Appendix A Simulation Settings

In this section we detail the simulation settings. Let $\mathbf{S}_{10} = \mathbf{S}_{11} = (1, O_1)$, $\mathbf{S}_{20} = (1, R_2, O_1, A_1, O_1A_1, O_2)$, $\mathbf{S}_{21} = (1, A_1, O_2)$, the data is generated sequentially according to: $O_1 \sim$Bern$(\frac{1}{2})$, $A_1|O_1 \sim$Bern$\big(\sigma\left(\mathbf{S}_{10}^{\intercal}\boldsymbol{\xi}_1^0\right)\big)$, $O_2|O_1, A_1, R_2 \sim \mathcal{N}((1, O_1, A_1, O_1A_1, R_2)^{\intercal}\boldsymbol{\delta}_1^0, 2)$, and $A_2|O_1, O_2, A_1, R_2 \sim$Bern$\big(\sigma\left(\mathbf{S}_{20}^{\intercal}\boldsymbol{\xi}_2^0 + \xi_{26}^0 O_2^2\right)\big)$. Setting (1) has continuous outcomes: $R_2|\mathbf{S}_1 \sim \mathcal{N}\left(\mathbf{S}_{10}^{\intercal}\boldsymbol{\beta}_1^0 + A_1(\mathbf{S}_{11}^{\intercal}\boldsymbol{\gamma}_1^0), 1\right)$,
$R_3|\mathbf{S}_2 \sim \mathcal{N}\left(\mathbf{S}_{20}^{\intercal}\boldsymbol{\beta}_2^0 + \beta_{27}^0 O_2^2 R_2 \sin\left(\frac{1}{O_2^2 R_2}\right) + A_2(\mathbf{S}_{21}^{\intercal}\boldsymbol{\gamma}_2^0), 2\right)$. Setting (2) has binary outcomes:
$\mathbb{P}(R_2 = 1|\mathbf{S}_1) = \sigma\left(\mathbf{S}_{10}^{\intercal}\boldsymbol{\beta}_1^0 + A_1(\mathbf{S}_{11}^{\intercal}\boldsymbol{\gamma}_1^0\right)$,
$\mathbb{P}(R_3 = 1|\mathbf{S}_2) = \sigma\left(\mathbf{S}_{20}^{\intercal}\boldsymbol{\beta}_2^0 + \beta_{27}^0 O_2^2 R_2 \sin\left(\frac{1}{O_2^2 R_2}\right) + A_2(\mathbf{S}_{21}^{\intercal}\boldsymbol{\gamma}_2^0\right)$,

To explore the method's performance under model miss-specification, we vary $\beta_{27}, \xi_{26} \in (-1, 1)$, and we fit models $Q_1(\mathbf{S}_1, A_1) = \mathbf{S}_{10}^{\intercal}\boldsymbol{\beta}_1^0 + A_1(\mathbf{S}_{11}^{\intercal}\boldsymbol{\gamma}_1^0)$, $Q_2(\check{\mathbf{S}}_2, A_2) = \check{\mathbf{S}}_{20}^{\intercal}\boldsymbol{\beta}_2^0 + A_2(\mathbf{S}_{21}^{\intercal}\boldsymbol{\gamma}_2^0)$ for the Q functions, $\pi_1(\mathbf{S}_1) = \sigma\left(\mathbf{S}_{10}^{\intercal}\boldsymbol{\xi}_1\right)$ and $\pi_2(\check{\mathbf{S}}_2) = \sigma\left(\mathbf{S}_{20}^{\intercal}\boldsymbol{\xi}_2\right)$ for the propensity scores. Datasets are generated using $(n, N) \in \{(135, 1272), (500, 10000)\}$. Parameters are consistent with [8]: $\boldsymbol{\xi}_1^0 = (0.3, -0.5)^{\intercal}$, $\boldsymbol{\beta}_1^0 = (3, 0, 0.1, -0.5)^{\intercal}$, $\boldsymbol{\delta}_1^0 = (0, 0.5, -0.75, 0.25, -.75)^{\intercal}$,
$\boldsymbol{\gamma}_2^0 = (0, 0.5, 0.1, -1, -0.1, 0, -.5)^{\intercal}$, $\boldsymbol{\beta}_2^0 = (3, 0, 0.1, -0.5, -0.5, 0, .1)^{\intercal}$, $\boldsymbol{\gamma}_2^0 = (1, 0.25, 0.5)^{\intercal}$, $\boldsymbol{\xi}_2^0 = (0, 0.5, 0.1, -1, -0.1)^{\intercal}$.

Appendix B Assumptions

Assumption B.1 *(a) Sample size for $\mathcal{U}$, and $\mathcal{L}$, are such that $n/N \longrightarrow 0$ as $N, n \longrightarrow \infty$, (b) $\check{\mathbf{S}}_t \in \mathcal{H}_t$, $\check{\mathbf{X}}_t \in \mathcal{X}_t$ have finite second moments and compact support in $\mathcal{H}_t \subset \mathbb{R}^{q_t}$, $\mathcal{X}_t \subset \mathbb{R}^{p_t}$ $t = 1, 2$ respectively (c) $\Sigma_1,\ \Sigma_2$ are nonsingular.*

Assumption B.2 *Define the following class of functions:*

$$\mathcal{Q}_t \equiv \left\{Q_t : \mathcal{X}_1 \mapsto \mathbb{R}|\boldsymbol{\theta}_1 \in \Theta_1 \subset \mathbb{R}^{p_t}\right\},\ t = 1, 2,$$

with Θ_1, Θ_2 open bounded sets, and p_1, p_2 fixed under (1). Suppose the population equations for the Q functions $\mathbb{E}[S_t^{\theta}(\boldsymbol{\theta}_t)] = \mathbf{0}, t = 1, 2$ have solutions $\bar{\boldsymbol{\theta}}_1, \bar{\boldsymbol{\theta}}_2$, where

$$S_2^{\theta}(\boldsymbol{\theta}_2) = \frac{\partial}{\partial \boldsymbol{\theta}_2^{\intercal}}\|R_3 - Q_2(\check{\mathbf{X}}_2; \boldsymbol{\theta}_2)\|_2^2,\ S_1^{\theta}(\boldsymbol{\theta}_1) = \frac{\partial}{\partial \boldsymbol{\theta}_1^{\intercal}}\|R_2^* - Q_1(\check{\mathbf{X}}_1; \boldsymbol{\theta}_1)\|_2^2.$$

The population minimizers satisfy $\bar{\boldsymbol{\theta}}_t \in \Theta_t$, $t = 1, 2$.

As discussed assuming model (1) are likely miss-specified, therefore we establish results for our doubly robust semi-supervised value function estimator. For this, define the following class of functions:

$$\mathcal{W}_t \equiv \left\{\pi_t : \mathcal{H}_t \mapsto \mathbb{R}|\boldsymbol{\theta}_1 \in \Theta_t, \boldsymbol{\xi}_t \in \Omega_t\right\},\ t = 1, 2,$$

under propensity score models $\pi_1,\ \pi_2$ in (2).

Assumption B.3 *Let the population equations $\mathbb{E}\left[S_t^{\xi}(\check{\mathbf{S}}_t; \boldsymbol{\Theta}_t)\right] = \mathbf{0}, t = 1, 2$ have solutions $\bar{\boldsymbol{\xi}}_1, \bar{\boldsymbol{\xi}}_2$, where*

$$S_t^{\xi}(\check{\mathbf{S}}_t; \boldsymbol{\Theta}_t) = \frac{\partial}{\partial \boldsymbol{\xi}_t}\log\left[\pi_t(\check{\mathbf{S}}_t; \boldsymbol{\xi}_t)^{I(d_t = A_t)}\{1 - \pi_t(\check{\mathbf{S}}_t; \boldsymbol{\xi}_t)\}^{(1 - I\{d_t = A_t\})}\right],\ t = 1, 2,$$

(i) Ω_1, Ω_2 are open, bounded sets and the population solutions satisfy $\bar{\boldsymbol{\xi}}_t \in \Omega_t, t = 1, 2$,
(ii) for $\bar{\boldsymbol{\xi}}_t, t = 1, 2$, $\inf_{\check{\mathbf{S}}_t \in \mathcal{H}_1} \pi_1(\check{\mathbf{S}}_t; \bar{\boldsymbol{\xi}}_t) > 0$,
(iii) for $t = 1, 2$,

$$\sup_{\boldsymbol{\xi}_t}\left\|\mathbb{P}_n S_t^{\xi}(\check{\mathbf{S}}_t; \boldsymbol{\Theta}_t) - \mathbb{E}\left[S_t^{\xi}(\check{\mathbf{S}}_t; \boldsymbol{\Theta}_t)\right]\right\|_{L_2(\mathbb{P})} \xrightarrow{p} 0,$$

$$\inf_{\boldsymbol{\xi}_t : d(\boldsymbol{\xi}_t, \bar{\boldsymbol{\xi}}_t) \geq \delta}\left\|\mathbb{E}\left[S_t^{\xi}(\check{\mathbf{S}}_t; \boldsymbol{\Theta}_t)\right]\right\|_{L_2(\mathbb{P})} > 0,\ \forall \delta > 0.$$

Assumption B.4 *Functions $m_2, m_{\omega_2}, m_{t\omega_2}$ $t = 2, 3$ are such that (i)* $\sup_{\vec{\mathbf{U}}} |m_s(\vec{\mathbf{U}})| < \infty$*,* $s \in \{2, \omega_2, 2\omega_2, 3\omega_2\}$ *and (ii) the estimated functions $\hat{m}_s$ satisfy (ii)* $\sup_{\vec{\mathbf{U}}} |\widehat{m}_s(\vec{\mathbf{U}}) - m_s(\vec{\mathbf{U}})| = o_{\mathbb{P}}(1)$*,* $s \in \{2, \omega_2, 2\omega_2, 3\omega_2\}$*.*

Assumption B.3 is standard for empirical process estimation [25]. In particular (iii) requires that the score converges to its population limit in $L_2(\mathbb{P})$ norm as defined in Section 3, and a well separated uniqueness condition for $\bar{\boldsymbol{\xi}}$. Assumption B.4 is the propensity score equivalent version of Assumption **??**. However note that for this to be satisfied we are relying on the positivity Assumption (ii) made in Section 2. Finally, from Assumption B.3, as we use maximum likelihood estimation for $\widehat{\boldsymbol{\xi}}$, there exists an influence function $\boldsymbol{\psi}^{\xi} : \mathcal{H} \mapsto \Omega$ such that $\sqrt{n}(\widehat{\boldsymbol{\xi}} - \bar{\boldsymbol{\xi}}) = n^{-1/2} \sum_{i=1}^{n} \boldsymbol{\psi}_i^{\xi} + o_{\mathbb{P}}(1)$ and $\mathbb{E}[\boldsymbol{\psi}^{\xi}] = 0$, $\mathbb{E}[(\boldsymbol{\psi}^{\xi})^{\intercal}\boldsymbol{\psi}^{\xi}] < \infty$. Further let $\boldsymbol{\psi}^{\theta} = (\boldsymbol{\psi}_1^{\intercal}, \boldsymbol{\psi}_2^{\intercal})^{\intercal}$ be the concatenation of the influence functions from Theorems **??** & **??**. Under assumptions B.1-B.4 we are now ready to state our theoretical justification for the value function estimator in equation (6), the proof can be found in Appendix **??**.

Appendix C Proof of Main Results

[Proof of Proposition 4.1]
By Theorem **??** in ... we have $\widehat{V}_{\mathsf{SSL}_{\mathsf{DR}}} - \mathbb{E}\left[\mathcal{V}_{\bar{\boldsymbol{\Theta}},\bar{\mu}}\right] = o_{\mathbb{P}}(1)$, therefore

$$\widehat{V}_{\mathsf{SSL}_{\mathsf{DR}}} - \bar{V} = \frac{1}{n}\sum_{i=1}^{n} \psi_{\mathcal{V}}^{ss}(\vec{\mathbf{U}}_i; \bar{\boldsymbol{\theta}}, \bar{\boldsymbol{\xi}}) + \mathbb{E}\left[\widehat{V}_{\mathsf{SSL}_{\mathsf{DR}}}\right] - \bar{V} + o_{\mathbb{P}}(1),$$

by Lemma C.1 we get

$$\mathbb{E}\left[\mathcal{V}_{\bar{\boldsymbol{\Theta}},\bar{\mu}}\right] - \bar{V} = \mathbb{E}\left[\left\{1 - \frac{\pi_1(\mathbf{S}_1)}{\pi_1(\mathbf{S}_1; \bar{\boldsymbol{\xi}})}\right\}\left\{Q_1^o(\mathbf{S}_1) - Q_1^o(\mathbf{S}_1; \bar{\boldsymbol{\theta}}_1)\right\}\right],$$

thus

$$\widehat{V}_{\mathsf{SSL}_{\mathsf{DR}}} - \bar{V} = \mathbb{E}\left[\left\{1 - \frac{\pi_1(\mathbf{S}_1)}{\pi_1(\mathbf{S}_1; \bar{\boldsymbol{\xi}})}\right\}\left\{Q_1^o(\mathbf{S}_1) - Q_1^o(\mathbf{S}_1; \bar{\boldsymbol{\theta}}_1)\right\}\right]$$

if either (1) or (2) are correct, then $\widehat{V}_{\mathsf{SSL}_{\mathsf{DR}}} - \bar{V} = o_{\mathbb{P}}(1)$.

Lemma C.1 *Let $\hat{Q}_t$, $\hat{\pi}_t$ $t = 1, 2$ be any functions which satisfy Assumptions B.2 and B.3 respectively and define*

$$\mathcal{V}_{\mathsf{SSL}_{\mathsf{DR}}}(\mathbf{S}) = \bar{Q}_1(\mathbf{S}_1; \widehat{\boldsymbol{\theta}}_1) + \frac{I(\widehat{d}_1 = A_1)}{\pi_1(\mathbf{S}_1; \widehat{\boldsymbol{\xi}}_1)}\left\{R_2 - \left[\hat{Q}_1(\mathbf{S}_1, A_1) - \bar{Q}_2(\check{\mathbf{S}}_2; \widehat{\boldsymbol{\theta}}_2)\right]\right\}$$
$$+ \frac{I(\widehat{d}_1 = A_1)I(\widehat{d}_2 = A_2)}{\pi_1(\mathbf{S}_1; \widehat{\boldsymbol{\xi}}_1)\pi_2(\check{\mathbf{S}}_2; \widehat{\boldsymbol{\xi}}_2)}\left\{R_3 - Q_2(\check{\mathbf{S}}_2, A_2; \widehat{\boldsymbol{\theta}}_2)\right\},$$

then the bias term is

$$\begin{aligned} Bias\left(\mathcal{V}_{\mathsf{SSL}_{\mathsf{DR}}}, \bar{V}\right) \equiv & \mathbb{E}\left[\mathcal{V}_{\mathsf{SSL}_{\mathsf{DR}}}\right] - \bar{V} \\ = & \mathbb{E}\left[\left\{1 - \frac{\pi_1^0(\mathbf{S}_1)}{\hat{\pi}_1(\mathbf{S}_1)}\right\}\left\{Q_1^{0*}(\mathbf{S}_1) - \hat{Q}_1^*(\mathbf{S}_1)\right\}\right] \\ & + \mathbb{E}\left[\frac{\pi_1^0(\mathbf{S}_1)}{\hat{\pi}_1(\mathbf{S}_1)}\left\{1 - \frac{\pi_2^0(\check{\mathbf{S}}_2)}{\hat{\pi}_2(\check{\mathbf{S}}_2)}\right\}\left\{Q_2^{0*}(\check{\mathbf{S}}_2) - \hat{\bar{Q}}_2(\check{\mathbf{S}}_2)\right\}\right], \end{aligned}$$

where $\bar{V} = \mathbb{E}[\mathbb{E}[R_2 + \mathbb{E}[R_3|\mathbf{S}_2, R_2, A_2 = \bar{d}_2]|\mathbf{S}_1, A_1 = \bar{d}_1]]$ *is the mean population value under the optimal treatment rule, and* $Q_1^{0*}(\mathbf{X}_1) = \mathbb{E}[R_2 + \mathbb{E}[R_3|\mathbf{S}_2, A_2 = \bar{d}_2(\mathbf{S}_2), R_2]|\mathbf{S}_1, A_1 = \bar{d}_1(\mathbf{S}_1)]$, $Q_2^{0*}(\check{\mathbf{S}}_2) = \mathbb{E}[R_3|\mathbf{S}_2, A_2 = \bar{d}_2(\mathbf{S}_2), R_2]$.

[Proof of Lemma C.1]

First note that from the reffitting step, using iterated expectations we have

$$\mathbb{E}\left[\bar{\mu}_2\right] = \mathbb{E}\left[\bar{\mu}_2 - Y_2\right] + \mathbb{E}[Y_2] = \mathbb{E}\left[\mathbb{E}[Y_2|\vec{\mathbf{U}}] - Y_2\right] + \mathbb{E}[Y_2] = \mathbb{E}[Y_2],$$

and similar for $\bar{\mu}_{\omega_2}, \bar{\mu}_{t\omega_2}, t=2,3$, therefore

$$
\begin{aligned}
\text{Bias}\left(\mathcal{V}_{\mathsf{SSL}_{\mathsf{DR}}}, \bar{V}\right) =& \mathbb{E}\left[\mathcal{V}_{\mathsf{SSL}_{\mathsf{DR}}}\right] - \mathbb{E}[\mathbb{E}[Y_2 + \mathbb{E}[Y_3|\mathbf{S}_2, Y_2, A_2 = \bar{d}_2]|\mathbf{S}_1, A_1 = \bar{d}_1]] \\
=& \mathbb{E}\left[Q_1^{0*}(\mathbf{S}_1) - \hat{Q}_1^*(\mathbf{S}_1)\right] - \mathbb{E}\left[\frac{I(\widehat{d}_1 = A_1)}{\hat{\pi}_1(\mathbf{S}_1)}\left\{Y_2 - \left[\hat{Q}_1(\mathbf{S}_1, A_1) - \hat{\bar{Q}}_2(\check{\mathbf{S}}_2)\right]\right\}\right] \\
&+ \mathbb{E}\left[\frac{I(\widehat{d}_1 = A_1) l_2(\widehat{d}_{2j}, A_{2j})}{\hat{\pi}_1(\mathbf{S}_1)\hat{\pi}_2(\check{\mathbf{S}}_2)}\left\{Y_3 - \hat{Q}_2(\check{\mathbf{S}}_2, A_2)\right\}\right].
\end{aligned}
$$

Adding and subtracting $Q_2^{*0}(\check{\mathbf{S}}_2)$,

$$
\begin{aligned}
= \mathbb{E}\left[Q_1^{0*}(\mathbf{S}_1) - \hat{Q}_1^*(\mathbf{S}_1)\right] - \mathbb{E}\Bigg[\frac{I(\widehat{d}_1 = A_1)}{\hat{\pi}_1(\mathbf{S}_1)}\Big\{& Y_2 + \mathbb{E}[Y_3|\mathbf{S}_2, \widehat{d}_2, Y_2] - \hat{Q}_1(\mathbf{S}_1, A_1) \\
&+ Q_2^{*0}(\check{\mathbf{S}}_2) - \hat{\bar{Q}}_2(\check{\mathbf{S}}_2)\Big\}\Bigg] \\
&+ \mathbb{E}\left[\frac{I(\widehat{d}_1 = A_1) I(\widehat{d}_2 = A_2)}{\hat{\pi}_1(\mathbf{S}_1)\hat{\pi}_2(\check{\mathbf{S}}_2)}\left\{Y_3 - \hat{Q}_2(\check{\mathbf{S}}_2, A_2)\right\}\right],
\end{aligned}
$$

using iterated expectations in the second and fourth terms:

$$
\begin{aligned}
=& \mathbb{E}\left[Q_1^{0*}(\mathbf{S}_1) - \hat{Q}_1^*(\mathbf{S}_1)\right] - \mathbb{E}\left[\mathbb{E}\left[\frac{I(\widehat{d}_1 = A_1)}{\hat{\pi}_1(\mathbf{S}_1)}\left\{Y_2 + \mathbb{E}[Y_3|\mathbf{S}_2, \bar{d}_2, Y_2] - \hat{Q}_1(\mathbf{S}_1, A_1)\right\}\middle|\mathbf{S}_1, A_1\right]\right] \\
&+ \mathbb{E}\left[\frac{I(\widehat{d}_1 = A_1)}{\hat{\pi}_1(\mathbf{S}_1)}\left\{Q_2^{*0}(\check{\mathbf{S}}_2) - \hat{\bar{Q}}_2(\check{\mathbf{S}}_2)\right\}\right] \\
&+ \mathbb{E}\left[\mathbb{E}\left[\frac{I(\widehat{d}_1 = A_1) l_2(\widehat{d}_{2j}, A_{2j})}{\hat{\pi}_1(\mathbf{S}_1)\hat{\pi}_2(\check{\mathbf{S}}_2)}\left\{Y_3 - \hat{Q}_2(\check{\mathbf{S}}_2, A_2)\right\}\middle|\mathbf{S}_2, A_2, Y_2\right]\right] \\
=& \mathbb{E}\left[Q_1^{0*}(\mathbf{S}_1) - \hat{Q}_1^*(\mathbf{S}_1)\right] - \mathbb{E}\left[\frac{I(\widehat{d}_1 = A_1)}{\hat{\pi}_1(\mathbf{S}_1)}\left\{\mathbb{E}\left[Y_2 + \mathbb{E}[Y_3|\mathbf{S}_2, \bar{d}_2, Y_2]\middle|\mathbf{S}_1, A_1\right] - \hat{Q}_1(\mathbf{S}_1, A_1)\right\}\right] \\
&+ \mathbb{E}\left[\frac{I(\widehat{d}_1 = A_1)}{\hat{\pi}_1(\mathbf{S}_1)}\left\{Q_2^{*0}(\check{\mathbf{S}}_2) - \hat{\bar{Q}}_2(\check{\mathbf{S}}_2)\right\}\right] \\
&+ \mathbb{E}\left[\frac{I(\widehat{d}_1 = A_1) l_2(\widehat{d}_{2j}, A_{2j})}{\hat{\pi}_1(\mathbf{S}_1)\hat{\pi}_2(\check{\mathbf{S}}_2)}\left\{\mathbb{E}\left[Y_3|\mathbf{S}_2, A_2, Y_2\right] - \hat{Q}_2(\check{\mathbf{S}}_2, A_2)\right\}\right].
\end{aligned}
$$

Using the definitions of $Q_t^0, t=1,2$:

$$
\begin{aligned}
=& \mathbb{E}\left[Q_1^0(\mathbf{S}_1) - \hat{Q}_1^*(\mathbf{S}_1)\right] - \mathbb{E}\left[\frac{I(\widehat{d}_1 = A_1)}{\hat{\pi}_1(\mathbf{S}_1)}\left\{Q_1^0(\mathbf{S}_1, A_1) - \hat{Q}_1(\mathbf{S}_1, A_1)\right\}\right] \\
&+ \mathbb{E}\left[\frac{I(\widehat{d}_1 = A_1)}{\hat{\pi}_1(\mathbf{S}_1)}\left\{Q_2^{*0}(\check{\mathbf{S}}_2) - \hat{\bar{Q}}_2(\check{\mathbf{S}}_2)\right\}\right] \\
&+ \mathbb{E}\left[\frac{I(\widehat{d}_1 = A_1) l_2(\widehat{d}_{2j}, A_{2j})}{\hat{\pi}_1(\mathbf{S}_1)\hat{\pi}_2(\check{\mathbf{S}}_2)}\left\{Q_2^0(\check{\mathbf{S}}_2, A_2) - \hat{Q}_2(\check{\mathbf{S}}_2, A_2)\right\}\right],
\end{aligned}
$$

assuming $A_1 \perp A_2|\mathbf{S}_2, Y_2$ using iterated expectations where we condition on $A_t = \bar{d}_t$:

$$
\begin{aligned}
=& \mathbb{E}\left[Q_1^{0*}(\mathbf{S}_1) - \hat{Q}_1^*(\mathbf{S}_1)\right] - \mathbb{E}\left[\frac{w_1^0(\mathbf{S}_1)}{\hat{\pi}_1(\mathbf{S}_1)}\left\{Q_1^{0*}(\mathbf{S}_1) - \hat{\bar{Q}}_1(\mathbf{S}_1)\right\}\right] \\
&+ \mathbb{E}\left[\frac{w_1^0(\mathbf{S}_1)}{\hat{\pi}_1(\mathbf{S}_1)}\left\{Q_2^{*0}(\check{\mathbf{S}}_2) - \hat{\bar{Q}}_2(\check{\mathbf{S}}_2)\right\}\right] \\
&+ \mathbb{E}\left[\frac{w_1^0(\mathbf{S}_1) w_2^0(\check{\mathbf{S}}_2)}{\hat{\pi}_1(\mathbf{S}_1)\hat{\pi}_2(\check{\mathbf{S}}_2)}\left\{Q_2^{*0}(\check{\mathbf{S}}_2) - \hat{\bar{Q}}_2(\check{\mathbf{S}}_2)\right\}\right]
\end{aligned}
$$

finally, factorizing common terms:

$$
\begin{aligned}
=&\mathbb{E}\left[\left\{1-\frac{\pi_1^0(\mathbf{S}_1)}{\hat{\pi}_1(\mathbf{S}_1)}\right\}\left\{Q_1^{0*}(\mathbf{S}_1)-\hat{Q}_1^{*}(\mathbf{S}_1)\right\}\right]\\
+&\mathbb{E}\left[\frac{\pi_1^0(\mathbf{S}_1)}{\hat{\pi}_1(\mathbf{S}_1)}\left\{1-\frac{\pi_2^0(\check{\mathbf{S}}_2)}{\hat{\pi}_2(\check{\mathbf{S}}_2)}\right\}\left\{Q_2^{*0}(\check{\mathbf{S}}_2)-\hat{\check{Q}}_2(\check{\mathbf{S}}_2)\right\}\right].
\end{aligned}
$$