

## A PROOFS

### A.1 ATE IDENTIFICATION AND SUFFICIENT CONDITIONS

*Proof.* Under Assumptions 1 and 2, the variables in  $X_i$  are a sufficient adjustment set, so that the ATE is identifiable as

$$\begin{aligned}
 \tau &= \mathbb{E}[Y(1) - Y(0)] \\
 &= \mathbb{E}[\mathbb{E}[Y(1) | X] - \mathbb{E}[Y(0) | X]] \\
 &= \mathbb{E}[\mathbb{E}[Y(1) | X, W = 1] - \mathbb{E}[Y(0) | X, W = 0]] && \text{Assumptions 1 and 2} \\
 &= \mathbb{E}[\mathbb{E}[Y | X, W = 1] - \mathbb{E}[Y | X, W = 0]] && \text{SUTVA} \\
 &= \mathbb{E}\left[\frac{WY}{e(X)} - \frac{(1-W)Y}{1-e(X)}\right] && \text{Assumption 3}
 \end{aligned}$$

□

### A.2 PROOF OF THEOREM 1

*Proof.* **Unbiasedness.** We have:

$$\begin{aligned}
 \mathbb{E}[\hat{\tau}^{\text{IPW}^*}] &= \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^n \left(\frac{W_i Y_i}{e(X_i)} - \frac{(1-W_i)Y_i}{1-e(X_i)}\right)\right] \\
 &= \frac{1}{n} \sum_{k=1}^K \mathbb{E}\left[\mathbb{E}\left[\sum_{i=1}^n \left(\frac{W_i Y_i}{e(X_i)} - \frac{(1-W_i)Y_i}{1-e(X_i)}\right) \cdot \mathbb{1}_{[H_i=k]} \mid n_1, \dots, n_K\right]\right] \\
 &= \frac{1}{n} \sum_{k=1}^K \mathbb{E}[n_k] \left(\mathbb{E}\left[\frac{W_i Y_i}{e(X_i)} \mid H_i = k\right] - \mathbb{E}\left[\frac{(1-W_i)Y_i}{1-e(X_i)} \mid H_i = k\right]\right) \\
 &= \sum_{k=1}^K \rho_k \left(\mathbb{E}\left[\frac{W_i Y_i}{e(X_i)} \mid H_i = k\right] - \mathbb{E}\left[\frac{(1-W_i)Y_i}{1-e(X_i)} \mid H_i = k\right]\right) \tag{6}
 \end{aligned}$$

Let us focus on the first term of the sum:

$$\begin{aligned}
\mathbb{E} \left[ \frac{W_i Y_i}{e(X_i)} \mid H_i = k \right] &= \mathbb{E} \left[ \frac{W_i Y_i(1)}{e(X_i)} \mid H_i = k \right] && \text{SUTVA} \\
&= \mathbb{E} \left[ \mathbb{E} \left[ \frac{W_i Y_i(1)}{e(X_i)} \mid X_i, H_i = k \right] \right] \\
&= \mathbb{E} \left[ \frac{\mathbb{E}[W_i \mid X_i, H_i = k] \mathbb{E}[Y_i(1) \mid X_i, H_i = k]}{e(X_i)} \right] && \text{Assumptions 1 and 2} \\
&= \mathbb{E} \left[ \frac{e_k(X_i) \mathbb{E}[Y_i(1) \mid X_i, H_i = k]}{e(X_i)} \right] \\
&= \mathbb{E} \left[ \frac{e_k(X_i) \mathbb{E}[Y_i(1) \mid X_i, H_i = k]}{\mathbb{E}[e_k(X_i) \mid X_i, H_i = k]} \right] && \text{Definition of } e(X_i) \\
&= \mathbb{E} \left[ \mathbb{E}[Y_i(1) \mid X_i, H_i = k] \mathbb{E} \left[ \frac{e_k(X_i)}{\mathbb{E}[e_k(X_i)]} \mid X_i \right] \right] \\
&= \mathbb{E}[Y_i(1) \mid H_i = k]
\end{aligned}$$

Similarly, we have  $\mathbb{E} \left[ \frac{(1-W_i)Y_i}{1-e(X_i)} \mid H_i = k \right] = \mathbb{E}[Y_i(0) \mid H_i = k]$ , so that  $\mathbb{E}[\hat{\tau}^{\text{IPW}^*}] = \sum_{k=1}^K \rho_k \tau_k$ , which concludes the proof of unbiasedness of the oracle multi-site centralized IPW estimator.

For the AIPW estimator, following the same steps as in eq. (6), we have

$$\begin{aligned}
\mathbb{E}[\hat{\tau}^{\text{AIPW}^*}] &= \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \left( \mu_1(X_i) - \mu_0(X_i) + \frac{W_i(Y_i - \mu_1(X_i))}{e(X_i)} - \frac{(1-W_i)(Y_i - \mu_0(X_i))}{1-e(X_i)} \right) \right] \\
&= \sum_{k=1}^K \rho_k \left( \mathbb{E}[\mu_1(X_i) - \mu_0(X_i) \mid H_i = k] \right. \\
&\quad \left. + \mathbb{E} \left[ \frac{W_i(Y_i - \mu_1(X_i))}{e(X_i)} \mid H_i = k \right] - \mathbb{E} \left[ \frac{(1-W_i)(Y_i - \mu_0(X_i))}{1-e(X_i)} \mid H_i = k \right] \right) \\
&= \sum_{k=1}^K \rho_k (\mathbb{E}[\tau(X_i) \mid H_i = k] + 0 + 0) \\
&= \tau
\end{aligned}$$

**Variance.** As we consider uniformly bounded potential outcomes, that is  $\forall w \in \{0, 1\}, \exists (L, U) \in \mathbb{R}^2, L < Y(w) < U$ , and Assumption 3, we have that  $\mathbb{E} \left[ \frac{Y_i(1)^2}{e(X_i)} \right] < \infty$  and  $\mathbb{E} \left[ \frac{Y_i(0)^2}{1-e(X_i)} \right] < \infty$ , so the quantities that follow are well defined:

$$\begin{aligned}
\mathbb{V} [\hat{\tau}^{\text{IPW}^*}] &= \mathbb{V} \left[ \frac{1}{n} \sum_{i=1}^n \left( \frac{W_i Y_i}{e(X_i)} - \frac{(1 - W_i) Y_i}{1 - e(X_i)} \right) \right] \\
&= \frac{1}{n^2} \sum_{i=1}^n \mathbb{V} \left[ \frac{W_i Y_i}{e(X_i)} - \frac{(1 - W_i) Y_i}{1 - e(X_i)} \right] && \text{i.i.d. observations} \\
&= \frac{1}{n^2} \sum_{i=1}^n \left( \mathbb{E} \left[ \left( \frac{W_i Y_i}{e(X_i)} - \frac{(1 - W_i) Y_i}{1 - e(X_i)} \right)^2 \right] - \tau^2 \right) && \text{unbiasedness} \\
&= \frac{1}{n^2} \sum_{i=1}^n \left( \mathbb{E} \left[ \left( \frac{W_i Y_i}{e(X_i)} \right)^2 \right] + \mathbb{E} \left[ \left( \frac{(1 - W_i) Y_i}{1 - e(X_i)} \right)^2 \right] - 2\mathbb{E}[0] - \tau^2 \right) \\
&= \frac{1}{n^2} \sum_{i=1}^n \left( \mathbb{E} \left[ \left( \frac{W_i Y_i}{e(X_i)} \right)^2 \right] + \mathbb{E} \left[ \left( \frac{(1 - W_i) Y_i}{1 - e(X_i)} \right)^2 \right] - \tau^2 \right)
\end{aligned}$$

$$\begin{aligned}
\mathbb{E} \left[ \left( \frac{W_i Y_i}{e(X_i)} \right)^2 \right] &= \mathbb{E} \left[ \mathbb{E} \left[ \left( \frac{W_i Y_i}{e(X_i)} \right)^2 \mid X_i \right] \right] \\
&= \mathbb{E} \left[ \frac{\mathbb{E} [W_i Y_i \mid X_i]^2}{e(X_i)^2} \right] \\
&= \mathbb{E} \left[ \frac{\mathbb{E} [W_i \mid X_i] \mathbb{E} [Y_i^2 \mid X_i, W_i = 1]}{e(X_i)^2} \right] && \text{Assumption 1} \\
&= \mathbb{E} \left[ \frac{\mathbb{E} [Y_i(1)^2 \mid X_i]}{e(X_i)} \right] && \text{SUTVA} \\
&= \mathbb{E} \left[ \frac{Y_i(1)^2}{e(X_i)} \right]
\end{aligned}$$

Similarly,  $\mathbb{E} \left[ \left( \frac{(1 - W_i) Y_i}{1 - e(X_i)} \right)^2 \right] = \mathbb{E} \left[ \frac{Y_i(0)^2}{1 - e(X_i)} \right]$ . Then, the variance of the oracle multi-site IPW estimator is

$$\mathbb{V} [\hat{\tau}^{\text{IPW}^*}] = \frac{1}{n} \left( \mathbb{E} \left[ \frac{Y_i(1)^2}{e(X_i)} \right] + \mathbb{E} \left[ \frac{Y_i(0)^2}{1 - e(X_i)} \right] - \tau^2 \right).$$

For the variance of the oracle multi-site centralized AIPW, we first notice that since  $\mathbb{E}[Y_i(w) - \mu_w(X_i) | X_i] = 0$  for treatment  $w$ , we have

$$\begin{aligned}
A &= \text{Cov} \left( \tau(X_i), \frac{W_i(Y_i - \mu_1(X_i))}{e(X_i)} - \frac{(1 - W_i)(Y_i - \mu_0(X_i))}{1 - e(X_i)} \right) \\
&= \mathbb{E} \left[ \tau(X_i) \left( \frac{W_i(Y_i - \mu_1(X_i))}{e(X_i)} - \frac{(1 - W_i)(Y_i - \mu_0(X_i))}{1 - e(X_i)} \right) \right] \\
&= \mathbb{E} \left[ \tau(X_i) \mathbb{E} \left[ \frac{W_i(Y_i - \mu_1(X_i))}{e(X_i)} | X_i \right] \right] - \mathbb{E} \left[ \tau(X_i) \mathbb{E} \left[ \frac{(1 - W_i)(Y_i - \mu_0(X_i))}{1 - e(X_i)} | X_i \right] \right] \\
&= \mathbb{E} \left[ \tau(X_i) \frac{e(X_i) \mathbb{E}[Y_i(1) - \mu_1(X_i) | X_i]}{e(X_i)} \right] - \mathbb{E} \left[ \tau(X_i) \frac{(1 - e(X_i)) \mathbb{E}[Y_i(0) - \mu_0(X_i) | X_i]}{1 - e(X_i)} \right] \\
&= 0
\end{aligned}$$

Then,

$$\begin{aligned}
\mathbb{V}[\hat{\tau}^{\text{AIPW}^*}] &= \mathbb{V} \left[ \frac{1}{n} \sum_{i=1}^n \left( \mu_1(X_i) - \mu_0(X_i) + \frac{W_i(Y_i - \mu_1(X_i))}{e(X_i)} - \frac{(1 - W_i)(Y_i - \mu_0(X_i))}{1 - e(X_i)} \right) \right] \\
&= \frac{1}{n^2} \sum_{i=1}^n \left( \mathbb{V}[\tau(X_i)] + \mathbb{V} \left[ \frac{W_i(Y_i - \mu_1(X_i))}{e(X_i)} \right] + \mathbb{V} \left[ \frac{(1 - W_i)(Y_i - \mu_0(X_i))}{1 - e(X_i)} \right] + 2A \right) \\
&= \frac{1}{n^2} \sum_{i=1}^n \left( \mathbb{V}[\tau(X_i)] + \mathbb{E} \left[ \left( \frac{(Y_i - \mu_1(X_i))^2}{e(X_i)} \right) \right] + \mathbb{E} \left[ \left( \frac{(Y_i - \mu_0(X_i))^2}{1 - e(X_i)} \right)^2 \right] \right) \\
&= \frac{1}{n} \left( \mathbb{V}[\tau(X_i)] + \mathbb{E} \left[ \left( \frac{(Y - \mu_1(X_i))^2}{e(X_i)} \right) \right] + \mathbb{E} \left[ \left( \frac{(Y - \mu_0(X_i))^2}{1 - e(X_i)} \right)^2 \right] \right)
\end{aligned}$$

Because these variances are well defined from what precedes, the Central Limit Theorem can be applied to  $\hat{\tau}^{\text{IPW}^*}$  and  $\hat{\tau}^{\text{AIPW}^*}$ , which gives the result in Theorem 1.  $\square$

## A.3 PROOF OF THEOREM 2

*Proof. Unbiasedness.* We first prove the unbiasedness of the local IPW estimators:

$$\begin{aligned}
\mathbb{E} \left[ \hat{\tau}_k^{\text{IPW}^*} \mid H_i = k \right] &= \mathbb{E} \left[ \frac{W_i Y_i}{e_k(X_i)} - \frac{(1 - W_i) Y_i}{1 - e_k(X_i)} \mid H_i = k \right] && \text{i.i.d.} \\
&= \mathbb{E} \left[ \frac{W_i Y_i(1)}{e_k(X_i)} - \frac{(1 - W_i) Y_i(0)}{1 - e_k(X_i)} \mid H_i = k \right] && \text{SUTVA} \\
&= \mathbb{E} \left[ \mathbb{E} \left[ \frac{W_i Y_i(1)}{e_k(X_i)} \mid H_i = k, X_i \right] - \mathbb{E} \left[ \frac{(1 - W_i) Y_i(0)}{1 - e_k(X_i)} \mid X_i, H_i = k, X_i \right] \right] \\
&= \mathbb{E} \left[ \frac{\mathbb{E} [W_i \mid H_i = k, X_i] \mathbb{E} [Y_i(1) \mid H_i = k, X_i]}{e_k(X_i)} \right. \\
&\quad \left. - \frac{\mathbb{E} [(1 - W_i) \mid H_i = k, X_i] \mathbb{E} [Y_i(0) \mid H_i = k, X_i]}{1 - e_k(X_i)} \right] && \text{Ass. 1} \\
&= \mathbb{E} [\mathbb{E} [Y_i(1) - Y_i(0) \mid H_i = k, X_i]] \\
&= \mathbb{E} [Y_i(1) - Y_i(0) \mid H_i = k] \\
&= \tau_k
\end{aligned}$$

This yields

$$\begin{aligned}
\mathbb{E} \left[ \hat{\tau}^{\text{meta-IPW}^*} \right] &= \mathbb{E} \left[ \sum_{k=1}^K \frac{n_k}{n} \hat{\tau}_k^{\text{IPW}^*} \right] \\
&= \sum_{k=1}^K \mathbb{E} \left[ \frac{n_k}{n} \mathbb{E} \left[ \hat{\tau}_k^{\text{IPW}^*} \mid H_i = k \right] \right] \\
&= \sum_{k=1}^K \mathbb{E} \left[ \frac{n_k}{n} \right] \tau_k \\
&= \sum_{k=1}^K \rho_k \tau_k \\
&= \tau
\end{aligned}$$

**Variance.** For the variance of the local ATEs, we follow the same steps as in the previous proof which yields

$$\begin{aligned}
\mathbb{V} \left[ \hat{\tau}_k^{\text{IPW}^*} \mid H = k \right] &= \frac{1}{n_k} \left( \mathbb{E} \left[ \frac{Y_i(1)^2}{e_k(X_i)} \mid H_i = k \right] - \mathbb{E} \left[ \frac{Y_i(0)^2}{1 - e_k(X_i)} \mid H_i = k \right] - \tau_k^2 \right) \\
&= \frac{1}{n_k} V_{k,i}
\end{aligned}$$

with  $V_{k,i} = \mathbb{E} \left[ \frac{Y_i(1)^2}{e_k(X_i)} \mid H_i = k \right] - \mathbb{E} \left[ \frac{Y_i(0)^2}{1 - e_k(X_i)} \mid H_i = k \right] - \tau_k^2$ . Finally, by Lemma 2, we have

$$\begin{aligned}
\mathbb{V} \left[ \hat{\tau}^{\text{meta-IPW}^*} \right] &= \mathbb{E} \left[ \mathbb{V} \left[ \hat{\tau}^{\text{meta-IPW}^*} \mid H = k \right] \right] + \mathbb{V} \left[ \mathbb{E} \left[ \hat{\tau}^{\text{meta-IPW}^*} \mid H = k \right] \right] \\
&= \frac{1}{n} \sum_{k=1}^K \rho_k V_{k,i} + \frac{1}{n} \mathbb{V} [\tau_H]
\end{aligned}$$

The proof for the meta AIPW estimator follows the same steps.  $\square$

#### A.4 DECOMPOSITION OF THE GLOBAL PROPENSITY SCORE

By the law of total probabilities,

$$\begin{aligned}
 e(x) &= \mathbb{P}(W_i = 1 \mid X_i) \\
 &= \sum_{k=1}^K \mathbb{P}(W_i = 1 \cap H_i = k \mid X_i = x) \\
 &= \sum_{k=1}^K \mathbb{P}(H_i = k \mid X_i) \mathbb{P}(W_i = 1 \mid X_i = x, H_i = k) \\
 &= \sum_{k=1}^K \mathbb{P}(H_i = k) \frac{P(X_i \mid H_i = k)}{P(X_i)} e_k(X_i) \quad (\text{DW}) \\
 &= \sum_{k=1}^K \mathbb{P}(H_i = k \mid X_i) e_k(X_i) \quad (\text{MW})
 \end{aligned}$$

#### A.5 PROOF OF THEOREM 3

*Proof.*

$$\begin{aligned}
 \hat{\tau}_{\text{IPW}}^{\text{fed}^*} &= \sum_{k=1}^K \frac{n_k}{n} \hat{\tau}_{\text{IPW}}^{\text{fed}(k)} \\
 &= \sum_{k=1}^K \frac{n_k}{n} \left( \frac{1}{n_k} \sum_{i=1}^{n_k} \left( \frac{W_i Y_i}{e(X_i)} - \frac{(1 - W_i) Y_i}{1 - e(X_i)} \right) \right) \\
 &= \frac{1}{n} \sum_{i=1}^n \left( \frac{W_i Y_i}{e(X_i)} - \frac{(1 - W_i) Y_i}{1 - e(X_i)} \right) \\
 &= \hat{\tau}^{\text{IPW}^*}
 \end{aligned}$$

We prove similarly that  $\hat{\tau}_{\text{AIPW}}^{\text{fed}^*} = \hat{\tau}^{\text{AIPW}^*}$ .  $\square$

#### A.6 PROOF OF THEOREM 4

We start by two technical lemmas.

**Lemma 1.** *Under Assumptions 1 and 2 we have,*

$$\mathbb{V}[\tau(X_i)] = \sum_{k=1}^K \rho_k \mathbb{V}[\tau(X_i) \mid H_i = k] + \mathbb{V}[\tau_{H_i}]$$

*Proof.*

$$\begin{aligned}
 \mathbb{V}[\tau(X_i)] &= \mathbb{E}[\mathbb{V}[\tau(X_i) \mid H_i = k]] + \mathbb{V}[\mathbb{E}[\tau(X_i) \mid H_i = k]] \\
 &= \sum_{k=1}^K \rho_k \mathbb{V}[\tau(X_i) \mid H_i = k] + \mathbb{V}\left[\sum_{k=1}^K \mathbb{1}_{[H_i=k]} \tau_{H_i}\right] \\
 &= \sum_{k=1}^K \rho_k \mathbb{V}[\tau(X_i) \mid H_i = k] + \mathbb{V}\left[\tau_{H_i} \sum_{k=1}^K \mathbb{1}_{[H_i=k]}\right] \\
 &= \sum_{k=1}^K \rho_k \mathbb{V}[\tau(X_i) \mid H_i = k] + \mathbb{V}[\tau_{H_i}]
 \end{aligned}$$

□

**Lemma 2.** For the general form of meta-analysis estimator  $\hat{\tau}^{\text{meta}} = \sum_{k=1}^K \frac{n_k}{n} \hat{\tau}_k$ , we have

$$\mathbb{V}[\hat{\tau}^{\text{meta}}] = \frac{1}{n} \sum_{k=1}^K \rho_k V_k + \frac{1}{n} \mathbb{V}[\tau_{H_i}]$$

where  $V_k = \mathbb{V}[\hat{\tau}_k \mid H_i = k]$  is the within-site variance of the ATE estimator in site  $k$  and  $\mathbb{V}[\tau_{H_i}] = \mathbb{V}[\mathbb{E}[Y_i(1) - Y_i(0) \mid H_i]]$  is the between-sites variance of the local ATEs.

*Proof.*

$$\begin{aligned}
 \mathbb{V}[\hat{\tau}^{\text{meta}}] &= \mathbb{V}\left[\sum_{k=1}^K \frac{n_k}{n} \hat{\tau}_k\right] \\
 &= \mathbb{E}\left[\mathbb{V}\left[\sum_{k=1}^K \frac{n_k}{n} \hat{\tau}_k \mid H_1, \dots, H_n\right]\right] + \mathbb{V}\left[\mathbb{E}\left[\sum_{k=1}^K \frac{n_k}{n} \hat{\tau}_k \mid H_1, \dots, H_n\right]\right] \\
 &= \mathbb{E}\left[\sum_{k=1}^K \frac{n_k}{n^2} V_k\right] + \mathbb{V}\left[\sum_{k=1}^K \frac{n_k}{n} \tau_k\right] = \frac{1}{n} \sum_{k=1}^K \rho_k V_k + \mathbb{V}\left[\frac{1}{n} \sum_{k=1}^K \sum_{i=1}^n \mathbb{1}_{[H_i=k]} \tau_{H_i}\right] \\
 &= \frac{1}{n} \sum_{k=1}^K \rho_k V_k + \mathbb{V}\left[\frac{1}{n} \sum_{i=1}^n \tau_{H_i} \sum_{k=1}^K \mathbb{1}_{[H_i=k]}\right] = \frac{1}{n} \sum_{k=1}^K \rho_k V_k + \mathbb{V}\left[\frac{1}{n} \sum_{i=1}^n \tau_{H_i}\right] \\
 &= \frac{1}{n} \sum_{k=1}^K \rho_k V_k + \frac{1}{n} \mathbb{V}[\tau_H],
 \end{aligned}$$

□

We can now prove Theorem 4.

*Proof.* We begin with IPW estimators, and then move to AIPW.

**IPW.** First, applying Jensen's inequality with the strictly convex function  $t \mapsto \frac{1}{t}$  in  $]0; 1[$  and summing-to-one weights  $\omega_k(X) = \mathbb{P}(H_i = k \mid X_i = X)$ , we have

$$\mathbb{E} \left[ \frac{Y_i(1)^2}{e(X_i)} \right] < \mathbb{E} \left[ \sum_{k=1}^K \omega_k(X) \frac{Y_i(1)^2}{e_k(X_i)} \right]$$

if  $\exists(k, k') \in [K], e_k(X_i) \neq e_{k'}(X_i)$ , and equality if  $\forall k \in [K], e_k(X_i) = e(X_i)$ , i.e. if the local propensity scores are all equal to one another. Then, with the same condition on strictness and equality,

$$\begin{aligned} \mathbb{E} \left[ \frac{Y_i(1)^2}{e(X_i)} \right] &\leq \mathbb{E} \left[ \sum_{k=1}^K \omega_k(X) \frac{Y_i(1)^2}{e_k(X_i)} \right] \\ &= \sum_{k=1}^K \mathbb{E} \left[ \mathbb{E} \left[ \mathbb{1}_{[H_i=k]} \mid X_i \right] \frac{Y_i(1)^2}{e_k(X_i)} \right] \\ &= \sum_{k=1}^K \mathbb{E} \left[ \mathbb{E} \left[ \mathbb{1}_{[H_i=k]} \frac{Y_i(1)^2}{e_k(X_i)} \mid X_i \right] \right] \\ &= \sum_{k=1}^K \mathbb{E} \left[ \mathbb{1}_{[H_i=k]} \frac{Y_i(1)^2}{e_k(X_i)} \right] \\ &= \sum_{k=1}^K \rho_k \mathbb{E} \left[ \frac{Y_i(1)^2}{e_k(X_i)} \mid H_i = k \right] \end{aligned}$$

Similarly,  $\mathbb{E} \left[ \frac{Y_i(0)^2}{1-e(X_i)} \right] \leq \sum_{k=1}^K \rho_k \mathbb{E} \left[ \frac{Y_i(0)^2}{1-e_k(X_i)} \mid H_i = k \right]$ . Then,

$$\begin{aligned} \mathbb{V} \left[ \hat{\tau}^{\text{fed-IPW}^*} \right] &= \frac{1}{n} \left( \mathbb{E} \left[ \frac{Y_i(1)^2}{e(X_i)} + \frac{Y_i(0)^2}{1-e(X_i)} \right] - \tau^2 \right) \\ &\leq \frac{1}{n} \left( \sum_{k=1}^K \rho_k \mathbb{E} \left[ \frac{Y_i(1)^2}{e_k(X_i)} + \frac{Y_i(0)^2}{1-e_k(X_i)} \mid H_i = k \right] - \tau^2 \right) \\ &\leq \frac{1}{n} \left( \sum_{k=1}^K \rho_k \underbrace{\left( \mathbb{E} \left[ \frac{Y_i(1)^2}{e_k(X_i)} + \frac{Y_i(0)^2}{1-e_k(X_i)} \mid H_i = k \right] - \tau_k^2 \right)}_{:=V_k} + \underbrace{\sum_{k=1}^K \rho_k \tau_k^2 - \tau^2}_{\mathbb{V}(\tau_H)} \right) \end{aligned}$$

On the other hand, by Lemma 2,

$$\mathbb{V} \left[ \hat{\tau}^{\text{meta-IPW}^*} \right] = \frac{1}{n} \sum_{k=1}^K \rho_k V_k + \frac{1}{n} \mathbb{V}[\tau_H].$$

Hence  $\mathbb{V} \left[ \hat{\tau}^{\text{fed-IPW}^*} \right] = \mathbb{V} \left[ \hat{\tau}^{\text{meta-IPW}^*} \right]$  if  $\forall k \in [K], e_k = e$ , and  $\mathbb{V} \left[ \hat{\tau}^{\text{fed-IPW}^*} \right] < \mathbb{V} \left[ \hat{\tau}^{\text{meta-IPW}^*} \right]$  if  $\exists(k, k') \in [K]^2, e_k \neq e_{k'}$ .



**AIPW.** Similarly to the IPW case, we have

$$\begin{aligned}\mathbb{E} \left[ \left( \frac{W_i(Y_i - \mu_1)}{e(X_i)} \right)^2 \right] &\leq \sum_{k=1}^K \rho_k \mathbb{E} \left[ \left( \frac{W_i(Y_i - \mu_1)}{e_k(X_i)} \right)^2 \mid H_i = k \right], \\ \mathbb{E} \left[ \left( \frac{(1 - W_i)(Y_i - \mu_0)}{1 - e(X_i)} \right)^2 \right] &\leq \sum_{k=1}^K \rho_k \mathbb{E} \left[ \left( \frac{(1 - W_i)(Y_i - \mu_0)}{1 - e_k(X_i)} \right)^2 \mid H_i = k \right],\end{aligned}$$

and with Lemma 1:

$$\mathbb{V}[\tau(X_i)] = \sum_{k=1}^K \rho_k \mathbb{V}[\tau(X_i) \mid H_i = k] + \mathbb{V}[\tau_{H_i}].$$

Then, using Lemma 2, we have the desired result.  $\square$

## A.7 PROOF OF THEOREM 5

*Proof.* Let  $f : t \mapsto \frac{1}{t(1-t)}$  with  $t \in ]0, 1[$ .  $f$  is convex. Then by Jensen's inequality with summing-to-one weights  $\omega_k(X) = \mathbb{P}(H_i = k \mid X_i = X)$ ,

$$\begin{aligned}\mathcal{O}_{\text{global}} &= \mathbb{E} \left[ f \left( \sum_{k=1}^K \omega_k(X) e_k(X) \right) \right] \\ &\leq \mathbb{E} \left[ \sum_{k=1}^K \omega_k(X) f(e_k(X)) \right] \\ &\leq \sum_{k=1}^K \mathbb{E} [\mathbb{E} [\mathbb{1}_{[H_i=k|X_i]}] f(e_k(X))] \\ &\leq \sum_{k=1}^K \mathbb{E} [\mathbb{E} [\mathbb{1}_{[H_i=k]} f(e_k(X)) \mid X_i]] \\ &\leq \sum_{k=1}^K \mathbb{E} [\mathbb{1}_{[H_i=k]} f(e_k(X))] \\ &\leq \sum_{k=1}^K \mathbb{P}(H_i = k) \mathbb{E} [f(e_k(X)) \mid H_i = k] \\ &\leq \sum_{k=1}^K \rho_k \mathcal{O}_k\end{aligned}$$

$\square$

## B FEDERATED LEARNING OF MEMBERSHIP WEIGHTS

We describe how to estimate membership weights using a general parametric classification model trained via Federated Averaging (FedAvg) (McMahan et al., 2017), without requiring access to individual-level data.

### B.1 GENERAL PARAMETRIC MODEL

Let  $\omega_k(X_i; \Theta)$  denote the predicted membership probability of site  $k$  for covariate vector  $X_i$ , where  $\Theta$  are the parameters of the model. For a generic parametric model,  $\Theta \in \mathbb{R}^p$  represents all trainable parameters,

and the membership-weight estimation problem can be formulated as the minimization of the empirical cross-entropy loss:

$$\ell(\Theta; \mathcal{D}) = -\frac{1}{|\mathcal{D}|} \sum_{i=1}^{|\mathcal{D}|} \sum_{k=1}^K \mathbb{1}_{\{H_i=k\}} \log(\omega_k(X_i; \Theta)),$$

where  $H_i^{\text{enc}}$  is the one-hot encoding of the site membership  $H_i$ .

This framework encompasses a broad class of models, including:

- Multinomial logistic regression:

$$\omega_k(X_i; \Theta) = \frac{\exp(\theta_k^\top X_i)}{\sum_{k'=1}^K \exp(\theta_{k'}^\top X_i)}, \quad \Theta = (\theta_1, \dots, \theta_K) \in \mathbb{R}^{d \times K}.$$

- Neural networks, where  $\Theta$  includes all weights and biases across layers, and  $\omega_k(\cdot)$  is the softmax output of the final layer.

## B.2 FEDERATED TRAINING PROCEDURE

Algorithm 1 summarizes the generic Federated Averaging procedure for training any differentiable parametric model for membership-weight estimation.

---

### Algorithm 1 Federated Learning of Membership Weights with a Parametric Model

---

```

1: Input:  $K$  sites,  $E$  local steps,  $\eta$  learning rate,  $T$  communication rounds,  $B$  batch size
2: Initialize global parameters  $\Theta_0$ 
3: for  $t = 1$  to  $T$  do
4:   for each client  $k \in [1, \dots, K]$  in parallel do
5:      $\Theta_{t+1}^{(k)} \leftarrow \text{LOCALUPDATE}(k, \Theta_t)$ 
6:   end for
7:    $\Theta_{t+1} \leftarrow \sum_{k=1}^K \frac{n_k}{n} \Theta_{t+1}^{(k)}$  // FedAvg aggregation
8: end for
9: LocalUpdate( $k, \Theta$ ):
10: for  $e = 1$  to  $E$  do
11:   Sample mini-batch  $\mathcal{B}_k$  of size  $B$  from local data  $\mathcal{D}_k$ 
12:   Compute predictions  $\omega_k(X_i; \Theta)$  for  $i \in \mathcal{B}_k$ 
13:   Compute gradient  $\nabla \ell(\Theta; \mathcal{B}_k)$  of cross-entropy loss
14:    $\Theta \leftarrow \Theta - \eta \nabla \ell(\Theta; \mathcal{B}_k)$ 
15: end for
16: return  $\Theta$ 

```

---

With a suitable choice of learning rate  $\eta$  and a moderate number of local steps  $E$  per round, Algorithm 1 converges to the same solution as centralized training as the number of communication rounds  $T \rightarrow \infty$  (Stich, 2019; Khaled et al., 2020; Li et al., 2019). The resulting federated parameter estimate  $\hat{\Theta}^{\text{fed}}$  yields the estimated membership weights:

$$\hat{\omega}_k(X_i) = \omega_k(X_i; \hat{\Theta}^{\text{fed}}), \quad k = 1, \dots, K.$$

### B.3 EXAMPLE: MULTINOMIAL LOGISTIC REGRESSION

When the model is multinomial logistic regression,  $\Theta$  is the matrix of regression coefficients, and the local gradient takes the closed form:

$$\nabla \ell(\Theta; \mathcal{B}_k) = \frac{1}{B} \sum_{i \in \mathcal{B}_k} X_i (\omega(X_i; \Theta) - H_i^{\text{enc}}),$$

where  $\omega(X_i; \Theta)$  is the vector of softmax probabilities for sample  $i$ .

## C SIMULATION DETAILS

The parameters common to all settings are shown in Table 1, where  $\gamma_2^{(\text{weak})} = [-2.5, -1, -0.15, -0.15, 0, -0.15, -1, -0.15, -0.15, 0]$  and  $\gamma_2^{(\text{good})} = [-.05, -.1, -.05, -.1, .05, -.1, -.05, -.1, .05, -.1]$ .

Parameter	Center 1	Center 2	Center 3
$d$	10		
$\mu_1(X)$	$\sum_{j=1}^5 \frac{j}{10} X_j^2 + \sum_{j=6}^{10} \frac{j}{10} X_j + X_9 * X_{10}$		
$\mu_0(X)$	$\sum_{j=1}^5 \frac{3j-10}{30} X_j^2 + \sum_{j=6}^{10} \frac{3j-10}{30} X_j + X_1 * X_{10}$		
$e_k$	Logistic( $x, \gamma_k$ )		
$\gamma_k$	$[-.25, .25, -.25, -.25, .25, -.25, -.25, .25, -.25, .25]$	No overlap: not logistic (only controls) Weak overlap: $\gamma_2^{(\text{weak})}$ Good overlap: $\gamma_2^{(\text{good})}$	$[.15, -.15, .15, -.15, .15, -.15, .15, -.15, .15, -.15]$

Table 1: Common simulation parameters.

DGP A-specific settings are shown in Table 2, where  $J_d$  is the  $d \times d$  matrix of ones, and  $I_d$  is the  $d \times d$  identity matrix.

Parameter	Center 1	Center 2	Center 3
$n_k$	650		
$\mathcal{D}_k$	$\mathcal{N}(\mu_k, \Sigma_k)$		
$\mu_k$	$(1, \dots, 1) \in \mathbb{R}^d$	$(1.5, 1.5, 1.5, 1, \dots, 1) \in \mathbb{R}^d$	$(2, 2, 2, 1, \dots, 1) \in \mathbb{R}^d$
$\Sigma_k$	$I_d + 0.5J_d$	$0.6I_d + 0.4J_d$	$3I_d + 0.3J_d$

Table 2: Simulation parameters specific to DGP A.

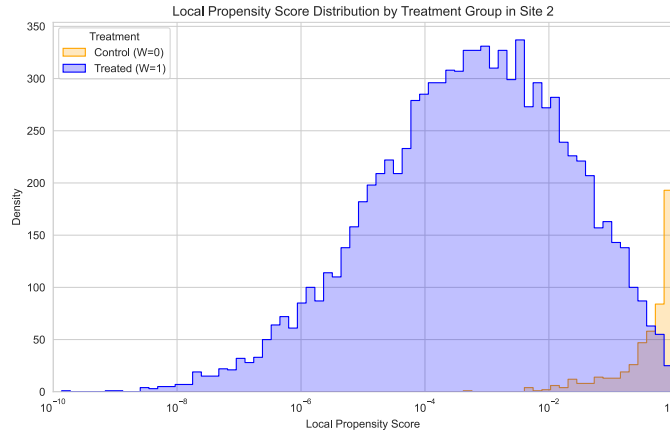
DGP B-specific settings are shown in Table 3.

Parameter	Value
$n$	4000
$\mathcal{D}$	$\frac{2}{3}\mathcal{N}(\mu_1, \Sigma_1) + \frac{1}{3}\mathcal{N}(\mu_2, \Sigma_2)$
$\mu_1$	$(0, \dots, 0) \in \mathbb{R}^d$
$\mu_2$	$(1.5, \dots, 1.5) \in \mathbb{R}^d$
$\Sigma_1$	$I_d$
$\Sigma_2$	$I_d + 0.5J_d$
$\mathbb{P}(H_i = k   X)$	$\text{Logistic}(x, \theta_k)$
$\theta_1$	$[-0.5, -0.5, 0.2, -0.5, -0.5, 0.2, -0.5, -0.5, 0.2, 0.2]$
$\theta_2$	$[0.5, 0.5, 0.2, 0.5, 0.5, 0.2, 0.5, 0.5, 0.2, 0.5]$
$\theta_3$	$[1, 1, 0.2, 0.2, 0.2, 0.2, 0.2, 0.2, 0.2, 0.2]$

Table 3: Simulation parameters specific to DGP B.

## D OVERLAP IMPROVEMENT

Echoing theorem 5, Figures 5 and 6 display the empirical distributions (on a log-scale) of the local propensity score in site 2 ( $e_2$ ) and of the global propensity score  $e$  in the *Poor local overlap* scenario (see Figure 2b for corresponding results). A good overlap is when both the propensity score distributions for the treated and control overlap, and are far from 0. We see that the poor overlap at site 2 (Figure 5), with values of  $e_2$  on the local data close to 0, is significantly improved at the global level (Figure 6).

Figure 5: Local overlap in site 2 for the *Poor local overlap* scenario (DGP A).

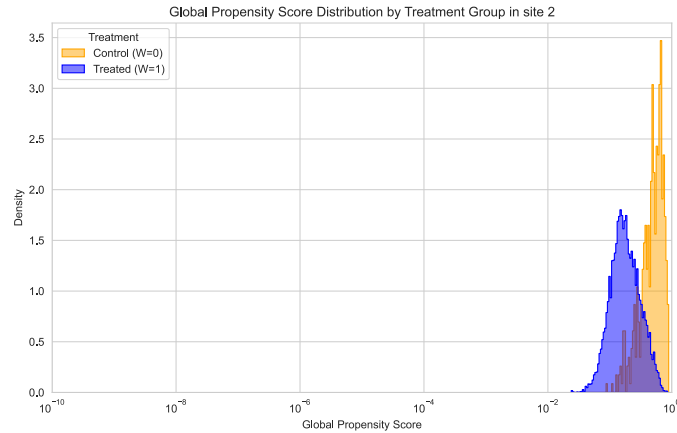


Figure 6: Global overlap for the *Poor local overlap* scenario (DGP A). We see a clear improvement compared to the local overlap in site 2 (Figure 5).

## E ADDITIONAL SIMULATIONS

### E.1 LARGE NUMBER OF SITES

We further investigate the robustness of our estimators as the number of sites increases. Specifically, we simulate  $K = 20$  sites under weak local overlap (corresponding to the setting of Figure 2b in the main paper) and report the bias and mean squared error (MSE) for both data-generating processes (DGP A and DGP B).

Table 4: Bias and mean squared error (MSE) under  $K = 20$  sites with weak local overlap. Bold values indicate unbiasedness or lowest MSE.

Method	Bias (DGP A)	MSE (DGP A)	Bias (DGP B)	MSE (DGP B)
Oracle IPW	0.00	0.298	0.00	0.130
Fed. IPW-MW (logistic)	0.48	0.294	<b>0.00</b>	<b>0.095</b>
Fed. IPW-MW (NN)	<b>0.00</b>	0.240	<b>0.00</b>	0.123
Fed. IPW-DW	<b>0.00</b>	<b>0.096</b>	0.50	0.450
Meta-SW IPW	-2.09	13.279	-1.77	7.977
Oracle AIPW	0.00	0.007	0.00	0.049
Fed. AIPW-MW (logistic)	0.79	0.639	<b>0.00</b>	<b>0.078</b>
Fed. AIPW-MW (NN)	<b>0.00</b>	0.153	<b>0.00</b>	0.084
Fed. AIPW-DW	<b>0.00</b>	<b>0.019</b>	0.25	0.194
Meta-SW AIPW	-1.24	1.775	-2.36	18.465

As shown in Table 4, estimators using neural network membership weights consistently achieve unbiasedness across both DGPs, although with slightly higher variance compared to their well-specified parametric counterparts. This is consistent with the behavior of more flexible models, which trade off a small increase

in variance for robustness to misspecification. When parametric assumptions are correctly specified (e.g., Gaussian density weights in DGP A, logistic membership weights in DGP B), these methods achieve the lowest MSE. Overall, the NN-based approach demonstrates robust consistency across DGPs and scales well with the number of sites.

## E.2 FULLY MODEL-AGNOSTIC ATE ESTIMATION

We additionally investigate the benefits of fully model-agnostic estimation for membership weights  $P(H | X)$ . In this setting, we use federated neural networks to estimate membership probabilities in a flexible way and pair them with non-parametric local propensity score estimators (e.g., Random Forests). This procedure does not require prior parametric knowledge about  $H | X$  or  $X | H$ , making it robust to a wide range of DGPs.

We simulate  $K = 3$  sites with  $d = 10$  covariates,  $n_k = 600$  observations per site, and  $X | H = k \sim D_k$  as a mixture of Gaussians with weak local overlap. The results are summarized in Table 5.

Method	Bias	MSE
Oracle IPW	0.00	0.116
<b>Fed. IPW-MW (NN MW + RF <math>e_k</math>)</b>	<b>0.00</b>	<b>0.075</b>
Meta-SW IPW (RF $e_k$ )	0.89	0.884
Oracle AIPW	0.00	0.030
<b>Fed. AIPW-MW (NN MW + RF <math>e_k</math>)</b>	<b>0.00</b>	<b>0.040</b>
Meta-SW AIPW (RF $e_k$ )	-0.36	0.180

Table 5: Bias and MSE under fully non-parametric estimation of membership weights and local propensity scores ( $K = 3$ , weak local overlap). Bold values indicate unbiased (or nearly unbiased) estimators.

As shown in Table 5, our federated approach with neural-network-based membership weights and random-forest local propensities yields nearly unbiased estimates with substantially lower MSE compared to meta-analysis estimators. This demonstrates that misspecifications in either the membership weights or the local propensity scores can be effectively mitigated by adopting fully non-parametric methods in a federated learning setting.

## E.3 ROBUSTNESS TO LOCAL MISSPECIFICATIONS

Similar to meta-analysis estimators, we argue that inconsistencies in local propensity models can be mitigated as  $K$  grows, provided that such inconsistencies are marginal relative to the number of correctly specified sites. To illustrate this, we simulate  $K$  sites under DGP B, each holding  $n_k = 60$  observations drawn from homogeneous covariate distributions and sharing the same true propensity model (logistic). We then intentionally misspecify the propensity model for site  $k = 1$ , setting  $\hat{e}_1(x) \equiv 0.1$  for all  $x$ .

Table 6 reports the bias and mean squared error (MSE) of several estimators under two scenarios:  $K = 3$  (few sites) and  $K = 200$  (many sites).

Method	Bias ( $K = 3$ )	MSE ( $K = 3$ )	Bias ( $K = 200$ )	MSE ( $K = 200$ )
Oracle IPW	0.00	0.272	0.00	0.002
Fed. IPW (logistic MW)	0.38	0.344	0.00	0.002
Meta IPW	2.61	7.902	0.03	0.003
Oracle AIPW	0.00	0.058	0.00	0.001
Fed. AIPW (logistic MW)	-0.33	0.325	-0.01	0.002
Meta AIPW	-0.51	0.481	-0.07	0.007

Table 6: Bias and mean squared error (MSE) under local misspecification of site  $k = 1$  propensity model.

With only  $K = 3$  sites, both federated and meta-analysis estimators exhibit substantial bias. However, as the number of sites increases, the influence of the single misspecified site becomes negligible, and the bias decreases markedly (e.g., from 0.38 to 0.00 for Federated IPW with logistic membership weights). As expected, the AIPW estimator retains its double-robustness property even under misspecification of the outcome model, leading to consistently low bias and MSE across settings.