=1}^n \psi_{D1}(w_i) + o_P(1) \overset{d}{\rightsquigarrow} N(0, \mathrm{Var}(\mathbb{IF}\{D_Q\})).$$

Then to see that

$$\sqrt{n}\left(\frac{\widehat{N}_Q}{\widehat{D}_Q} - \frac{N_Q}{D_Q}\right) = \frac{1}{\sqrt{n}}\sum_{i=1}^n \left(\frac{1}{D_Q}\psi_{N1}(w_i) - \frac{N_Q}{D_Q^2}\psi_{D1}(w_i)\right) \overset{d}{\rightsquigarrow} N(0, \mathrm{Var}(\mathbb{IF}\{N_Q/D_Q\})),$$

note that $\mathbb{IF}\{N_Q/D_Q\}$ is composed of $\mathbb{IF}\{N_Q\}$ and $\mathbb{IF}\{D_Q\}$ the same way that ψ_1 is composed of ψ_{N1} and ψ_{D_Q}:

$$\mathbb{IF}\{N_Q/D_Q\} = \frac{D_Q\mathbb{IF}\{N_Q\} - N_Q\mathbb{IF}\{D_Q\}}{D_Q^2} = [1/D_Q, -N_Q/D_Q^2]\begin{bmatrix}\mathbb{IF}\{N_Q\}\\ \mathbb{IF}\{D_Q\}\end{bmatrix}$$

which is the same as for $\hat{\theta}_Q = \widehat{N}_Q/\widehat{D}_Q$ as in Lemma EC.1:

$$\sqrt{n}(\widehat{N}_Q/\widehat{D}_Q - N_Q/D_Q) = [1/D_Q, -N_Q/D_Q^2]\begin{bmatrix}\sum_{i=1}^n \psi_{N1}(w_i)/\sqrt{n} + o_P(1)\\ \sum_{i=1}^n \psi_{D1}(w_i)/\sqrt{n} + o_P(1)\end{bmatrix}.$$

The results for the decomposition terms in Equation (1) follow similarly.

Appendix D: Additional details for estimation algorithm

D.1. Data splits and cross-fitting

As is standard practice, performance for the model $f(\cdot)$ is not evaluated on training data so that we measure loss degradation only from distribution shift, rather than also from overfitting. Thus, the data from training distribution P has separate training and validation splits, the evaluated model f is trained on the training split, and evaluated on the validation split. The evaluated model is also evaluated on the data from target distribution Q (which consists of only one split). We measure the change in performance on Q vs on the validation split of P.

The practice of evaluating models on data separate from the data they were trained on also applies to the domain classifier classifier $\hat{\pi}(x)$. One natural option that uses all of the available data is called cross-fitting (Zheng and van der Laan 2011, Chernozhukov et al. 2018, Yadlowsky 2022) in which the data are first split into K folds. Then, for each fold k, the domain classifier is learned on the union of all the other folds, and finally evaluated on fold k.

There are various ways to set up the data splits. We use two different ways of splitting data in Sections 4.1 and 4.2 and describe them in Appendix F.1.2 and F.3.2, respectively.

D.2. Data re-use and overfitting to the test set

It's possible that adjusting a model in response to its performance on the test set (e.g. by adding new covariates, or by modifying the training procedure to reweight training data, etc) may affect the validity of the inferences for model performance, especially if the test set is reused aggressively.

Practical ways to deal with data re-use (e.g. collecting new test data for evaluation, holding out test data for future evaluations, etc), are an interesting direction for future work. At the same time, to our understanding, it is common to reuse existing training and test data in settings where data are limited (which is a setting we focus on in our examples), and we expect the effect of reusing data on inferences for test loss to be negligible in most reasonable use cases.

D.3. Practical considerations for learning the domain classifier

To estimate $\hat{\pi}(x) = \mathbb{P}(T = 1|X = x)$, we train classifier on samples from P_X and Q_X using logistic loss, since the true $\pi(x)$ minimizes the expected logistic loss, and furthermore minimizers of the logistic loss produce values in $[0, 1]$, in contrast to squared loss.

To perform model checking, since the auxiliary domain classifier is analogous to a propensity score in causal inference, the usual checks on propensity scores can be used to check validity of the domain classifier. These include checks for balance, overlap, and calibration (Imbens and Rubin 2015, Gutman et al. 2022). It is also useful to check moments: for example, since $\mathbb{E}[\pi(X)] = \mathbb{P}(T = 1) = \mathbb{E}[T]$, one should confirm that the sample mean of $\hat{\pi}(X)$ is also close to the sample mean of T.

D.4. Practical recommendations in settings with low overlap between P_X, Q_X

Note that our decomposition in Section 2 relies on the existence of a shared distribution S_X. If there are no values of X over which $P_X > 0$ and $Q_X > 0$, then there are no values over which to compare $P_{Y|X}$ and $Q_{Y|X}$. In practice, one way to assess whether this problem is occurring is to look at the classifier probabilities of a classifier classifying between P_X and Q_X: if such a classifier distinguishes between P_X and Q_X correctly and with high confidence, then that may indicate a

lack of values of X that are shared between P_X and Q_X. Because of this, choice of covariates X is critical: for example, you would not want to include a covariate that is a dataset identifier that takes on one value for P and a separate value for Q.

Note that while the decomposition only requires the existence of S_X, in practice, a limited shared distribution over X may result in estimated decomposition terms with higher variance. Recall from Algorithm 1, we estimate $\mathbb{E}_{S_X}[R_P(X)], \mathbb{E}_{S_X}[R_Q(X)]$ as weighted sums:

$$\mathbb{E}_{S_X}[R_P(X)] \approx \frac{\sum_{i=1}^{n_P} \ell(f(X_i), Y_i) w_P(\widehat{\pi}(X_i), \widehat{\alpha})}{\sum_{i=1}^{n_P} w_P(\widehat{\pi}(X_i), \widehat{\alpha})}$$
$$\mathbb{E}_{S_X}[R_Q(X)] \approx \frac{\sum_{j=1}^{n_Q} \ell(f(X_j), Y_j) w_Q(\widehat{\pi}(X_j), \widehat{\alpha})}{\sum_{j=1}^{n_Q} w_Q(\widehat{\pi}(X_j), \widehat{\alpha})}.$$

In an extreme case of limited overlap, if weights w_P are 0 except for a small subset of X's, and w_Q are also 0 except for a small subset of X's, then $\mathbb{E}_{S_X}[R_P(X)], \mathbb{E}_{S_X}[R_Q(X)]$ are each weighted means over small subsets of P, Q; these means can have high variance due to having a small (effective) sample size.

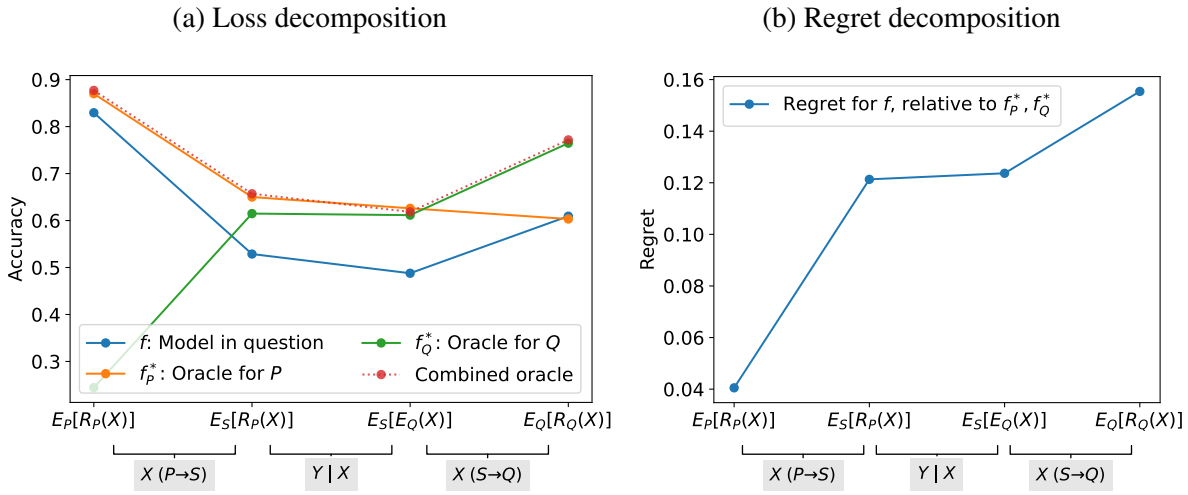
Appendix E: Regret example

In this section, we provide a case study in which we look at regret. The task is the same as in Section 4.1, in which we predict whether an individual is employed (Y) based on their (tabular) census data (X), using the dataset from Ding et al. (2021).

Here, we consider a model f that was trained on the full population from 2018, and we would like to understand why the model's performance is different between two chosen subpopulations, which happen to correspond to the group of people of ages ≤ 25 (P) and the group of people of ages ≥ 18 (Q). In practice, P and Q could be any two subpopulations of interest with covariate overlap (so that S_X exists). We aim to understand this difference in model performance between P and Q by using the regret decomposition in Section 2.4. We train f_P^* on P and f_Q^* on Q, and obtain the loss decompositions from Section 2 for f, f_P^*, f_Q^* in Figure EC.1a. We also use Section 2.4 to obtain regret decompositions, which we display in Figure EC.1b. The regret decomposition for the "combined oracle" in Figure EC.1a consists of the terms $\mathbb{E}_P[R_P^*(X)], \mathbb{E}_S[R_P^*(X)], \mathbb{E}_S[R_Q^*(X)], \mathbb{E}_Q[R_Q^*(X)]$.

Observe that by construction, the shared distribution S_X should approximately consist of people of ages 18 through 25. Observe that the loss decomposition for f has a U-shape, indicating worse performance due to X shift for $P \to S$, but better performance due to X shift for $S \to Q$. This tells us that f performs worse on the S, compared to both P and Q. This is expected: people under the

Figure EC.1 Loss and regret decomposition in Section E



age of 18 are generally not employed (so that employment is easy to predict), and it may be difficult to predict whether people of age 18-25 are employed as they may or may not be in school.

In contrast, the regret decomposition tells us that the performance of f worsens relative to the oracle (f_P^*) over the X shift for $P \to S$, and also worsens relative to the oracle (f_Q^*) over the X shift for $S \to Q$. Here, as we use accuracy, which can also be thought of as 1 minus squared loss, if we additionally treat f_P^* and f_Q^* as $\mathbb{E}_P[Y \mid X]$ and $\mathbb{E}_Q[Y \mid X]$, then as discussed in Section 2.4, regret captures the degree to which f moves away from oracles f_P^* and f_Q^*, rather than also capturing changes in irreducible error. Thus, this regret decomposition tells us that f has moved away from f_P^* over the X shift for $P \to S$, and also away from f_Q^* over the X shift for $S \to Q$.

See additional experiment details in Appendix F.2.

Appendix F: Additional details for experiments

F.1. Adult dataset experiment details

F.1.1. Models We use random forest classifiers using default settings from the `sklearn` python package to fit both the employment model and the domain classifier $\hat{\pi}(x)$. The employment model has parameter `max_depth` set to 2.

F.1.2. Data splits Each of train and target datasets are split into an 80%-20% split. The model to be evaluated is trained on the 80% split of the train dataset. The domain classifier is trained and the decomposition is evaluated on the 20% split of the train dataset and all of the target dataset, using cross-fitting with 3 splits as described in Appendix D.1. This way, both the model to be evaluated and the domain classifier are evaluated on data on which they are not trained. Note that after the models are adjusted, the data on which the models are evaluated remain the same (Appendix D.2).

F.1.3. Additional data processing The X features are all of those in the Adult dataset unless otherwise specified, but with some of them discretized and made into a binary encoding:

- SCHL (educational attainment) was made into the following binary outcomes:
  - —Whether someone finished high school
  - —Whether someone finished college
  - —Whether someone finished a post-grad degree
- MIL (military) was made into the following:
  - —Whether someone is active military
  - —Whether someone is a veteran
- CIT (citizenship) was made into the following:
  - —Whether someone is born in the US
  - —Whether someone is born in a US territory
  - —Whether someone has American parents
  - —Whether someone is naturalized
  - —Whether someone is not a citizen
- MIG (mobility) was made into the following:
  - —Whether someone moved residence
- MAR (marriage) was made into the following:
  - —Whether someone is married

F.2. Regret experiments (Appendix E)

Details are the same as in Appendix F.1 unless otherwise specified.

F.2.1. Models The models here are trained the same as in Appendix F.1, except that f_P^*, f_Q^* are also random forest classifiers, but with `max_depth` set to 6, so that they are better oracles.

F.2.2. Data splits We use an 80%-20% split on each of P and Q. f is trained on the 80% of both P and Q. f_P^* is trained on the 80% of P, f_Q^* is trained on the 80% of Q. Then $\hat{\pi}$ is trained on the remaining 20% of both P and Q, wiith cross-fitting with 3 splits. The decomposition is also evaluated on these 20% of P and Q, with these same splits.

F.3. FMoW-wilds experiment details

F.3.1. Models We use a pre-trained DANN neural network model downloaded from the WILDS leaderboard, corresponding to seed 0. The domain classifiers are logistic regressions on top of last layer features from the pre-trained neural network model, with logistic regressions trained using

`sklearn` (Pedregosa et al. 2011). Reported bootstrap standard deviations in Figure 6 are calculated by fixing the DANN neural network model, bootstrap resampling data for the decomposition, and re-calculating the decomposition for each resampled dataset.

F.3.2. Data splits The DANN neural network models being evaluated have been trained on the `training` data split, and also using the covariates from the `test-unlabeled` data split, which should be drawn from the same distribution as the `test` data split. The dataset also includes `id_test` and `id_val` splits, which should be drawn from the same distribution as `train`. Similar to Appendix F.1.2, we only want to evaluate models (the DANN neural network, and also the domain classifier) on datasets on which they were not trained, so we perform the decomposition on the union of `id_test` and `id_val` as P, and `test` as Q, and also use cross-fitting.

Appendix G: Shared space with lowest worst-case variance, assuming known $\pi(x), \alpha$

For a given P_X and Q_X, one might wonder how to choose a shared S_X so that the estimator $\hat{\theta}_Q - \hat{\theta}_P$ from Equation 17 attributing changes in performance to $Y \mid X$ shift has the lowest variance, for the worst-case $P_{Y|X}, Q_{Y|X}, f$, and ℓ for bounded ℓ. We use notation similar to Section 3 as follows:

Consider P_X with density p and Q_X with density q, and let $\mathbb{P}$ denote a mixture of P and Q, where a random variable $T \in \{0,1\}$ denotes when a data point is drawn from Q, rather than P. Let $\alpha = \mathbb{P}[T=1]$ (the proportion of the mixture from Q) so that $\mathbb{P}_X$ has density $\alpha q(x) + (1-\alpha)p(x)$. Let m denote the marginal distribution of X in $\mathbb{P}_X$. Let S_X be the shared distribution over X, with density s.

Recall that our estimator for $Y \mid X$ shift from Equation 17 is

$$\hat{\theta}_Q - \hat{\theta}_P = \frac{1}{N_Q}\sum_{i=1}^{N_Q} \ell(Y_i, f(X_i)) w_Q(X_i) - \frac{1}{N_P}\sum_{i=1}^{N_P} \ell(Y_i, f(X_i)) w_P(X_i) \tag{EC.6}$$

where the first term is a sum over samples from Q, the second term is a sum over samples from P, and

$$w_P(x) := \frac{s(x)}{p(x)} \quad \text{and} \quad w_Q(x) := \frac{s(x)}{q(x)}.$$

Given P_X, Q_X, we want to find the best S_X (via w_P, w_Q) in terms of variance for the estimator in (EC.6). We consider the worst case $P_{Y|X}, Q_{Y|X}, f, \ell$ for bounded ℓ, and also assume that the weights w_P, w_Q are perfectly known (via perfectly known π, α). We obtain the following best S_X in Theorem EC.2.

THEOREM EC.2. *For fixed P_X, Q_X, the shared distribution S_X producing the lowest variance for the estimator in (EC.6), in the worst case over $P_{Y|X}, Q_{Y|X}, f, \ell$ for bounded ℓ, and assuming the weights w_P, w_Q are perfectly known (via perfectly known π, α), i.e. the solution to*

$$\inf_{S_X} \sup_{\substack{P_{Y|X}, Q_{Y|X}, f, \\ \ell \text{ bounded}}} \alpha \mathrm{Var}_Q[\ell(Y, f(X)) w_Q(X)] + (1-\alpha)\mathrm{Var}_P[\ell(Y, f(X)) w_P(X)],$$

has density

$$s(x) \propto \frac{p(x)q(x)}{\alpha^3 q(x) + (1-\alpha)^3 p(x)}. \tag{EC.7}$$

Note that this S_X is similar to Equation 2 with density $\propto \frac{p(x)q(x)}{p(x)+q(x)}$ (they are identical when $\alpha = 0.5$), but here the denominator has a stronger dependence on α. We do not recommend this shared space distribution: see Appendix H where we empirically compare different shared space distributions.

Proof Without loss of generality, if we assume that ℓ is bounded between $[-1, 1]$, then this quantity above is upper bounded by

$$\inf_{S_X} \alpha \mathrm{Var}_Q[Z w_Q(X)] + (1-\alpha)\mathrm{Var}_P[Z w_P(X)]$$

where $\ell(Y, f(X))$ has been replaced by Z, where $Z \overset{iid}{\sim} 2\text{Bern}(0.5) - 1$ is drawn independently between P and Q. This upper bound is attained when, for example, $\ell(y_1, y_2)$ is 1 if $y_1 = y_2$ and -1 otherwise, $f(x) := 0$ for all x, and $Y \overset{iid}{\sim} \text{Bern}(0.5)$ independently for both P and Q, so that Z has the distribution above.

We can re-express the objective above as

$$\inf_{S_X} \alpha \mathbb{E}_Q[w_Q(X)^2] + (1-\alpha)\mathbb{E}_P[w_P(X)^2].$$

To proceed, we rewrite the above as an expectation over $\mathbb{P}$, where we let m denote the marginal density of X in $\mathbb{P}_X$:

$$\inf_{S_X} \mathbb{E}\left[w_Q(X)^2 \alpha \frac{q(X)}{m(X)} + w_P(X)^2 (1-\alpha) \frac{p(X)}{m(X)} \right].$$

Now recall that $w_Q(x) = \frac{s(x)}{q(x)}$ and $w_P(x) = \frac{s(x)}{p(x)}$ to rewrite the above as

$$\inf_{S_X} \mathbb{E}\left[\left(\frac{s(X)}{q(X)}\right)^2 \alpha \frac{q(X)}{m(X)} + \left(\frac{s(X)}{p(X)}\right)^2 (1-\alpha) \frac{p(X)}{m(X)} \right].$$

Next, we will then rewrite the objective above using

$$\pi(x) := \mathbb{P}[T = 1 \mid X = x] = \frac{\alpha q(x)}{m(x)} = \frac{\alpha q(x)}{\alpha q(x) + (1-\alpha)p(x)}$$

so that

$$\pi(x) = \frac{\alpha q(x)}{m(x)}, \quad 1 - \pi(x) = \frac{(1-\alpha)p(x)}{m(x)}.$$

Now we rewrite the objective from before as follows:

$$\begin{aligned}
&\inf_{S_X} \ \mathbb{E}\left[\left(\frac{s(X)}{q(X)}\right)^2 \pi(X) + \left(\frac{s(X)}{p(X)}\right)^2 (1-\pi(X))\right] \\
&= \inf_{S_X} \ \mathbb{E}\left[\left(\frac{s(X)}{m(X)}\right)^2 \left(\frac{m(X)}{q(X)}\right)^2 \pi(X) + \left(\frac{s(X)}{m(X)}\right)^2 \left(\frac{m(X)}{p(X)}\right)^2 (1-\pi(X))\right] \\
&= \inf_{S_X} \ \mathbb{E}\left[\left(\frac{s(X)}{m(X)}\right)^2 \left(\frac{1}{\alpha^2\pi(X)} + \frac{1}{(1-\alpha)^2(1-\pi(X))}\right)\right] \\
&= \inf_{S_X} \ \mathbb{E}\left[\left(\frac{s(X)}{m(X)}\right)^2 \left(\frac{\alpha^2\pi(X) + (1-\alpha^2)(1-\pi(X))}{\alpha^2\pi(X)(1-\alpha)^2(1-\pi(X))}\right)\right].
\end{aligned} \tag{EC.8}$$

Now we can use Cauchy-Schwarz:

$$\mathbb{E}[C/D] \le \mathbb{E}[C^2]\mathbb{E}[1/D^2]$$

where we let

$$\begin{aligned}
C &:= \frac{s(X)}{m(X)} \sqrt{\frac{\alpha^2\pi(X) + (1-\alpha^2)(1-\pi(X))}{\alpha^2\pi(X)(1-\alpha)^2(1-\pi(X))}} \\
D &:= \sqrt{\frac{\alpha^2\pi(X) + (1-\alpha^2)(1-\pi(X))}{\alpha^2\pi(X)(1-\alpha)^2(1-\pi(X))}}
\end{aligned}$$

and note that Equation EC.8 is $\mathbb{E}[C^2]$. Note also that equality holds in Cauchy-Schwarz when $C^2 \propto 1/D^2$, so that the objective in Equation EC.8, which is also $\mathbb{E}[C^2]$, is minimized when

$$\frac{s(X)}{m(X)} \propto \frac{\alpha^2\pi(X)(1-\alpha)^2(1-\pi(X))}{\alpha^2\pi(X) + (1-\alpha^2)(1-\pi(X))}.$$

Move $\pi(x)$ to the RHS, substitute $\pi(x) = \alpha\frac{q(x)}{m(x)}$ and $1-\pi(x) = (1-\alpha)\frac{p(x)}{m(x)}$, and then rearrange to obtain

$$s(x) \propto \frac{\alpha^2\pi(x)(1-\alpha)^2(1-\pi(x))}{\alpha^2\pi(x) + (1-\alpha)^2(1-\pi(x))} m(x)$$

$$= \frac{\alpha^3 \frac{q(x)}{m(x)}(1-\alpha)^3 \frac{p(x)}{m(x)}}{\alpha^3 \frac{q(x)}{m(x)} + (1-\alpha)^3 \frac{q(x)}{m(x)}} m(x)$$
$$= \frac{\alpha^3(1-\alpha)^3 p(x)q(x)}{\alpha^3 q(x) + (1-\alpha)^3 p(x)}$$

so that

$$s(x) \propto \frac{p(x)q(x)}{\alpha^3 q(x) + (1-\alpha)^3 p(x)}.$$

Q.E.D.

Appendix H: Alternative definitions of shared space

The shared space S_X we use in most of this work has density

$$s_X(x) \propto \frac{p_X(x)q_X(x)}{p_X(x) + q_X(x)}$$

as first defined in Equation (2), so that it has support only where both p_X and q_X has support, and also higher (resp. lower) density when $p_X(x)$ and $q_X(x)$ have higher (resp. lower) density. As mentioned in Equation (3), there are other definitions of S_X that have similar properties. We repeat the experiments Section 4.1 with these alternative definitions of S_X and show that the decompositions are generally not too sensitive to the specific choice of S_X in Figure EC.2, assuming a reasonable choice of clipping threshold where relevant. We also compare with the best worst-case S_X from Appendix G; we do not recommend it, as worst-case assumptions may be inappropriate in real-world settings, as demonstrated in e.g. Figure EC.2c.

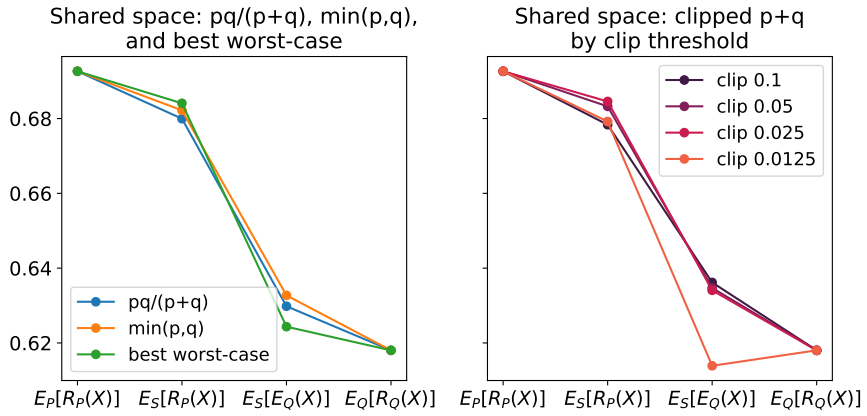
Appendix I: Additional experiment plots

I.1. Predicting employment from census data

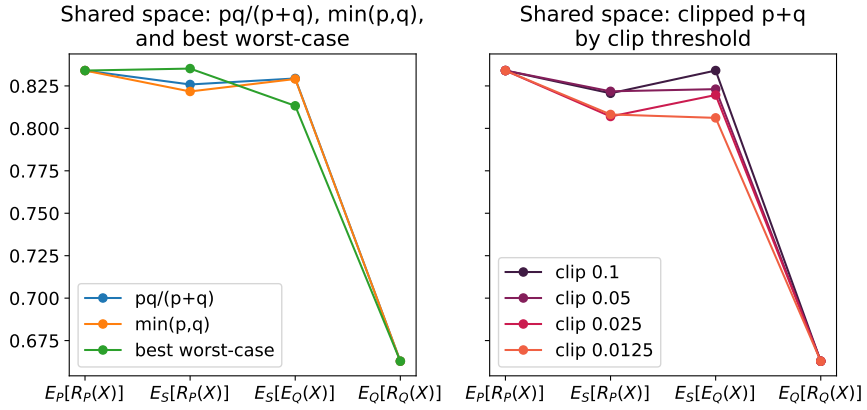
In Figure EC.3 we include additional figures for Section 4.1.

I.2. Diagnosing failures of domain-adaptation methods

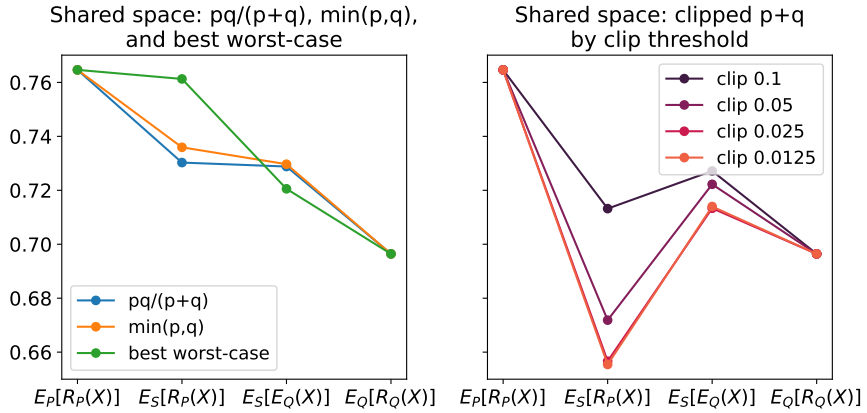
In Figure EC.4 we include additional figures for Section 4.2.



(a) $Y|X$ shift: original model trained on West Virginia and evaluated on Maryland

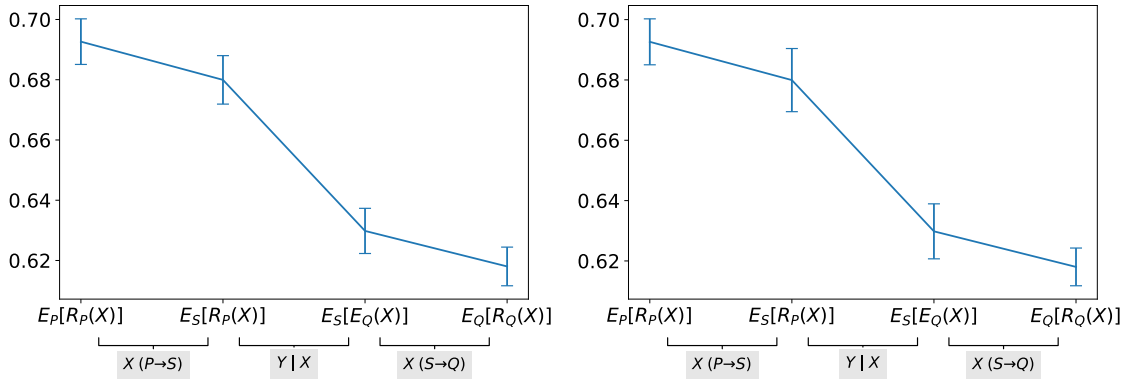


(b) X shift: model trained on only age $\leq$25 and evaluated on general population

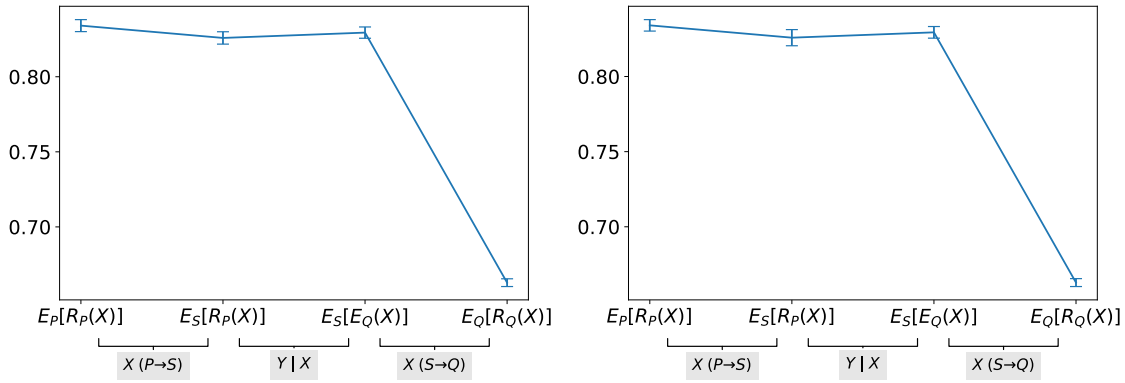


(c) X shift: model trained on over-sampling age ≤ 25 and evaluated on general population

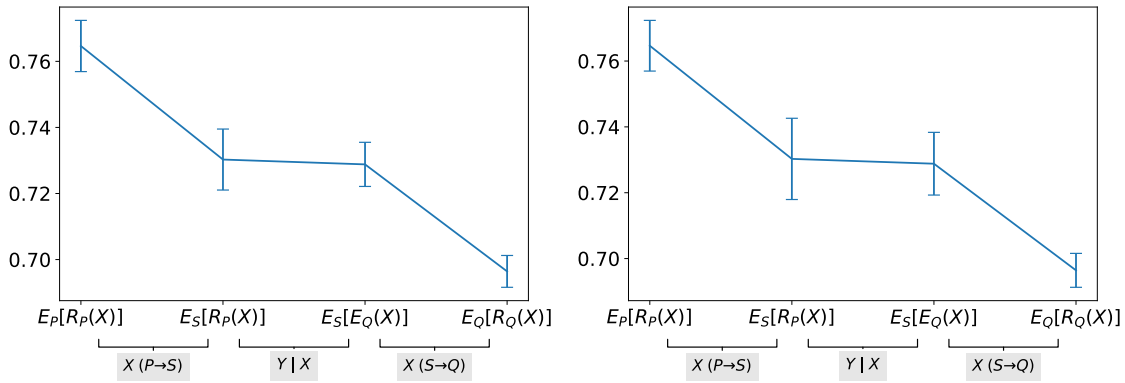
Figure EC.2 Comparison of different shared spaces for experiments in Section 4.1. On the Y axis is accuracy; note that all results hold if we replace $\ell(\cdot)$ with accuracy. The leftmost plots correspond to Equations (2), (3a), and (EC.7), while the rightmost plots correspond to Equation (3b). Because of the need to choose a threshold for Equation (3b) and the sensitivity of decompositions to the threshold value, we do not recommend it. Because the best worst-case decomposition in Figure EC.2c does not reflect the ground-truth X shift, we do not recommend Equation (EC.7).



(a) $Y|X$ shift: original model trained on West Virginia and evaluated on Maryland. Left: CI from 500 bootstrap resamples; right: CI from influence function



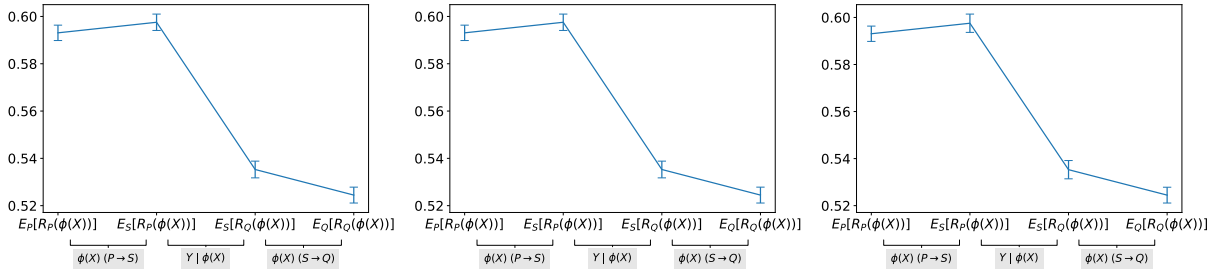
(b) X shift: model trained on only age $\leq$25 and evaluated on general population. Left: CI from 500 bootstrap resamples; right: CI from influence function



(c) X shift: model trained on over-sampling age ≤ 25 and evaluated on general population. Left: CI from 500 bootstrap resamples; right: CI from influence function

Figure EC.3 Decomposition plots from Section 4.1 with confidence intervals. Left: confidence intervals are ± 1 standard deviation of the estimand over 500 bootstrap resamples. In each resampling, the prediction model and the propensity model are re-fit. Right: confidence intervals are ± 1 standard deviation of the estimand using influence function-based calculations as in Section 5.

Figure EC.4 Decomposition plots from Section 4.2 with confidence intervals, for DANN. In each resampling, the propensity model is re-fit but the prediction model is unchanged. Left: confidence intervals are ± 1 standard deviation of the estimand over 500 bootstrap resamples. Middle: confidence intervals are ± 1 standard deviation of the estimand using influence function-based calculations as in Section 5, where $\hat{\mu}_Q, \hat{\mu}_P$ are linear regressions on features $\phi(X)$. Right: confidence intervals are ± 1 standard deviation of the estimand using influence function-based calculations as in Section 5, where $\hat{\mu}_Q, \hat{\mu}_P$ are XGBoost models on features $\phi(X)$.



References

Chernozhukov V, Chetverikov D, Demirer M, Duflo E, Hansen C, Newey W, Robins J (2018) Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal* 21(1):C1–C68.

Ding F, Hardt M, Miller J, Schmidt L (2021) Retiring adult: New datasets for fair machine learning. *Advances in Neural Information Processing Systems 34* 34.

Gutman R, Karavani E, Shimoni Y (2022) Propensity score models are better when post-calibrated. *arXiv:2211.01221 [stat.ML]* URL https://arxiv.org/abs/2211.01221.

Imbens G, Rubin D (2015) *Causal Inference for Statistics, Social, and Biomedical Sciences* (Cambridge University Press).

Kennedy EH (2022) Semiparametric doubly robust targeted double machine learning: a review. *arXiv:2203.06469 [stat.ME]* .

Newey WK, McFadden D (1994) Large sample estimation and hypothesis testing. *Handbook of Econometrics* 4:2111–2245.

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E (2011) Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12:2825–2830.

Yadlowsky S (2022) On cross-fitting with plug-in estimators. URL https://www.syadlowsky.com/blog/semiparametric/2022/10/24/on-cross-fitting-with-plug-in-estimators.html.

Zheng W, van der Laan MJ (2011) *Cross-Validated Targeted Minimum-Loss-Based Estimation*, 459–474 (New York, NY: Springer New York), ISBN 978-1-4419-9782-1, URL http://dx.doi.org/10.1007/978-1-4419-9782-1_27.