Supplementary material for: Double machine learning and design in batch adaptive experiments

Harrison H. Li
Stanford University

Art B. Owen
Stanford University

All numbered equations correspond to those in the main text.

A Nonstationary batches

Assumption 2.1 in the main text supposes the distribution P^S of the covariates and potential outcomes $S_{ti} = (X_{ti}, Y_{ti}(0), Y_{ti}(1))$ is stationary across batches $t = 1, \dots, T$. Here we relax that assumption. Let P_t^S be the distribution of the vector S_{ti} in batch $t = 1, \dots, T$, now allowed to vary across batches (in the main text it is assumed that $P_1^S = \cdots = P_T^S = P^S$). Then we have the following relaxation of Assumption 2.1:

Assumption A.1 (Relaxation of Assumption 2.1)**.** *For some fixed number of batches $T \geqslant 2$, the vectors*

$$S_{ti} = (X_{ti}, Y_{ti}(0), Y_{ti}(1)), \quad 1 \leqslant t \leqslant T, \quad 1 \leqslant i \leqslant N_t$$

are mutually independent such that for each batch $t = 1, \dots, T$, we have $S_{ti} \sim P_t^S$. Furthermore, the sample sizes N_t satisfy (3)*, and the vector $W_{ti} = (X_{ti}, Z_{ti}, Y_{ti})$ is observed where the outcomes Y_{ti} satisfy the SUTVA assumption* (1)*.*

Now letting P_0^S be the mixture distribution $\sum_{t=1}^T \kappa_t P_t^S$, we introduce the notation $\mathbb{E}_{t,e}[f(W)]$, which denotes an expectation under the distribution $P_{t,e} = P_{t,e}^W$ on $W = (X, Z, Y)$ induced by $S = (X, Y(0), Y(1)) \sim P_t^S$ and $Z \mid X \sim \mathrm{Bern}(e(X))$ for any propensity $e(\cdot)$ and $t = 0, 1, \dots, T$. The notation P_t^X refers to the corresponding marginal distribution of the covariates X. Then the score equation (5) will be generalized to

$$\mathbb{E}_{t,e}[s(W; \theta_0, \nu_0, e')] = 0, \quad \forall e, e' \in \mathcal{F}_\gamma, \quad t = 1, \dots, T, \tag{S1}$$

which we will require to identify θ_0 in each batch:

Assumption A.2 (Relaxation of Assumption 2.2)**.** *The estimand $\theta_0 \in \mathbb{R}^p$ of interest satisfies* (S1) *for some $\gamma \in [0, 1/2)$, some nuisance parameters ν_0 lying in a known convex set $\mathcal{N}$, and some score $s(\cdot)$ satisfying* (4)*.*

Equation (S1) encodes a requirement that the same parameters θ_0 and ν_0 satisfy the score equations for all batches $t = 1, \dots, T$; to ensure this, we require those parameters to be stationary across batches. For instance, for ATE estimation with the score $s_{\mathrm{AIPW}}(\cdot)$, we require the conditional mean functions $\mathbb{E}[Y(z) \mid X = x] = m_0(z, x)$ to be stationary across batches. For estimation under the partially linear model with the score s_{EPL}, we also require the outcome variance functions $\mathrm{Var}_t(Y(z) \mid X = x) = v_0(z, x)$ to remain stationary. However, in both cases the covariate distribution can otherwise vary arbitrarily across batches, as can higher moments of the conditional distributions of the potential outcomes $Y(z)$ given the covariates X.

Due to the possibility of covariate shift, the relevant mixture propensity scores are now

$$e_{0,N}(x) = \sum_{t=1}^{T} \frac{N_t}{N} e_t(x) \frac{\mathrm{d}P_t^X}{\mathrm{d}P_0^X}(x) \quad \text{and} \quad e_0(x) = \sum_{t=1}^{T} \kappa_t e_t(x) \frac{\mathrm{d}P_t^X}{\mathrm{d}P_0^X}(x). \tag{S2}$$

These definitions generalize (9) and (10). Here P_t^X denotes the marginal distribution of X when $S \sim P_t^S$. When there is no covariate shift, we have $\mathrm{d}P_t^X/\mathrm{d}P_0^X(x) = 1$ for all x and we recover (9) and (10) in the main text. The expressions in (S2) are derived using Bayes' rule as the conditional probability that $Z = 1$ given $X = x$ when (X, Z, Y) is drawn uniformly at random from the pooled collection of observations $\{W_{ti} \mid 1 \leqslant t \leqslant T, 1 \leqslant i \leqslant N_t\}$ in a non-adaptive batch experiment with propensities $e_1(\cdot), \ldots, e_T(\cdot)$.

Where indicated, we are able to generalize various results in Sections 3 and 4. The generalized results are as stated in the main text, if we make the following changes to the notation and assumptions:

1. Assumptions 2.1 and 2.2 are replaced by Assumptions A.1. and A.2, respectively
2. Any references to the mixture propensities in $e_{0,N}(\cdot)$ and $e_0(\cdot)$ correspond to the more general definitions in (S2), rather than (9) and (10).
3. Any expectations of the form $\mathbb{E}[f(W)]$ without subscripts are interpreted as being taken under the distribution P_{0,e_0} on W.
4. Any expectations of the form $\mathbb{E}_t[f(W)], t = 0, 1, \ldots, T$ are interpreted as being taken under the distribution P_{t,e_t} on W, and any expectations of the form $\mathbb{E}_{0,N}[f(W)]$ are interpreted as being taken under the distribution $P_{0,e_{0,N}}$ on W.

B Technical lemmas

Here we give some technical lemmas used in our proofs.

B.1 Asymptotics

Definition B.1. For any sequence of random vectors $\{X_n : n \geqslant 1\}$ and constants $a_n \downarrow 0$, we write $X_n = O_p(a_n)$ if $\lim_{M\to\infty} \limsup_{n\to\infty} \Pr(\|X_n\| > M a_n) = 0$. We write $X_n = o_p(a_n)$ if for every $M > 0$, $\limsup_{n\to\infty} \Pr(\|X_n\| > M a_n) = 0$.

Lemma B.1. *Let X_n be a sequence of random vectors and $\{\mathcal{F}_n, n \geqslant 1\}$ be a sequence of σ-algebras such that $\mathbb{E}\big[\|X_n\| \mid \mathcal{F}_n\big] = o_p(1)$. Then $X_n = o_p(1)$.*

Proof of Lemma B.1. Fixing $M > 0$, we have $M\mathbf{1}(\|X_n\| > M) \leqslant \|X_n\|$ for all n. Taking conditional expectations given $\mathcal{F}_n$ on both sides we have

$$P\big(\|X_n\| > M \mid \mathcal{F}_n\big) \leqslant M^{-1}\mathbb{E}\big[\|X_n\| \mid \mathcal{F}_n\big]. \tag{S3}$$

Thus if $\mathbb{E}\big[\|X_n\| \mid \mathcal{F}_n\big] = o_p(1)$ we have $\Pr(\|X_n\| > M \mid \mathcal{F}_n) = o_p(1)$ as well. But $\Pr(\|X_n\| > M \mid \mathcal{F}_n)$ is uniformly bounded so its expectation converges to zero, i.e., $\Pr(\|X_n\| > M) = o(1)$. Since $M > 0$ was arbitrary we conclude that $X_n = o_p(1)$. $\square$

Lemma B.2. *$X_n = O_p(a_n)$ if and only if for every sequence $b_n \uparrow \infty$ we have* $\Pr(\|X_n\| > b_n a_n) \to 0$ *as $n \to \infty$.*

Proof of Lemma B.2. Fix $b_n \uparrow \infty$ and $\epsilon > 0$. If $X_n = O_p(a_n)$ then there exists $M < \infty$ such that $\limsup_{n\to\infty} \Pr(\|X_n\| > Ma_n) < \epsilon$. Since $b_n > M$ eventually we conclude $\limsup_{n\to\infty} \Pr(\|X_n\| > b_n a_n) < \epsilon$ as well. With $\epsilon > 0$ arbitrary, the result follows. Conversely now suppose we do not have $X_n = O_p(a_n)$. Then there exists $\epsilon > 0$ such that $\limsup_{n\to\infty} \Pr(\|X_n\| > Ma_n) \geqslant \epsilon$ for all $M < \infty$. Defining $n_0 = 1$, this ensures that for each $k = 1, 2, \ldots$, there exists $n_k > n_{k-1}$ so that $\Pr(\|X_{n_k}\| > ka_n) \geqslant \epsilon$. But then for $b_n = \max\{k \geqslant 0 : n_k \leqslant n\}$, we have $b_n \uparrow \infty$ yet $\Pr(\|X_n\| > b_n a_n) \geqslant \epsilon$ for all $n \in n_1, n_2, \ldots$ so $\Pr(\|X_n\| > b_n a_n)$ does not converge to 0 as $n \to \infty$. $\square$

We now state an important result in empirical process theory in our proof of Lemma 5.2 above. For a metric space $(\mathcal{M}, d)$, let $B_\epsilon(m_0) = \{m \in \mathcal{M} \mid d(m, m_0) \leqslant \epsilon\}$ be the ϵ-ball around $m_0 \in \mathcal{M}$. For any set $\mathcal{S} \subseteq \mathcal{M}$, the ϵ covering number $\mathcal{N}(\epsilon, \mathcal{S}, d)$ of $\mathcal{S}$ is then defined as the smallest number of ϵ-balls in $\mathcal{M}$ whose union contains $\mathcal{S}$. For a subset $\mathcal{F}$ of the space $L^2(\mathcal{X}; P)$ of P-square integrable real-valued functions on $\mathcal{X}$ with $\sup_{f\in\mathcal{F}} \|f\|_\infty \leqslant \bar{F} < \infty$, control over the logarithm of the covering numbers of $\mathcal{F}$ (the *metric entropy*) over a variety of radii ϵ under the random metric $L^2(P_n)$ given by $L^2(P_n)(f_1, f_2) = \|f_1 - f_2\|_{2,P_n} = (\int (f_1(x) - f_2(x))^2 \mathrm{d}P_n(x))^{1/2}$ implies control of the empirical process $\sup_{f\in\mathcal{F}} |(P_n - P)f|$ where $Qf := \int f(x)\mathrm{d}Q(x)$ for any measure Q on $\mathcal{X}$. Here P_n is the empirical probability measure on observations $X_1, \ldots, X_n \overset{\text{iid}}{\sim} P$. This result is due to a "chaining" argument of [1]. We restate a more direct version of this result below, which is Lemma A.4 of [2].

Lemma B.3. *For a class $\mathcal{F}$ of P-measurable functions $f : \mathcal{X} \to \mathbb{R}$ with $\sup_{f\in\mathcal{F}} \|f\|_\infty \leqslant \bar{F}$, there exists a universal constant $K < \infty$ such that for P_n the empirical distribution of $X_1, \ldots, X_n \overset{\text{iid}}{\sim} P$,*

$$\sup_{f\in\mathcal{F}} \big|(P_n - P)f\big| \leqslant K\bar{F}n^{-1/2} \int_0^1 \sqrt{\log \mathcal{N}(\epsilon, \mathcal{F}, L^2(P_n))}\, \mathrm{d}\epsilon.$$

Remark B.1. Often, control of the right-hand side in the previous display is shown by controlling

$$\sup_Q \int_0^1 \sqrt{\log \mathcal{N}(\epsilon, \mathcal{F}, L^2(Q))}\, \mathrm{d}\epsilon$$

where the supremum is taken over all finitely supported probability measures Q. See Sections 2.5 and 2.6 of [3] for further discussion.

We also make use of the following elementary results on covering numbers.

Lemma B.4. *Let $\mathcal{E}$ be a collection of functions contained in the class $\mathcal{F}_0$. Let $\mathcal{G} = \{x \mapsto g(e(x), \eta(x)) \mid e \in \mathcal{E}\}$ for some $\eta : \mathcal{X} \to \mathcal{W}$ and $g : [0,1] \times \mathcal{W} \to \mathbb{R}$ with $g(\cdot, w)$ continuous on $[0,1]$ and $\sup_{k\in[0,1],w\in\mathcal{W}} |g'(k,w)| \leqslant C$ for some $C < \infty$, where $g'(\cdot,\cdot)$ denotes the partial derivative of $g(\cdot,\cdot)$ with respect to the first argument. Then for all probability distributions P on $\mathcal{X}$ and $\epsilon > 0$, we have $\mathcal{N}(C\epsilon, \mathcal{G}, L^2(P)) \leqslant \mathcal{N}(\epsilon, \mathcal{E}, L^2(P))$.*

Proof of Lemma B.4. Fix $\epsilon > 0$ and a probability distribution P on $\mathcal{X}$. For each $e \in \mathcal{E}$ let $h^{(e)}(x) = g(e(x), \eta(x))$ for each $x \in \mathcal{X}$, so that $h^{(e)} \in \mathcal{G}$. Suppose $\{e_1, \ldots, e_N\}$ is an ϵ-cover of $\mathcal{E}$ in the $L^2(P)$ norm. WLOG we can assume that each e_k is a member of $\mathcal{F}_\gamma$. Then for each $k = 1, \ldots, N$, by the uniform bound on g' we have

$$\|h^{(e)} - h^{(e_k)}\|_{2,P}^2 = \mathbb{E}_P\big[|g(e(X), \eta(X)) - g(e_k(X), \eta(X))|^2\big] \leqslant C^2 \|e - e_k\|_{2,P}^2$$

so that $\{g^{(e_1)}, \ldots, g^{(e_N)}\}$ is a $C\epsilon$ cover of $\mathcal{G}$ in the $L^2(P)$ norm. □

Lemma B.5. *Let $\mathcal{E}$ be a collection of functions contained in the class $\mathcal{F}_0$. Define $\mathcal{E}_2^- = \{(f-g)^2 : f \in \mathcal{E}, g \in \mathcal{E}\}$. Then for every $\epsilon > 0$ and probability measure P on $\mathcal{X}$ we have*

$$\mathcal{N}(\epsilon; \mathcal{E}_2^-, L^2(P)) \leqslant \mathcal{N}\Big(\frac{\epsilon}{4}; \mathcal{E}, L^2(P)\Big)^2. \tag{S4}$$

Proof. Fix $\epsilon > 0$ and a probability distribution P on $\mathcal{X}$. Define the collection $\mathcal{E}^- = \{f - g : f \in \mathcal{E}, g \in \mathcal{E}\}$. Suppose $\{f_1, \ldots, f_N\}$ is a $\epsilon/4$ cover of $\mathcal{E}$ in the $L^2(P)$ norm. Then the collection $D = \{d_{ij} = f_i - f_j : 1 \leqslant i, j \leqslant N\}$ is a $\epsilon/2$ cover of $\mathcal{E}^-$ in the $L^2(P)$ norm, since for any $f - g \in \mathcal{E}^-$ there exist i, j such that $\|f - f_i\|_{2,P} \vee \|g - f_j\|_{2,P} \leqslant \epsilon/4$ and so

$$\|(f-g) - d_{ij}\|_{2,P} \leqslant \|f - f_i\|_{2,P} + \|g - f_j\|_{2,P} \leqslant \frac{\epsilon}{2}$$

showing that $\mathcal{N}(\epsilon/2, \mathcal{E}^-, L^2(P)) \leqslant (\mathcal{N}(\epsilon/4, \mathcal{E}, L^2(P)))^2$. But by applying Lemma B.4 with $\mathcal{E} = \mathcal{E}^-$ and $g(e, w) = e^2$ (hence we can take $C = 2$) we have $\mathcal{N}(\epsilon, \mathcal{E}_2^-, L^2(P)) \leqslant \mathcal{N}(\epsilon/2, \mathcal{E}^-, L^2(P))$ for all $\epsilon > 0$. Chaining together the inequalities preceding two sentences establishes (S4). □

B.2 Miscellaneous

Here we have some standalone technical lemmas. Their proofs do not depend on any of our other results.

Lemma B.6. *Suppose X and Y are mean zero random vectors in $\mathbb{R}^p$ with finite second moments where* $\mathrm{Cov}(Y)$ *has full rank. Then*

$$\mathrm{Cov}(X) \succcurlyeq \mathrm{Cov}(X, Y)(\mathrm{Cov}(Y))^{-1}\mathrm{Cov}(X, Y)^\top.$$

Proof of Lemma B.6. For any matrix $A \in \mathbb{R}^{p\times p}$ we have $(X + AY)(X + AY)^\top \succcurlyeq 0$, hence

$$\mathbb{E}[(X + AY)(X + AY)^\top] = \mathrm{Cov}(X) + \mathrm{Cov}(X, Y)A^\top + A\mathrm{Cov}(X, Y)^\top + A\mathrm{Cov}(Y)A^\top \succcurlyeq 0$$

Taking $A = -\mathrm{Cov}(X, Y)(\mathrm{Cov}(Y))^{-1}$ yields the desired result. □

Lemma B.7. *If $Y = \sum_{i=1}^n Y_i$ is a random vector where $Y_1, \dots, Y_n$ are independent with mean 0 and finite second moments, then $\mathbb{E}[\|Y\|^2] = \sum_{i=1}^n \mathbb{E}[\|Y_i\|^2]$.*

Proof of Lemma B.7. By assumption we have $\mathbb{E}[Y_i^\top Y_j] = \mathbb{E}[Y_i]^\top \mathbb{E}[Y_j] = 0$ for $i \neq j$, so

$$\mathbb{E}[\|Y\|^2] = \mathbb{E}\left[\left\|\sum_{i=1}^n Y_i\right\|^2\right] = \mathbb{E}\left[\left(\sum_{i=1}^n Y_i\right)^\top \left(\sum_{i=1}^n Y_i\right)\right] = \sum_{i=1}^n \mathbb{E}[Y_i^\top Y_i] = \sum_{i=1}^n \mathbb{E}[\|Y_i\|^2].$$

□

C Proofs

Here we collect proofs of the formal results stated in the main text; some are generalized to allow for some nonstationarities across batches, as described in Appendix A.

C.1 Proof of Proposition 3.1: CLT for oracle $\hat{\theta}^*$

The CLT of Proposition 3.1 for our oracle pooled estimator $\hat{\theta}^*$ holds under the numbered generalizations in Appendix A without further restrictions; here we prove this more general proposition. It is helpful to begin by noting that $P_{0,e_0} = \sum_{t=1}^t \kappa_t P_{t,e_t}$, which shows, for instance, that for any P_0-integrable function f

$$\mathbb{E}_t[|f(W)|] = \mathbb{E}_0\left[|f(W)|\frac{\mathrm{d}P_{t,e_t}}{\mathrm{d}P_{0,e_0}}(W)\right] \leqslant \kappa_t^{-1}\mathbb{E}_0[|f(W)|]$$

By score linearity (4) we can write

$$\sqrt{N}(\hat{\theta}^* - \theta_0) = -\left(\frac{1}{N}\sum_{t=1}^T\sum_{i=1}^{N_t} s_a(W_{ti};\nu_0, e_{0,N})\right)^{-1}\left(\frac{1}{\sqrt{N}}\sum_{t=1}^T\sum_{i=1}^{N_t} s(W_{ti};\theta_0,\nu_0, e_{0,N})\right) \tag{S5}$$

whenever $\hat{\theta}^*$ exists. Defining

$$r_N = \sum_{t=1}^T \frac{N_t}{N}\frac{1}{N_t}\sum_{i=1}^{N_t}(s_a(W_{ti};\nu_0, e_{0,N}) - s_a(W_{ti};\nu_0, e_0)).$$

we have from (3) and the law of large numbers that

$$\begin{aligned}
\frac{1}{N}\sum_{t=1}^T\sum_{i=1}^{N_t} s_a(W_{ti};\nu_0, e_{0,N}) &= \sum_{t=1}^T \frac{N_t}{N}\frac{1}{N_t}\sum_{i=1}^{N_t} s_a(W_{ti};\nu_0, e_{0,N}) \\
&= r_N + \sum_{t=1}^T \frac{N_t}{N}(\mathbb{E}_t[s_a(W;\nu_0, e_0)] + o_p(1)) \\
&= \sum_{t=1}^T \kappa_t \mathbb{E}_t[s_a(W;\nu_0, e_0)] + o_p(1) \\
&= \mathbb{E}_0\left[\sum_{t=1}^T \kappa_t \frac{\mathrm{d}P_t}{\mathrm{d}P_0}(W)s_a(W;\nu_0, e_0)\right] + o_p(1) \\
&= \mathbb{E}_0[s_a(W;\nu_0, e_0)] + o_p(1). \qquad (S6)
\end{aligned}$$

Then the third equality follows because

$$\|r_N\| \leqslant \sum_{t=1}^{T} \frac{N_t}{N} \frac{1}{N_t} \sum_{i=1}^{N_t} \|s_a(W_{ti}; \nu_0, e_{0,N}) - s_a(W_{ti}; \nu_0, e_0)\| = o_p(1)$$

by Lemma B.1 and condition 1 of the Proposition, noting that

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{N_t} \sum_{i=1}^{N_t} \|s_a(W_{ti}; \nu_0, e_{0,N}) - s_a(W_{ti}; \nu_0, e_0)\|\right] &\leqslant \kappa_t^{-1} \mathbb{E}_0[\|s_a(W; \nu_0, e_{0,N}) - s_a(W; \nu_0, e_0)\|] \\
&\leqslant \kappa_t^{-1} (\mathbb{E}_0[\|s_a(W; \nu_0, e_{0,N}) - s_a(W; \nu_0, e_0)\|^2])^{1/2} \\
&\leqslant \kappa_t^{-1} \delta_N
\end{aligned}$$

for $t = 1, \ldots, T$. Invertibility of $\mathbb{E}_0[s_a]$ from condition 2 ensures that $\hat{\theta}^*$ is well-defined with probability tending to 1 by (S6).

Next, we fix $c \in \mathbb{R}^p$ with $\|c\| = 1$ and a batch $t \in \{1, \ldots, T\}$. Define

$$U_{N,t,i} = \frac{c^\top s(W_{ti}; \theta_0, \nu_0, e_{0,N})}{(N_t \cdot c^\top V_{t,N} c)^{1/2}}, \quad i = 1, \ldots, N_t$$

where $V_{t,N} = \mathbb{E}_t[s(W; \theta_0, \nu_0, e_{0,N})^{\otimes 2}]$. Evidently the random variables $U_{N,1}, \ldots, U_{N,N_t}$ are independent, with $\mathbb{E}[U_{N,t,i}] = 0$ for all i by (S1), and $\sum_{i=1}^{N_t} \mathbb{E}[U_{N_t,i}^2] = 1$. Furthermore we have

$$\lim_{N_t \to \infty} \sum_{i=1}^{N_t} \mathbb{E}[|U_{N,t,i}|^q] \leqslant N_t \cdot \frac{\mathbb{E}_t[\|s(W; \theta_0, \nu_0, e_{0,N})\|^q]}{(N_t \cdot c^\top V_{t,N} c)^{q/2}} = O(N_t^{1-q/2}) = o(1)$$

for all sufficiently large N and some $q > 2$ by condition 3. Then by the Lyapunov CLT we have

$$\sum_{i=1}^{N_t} U_{N,t,i} = c^\top \left((c^\top V_{t,N} c)^{-1/2} \frac{1}{\sqrt{N_t}} \sum_{i=1}^{N_t} s(W_{ti}; \theta_0, \nu_0, e_{0,N}) \right) \xrightarrow{\mathrm{d}} \mathcal{N}(0, 1).$$

$$V_{t,N} = \mathbb{E}_t[s(W; \theta_0, \nu_0, e_{0,N})^{\otimes 2}] \to \mathbb{E}_t[s(W; \theta_0, \nu_0, e_0)^{\otimes 2}] \equiv V_t$$

as $N \to \infty$. Therefore

$$c^\top \left(\frac{1}{\sqrt{N_t}} \sum_{i=1}^{N_t} s(W_{ti}; \theta_0, \nu_0, e_{0,N}) \right) \xrightarrow{\mathrm{d}} \mathcal{N}(0, c^\top V_t c)$$

and since c was arbitrary,

$$\frac{1}{\sqrt{N_t}} \sum_{i=1}^{N_t} s(W_{ti}; \theta_0, \nu_0, e_{0,N}) \xrightarrow{\mathrm{d}} \mathcal{N}(0, V_t), \quad t = 1, \ldots, T.$$

With the left-hand side of the preceding display independent across batches $t = 1, \ldots, T$, we have

$$\begin{aligned}
\frac{1}{\sqrt{N}} \sum_{t=1}^{T} \sum_{i=1}^{N_t} s(W_{ti}; \theta_0, \nu_0, e_{0,N}) &= \sum_{t=1}^{T} \sqrt{\frac{N_t}{N}} \frac{1}{\sqrt{N_t}} \sum_{i=1}^{N_t} s(W_{ti}; \theta_0, \nu_0, e_{0,N}) \\
&\xrightarrow{\mathrm{d}} \mathcal{N}\left(0, \sum_{t=1}^{T} \kappa_t V_t\right). \qquad \text{(S7)}
\end{aligned}$$

With

$$\sum_{t=1}^{T} \kappa_t V_t = \sum_{t=1}^{T} \kappa_t \mathbb{E}_t[s(W;\theta_0,\nu_0,e_0)^{\otimes 2}] = \mathbb{E}_0\left[\sum_{t=1}^{T} \kappa_t \frac{\mathrm{d}P_t}{\mathrm{d}P_0}(W) s(W;\theta_0,\nu_0,e_0)^{\otimes 2}\right]$$
$$= \mathbb{E}_0[s(W;\theta_0,\nu_0,e_0)^{\otimes 2}]$$

the result of the Proposition follows by (S5) and (S6).

C.2 Proof of Corollary 3.1: CLT for $\hat{\theta}^*_{\mathrm{AIPW}}$

Corollary 3.1, which applies Proposition 3.1 to prove a CLT for the oracle estimator $\hat{\theta}^*_{\mathrm{AIPW}}$ of $\theta_{0,\mathrm{ATE}}$, holds under the numbered generalizations of Appendix A under one additional condition: that the mean functions are stationary, meaning $\mathbb{E}_t[Y(z) \mid X = x] = m_0(z,x)$ for all $t = 1,\dots,T$, $z = 0,1$, and $x \in \mathcal{X}$. This condition is needed to ensure Assumption A.2 holds, as discussed in Appendix A.

Our proof proceeds by showing that the conditions of the Corollary imply the conditions of generalized Proposition 3.1 proven in the previous section with $\theta_0 = \theta_{0,\mathrm{ATE}}$ and $s = s_{\mathrm{AIPW}}(\cdot)$. That is, first we show Assumption A.2 is satisfied with $\theta_0 = \theta_{0,\mathrm{ATE}}$ and $s = s_{\mathrm{AIPW}}(\cdot)$. Then we show the three numbered conditions in Proposition 3.1.

First, for brevity let $\nu_0 = \nu_{0,\mathrm{AIPW}}(\cdot) = (m_0(0,\cdot), m_0(1,\cdot))$, and note that for any e, $e' \in \mathcal{F}_\gamma$ we have

$$\begin{aligned}
\mathbb{E}_{t,e}[s_{\mathrm{AIPW}}(W;\theta_{0,\mathrm{ATE}},\nu_0,e')] &= \mathbb{E}_{t,e}[m_0(1,X) - m_0(0,X) - \theta_{0,\mathrm{ATE}}] \\
&\quad + \mathbb{E}_{t,e}\left[\frac{Z(Y(1) - m_0(1,X))}{e'(X)} - \frac{(1-Z)(Y(0) - m_0(0,X))}{1 - e'(X)}\right] \\
&= \mathbb{E}_{t,e}\left[\frac{e(X)}{e'(X)} \mathbb{E}_{t,e}[Y(1) - m_0(1,X) \mid X]\right] \\
&\quad - \mathbb{E}_{t,e}\left[\frac{1-e(X)}{1-e'(X)} \mathbb{E}_{t,e}[Y(0) - m_0(0,X) \mid X]\right] \\
&= 0
\end{aligned}$$

using unconfoundedness and stationarity of the mean function. All the necessary expectations exist by our assumption that $\gamma > 0$. Hence Assumption A.2 is satisfied. Next we show the conditions of Proposition 3.1:

1. Trivially we have

$$\mathbb{E}_0[|s_{\mathrm{AIPW},a}(W;\nu_0,e_{0,N}) - s_{\mathrm{AIPW},a}(W;\nu_0,e_0)|^2] = \mathbb{E}_0[|-1-(-1)|^2] = 0.$$

Next we compute the following for each $(\nu, e) \in \mathcal{N} \times \mathcal{F}_\gamma$:

$$\begin{aligned}
&s_{\mathrm{AIPW}}(W;\theta_0,\nu,e) - s_{\mathrm{AIPW}}(W;\theta_0,\nu_0,e_0)\\
&= \left(1 - \frac{Z}{e_0(X)}\right)(m(1,X) - m_0(1,X))\\
&\quad + \left(1 - \frac{1-Z}{1-e_0(X)}\right)(m_0(0,X) - m(0,X))\\
&\quad + Z(Y - m(1,X))(e(X)^{-1} - e_0(X)^{-1})\\
&\quad - (1-Z)(Y - m(0,X))((1-e(X))^{-1} - (1-e_0(X))^{-1}).
\end{aligned}$$

Plugging in $(\nu, e) = (\nu_0, e_{0,N})$ gives, by Minkowski's inequality, that

$$(\mathbb{E}_0[|s_{\mathrm{AIPW}}(W;\theta_0,\nu_0,e_{0,N}) - s_{\mathrm{AIPW}}(W;\theta_0,\nu_0,e_0)|^2])^{1/2} \leqslant A_0 + B_0$$

where

$$\begin{aligned}
A_0 &= (\mathbb{E}_0[Z^2(Y(1) - m_0(1,X))^2(e_{0,N}(X)^{-1} - e_0(X)^{-1})^2])^{1/2}\\
&\leqslant \gamma^{-2}(\mathbb{E}_0[(Y(1) - m_0(1,X))^2(e_{0,N}(X) - e_0(X))^2])^{1/2}\\
&= \gamma^{-2}(\mathbb{E}_0[(e_{0,N}(X) - e_0(X))^2 v_0(1,X)])^{1/2}\\
&\leqslant C\gamma^{-2}\|e_{0,N} - e_0\|_{2,P_0^X}.
\end{aligned}$$

By an analogous computation

$$\begin{aligned}
B_0 &= (\mathbb{E}_0[(1-Z)^2(Y(0) - m_0(0,X))^2((1-e_{0,N}(X))^{-1} - (1-e_0(X))^{-1})^2])^{1/2}\\
&\leqslant C\gamma^{-2}\|e_{0,N} - e_0\|_{2,P_0^X}.
\end{aligned}$$

The result now follows because

$$\begin{aligned}
|e_{0,N}(x) - e_0(x)| &= \left|\sum_{t=1}^{T}\left(\frac{N_t}{N} - \kappa_t\right)e_t(x)\frac{\mathrm{d}P_t^X}{\mathrm{d}P_0^X}(x)\right|\\
&\leqslant \left(\sup_{1\leqslant t\leqslant T}\kappa_t^{-1}\left|\frac{N_t}{N} - \kappa_t\right|\right)\sum_{1\leqslant t\leqslant T} e_t(x)\\
&\leqslant T\left(\sup_{1\leqslant t\leqslant T}\kappa_t^{-1}\left|\frac{N_t}{N} - \kappa_t\right|\right) = o(1)
\end{aligned}$$

and hence

$$\|e_{0,N} - e_0\|_{2,P_0^X}^2 \leqslant \sup_{x\in\mathcal{X}}|e_{0,N}(x) - e_0(x)|^2 = o(1).$$

This bound only uses the fact that propensities are bounded between 0 and 1, so

$$\sup_{x\in\mathcal{X}, e_1(\cdot),\dots,e_T(\cdot)\in\mathcal{F}_0}|e_{0,N}(x) - e_0(x)| = o(1). \tag{S8}$$

2. Evidently $\mathbb{E}_0[s_{\mathrm{AIPW},a}(W;\nu_0,e_0)] = -1$ is invertible. Additionally for $z = 0, 1$ we have $\mathbb{E}[Y(z)^2] \leqslant C$ by the moment conditions in Assumption 3.1 so

$$\mathbb{E}_0[m_0(z,X)^2] = \mathbb{E}_0[(\mathbb{E}_0(Y(z) \mid X))^2] \leqslant \mathbb{E}_0[Y(z)^2] < \infty.$$

Now $s_{\mathrm{AIPW}}(W;\theta_{0,\mathrm{ATE}},\nu_{0,\mathrm{AIPW}},e_0)$ is the sum of the following terms:

$$\begin{aligned}
&m_0(1,X)\left(1 - \frac{Z}{e_0(X)}\right),\\
&m_0(0,X)\left(\frac{1-Z}{1-e_0(X)} - 1\right) - \theta_{0,\mathrm{ATE}}, \quad \text{and}\\
&Y\left(\frac{Z}{e_0(X)} - \frac{1-Z}{1-e(X)}\right).
\end{aligned}$$

These are all square integrable, because Z, $(e_0(X))^{-1}$, and $(1-e_0(X))^{-1}$ are all uniformly bounded.

3. From $(\mathbb{E}_0[|Y(z)|^q])^{1/q} \leqslant C$ for $z = 0, 1$ by Assumption 3.1, we have

$$\mathbb{E}_0[|m_0(z,X)|^q] = \mathbb{E}[|\mathbb{E}_0[Y(z) \mid X]|^q] \leqslant \mathbb{E}_0[|Y(z)|^q] \leqslant C^q.$$

Then each of

$$\begin{aligned}
&\mathbb{E}_0\left[|m_0(1,X)|^q\left|1 - \frac{Z}{e_0(X)}\right|^q\right],\\
&\mathbb{E}_0\left[|m_0(0,X)|^q\left|\frac{1-Z}{1-e_0(X)} - 1\right|^q\right], \quad \text{and}\\
&\mathbb{E}_0\left[|Y|^q\left|\frac{Z}{e_0(X)} - \frac{1-Z}{1-e(X)}\right|^q\right]
\end{aligned}$$

is at most $[C(1+\gamma^{-1})]^q$ and the desired condition holds by Minkowski's inequality.

Now we can apply Proposition 3.1, to conclude $\sqrt{N}(\hat{\theta}^* - \theta_{0,\mathrm{ATE}}) \xrightarrow{\mathrm{d}} \mathcal{N}(0, V_0)$ where

$$\begin{aligned}
V_0 &= \mathbb{E}_0[s_{\mathrm{AIPW}}(W;\theta_{0,\mathrm{ATE}},\nu_{0,\mathrm{AIPW}},e_0)^2]\\
&= \mathbb{E}_0\left[\left((\tau_0(X) - \theta_{0,\mathrm{ATE}}) + \frac{Z(Y(1) - m_0(1,X))}{e_0(X)} - \frac{(1-Z)(Y(0) - m_0(0,X))}{1-e_0(X)}\right)^2\right]\\
&= \mathbb{E}_0[(\tau_0(X) - \theta_{0,\mathrm{ATE}})^2] + \mathbb{E}_0\left[\frac{Z(Y(1) - m_0(1,X))^2}{(e_0(X))^2}\right] + \mathbb{E}_0\left[\frac{(1-Z)(Y(0) - m_0(0,X))^2}{(1-e_0(X))^2}\right]\\
&= \mathbb{E}_0[(\tau_0(X) - \theta_{0,\mathrm{ATE}})^2] + \mathbb{E}_0\left[\frac{1}{e_0(X)} \cdot \mathbb{E}[(Y(1) - m_0(1,X))^2 \mid X]\right]\\
&+ \mathbb{E}_0\left[\frac{1}{1-e_0(X)} \cdot \mathbb{E}[(Y(0) - m_0(0,X))^2 \mid X]\right]\\
&= V_{0,\mathrm{AIPW}}
\end{aligned}$$

The third equality in the preceding display follows by noting that the three cross terms in the expansion of the square have mean zero. The first two vanish by conditioning on X and the third because $Z(1-Z) = 0$.

C.3 Proof of Corollary 3.2: CLT for $\hat{\theta}^*_{\mathrm{EPL}}$

Here we prove Corollary 3.2, the CLT for the partial linear estimator $\hat{\theta}^*_{\mathrm{EPL}}$ of the regression parameter $\theta_{0,\mathrm{PL}}$ under the linear treatment effect assumption (6). This Corollary holds under the numbered generalizations of Appendix A with the additional condition that the mean and variance functions are stationary. That is, we have $\mathbb{E}_t[Y(z) \mid X = x] = m_0(z,x)$ and $\mathrm{Var}_t(Y(z) \mid X = x) = v_0(z,x)$ for all $t = 1,\ldots,T$, $z = 0,1$, and $x \in \mathcal{X}$. This condition is needed to ensure Assumption A.2 holds.

As in the proof of Corollary 3.1, we first show that Assumption A.2 holds with $\gamma = 0$, estimand $\theta_0 = \theta_{0,\mathrm{PL}}$, score $s(\cdot) = s_{\mathrm{EPL}}(\cdot)$, and nuisance functions $\nu_0 = \nu_{0,\mathrm{EPL}}(\cdot) = (m_0(0,\cdot), v_0(0,\cdot), v_0(1,\cdot))$ lying in the nuisance set $\mathcal{N} = \mathcal{N}_{\mathrm{EPL}}$. Then we show that the three numbered conditions in Proposition 3.1 hold.

Fix $e(\cdot), e'(\cdot) \in \mathcal{F}_0$. For each $t = 0,1,\ldots,T$,

$$\begin{aligned}\mathbb{E}_{t,e}[Y \mid X, Z = 0] &= \mathbb{E}_{t,e}[Y(0) \mid X] = m_0(0,X), \quad \text{and}\\ \mathbb{E}_{t,e}[Y \mid X, Z = 1] &= \mathbb{E}_{t,e}[Y(1) \mid X] = m_0(1,X)\end{aligned}$$

hold by the unconfoundedness and SUTVA assumptions. Hence by (6),

$$\mathbb{E}_{t,e}[Y \mid X, Z] = m_0(0,X) + Z\psi(X)^\top\theta_0. \tag{S9}$$

Thus for any $e'(\cdot) \in \mathcal{F}_\gamma$,

$$\begin{aligned}\mathbb{E}_{t,e}[s(W;\theta_{0,\mathrm{PL}},\nu_0,e')] &= \mathbb{E}_{t,e}[w(X;\nu_0,e')(Z - e'(X))(Y - m_0(0,X) - Z\psi(X)^\top\theta_{0,\mathrm{PL}})\psi(X)]\\ &= 0\end{aligned}$$

after conditioning on (X,Z) and applying (S9). Integrability is not a concern because $w(X;\nu_0,e') \leqslant c^{-1}$ for $\nu_0 \in \mathcal{N}_{\mathrm{EPL}}$. Thus Assumption A.2 is satisfied.

Now we consider the numbered conditions of Proposition 3.1 in turn.

1. Because the predictor variables $\psi(X)$ satisfy $\|\psi(X)\| \leqslant C$

$$\begin{aligned}(\mathbb{E}_0[\|s_a(W;\nu_0,e_{0,N}) - s_a(W;\nu_0,e_0)\|^2])^{1/2} &= (\mathbb{E}_0[\|Z\Delta(X,Z)\psi(X)\psi(X)^\top\|^2])^{1/2}\\ &\leqslant C^2(\mathbb{E}_0[\Delta(X,Z)^2])^{1/2}\end{aligned}$$

where

$$\begin{aligned}\Delta(X,Z) &= w(X;\nu_0,e_0)(Z - e_0(X)) - w(X;\nu_0,e_{0,N})(Z - e_{0,N}(X))\\ &= (w(X;\nu_0,e_0) - w(X;\nu_0,e_{0,N}))(Z - e_0(X)) + w(X;\nu_0,e_{0,N})(e_{0,N}(X) - e_0(X)).\end{aligned}$$

This Δ satisfies

$$\begin{aligned}(\mathbb{E}_0[\Delta(X,Z)^2])^{1/2} &\leqslant (\mathbb{E}_0[(w(X;\nu_0,e_0) - w(X;\nu_0,e_{0,N}))^2(Z - e_0(X))^2])^{1/2}\\ &\quad + (\mathbb{E}_0[w^2(X;\nu_0,e_{0,N})(e_{0,N}(X) - e_0(X))^2])^{1/2}.\end{aligned} \tag{S10}$$

Now $c \leqslant v_0(z,x) \leqslant \mathbb{E}[Y(z)^2 \mid X = x] \leqslant C$ for all $z = 0,1$ and $x \in \mathcal{X}$, so for any propensity $e(\cdot)$ we have

$$C^{-1} \leqslant w(X;\nu_0,e) = (v_0(0,X)e(X) + v_0(1,X)(1 - e(X)))^{-1} \leqslant c^{-1}$$

and

$$\sup_{e\in[0,1]}\left|\frac{\partial}{\partial e}\frac{1}{v_0(0,X)e+v_0(1,X)(1-e)}\right| = \sup_{e\in[0,1]}\left|\frac{v_0(0,X)-v_0(1,X)}{(v_0(0,X)e+v_0(1,X)(1-e))^2}\right|$$
$$\leqslant \frac{2C}{c^2}.$$

Applying these latter two facts to equation (S10) yields

$$(\mathbb{E}_0[\Delta(X,Z)^2])^{1/2} \leqslant \left(\frac{2C}{c^2}+c^{-1}\right)\|e_{0,N}-e_0\|_{2,P_0^X} = o(1)$$

by (S8). Similarly,

$$\begin{aligned}
&(\mathbb{E}_0[\|s(W;\theta_{0,\mathrm{PL}},\nu_0,e_{0,N})-s(W;\theta_{0,\mathrm{PL}},\nu_0,e_0)\|^2])^{1/2}\\
&=(\mathbb{E}_0[\|\Delta(X,Z)(Y-m_0(0,X)-Z\psi(X)^\top\theta_0)\psi(X)\|^2])^{1/2}\\
&\leqslant C(\mathbb{E}_0[\Delta(X,Z)^2v_0(Z,X)])^{1/2}\\
&\leqslant C^{3/2}(\mathbb{E}_0[\Delta(X,Z)^2])^{1/2}\\
&=o(1).
\end{aligned}$$

2. For invertibility of $\mathbb{E}_0[s_a]$, we compute

$$\begin{aligned}
\mathbb{E}_0[s_a(W;\nu_0,e_0)] &= \mathbb{E}_0[-w(X;\nu_0,e_0)(Z-e_0(X))Z\psi(X)\psi(X)^\top]\\
&=\mathbb{E}_0[-w(X;\nu_0,e_0)e_0(X)(1-e_0(X))\psi(X)\psi(X)^\top]\\
&\preccurlyeq -C^{-1}\mathbb{E}_0[e_0^2(X)(1-e_0^2(X))\psi(X)\psi(X)^\top].
\end{aligned}$$

With $\mathbb{E}_0[e_0(X)(1-e_0(X))\psi(X)\psi(X)^\top]$ positive definite by assumption, $\mathbb{E}_0[s_a(W;\nu_0,e_0)]$ is strictly negative definite and hence invertible.

For boundedness of $\mathbb{E}_0[\|s\|^2]$ with $e=e_0$, we compute

$$\begin{aligned}
\mathbb{E}_0[\|s(W;\theta_0,\nu_0,e_0)\|^2] &= \mathbb{E}_0[w(X;\nu_0,e_0)^2(Z-e_0(X))^2(Y-m_0(0,X)-Z\psi(X)^\top\theta_0)^2\|\psi(X)\|^2]\\
&\leqslant \frac{C^2}{c^2}\mathbb{E}_0[(Y-m_0(0,X)-Z\psi(X)^\top\theta_0)^2]\\
&=\frac{C^2}{c^2}\mathbb{E}_0[v_0(Z,X)]\leqslant\frac{C^3}{c^2}
\end{aligned}$$

3. For boundedness of $\mathbb{E}_0[\|s\|^q]$ with $e=e_{0,N}$,

$$\begin{aligned}
&(\mathbb{E}_0[\|s(W;\theta_0,\nu_0,e_{0,N})\|^q])^{1/q}\\
&\quad=\Big(\mathbb{E}_0[|w(X;\nu_0,e_{0,N})|^q|Z-e_{0,N}(X)|^q|Y-m_0(0,X)-Z\psi(X)^\top\theta_0|^q\|\psi(X)\|^q]\Big)^{1/q}\\
&\quad\leqslant\frac{C}{c}(\mathbb{E}_0[|Y-m_0(0,X)-Z\psi(X)^\top\theta_0|^q])^{1/q}\\
&\quad\leqslant\frac{4C^2}{c}
\end{aligned}$$

where the final inequality follows from Minkowski's inequality and Jensen's inequality as below:

$$
\begin{aligned}
(\mathbb{E}_0[|Y - \mathbb{E}_0(Y \mid X, Z)|^q])^{1/q} &\leqslant (\mathbb{E}_0[|Y|^q])^{1/q} + (\mathbb{E}_0[|\mathbb{E}_0[Y \mid X, Z]|^q])^{1/q} \\
&\leqslant (\mathbb{E}_0[|Y|^q])^{1/q} + (\mathbb{E}_0[\mathbb{E}_0[|Y|^q \mid X, Z]])^{1/q} \\
&= 2(\mathbb{E}_0[|Y|^q])^{1/q} \\
&\leqslant 2[(\mathbb{E}_0[|Y(0)|^q])^{1/q} + (\mathbb{E}_0[|Y(1)|^q])^{1/q}] \\
&\leqslant 4C \text{ by Assumption 3.2}
\end{aligned}
$$

We now compute

$$
\begin{aligned}
\mathbb{E}_0[s_a(W; \nu_0, e_0)] &= \mathbb{E}_0[-w(X; \nu_0, e_0) Z(Z - e_0(X))\psi(X)\psi(X)^\top] \\
&= \mathbb{E}_0[-w(X; \nu_0, e_0)(1 - e_0(X))e_0(X)\psi(X)\psi(X)^\top] \\
&= -\mathbb{E}_0\left[\frac{e_0(X)(1 - e_0(X))}{v_0(0, X)e_0(X) + v_0(1, X)(1 - e_0(X))}\psi(X)\psi(X)^\top\right]
\end{aligned}
$$

and

$$
\begin{aligned}
\mathbb{E}_0[s(W; \theta_0, \nu_0, e_0)^{\otimes 2}] &= \mathbb{E}_0[w^2(X; \nu_0, e_0)(Z - e_0(X))^2(Y - m_0(0, X) - Z\psi(X)^\top\theta_0)^2\psi(X)\psi(X)^\top] \\
&= \mathbb{E}_0[w^2(X; \nu_0, e_0)(Z - e_0(X))^2 v_0(Z, X)\psi(X)\psi(X)^\top] \\
&= \mathbb{E}_0[w^2(X; \nu_0, e_0)e_0^2(X)v_0(0, X)(1 - e_0(X))\psi(X)\psi(X)^\top] \\
&\quad + \mathbb{E}_0[w^2(X; \nu_0, e_0)(1 - e_0(X))^2 v_0(1, X)e_0(X)\psi(X)\psi(X)^\top] \\
&= \mathbb{E}_0[e_0(X)(1 - e_0(X))w^2(X; \nu_0, e_0)(v_0(0, X)e_0(X) + v_0(1, X)(1 - e_0(X)))\psi(X)\psi(X) \\
&= -\mathbb{E}_0[s_a(W; \nu_0, e_0)].
\end{aligned}
$$

Finally, we apply Proposition 3.1 to conclude $\sqrt{N}(\hat{\theta}^* - \theta_{0,\mathrm{PL}}) \to \mathcal{N}(0, V_0)$ where

$$
\begin{aligned}
V_0 &= (\mathbb{E}_0[s_a(W; \nu_0, e_0)])^{-1}(\mathbb{E}_0[s(W; \theta_0, \nu_0, e_0)^{\otimes 2}])(\mathbb{E}_0[s_a(W; \nu_0, e_0)])^{-1} \\
&= -(\mathbb{E}_0[s_a(W; \nu_0, e_0)])^{-1} \\
&= V_{0,\mathrm{EPL}}.
\end{aligned}
$$

C.4 Proof of Theorem 3.1: Pooling dominates linear aggregation

Here we prove that the oracle pooled estimators $\hat{\theta}^*_{\mathrm{AIPW}}$ and $\hat{\theta}^*_{\mathrm{EPL}}$ dominate the best linearly aggregated estimators $\hat{\theta}^{*,(\mathrm{LA})}_{\mathrm{AIPW}}$ and $\hat{\theta}^{*,(\mathrm{LA})}_{\mathrm{EPL}}$ for estimating $\theta_{0,\mathrm{ATE}}$ and $\theta_{0,\mathrm{PL}}$, respectively. For $\theta_{0,\mathrm{PL}}$, Theorem 3.1 generalizes to nonstationary batches as described in Appendix A. But for $\theta_{0,\mathrm{ATE}}$, the Theorem does not necessarily hold under the nonstationarities of Appendix A. For example, suppose that $T = 2$, P_1^X is uniform on $(0, 1)$, and P_2^X is the distribution with density $2x$ on $(0, 1)$ with respect to Lebesgue measure. Let $v_0(z, x) = x$ and suppose $e_1(x) = e_2(x) = \kappa_1 = 1/2$ for each $x \in \mathcal{X}$. Then $V_{0,\mathrm{AIPW}} = 7/3 > V^{(\mathrm{LA})}_{\mathrm{AIPW}} = 16/7$. We can however generalize the ATE result to the case of a multivariate outcome variable, i.e., $Y(1), Y(0) \in \mathbb{R}^q$, where the conditional mean and variance functions $m_0(\cdot)$ and $v_0(\cdot)$ take values in $\mathbb{R}^q$ and $\mathbb{S}^q_+$, respectively.

We begin by showing the result for ATE estimation. We have $V_{\mathrm{AIPW}}^{(\mathrm{LA})} = \left(\sum_{t=1}^{T} \kappa_t V_{t,\mathrm{AIPW}}^{-1}\right)^{-1}$ where we compute

$$V_{t,\mathrm{AIPW}} = \mathbb{E}\left[\frac{v_0(1,X)}{e_t(X)} + \frac{v_0(0,X)}{1-e_t(X)} + (\tau_0(X)-\theta_0)^{\otimes 2}\right], \quad t = 1,\ldots,T$$

by applying Corollary 3.1 to the observations in each single batch $t \in \{1,\ldots,T\}$. Letting

$$\begin{aligned} h(e) &= \left(\mathbb{E}\left[\frac{v_0(1,X)}{e(X)} + \frac{v_0(0,X)}{1-e(X)} + (\tau_0(X)-\theta_0)^{\otimes 2}\right]\right)^{-1} \\ &= \left(\mathbb{E}\left[\frac{v_0(1,X)(1-e(X)) + v_0(0,X)e(X)}{e(X)(1-e(X))} + (\tau_0(X)-\theta_0)^{\otimes 2}\right]\right)^{-1} \end{aligned}$$

for each $e \in \mathcal{F}_\gamma$, we have

$$\begin{aligned} \left(V_{\mathrm{AIPW}}^{(\mathrm{LA})}\right)^{-1} &= \sum_{t=1}^{T} \kappa_t \left(\mathbb{E}\left[\frac{v_0(1,X)}{e_t(X)} + \frac{v_0(0,X)}{1-e_t(X)} + (\tau_0(X)-\theta_0)^{\otimes 2}\right]\right)^{-1} \\ &= \sum_{t=1}^{T} \kappa_t h(e_t), \quad \text{versus} \\ V_{0,\mathrm{AIPW}}^{-1} &= \left(\mathbb{E}\left[\frac{v_0(1,X)}{e_0(X)} + \frac{v_0(0,X)}{1-e_0(X)} + (\tau_0(X)-\theta_0)^{\otimes 2}\right]\right)^{-1} \\ &= h(e_0). \end{aligned}$$

Thus, to prove the desired result, it suffices to show $h(\cdot)$ is a concave (matrix-valued) function on $\mathcal{F}_\gamma$. To that end, fix $e_1, e_2 \in \mathcal{F}_\gamma$ and $\lambda \in (0,1)$. It suffices to show that $g(\lambda) = h(e_1 + \lambda(e_2 - e_1))$ is concave on $[0,1]$, i.e., that $g(\lambda) \succcurlyeq \lambda g(e_2) + (1-\lambda) g(e_1)$ for each $\lambda \in [0,1]$. Because $g(\cdot)$ is continuous on $[0,1]$, we need only show that $g''(\lambda) \preccurlyeq 0$ for each $\lambda \in (0,1)$. Letting $e_\lambda = e_1 + \lambda(e_2 - e_1)$ and $V_\lambda = h(e_\lambda)^{-1}$, we obtain

$$g''(\lambda) = 2h(e_\lambda)[B_\lambda h(e_\lambda) B_\lambda - C_\lambda] h(e_\lambda) \tag{S11}$$

for

$$\begin{aligned} B_\lambda &= \mathbb{E}\left[(e_2(X) - e_1(X))\left(\frac{v_0(0,X)}{(1-e_\lambda(X))^2} - \frac{v_0(1,X)}{(e_\lambda(X))^2}\right)\right], \quad \text{and} \\ C_\lambda &= \mathbb{E}\left[(e_2(X) - e_1(X))^2\left(\frac{v_0(1,X)}{(e_\lambda(X))^3} + \frac{v_0(0,X)}{(1-e_\lambda(X))^3}\right)\right]. \end{aligned}$$

To get this, we reversed the order of differentiation and expectation, which is justified by the regularity conditions in Assumption 3.1.

Since $h(e_\lambda)$ is symmetric and appears on the left and right in (S11), it suffices to show that $B_\lambda h(e_\lambda) B_\lambda \preccurlyeq C_\lambda$. For this we note that for $\tilde{Z}$ conditionally independent of $(Y(0), Y(1))$ given X with $\tilde{Z} \mid X \sim \mathrm{Bern}(e_\lambda(X))$, we have $h(e_\lambda) = (\mathrm{Cov}(b_\lambda))^{-1}$, $B_\lambda = \mathrm{Cov}(c_\lambda, b_\lambda) = B_\lambda'$, and $C_\lambda = \mathrm{Cov}(c_\lambda)$ where

$$b_\lambda = \frac{(1-\tilde{Z})(Y(0)-m_0(0,X))}{1-e_\lambda(X)} - \frac{\tilde{Z}(Y(1)-m_0(1,X))}{e_\lambda(X)} - (\tau_0(X)-\theta_0)$$
$$c_\lambda = (e_2(X)-e_1(X))\Big(\frac{\tilde{Z}(Y(1)-m_0(1,X))}{(e_\lambda(X))^2} + \frac{(1-\tilde{Z})(Y(0)-m_0(0,X))}{(1-e_\lambda(X))^2}\Big)$$

are mean zero random vectors in $\mathbb{R}^d$. Applying Lemma B.6 completes the proof for ATE.

Now we consider the partially linear model, allowing for the generalizations in Appendix A, as discussed above. We have $V_{\mathrm{EPL}}^{(\mathrm{LA})} = \big(\sum_{t=1}^T \kappa_t V_{t,\mathrm{EPL}}^{-1}\big)^{-1}$ where

$$V_{t,\mathrm{EPL}} = \bigg(\mathbb{E}_t\bigg[\frac{e_t(X)(1-e_t(X))}{v_0(0,X)e_t(X)+v_0(1,X)(1-e_t(X))}\psi(X)\psi(X)^\top\bigg]\bigg)^{-1}$$

by applying Corollary 3.2 to the observations in each single batch $t \in \{1,\dots,T\}$. Then

$$\begin{aligned}
\big(V_{\mathrm{EPL}}^{(\mathrm{LA})}\big)^{-1} &= \sum_{t=1}^T \kappa_t V_{t,\mathrm{EPL}}^{-1}\\
&= \sum_{t=1}^T \mathbb{E}_t\bigg[\kappa_t \frac{e_t(X)(1-e_t(X))}{v_0(0,X)e_t(X)+v_0(1,X)(1-e_t(X))}\psi(X)\psi(X)^\top\bigg]\\
&= \mathbb{E}_0\bigg[\sum_{t=1}^T \kappa_t \frac{\mathrm{d}P_t^X}{\mathrm{d}P_0^X}(X)\frac{e_t(X)(1-e_t(X))}{v_0(0,X)e_t(X)+v_0(1,X)(1-e_t(X))}\psi(X)\psi(X)^\top\bigg]\\
&\preccurlyeq \mathbb{E}_0\bigg[\frac{e_0(X)(1-e_0(X))}{v_0(0,X)e_0(X)+v_0(1,X)(1-e_0(X))}\psi(X)\psi(X)^\top\bigg]\\
&= (V_{0,\mathrm{EPL}})^{-1}
\end{aligned}$$

where the inequality follows from the fact that for each $x \in \mathcal{X}$, we have $\sum_{t=1}^T \kappa_t \mathrm{d}P_t^X/\mathrm{d}P_0^X(x) = 1$ and the map

$$e \mapsto \frac{e(1-e)}{v(0,x)e+v(1,x)(1-e)}\psi(x)\psi(x)^\top$$

is concave on $e \in [0,1]$.

C.5 Proof of Theorem 4.1: CLT for feasible $\hat{\theta}$ in a CSBAE

Here we prove the CLT for the feasible estimator $\hat{\theta}$ in a CSBAE. It holds under the nonstationarity conditions described in Appendix A.

We begin by constructing the non-adaptive batch experiment in the statement of the theorem. Define the counterfactual treatment indicators via

$$\tilde{Z}_{ti} = \mathbf{1}(U_{ti} \leqslant e_t(X_{ti})), \quad t = 1,\dots,T, \quad i = 1,\dots,N_t$$

for $U_{ti} \overset{\mathrm{iid}}{\sim} \mathbb{U}(0,1)$. Then the observations in the counterfactual non-adaptive batch experiment are the vectors $\tilde{W}_{ti} = (\tilde{Y}_{ti}, X_{ti}, \tilde{Z}_{ti})$, where

$$\tilde{Y}_{ti} = Y_{ti}(\tilde{Z}_{it}) = \tilde{Z}_{it}Y_{ti}(1) + (1-\tilde{Z}_{it})Y_{ti}(0).$$

The corresponding oracle estimator $\hat{\theta}^*$ from (11) is then

$$\hat{\theta}^* = -\Big(\frac{1}{N}\sum_{k=1}^{K}\sum_{(t,i)\in\mathcal{I}_k} s_a(\tilde{W}_{ti};\nu_0,e_{0,N})\Big)^{-1}\Big(\frac{1}{N}\sum_{k=1}^{K}\sum_{(t,i)\in\mathcal{I}_k} s_b(\tilde{W}_{ti};\nu_0,e_{0,N})\Big).$$

Because $(\nu_0, e_{0,N}) \in \mathcal{T}_N$, conditions 1, 2, and 3 of Proposition 3.1 are satisfied for this counterfactual non-adaptive batch experiment by equations (23), (24), and (26) along with condition (a) of Assumption 4.1. Hence the oracle CLT $\sqrt{n}(\hat{\theta}^* - \theta_0) \xrightarrow{\mathrm{d}} \mathcal{N}(0, V_0)$ holds with V_0 as in the conclusion of Proposition 3.1.

It remains to show that

$$\hat{\theta} = \hat{\theta}^* + o_p(N^{-1/2})$$

where

$$\hat{\theta} = -\Big(\frac{1}{N}\sum_{k=1}^{K}\sum_{(t,i)\in\mathcal{I}_k} s_a(W_{ti};\hat{\nu}^{(-k)},\hat{e}^{(-k)})\Big)^{-1}\Big(\frac{1}{N}\sum_{k=1}^{K}\sum_{(t,i)\in\mathcal{I}_k} s_b(W_{ti};\hat{\nu}^{(-k)},\hat{e}^{(-k)})\Big)$$

as in (19). To do so, we show that the following intermediate quantity

$$\tilde{\theta} = -\Big(\frac{1}{N}\sum_{k=1}^{K}\sum_{(t,i)\in\mathcal{I}_k} s_a(\tilde{W}_{ti};\hat{\nu}^{(-k)},\hat{e}^{(-k)})\Big)^{-1}\Big(\frac{1}{N}\sum_{k=1}^{K}\sum_{(t,i)\in\mathcal{I}_k} s_b(\tilde{W}_{ti};\hat{\nu}^{(-k)},\hat{e}^{(-k)})\Big)$$

satisfies both $\tilde{\theta} = \hat{\theta}^* + o_p(N^{-1/2})$ and $\hat{\theta} = \tilde{\theta} + o_p(N^{-1/2})$.

C.5.1 Showing that $\tilde{\theta} = \hat{\theta}^* + o_p(N^{-1/2})$

We first show $\tilde{\theta} = \hat{\theta}^* + o_p(N^{-1/2})$. By score linearity (4), we can write

$$N^{1/2}(\tilde{\theta} - \hat{\theta}^*) = N^{1/2}(\tilde{\theta} - \theta_0) - N^{1/2}(\hat{\theta}^* - \theta_0) = A_1B_1 + A_2B_2$$

where

$$\begin{aligned}
A_1 &= \Big(\frac{1}{N}\sum_{k=1}^{K}\sum_{(t,i)\in\mathcal{I}_k} s_a(\tilde{W}_{ti};\nu_0,e_{0,N})\Big)^{-1} - \Big(\frac{1}{N}\sum_{k=1}^{K}\sum_{(t,i)\in\mathcal{I}_k} s_a(\tilde{W}_{ti};\hat{\nu}^{(-k)},\hat{e}^{(-k)})\Big)^{-1},\\
B_1 &= \frac{1}{\sqrt{N}}\sum_{k=1}^{K}\sum_{(t,i)\in\mathcal{I}_k} s(\tilde{W}_{ti};\theta_0,\nu_0,e_{0,N}),\\
A_2 &= \Big(\frac{1}{N}\sum_{k=1}^{K}\sum_{(t,i)\in\mathcal{I}_k} s_a(\tilde{W}_{ti};\hat{\nu}^{(-k)},\hat{e}^{(-k)})\Big)^{-1} \quad \text{and}\\
B_2 &= \frac{1}{\sqrt{N}}\sum_{k=1}^{K}\sum_{(t,i)\in\mathcal{I}_k} [s(\tilde{W}_{ti};\theta_0,\nu_0,e_{0,N}) - s(\tilde{W}_{ti};\theta_0,\hat{\nu}^{(-k)},\hat{e}^{(-k)})].
\end{aligned}$$

We will prove that $A_1 = o_p(1)$, $B_1 = O_p(1)$, $A_2 = O_p(1)$, and $B_2 = o_p(1)$.

To show $A_1 = o_p(1)$, for each fold $k \in \{1, \ldots, K\}$ let $N_k = \sum_{t=1}^{T} n_{t,k} = N/K + O(1)$ be the total number of observations in fold k across all batches $t = 1, \ldots, T$. Also define the quantity

$$\tilde{A}_1^{(k)} = \frac{1}{N_k} \sum_{(t,i)\in\mathcal{I}_k} [s_a(\tilde{W}_{ti}; \nu_0, e_{0,N}) - s_a(\tilde{W}_{ti}; \hat{\nu}^{(-k)}, \hat{e}^{(-k)})] = \sum_{t=1}^{T} \frac{n_{t,k}}{N_k} \tilde{A}_{1,t}^{(k)}$$

where

$$\tilde{A}_{1,t}^{(k)} = \frac{1}{n_{t,k}} \sum_{i:(t,i)\in\mathcal{I}_k} s_a(\tilde{W}_{ti}; \nu_0, e_{0,N}) - s_a(\tilde{W}_{ti}; \hat{\nu}^{(-k)}, \hat{e}^{(-k)}).$$

For each $k = 1, \ldots, K$, let $\mathcal{E}_{N,k}$ be the event that $(\hat{\nu}^{(-k)}, \hat{e}^{(-k)}) \in \mathcal{T}_N$. This $\mathcal{E}_{N,k}$ is $\mathcal{S}^{(-k)}$ measurable and $\Pr(\mathcal{E}_{N,k}) \to 1$ as $N \to \infty$ by assumption. Then for all sufficiently large N,

$$\begin{aligned}
&\mathbb{E}\big[\|\tilde{A}_{1,t}^{(k)}\mathbf{1}(\mathcal{E}_{N,k})\| \mid \mathcal{S}^{(-k)}\big] \\
&\leqslant \mathbf{1}(\mathcal{E}_{N,k})\bigg(\frac{1}{n_{t,k}} \sum_{i:(t,i)\in I_j} \mathbb{E}\big[\|s_a(\tilde{W}_{ti}; \hat{\nu}^{(-k)}, \hat{e}^{(-k)}) - s_a(\tilde{W}_{ti}; \nu_0, e_{0,N})\| \mid \mathcal{S}^{(-k)}\big]\bigg) \\
&\leqslant \sup_{(\nu,e)\in\mathcal{T}_N} \mathbb{E}_t\big[\|s_a(W; \nu, e) - s_a(W; \nu_0, e_{0,N})\|\big] \\
&\leqslant \kappa_t^{-1} \sup_{(\nu,e)\in\mathcal{T}_N} \mathbb{E}_0\big[\|s_a(W; \nu, e) - s_a(W; \nu_0, e_0)\|\big] + \kappa_t^{-1}\mathbb{E}_0[\|s_a(W; \nu_0, e_{0,N}) - s_a(W; \nu_0, e_0)\|] \\
&\leqslant 2\kappa_t^{-1}\delta_N
\end{aligned}$$

by equation (23). The second inequality above uses the fact that $\hat{\nu}^{(-k)}$, $\hat{e}^{(-k)}$ are nonrandom given $\mathcal{S}^{(-k)}$, but the vectors $\tilde{W}_{ti}$ in fold k are independent of $\mathcal{S}^{(-k)}$ and i.i.d. from P_t; the third inequality follows because

$$\frac{\mathrm{d}P_t}{\mathrm{d}P_0}(w) \leqslant \kappa_t^{-1} \tag{S12}$$

for all w by the definitions of P_t and P_0. Then $\tilde{A}_{1,t}^{(k)} = o_p(1)$ by Lemma B.1. This immediately shows $\tilde{A}_1^{(k)} = o_p(1)$ for all folds $k = 1, \ldots, K$ and so

$$\tilde{A}_1 = \sum_{k=1}^{K} \frac{N_k}{N} \tilde{A}_1^{(k)} = o_p(1), \tag{S13}$$

too. Now recall the identity

$$\|A^{-1} - B^{-1}\| = \|A^{-1}(A - B)B^{-1}\| \leqslant \|A^{-1}\|\|A - B\|\|B^{-1}\| \tag{S14}$$

for any invertible square matrices A, B of the same size. By (S6) we know that

$$\frac{1}{N} \sum_{k=1}^{K} \sum_{(t,i)\in\mathcal{I}_k} s_a(\tilde{W}_{ti}; \nu_0, e_{0,N}) = \mathbb{E}_0[s_a(W; \nu_0, e_0)] + o_p(1) \tag{S15}$$

and hence

$$\left(\frac{1}{N}\sum_{k=1}^{K}\sum_{(t,i)\in\mathcal{I}_k} s_a(\tilde{W}_{ti};\nu_0,e_{0,N})\right)^{-1} \quad \text{and} \quad \left(\frac{1}{N}\sum_{k=1}^{K}\sum_{(t,i)\in\mathcal{I}_k} s_a(\tilde{W}_{ti};\hat{\nu}^{(-k)},\hat{e}^{(-k)})\right)^{-1}$$

are both $O_p(1)$. The preceding display along with (S13) and (S14) show $A_1 = o_p(1)$.

Equations (S13) and (S15) immediately imply that $A_2 = O_p(1)$, while (S7) implies that $B_1 = O_p(1)$. It remains to show $B_2 = o_p(1)$. To that end we consider the quantity

$$\begin{aligned} B_2^{(k)} &= \frac{1}{\sqrt{N_k}}\sum_{(t,i)\in\mathcal{I}_k} s(\tilde{W}_{ti};\theta_0,\hat{\nu}^{(-k)},\hat{e}^{(-k)}) - s(\tilde{W}_{ti};\theta_0,\nu_0,e_{0,N}) \\ &= \bar{B}_2^{(k)} + \sum_{t=1}^{T}\sqrt{\frac{n_{t,k}}{N_k}}\tilde{B}_{2,t}^{(k)} \end{aligned}$$

where for each $t = 1,\ldots,T$

$$\begin{aligned} \bar{B}_2^{(k)} &= \sum_{t=1}^{T}\sqrt{\frac{n_{t,k}}{N_k}}\frac{1}{\sqrt{n_{t,k}}}\sum_{i:(t,i)\in\mathcal{I}_k}\mathbb{E}[s(\tilde{W}_{ti};\theta_0,\hat{\nu}^{(-k)},\hat{e}^{(-k)}) - s(\tilde{W}_{ti};\theta_0,\nu_0,e_{0,N}) \mid \mathcal{S}^{(-k)}], \quad \text{and} \\ \tilde{B}_{2,t}^{(k)} &= \frac{1}{\sqrt{n_{t,k}}}\sum_{i:(t,i)\in\mathcal{I}_k} s(\tilde{W}_{ti};\theta_0,\hat{\nu}^{(-k)},\hat{e}^{(-k)}) - s(\tilde{W}_{ti};\theta_0,\nu_0,e_{0,N}) \\ &\quad - \mathbb{E}[s(\tilde{W}_{ti};\theta_0,\hat{\nu}^{(-k)},\hat{e}^{(-k)}) - s(\tilde{W}_{ti};\theta_0,\nu_0,e_{0,N}) \mid \mathcal{S}^{(-k)}]. \end{aligned}$$

To see that $\bar{B}_2^{(k)} = o_p(1)$, we write

$$\begin{aligned} \bar{B}_2^{(k)} &= N_k^{1/2}\sum_{t=1}^{T}\frac{n_{t,k}}{N_k}\int s(w;\theta_0,\hat{\nu}^{(-k)},\hat{e}^{(-k)}) - s(w;\theta_0,\nu_0,e_{0,N})\mathrm{d}P_t(w) \\ &= N_k^{1/2}\int s(w;\theta_0,\hat{\nu}^{(-k)},\hat{e}^{(-k)}) - s(w;\theta_0,\nu_0,e_{0,N})\mathrm{d}P_{0,N}(w) \\ &\quad + N_k^{1/2}\sum_{t=1}^{T}\left(\frac{n_{t,k}}{N_k} - \frac{N_t}{N}\right)\int s(w;\theta_0,\hat{\nu}^{(-k)},\hat{e}^{(-k)}) - s(w;\theta_0,\nu_0,e_{0,N})\mathrm{d}P_t(w). \end{aligned}$$

Now

$$f_N^{(k)}(\lambda) = \int s(w;\theta_0,\nu_0+\lambda(\hat{\nu}^{(-k)}-\nu_0),e_{0,N}+\lambda(\hat{e}^{(-k)}-e_{0,N})) - s(w;\theta_0,\nu_0,e_{0,N})\mathrm{d}P_{0,N}(w)$$

is twice continuously differentiable on $[0,1]$ by regularity. Hence by Taylor's theorem

$$\|f_N^{(k)}(1) - f_N^{(k)}(0) - f_N^{(k)\prime}(0)\| \leqslant \frac{1}{2}\sup_{\lambda\in(0,1)}\|f_N^{(k)\prime\prime}(\lambda)\|. \tag{S16}$$

By (22) we know that

$$\begin{aligned} \sup_{\lambda\in(0,1)}\|f_N^{(k)\prime\prime}(\lambda)\|\mathbf{1}(\mathcal{E}_{N,k}) &\leqslant \sup_{\lambda\in(0,1)}\sup_{(\nu,e)\in\mathcal{T}_N}\left\|\frac{\partial^2}{\partial\lambda^2}\mathbb{E}_{0,N}[s(W;\theta_0,\nu_0+\lambda(\nu-\nu_0),e_{0,N}+\lambda(e-e_{0,N}))]\right\| \\ &\leqslant N^{-1/2}\delta_N. \end{aligned}$$

With $f_N^{(k)}(0) = 0$, by (21) and (S16) we conclude that that

$$\|f_N^{(k)}(1)\| \mathbf{1}(\mathcal{E}_{N,K}) \leqslant \frac{3}{2} N^{-1/2} \delta_N.$$

Recalling that $\Pr(\mathcal{E}_{N,k}) \to 1$, we have $\|f_N^{(k)}(1)\| = o_p(N^{-1/2})$ and

$$\left\| N_k^{1/2} \int s(w; \theta_0, \hat{\nu}^{(-k)}, \hat{e}^{(-k)}) - s(w; \theta_0, \nu_0, e_{0,N}) \, \mathrm{d}P_{0,N}(w) \right\| = N_k^{1/2} \|f_N^{(k)}(1)\| = o_p(1).$$

Using $N_t/N \to \kappa_t$ from (3) we have

$$\begin{aligned}
&\left\| N_k^{1/2} \sum_{t=1}^{T} \left(\frac{n_{t,k}}{N_k} - \frac{N_t}{N} \right) \int s(w; \theta_0, \hat{\nu}^{(-k)}, \hat{e}^{(-k)}) - s(w; \theta_0, \nu_0, e_{0,N}) \, \mathrm{d}P_t(w) \right\| \\
&\leqslant N_k^{1/2} \sum_{t=1}^{T} \left| \frac{n_{t,k}}{N_k} - \frac{N_t}{N} \right| \cdot \frac{N}{N_t} \|f_N^{(k)}(1)\| \\
&= o_p(1)
\end{aligned}$$

as

$$\frac{\mathrm{d}P_t}{\mathrm{d}P_{0,N}}(w) \leqslant \frac{N}{N_t}, \quad t = 1, \ldots, T, \quad w \in \mathcal{W}. \tag{S17}$$

Thus we have shown $\bar{B}_2^{(k)} = o_p(1)$. Finally, for each batch $t = 1, \ldots, T$, the quantity $\tilde{B}_{2,t}^{(k)}$ is a sum of $n_{t,k}$ random variables that are i.i.d. and mean 0 conditional on $\mathcal{S}^{(-k)}$. Thus by Lemma B.7, we have

$$\begin{aligned}
\mathbb{E}\big[\|\tilde{B}_{2,t}^{(k)}\|^2 \mid \mathcal{S}^{(-k)}\big] &= \sum_{i:(t,i) \in \mathcal{I}_k} \frac{1}{n_{t,k}} \bar{\mathbb{E}}\big[\|r_{ti}^{(k)}\|^2\big] \\
&\leqslant \sum_{i:(t,i) \in \mathcal{I}_k} \frac{1}{n_{t,k}} \mathbb{E}\big[\|s(W_{ti}; \theta_0, \hat{\nu}^{(-k)}, \hat{e}^{(-k)}) - s(W_{ti}; \theta_0, \nu_0, e_{0,N})\|^2 \mid \mathcal{S}^{(-k)}\big]
\end{aligned}$$

where for each i such that $(t, i) \in \mathcal{I}_k$ we've defined

$$\begin{aligned}
r_{ti}^{(k)} &= s(W_{ti}; \theta_0, \hat{\nu}^{(-k)}, \hat{e}^{(-k)}) - s(\tilde{W}_{ti}; \theta_0, \nu_0, e_{0,N}) \\
&\quad - \mathbb{E}[s(\tilde{W}_{ti}; \theta_0, \hat{\nu}^{(-k)}, \hat{e}^{(-k)}) - s(\tilde{W}_{ti}; \theta_0, \nu_0, e_{0,N}) \mid \mathcal{S}^{(-k)}]
\end{aligned} \tag{S18}$$

and used the basic variance inequality

$$\mathbb{E}\big[\|X - \mathbb{E}[X \mid \mathcal{S}]\|^2\big] \leqslant \mathbb{E}\big[\|X\|^2 \mid \mathcal{S}\big] \tag{S19}$$

for any random vector X and σ-algebra $\mathcal{S}$. Hence by (S12), Minkowski's inequality, and then (24), we have

$$\begin{aligned}
\big(\mathbb{E}\big[\|\tilde{B}_{2,t}^{(k)}\|^2 \mid \mathcal{S}^{(-k)}\big]\big)^{1/2} \mathbf{1}(\mathcal{E}_{N,k}) &\leqslant \sup_{(\nu,e) \in \mathcal{T}_N} \big(\mathbb{E}_t\big[\|s(W; \theta_0, \nu, e) - s(W; \theta_0, \nu_0, e_{0,N})\|^2\big]\big)^{1/2} \\
&\leqslant \kappa_t^{-1/2} \sup_{(\nu,e) \in \mathcal{T}_N} \big(\mathbb{E}_{0,N}\big[\|s(W; \theta_0, \nu, e) - s(W; \theta_0, \nu_0, e_0)\|^2\big]\big)^{1/2} \\
&\quad + \kappa_t^{-1/2} \big(\mathbb{E}_{0,N}\big[\|s(W; \theta_0, \nu_0, e_{0,N}) - s(W; \theta_0, \nu_0, e_0)\|^2\big]\big)^{1/2} \\
&\leqslant 2\kappa_t^{-1/2} \delta_N.
\end{aligned}$$

Thus $\tilde{B}_{2,t}^{(k)} = o_p(1)$ by Lemma B.1. With

$$-B_2 = \sum_{k=1}^{K} \sqrt{\frac{N_k}{N}} B_2^{(k)} = \sum_{k=1}^{K} \sqrt{\frac{N_k}{N}} \left(\bar{B}_2^{(k)} + \sum_{t=1}^{T} \sqrt{\frac{n_{t,k}}{N_k}} \tilde{B}_{2,t}^{(k)} \right)$$

we conclude that $B_2 = o_p(1)$, as desired. This establishes that $\hat{\theta}^* = \tilde{\theta} + o_p(N^{-1/2})$.

C.5.2 Showing that $\hat{\theta} = \hat{\theta}^* + o_p(N^{-1/2})$

To establish that $\hat{\theta} = \tilde{\theta} + o_p(N^{-1/2})$, similar to above we write

$$N^{1/2}(\hat{\theta} - \tilde{\theta}) = N^{1/2}(\hat{\theta} - \theta_0) - N^{1/2}(\tilde{\theta} - \theta_0) = A_3 B_3 + A_4 B_4$$

where

$$\begin{aligned}
A_3 &= \left(\frac{1}{N} \sum_{k=1}^{K} \sum_{(t,i)\in\mathcal{I}_k} s_a(\tilde{W}_{ti}; \hat{\nu}^{(-k)}, \hat{e}^{(-k)}) \right)^{-1} - \left(\frac{1}{N} \sum_{k=1}^{K} \sum_{(t,i)\in\mathcal{I}_k} s_a(W_{ti}; \hat{\nu}^{(-k)}, \hat{e}^{(-k)}) \right)^{-1}, \\
B_3 &= \frac{1}{\sqrt{N}} \sum_{k=1}^{K} \sum_{(t,i)\in\mathcal{I}_k} s(\tilde{W}_{ti}; \theta_0, \hat{\nu}^{(-k)}, \hat{e}^{(-k)}), \\
A_4 &= \left(\frac{1}{N} \sum_{k=1}^{K} \sum_{(t,i)\in\mathcal{I}_k} s_a(W_{ti}; \hat{\nu}^{(-k)}, \hat{e}^{(-k)}) \right)^{-1}, \quad \text{and} \\
B_4 &= \frac{1}{\sqrt{N}} \sum_{k=1}^{K} \sum_{(t,i)\in\mathcal{I}_k} [s(\tilde{W}_{ti}; \theta_0, \hat{\nu}^{(-k)}, \hat{e}^{(-k)}) - s(W_{ti}; \theta_0, \hat{\nu}^{(-k)}, \hat{e}^{(-k)})].
\end{aligned}$$

We will show that $A_3 = o_p(1)$, $B_3 = O_p(1)$, $A_4 = O_p(1)$, and $B_4 = o_p(1)$. First we define

$$\tilde{A}_3 = \sum_{k=1}^{K} \frac{N_k}{N} \tilde{A}_3^{(k)}$$

where the term $\tilde{A}_3^{(k)}$ for each fold is decomposed into a sum over batches $t = 1, \ldots, T$:

$$\begin{aligned}
\tilde{A}_3^{(k)} &= \frac{1}{N_k} \sum_{(t,i)\in\mathcal{I}_k} [s_a(\tilde{W}_{ti}; \hat{\nu}^{(-k)}, \hat{e}^{(-k)}) - s_a(W_{ti}; \hat{\nu}^{(-k)}, \hat{e}^{(-k)})] \\
&= \frac{1}{N_k} \sum_{(t,i)\in\mathcal{I}_k} (\tilde{Z}_{ti} - Z_{ti})(s_a(W_{ti}(1); \hat{\nu}^{(-k)}, \hat{e}^{(-k)}) - s_a(W_{ti}(0); \hat{\nu}^{(-k)}, \hat{e}^{(-k)})) \\
&= \sum_{t=1}^{T} \frac{n_{t,k}}{N_k} \tilde{A}_{3,t}^{(k)}.
\end{aligned}$$

Here

$$\tilde{A}_{3,t}^{(k)} := \frac{1}{n_{t,k}} \sum_{i:(t,i)\in\mathcal{I}_k} (\tilde{Z}_{ti} - Z_{ti})(s_a(W_{ti}(1); \hat{\nu}^{(-k)}, \hat{e}^{(-k)}) - s_a(W_{ti}(0); \hat{\nu}^{(-k)}, \hat{e}^{(-k)}))$$

satisfies

$$\begin{aligned}\|\tilde{A}_{3,t}^{(k)}\| &\leqslant \frac{1}{n_{t,k}} \sum_{i:(t,i)\in\mathcal{I}_k} \mathbf{1}(Z_{ti} \neq \tilde{Z}_{ti}) \cdot \|s_a(W_{ti}(1); \hat{\nu}^{(-k)}, \hat{e}^{(-k)}) - s_a(W_{ti}(0); \hat{\nu}^{(-k)}, \hat{e}^{(-k)})\| \\ &\leqslant \sqrt{\frac{1}{n_{t,k}} \sum_{i:(t,i)\in\mathcal{I}_k} \|s_a(W_{ti}(1); \hat{\nu}^{(-k)}, \hat{e}^{(-k)}) - s_a(W_{ti}(0); \hat{\nu}^{(-k)}, \hat{e}^{(-k)})\|^2} \\ &\quad \cdot \sqrt{\frac{1}{n_{t,k}} \sum_{i:(t,i)\in\mathcal{I}_k} \mathbf{1}(Z_{ti} \neq \tilde{Z}_{ti})}\end{aligned}$$

by the Cauchy-Schwarz inequality. Recalling the σ-algebra $\mathcal{S}_t^{X,(k)}$ in the definition of a CSBAE (Definition 4.1), we have

$$\Pr(Z_{ti} \neq \tilde{Z}_{ti} \mid \mathcal{S}_t^{X,(k)}) = |\hat{e}_t^{(k)}(X_{ti}) - e_t(X_{ti})|$$

and so by (18)

$$\begin{aligned}\mathbb{E}\left[\frac{1}{n_{t,k}} \sum_{i:(t,i)\in\mathcal{I}_k} \mathbf{1}(Z_{ti} \neq \tilde{Z}_{ti}) \;\middle|\; \mathcal{S}_t^{X,(k)}\right] &= \frac{1}{n_{t,k}} \sum_{i:(t,i)\in\mathcal{I}_k} |\hat{e}_t^{(k)}(X_{ti}) - e_t(X_{ti})| \\ &\leqslant \|\hat{e}_t^{(k)} - e_t\|_{2,P_{N,t}^{X,(k)}} \\ &= o_p(1).\end{aligned}$$

Hence

$$\sqrt{\frac{1}{n_{t,k}} \sum_{i:(t,i)\in\mathcal{I}_k} \mathbf{1}(Z_{ti} \neq \tilde{Z}_{ti})} = o_p(1)$$

by Lemma B.1. Next, for $z = 0, 1$ and $(t, i) \in \mathcal{I}_k$, by Jensen's inequality and (S12) we have

$$\begin{aligned}\big(\mathbb{E}\big[\|s_a(W_{ti}(z); \hat{\nu}^{(-k)}, \hat{e}^{(-k)})\|^2 \mathbf{1}(\mathcal{E}_{N,k}) \mid \mathcal{S}^{(-k)}\big]\big)^{q/2} &\leqslant \sup_{(\nu,e)\in\mathcal{T}_N} (\mathbb{E}_t\big[\|s_a(W(z); \nu, e)\|^2\big])^{q/2} \\ &\leqslant \kappa_t^{-1} \sup_{(\nu,e)\in\mathcal{T}_N} \mathbb{E}_0[\|s_a(W(z); \nu, e)\|^q] \\ &\leqslant \kappa_t^{-1} C^q\end{aligned}$$

by (26). By Markov's inequality, for each batch $t = 1, \ldots, T$ and fold $k = 1, \ldots, K$,

$$\sqrt{\frac{1}{n_{t,k}} \sum_{i:(t,i)\in\mathcal{I}_k} \|s_a(W_{ti}(z); \hat{\nu}^{(-k)}, \hat{e}^{(-k)})\|^2} = O_p(1)$$

holds for $z = 0, 1$. Applying Minkowski's inequality with the $L^2_{P_{N,t}^X}$ norm shows

$$\sqrt{\frac{1}{n_{t,k}} \sum_{i:(t,i)\in\mathcal{I}_k} \|s_a(W_{ti}(1); \hat{\nu}^{(-k)}, \hat{e}^{(-k)}) - s_a(W_{ti}(0); \hat{\nu}^{(-k)}, \hat{e}^{(-k)})\|^2} = O_p(1).$$

We conclude $\tilde{A}_{3,t}^{(k)} = o_p(1)$, hence $\tilde{A}_3^{(k)} = o_p(1)$ and $\tilde{A}_3 = o_p(1)$ as well. Now we recall

$$\begin{aligned}\frac{1}{N}\sum_{k=1}^{K}\sum_{(t,i)\in\mathcal{I}_k} s_a(\tilde{W}_{ti};\hat{\nu}^{(-k)},\hat{e}^{(-k)}) &= \frac{1}{N}\sum_{k=1}^{K}\sum_{(t,i)\in\mathcal{I}_k} s_a(\tilde{W}_{ti};\nu_0,e_{0,N}) + o_p(1) \quad \text{(by (S13))}\\ &= \mathbb{E}_0[s_a(W;\nu_0,e_0)] + o_p(1) \quad \text{(by (S15))}.\end{aligned}$$

Then

$$\begin{aligned}\frac{1}{N}\sum_{k=1}^{K}\sum_{(t,i)\in\mathcal{I}_k} s_a(W_{ti};\hat{\nu}^{(-k)},\hat{e}^{(-k)}) &= \frac{1}{N}\sum_{k=1}^{K}\sum_{(t,i)\in\mathcal{I}_k} s_a(\tilde{W}_{ti};\hat{\nu}^{(-k)},\hat{e}^{(-k)}) - \tilde{A}_3\\ &= \mathbb{E}_0[s_a(W;\nu_0,e_0)] + o_p(1)\end{aligned}$$

as well. Thus we have both

$$\begin{aligned}\left(\frac{1}{N}\sum_{k=1}^{K}\sum_{(t,i)\in\mathcal{I}_k} s_a(\tilde{W}_{ti};\hat{\nu}^{(-k)},\hat{e}^{(-k)})\right)^{-1} &= (\mathbb{E}_0[s_a(W;\nu_0,e_0)])^{-1} + o_p(1), \quad \text{and}\\ \left(\frac{1}{N}\sum_{k=1}^{K}\sum_{(t,i)\in\mathcal{I}_k} s_a(W_{ti};\hat{\nu}^{(-k)},\hat{e}^{(-k)})\right)^{-1} &= (\mathbb{E}_0[s_a(W;\nu_0,e_0)])^{-1} + o_p(1). \end{aligned} \tag{S20}$$

Subtracting these two equations gives $A_3 = o_p(1)$.

Next, equation (S20) immediately provides $A_4 = O_p(1)$, while $B_3 = B_1 - B_2 = O_p(1)$ as shown above. Thus it only remains to show that $B_4 = o_p(1)$. We write

$$B_4 = \sum_{k=1}^{K}\sqrt{\frac{N_k}{N}}\Big(B_4^{(k),1} + B_4^{(k),2} + B_4^{(k),3}\Big)$$

where

$$B_4^{(k),1} = \sum_{t=1}^{T}\sqrt{\frac{n_{t,k}}{N_k}}\frac{1}{\sqrt{n_{t,k}}}\sum_{i:(t,i)\in\mathcal{I}_k}(\tilde{Z}_{ti}-Z_{ti})\big(s(W_{ti}(1);\theta_0,\nu_0,e_{0,N}) - s(W_{ti}(0);\theta_0,\nu_0,e_{0,N})\big),$$

$$\begin{aligned}B_4^{(k),2} &= N_k^{1/2}\sum_{t=1}^{T}\frac{n_{t,k}}{N_k}\frac{1}{n_{t,k}}\sum_{i:(t,i)\in\mathcal{I}_k}(\tilde{Z}_{ti}-Z_{ti})\mathbb{E}[s(W_{ti}(1);\theta_0,\hat{\nu}^{(-k)},\hat{e}^{(-k)}) - s(W_{ti}(1);\theta_0,\nu_0,e_{0,N}) \mid \mathcal{S}^{(-k)},X_{ti}]\\ &\quad - N_k^{1/2}\sum_{t=1}^{T}\frac{n_{t,k}}{N_k}\frac{1}{n_{t,k}}\sum_{i:(t,i)\in\mathcal{I}_k}(\tilde{Z}_{ti}-Z_{ti})\mathbb{E}[s(W_{ti}(0);\theta_0,\hat{\nu}^{(-k)},\hat{e}^{(-k)}) - s(W_{ti}(0);\theta_0,\nu_0,e_{0,N}) \mid \mathcal{S}^{(-k)},X_{ti}]\end{aligned}$$

and

$$\begin{aligned}B_4^{(k),3} &= \sum_{t=1}^{T}\sqrt{\frac{n_{t,k}}{N_k}}\frac{1}{\sqrt{n_{t,k}}}\sum_{i:(t,i)\in\mathcal{I}_k}(\tilde{Z}_{ti}-Z_{ti})\big(s(W_{ti}(1);\theta_0,\hat{\nu}^{(-k)},\hat{e}^{(-k)}) - s(W_{ti}(0);\theta_0,\hat{\nu}^{(-k)},\hat{e}^{(-k)})\big)\\ &\quad - B_4^{(k),1} - B_4^{(k),2}.\end{aligned}$$

It suffices to prove that for each $k = 1, \dots, K$, the terms $B_4^{(k),1}$, $B_4^{(k),2}$, and $B_4^{(k),3}$ are all asymptotically negligible. First, we have

$$\begin{aligned}
&\mathbb{E}\left[(\tilde{Z}_{ti} - Z_{ti})(s(W_{ti}(1); \theta_0, \nu_0, e_{0,N}) - s(W_{ti}(0); \theta_0, \nu_0, e_{0,N})) \mid \mathcal{S}_t^{X,(k)}\right] \\
&= (e_t(X_{ti}) - \hat{e}_t^{(k)}(X_{ti}))\mathbb{E}[s(W_{ti}(1); \theta_0, \nu_0, e_{0,N}) - s(W_{ti}(0); \theta_0, \nu_0, e_{0,N}) \mid \mathcal{S}_t^{X,(k)}] \\
&= (e_t(X_{ti}) - \hat{e}_t^{(k)}(X_{ti}))\mathbb{E}[s(W_{ti}(1); \theta_0, \nu_0, e_{0,N}) - s(W_{ti}(0); \theta_0, \nu_0, e_{0,N}) \mid X_{ti}] \\
&= 0.
\end{aligned}$$

We briefly justify each of the equalities in the preceding display:

- The first equality holds because given $\mathcal{S}_t^{X,(k)}$, the only randomness in $(Z_{ti}, \tilde{Z}_{ti})$ for any $(t, i) \in \mathcal{I}_k$ is in U_{ti}, which is independent of $(W_{ti}(0), W_{ti}(1))$, $\mathcal{S}^{(-k)}$, and $\mathcal{S}_t^{X,(k)}$; hence

$$\begin{aligned}
&(Z_{ti}, \tilde{Z}_{ti}) \perp\!\!\!\perp (W_{ti}(0), W_{ti}(1)) \mid \mathcal{S}_t^{X,(k)} \quad \text{and} \\
&(Z_{ti}, \tilde{Z}_{ti}) \perp\!\!\!\perp (W_{ti}(0), W_{ti}(1)) \mid \mathcal{S}^{(-k)}, \mathcal{S}_t^{X,(k)}.
\end{aligned} \tag{S21}$$

- The second equality holds because the vectors $\{S_{ti}, 1 \leqslant t \leqslant T, 1 \leqslant i \leqslant N_t\}$ are mutually independent (Assumption A.1).

- The third equality follows directly from (20).

For batches $t = 1, \dots, T$, the vectors $\{(\tilde{Z}_{ti}, Z_{ti}, W_{ti}(1), W_{ti}(0)), 1 \leqslant i \leqslant N_t\}$ are conditionally independent given $\mathcal{S}_t^{X,(k)}$. Defining

$$\mathcal{I}_{4,t}^{(k)} = \frac{1}{\sqrt{n_{t,k}}} \sum_{i:(t,i) \in \mathcal{I}_k} (\tilde{Z}_{ti} - Z_{ti})(s(W_{ti}(1); \theta_0, \nu_0, e_{0,N}) - s(W_{ti}(0); \theta_0, \nu_0, e_{0,N}))$$

and letting p be the real solution to $p^{-1} + 2q^{-1} = 1$, we apply Lemma B.7 and Holder's inequality to get

$$\begin{aligned}
&\mathbb{E}[\|\mathcal{I}_{4,t}^{(k)}\|^2 \mid \mathcal{S}_t^{X,(k)}] \\
&= \frac{1}{n_{t,k}} \sum_{i:(t,i) \in \mathcal{I}_k} \mathbb{E}[(\tilde{Z}_{ti} - Z_{ti})^2 \|s(W_{ti}(1); \theta_0, \nu_0, e_{0,N}) - s(W_{ti}(0); \theta_0, \nu_0, e_{0,N})\|^2 \mid \mathcal{S}_t^{X,(k)}] \\
&= \frac{1}{n_{t,k}} \sum_{i:(t,i) \in \mathcal{I}_k} |\hat{e}_t^{(k)}(X_{ti}) - e_t(X_{ti})| \mathbb{E}\left[\|s(W_{ti}(1); \theta_0, \nu_0, e_{0,N}) - s(W_{ti}(0); \theta_0, \nu_0, e_{0,N})\|^2 \mid \mathcal{S}_t^{X,(k)}\right]
\end{aligned}$$

using the conditional independence in (S21) once again. Then

$$\begin{aligned}
\mathbb{E}[\|\mathcal{I}_{4,t}^{(k)}\|^2 \mid \mathcal{S}_t^{X,(k)}] \leqslant &\left(\frac{1}{n_{t,k}} \sum_{i:(t,i) \in \mathcal{I}_k} |\hat{e}_t^{(k)}(X_{ti}) - e_t(X_{ti})|^p \right)^{1/p} \\
&\cdot \left(\frac{1}{n_{t,k}} \sum_{i:(t,i) \in \mathcal{I}_k} \left(\mathbb{E}[\|s(W_{ti}(1); \theta_0, \nu_0, e_{0,N}) - s(W_{ti}(0); \theta_0, \nu_0, e_{0,N})\|^2 \mid \mathcal{S}_t^{X,(k)}] \right)^{q/2} \right)^{2/q}.
\end{aligned}$$

We can assume WLOG that $q < 4$ so that $p > 2$. Then with $0 \leqslant |\hat{e}_t^{(j)}(x) - e_t(x)| \leqslant 1$ for all x, we conclude

$$\frac{1}{n_{t,k}} \sum_{i:(t,i)\in\mathcal{I}_k} \left|\hat{e}_t^{(k)}(X_{ti}) - e_t(X_{ti})\right|^p \leqslant \frac{1}{n_{t,k}} \sum_{i:(t,i)\in\mathcal{I}_k} \left(\hat{e}_t^{(k)}(X_{ti}) - e_t(X_{ti})\right)^2 = o_p(1) \tag{S22}$$

by (18). Next, from Jensen's inequality

$$\begin{aligned}
&\frac{1}{n_{t,k}} \sum_{i:(t,i)\in\mathcal{I}_k} \Big(\mathbb{E}\big[\|s(W_{ti}(1);\theta_0,\nu_0,e_{0,N}) - s(W_{ti}(0);\theta_0,\nu_0,e_{0,N})\|^2 \mid \mathcal{S}_t^{X,(k)}\big]\Big)^{q/2} \\
&\leqslant \frac{1}{n_{t,k}} \sum_{i:(t,i)\in\mathcal{I}_k} \mathbb{E}\left[\|s(W_{ti}(1);\theta_0,\nu_0,e_{0,N}) - s(W_{ti}(0);\theta_0,\nu_0,e_{0,N})\|^q \mid \mathcal{S}_t^{X,(k)}\right] \\
&= \mathbb{E}_t\big[\|s(W(1);\theta_0,\nu_0,e_{0,N}) - s(W(0);\theta_0,\nu_0,e_{0,N})\|^q \mid X\big].
\end{aligned}$$

But by Minkowski's inequality, (S12), and (26), we see

$$\begin{aligned}
&\big(\mathbb{E}_t[\|s(W(1);\theta_0,\nu_0,e_{0,N}) - s(W(0);\theta_0,\nu_0,e_{0,N})\|^q]\big)^{1/q} \\
&\leqslant \big(\mathbb{E}_t[\|s(W(1);\theta_0,\nu_0,e_{0,N})\|^q]\big)^{1/q} + \big(\mathbb{E}_t[\|s(W(0);\theta_0,\nu_0,e_{0,N})\|^q]\big)^{1/q} \\
&\leqslant \kappa_t^{-1/q}[(\mathbb{E}_0[\|s(W(1);\theta_0,\nu_0,e_{0,N})\|^q])^{1/q} + (\mathbb{E}_0[\|s(W(0);\theta_0,\nu_0,e_{0,N})\|^q])^{1/q}] \\
&\leqslant 2C\kappa_t^{-1/q}.
\end{aligned}$$

We conclude by Markov's inequality that

$$\frac{1}{n_{t,k}} \sum_{i:(t,i)\in\mathcal{I}_k} \Big(\mathbb{E}[\|s(W_{ti}(1);\theta_0,\nu_0,e_{0,N}) - s(W_{ti}(0);\theta_0,\nu_0,e_{0,N})\|^2 \mid \mathcal{S}_t^{X,(k)}]\Big)^{q/2} = O_p(1).$$

Along with (S22) we can conclude that $\mathcal{I}_{4,t}^{(k)} = o_p(1)$ by Lemma B.1. Then also

$$B_4^{(k),1} = \sum_{t=1}^{T} \sqrt{\frac{n_{t,k}}{N_k}} \mathcal{I}_{4,t}^{(k)} = o_p(1)$$

Next, for $z = 0, 1$ define

$$B_4^{(k),2}(z) = N_k^{1/2} \sum_{t=1}^{T} \frac{n_{t,k}}{N_k} \frac{1}{n_{t,k}} \sum_{i:(t,i)\in\mathcal{I}_k} (\tilde{Z}_{ti} - Z_{ti})\mathbb{E}\big[s(W_{ti}(z);\theta_0,\hat{\nu}^{(-k)},\hat{e}^{(-k)}) - s(W_{ti}(z);\theta_0,\nu_0,e_{0,N}) \mid \mathcal{S}^{(-k)}, X_{ti}\big]$$

so that $B_4^{(k),2} = B_4^{(k),2}(1) - B_4^{(k),2}(0)$. Then by the triangle inequality $\|B_4^{(k),2}(z)\|$ is no larger than

$$N_k^{1/2} \sum_{t=1}^{T} \frac{n_{t,k}}{N_k} \frac{1}{n_{t,k}} \sum_{i:(t,i)\in\mathcal{I}_k} |\tilde{Z}_{ti} - Z_{ti}| \cdot \|\mathbb{E}[s(W_{ti}(z);\theta_0,\hat{\nu}^{(-k)},\hat{e}^{(-k)}) - s(W_{ti}(z);\theta_0,\nu_0,e_{0,N}) \mid \mathcal{S}^{(-k)}, X_{ti}]\|.$$

Taking conditional expectations of both sides yields

$$
\begin{aligned}
&\mathbb{E}\big[\|B_4^{(k),2}(z)\| \mid \mathcal{S}^{(-k)}, \mathcal{S}_t^{X,(k)}\big] \\
&\leqslant N_k^{1/2} \sum_{t=1}^{T} \frac{n_{t,k}}{N_k} \frac{1}{n_{t,k}} \sum_{i:(t,i)\in\mathcal{I}_k} \big|\hat{e}_t^{(k)}(X_{ti}) - e_t(X_{ti})\big| \\
&\qquad\qquad \times \big\|\mathbb{E}[s(W_{ti}(z);\theta_0,\hat{\nu}^{(-k)},\hat{e}^{(-k)}) - s(W_{ti}(z);\theta_0,\nu_0,e_{0,N}) \mid \mathcal{S}^{(-k)}, X_{ti}]\big\| \\
&\leqslant N_k^{1/2} \sum_{t=1}^{T} \frac{n_{t,k}}{N_k} \cdot S_t^{(-k)}(z) \cdot \sqrt{\frac{1}{n_{t,k}} \sum_{i:(t,i)\in\mathcal{I}_k} (\hat{e}_t^{(k)}(X_{ti}) - e_t(X_{ti}))^2}
\end{aligned}
$$

by Cauchy-Schwarz. We have by (18) that

$$
\sqrt{\frac{1}{n_{t,k}} \sum_{i:(t,i)\in\mathcal{I}_k} (\hat{e}_t^{(k)}(X_{ti}) - e_t(X_{ti}))^2} = O_p(N^{-1/4}),
$$

Then by equation (27), we get $B_4^{(k),2}(z) = o_p(1)$ for $z = 0, 1$, and so $B_4^{(k),2} = o_p(1)$ as well.

Finally, we write

$$
B_4^{(k),3} = \sum_{t=1}^{T} \sqrt{\frac{n_{t,k}}{N_k}} B_{4,t}^{(k),3}
$$

for

$$
B_{4,t}^{(k),3} = \frac{1}{\sqrt{n_{t,k}}} \sum_{i:(t,i)\in\mathcal{I}_k} (\tilde{Z}_{ti} - Z_{ti}) \Delta_{ti}^{(-k)},
$$

where

$$
\begin{aligned}
\Delta_{ti}^{(-k)} &= (s(W_{ti}(1);\theta_0,\hat{\nu}^{(-k)},\hat{e}^{(-k)}) - s(W_{ti}(0);\theta_0,\hat{\nu}^{(-k)},\hat{e}^{(-k)})) \\
&\quad - (s(W_{ti}(1);\theta_0,\nu_0,e_{0,N}) - s(W_{ti}(0);\theta_0,\nu_0,e_{0,N})) \\
&\quad - \mathbb{E}[s(W_{ti}(1);\theta_0,\hat{\nu}^{(-k)},\hat{e}^{(-k)}) - s(W_{ti}(0);\theta_0,\hat{\nu}^{(-k)},\hat{e}^{(-k)}) \mid \mathcal{S}^{(-k)}, X_{ti}] \\
&\quad - \mathbb{E}[s(W_{ti}(1);\theta_0,\nu_0,e_0) - s(W_{ti}(0);\theta_0,\nu_0,e_0) \mid \mathcal{S}^{(-k)}, X_{ti}].
\end{aligned}
$$

For each $(t,i) \in \mathcal{I}_k$, we have

$$
(W_{ti}(0), W_{ti}(1)) \perp\!\!\!\perp \mathcal{S}_t^{X,(k)} \mid \mathcal{S}^{(-k)}, X_{ti} \tag{S23}
$$

since given $\mathcal{S}^{(-k)}$ and X_{ti}, the only remaining randomness in $(W_{ti}(0), W_{ti}(1))$ is in the potential outcomes $(Y_{ti}(0), Y_{ti}(1))$. These are independent of both $\mathcal{S}^{(-k)}$ and $\{X_{tj} \mid j \neq i\}$, the covariates of the other subjects in batch t. Thus $\mathbb{E}[\Delta_{ti}^{(-k)} \mid \mathcal{S}^{(-k)}, \mathcal{S}_t^{X,(k)}] = 0$, and by (S21) we have

$$
\mathbb{E}[B_{4,t}^{(k),3} \mid \mathcal{S}^{(-k)}, \mathcal{S}_t^{X,(k)}] = 0.
$$

Then applying Lemma B.7 and equations (S21) and (S19),

$$\begin{aligned}
\mathbb{E}[\|B_{4,t}^{(k),3}\|^2 \mid \mathcal{S}^{(-k)}, \mathcal{S}_t^{X,(k)}] &= \frac{1}{n_{t,k}} \sum_{i:(t,i)\in\mathcal{I}_k} \mathbb{E}[(\tilde{Z}_{ti} - Z_{ti})^2 \|\Delta_{ti}^{(-k)}\|^2 \mid \mathcal{S}^{(-k)}, \mathcal{S}_t^{X,(k)}] \\
&= \frac{1}{n_{t,k}} \sum_{i:(t,i)\in\mathcal{I}_k} |\hat{e}_t^{(k)}(X_{ti}) - e_t(X_{ti})| \mathbb{E}[\|\Delta_{ti}^{(-k)}\|^2 \mid \mathcal{S}^{(-k)}, \mathcal{S}_t^{X,(k)}] \\
&\leqslant \frac{1}{n_{t,k}} \sum_{i:(t,i)\in\mathcal{I}_k} |\hat{e}_t^{(k)}(X_{ti}) - e_t(X_{ti})| \mathbb{E}[\|\tilde{\Delta}_{ti}^{(-k)}\|^2 \mid \mathcal{S}^{(-k)}, \mathcal{S}_t^{X,(k)}] \\
&\leqslant \left(\frac{1}{n_{t,k}} \sum_{i:(t,i)\in\mathcal{I}_k} |\hat{e}_t^{(k)}(X_{ti}) - e_t(X_{ti})|^p\right)^{1/p} \Gamma_t^{(k)}
\end{aligned}$$

by Holder's inequality, where p is as above for

$$\begin{aligned}
\Gamma_t^{(k)} &= \left(\frac{1}{n_{t,k}} \sum_{i:(t,i)\in\mathcal{I}_k} (\mathbb{E}[\|\tilde{\Delta}_{ti}^{(-k)}\|^2 \mid \mathcal{S}^{(-k)}, \mathcal{S}_t^{X,(k)}])^{q/2}\right)^{2/q}, \quad \text{with} \\
\tilde{\Delta}_{ti}^{(-k)} &= (s(W_{ti}(1); \theta_0, \hat{\nu}^{(-k)}, \hat{e}^{(-k)}) - s(W_{ti}(0); \theta_0, \hat{\nu}^{(-k)}, \hat{e}^{(-k)})) \\
&\quad - (s(W_{ti}(1); \theta_0, \nu_0, e_{0,N}) - s(W_{ti}(0); \theta_0, \nu_0, e_{0,N})).
\end{aligned}$$

By Jensen's inequality

$$(\Gamma_t^{(k)})^{q/2} \leqslant \frac{1}{n_{t,k}} \sum_{i:(t,i)\in\mathcal{I}_k} \mathbb{E}\big[\|\tilde{\Delta}_{ti}^{(-k)}\|^q \mid \mathcal{S}^{(-k)}, \mathcal{S}_t^{X,(k)}\big].$$

Taking conditional expectations, we get

$$\mathbb{E}\big[(\Gamma_t^{(k)})^{q/2} \mid \mathcal{S}^{(-k)}\big] = \frac{1}{n_{t,k}} \sum_{i:(t,i)\in\mathcal{I}_k} \mathbb{E}\big[\|\tilde{\Delta}_{ti}^{(-k)}\|^q \mid \mathcal{S}^{(-k)}\big].$$

Then

$$\begin{aligned}
&\Big(\mathbb{E}[(\Gamma_t^{(k)})^{q/2} \mid \mathcal{S}^{(-k)}]\Big)^{1/q} \mathbf{1}(\mathcal{E}_{N,k}) \\
&\leqslant \sup_{(\nu,e)\in\mathcal{T}_N} (\mathbb{E}_t[\|(s(W(1);\theta_0,\nu,e) - s(W(0);\theta_0,\nu,e)) - (s(W(1);\theta_0,\nu_0,e_{0,N}) - s(W(0);\theta_0,\nu_0,e_{0,N}))\|^q])^{1/q} \\
&\leqslant \sup_{(\nu,e)\in\mathcal{T}_N} \sum_{z\in\{0,1\}} (\mathbb{E}_t[\|s(W(z);\theta_0,\nu,e)\|^q])^{1/q} + (\mathbb{E}_t[\|s(W(z);\theta_0,\nu_0,e_{0,N})\|^q])^{1/q} \\
&\leqslant 4C\kappa_t^{-1/q}
\end{aligned}$$

by (S12) and moment boundedness. We conclude that $\mathbb{E}[(\Gamma_t^{(k)})^{q/2} \mid \mathcal{S}^{(-k)}]\mathbf{1}(\mathcal{E}_{N,k})$ is uniformly bounded. Recalling that $\Pr(\mathcal{E}_{N,k}) \to 1$, Markov's inequality then ensures $\Gamma_t^{(k)} = O_p(1)$. In view of (S22), we then have $\mathbb{E}[\|B_{4,t}^{(k),3}\|^2 \mid \mathcal{S}^{(-k)}, \mathcal{S}_t^{X,(k)}] = o_p(1)$. This implies $B_{4,t}^{(k),3} = o_p(1)$ by Lemma B.1, and hence $B_4^{(k),3} = o_p(1)$. This establishes that $\hat{\theta} = \tilde{\theta} + o_p(N^{-1/2})$ and completes the proof.

C.6 Proof of Corollary 4.1: CLT for feasible $\hat{\theta}_{\text{AIPW}}$ in a CSBAE

Corollary 4.1, which shows the feasible estimator $\hat{\theta}_{\text{AIPW}}$ satisfies a CLT for estimating $\hat{\theta}_{0,\text{ATE}}$ in a CSBAE, holds under the numbered generalizations at the end of Appendix A, subject to the additional requirement that the mean functions are stationary, as in our generalization of Corollary 3.1 to nonstationary batches. We prove this more general result. The proof of our generalized Corollary 3.1 shows that Assumption A.2 is satisfied under the moment bounds in Assumption 3.1 with $\theta_0 = \theta_{0,\text{ATE}}$, $s(\cdot) = s_{\text{AIPW}}(\cdot)$, $\nu_0 = \nu_{0,\text{AIPW}} \in \mathcal{N} = \mathcal{N}_{\text{AIPW}}$, and any $\gamma \in (0, 1/2)$. It remains to show that the further conditions of Assumption 4.1, namely (a), (b), (c) and equations (21) through (26), are satisfied. Then Corollary 4.1 follows by Theorem 4.1.

Condition (a) Invertibility of $\mathbb{E}_0[\theta_0(W;\nu_0,e_0)]$ and existence of $\mathbb{E}_0[s(W;\theta_0,\nu_0,e_0)^2]$ follow from Assumption 3.1, as shown in the proof of Corollary 3.1.

Condition (b) Fix $\lambda \in [0,1]$ and $(\nu, e) \in \mathcal{T} = \mathcal{N} \times \mathcal{F}_\gamma$, where $\nu(\cdot) = (m(0,\cdot), m(1,\cdot))$. Consider

$$\begin{aligned}
f_N(\lambda) &:= s(W;\theta_0,\nu_0+\lambda(\nu-\nu_0), e_{0,N}+\lambda(e-e_{0,N})) \\
&= m_\lambda(1,X) - m_\lambda(0,X) + \frac{Z(Y-m_\lambda(1,X))}{e_{\lambda,N}(x)} - \frac{(1-Z)(Y-m_\lambda(0,X))}{1-e_{\lambda,N}(x)}
\end{aligned}$$

where $m_\lambda(z,x) := m_0(z,x) + \lambda(m(z,x)-m_0(z,x))$ and $e_{\lambda,N}(x) := e_{0,N}(x) + \lambda(e(x) - e_{0,N}(x))$. Taking two derivatives with respect to λ we get

$$\begin{aligned}
f_N'(\lambda) &= (m(1,X)-m_0(1,X)) - (m(0,X)-m_0(0,X)) \\
&\quad + Z\left(\frac{m_0(1,X)-m(1,X)}{e_{\lambda,N}(X)} - \frac{(Y-m_\lambda(1,X))(e(X)-e_{0,N}(X))}{e_{\lambda,N}(X)^2}\right) \\
&\quad - (1-Z)\left(\frac{m_0(0,X)-m(0,X)}{1-e_{\lambda,N}(X)} - \frac{(Y-m_\lambda(0,X))(e_{0,N}(X)-e(X))}{(1-e_{\lambda,N}(X))^2}\right), \quad \text{and} \\
f_N''(\lambda) &= 2Z\left(\frac{(Y-m_\lambda(1,X))(e(X)-e_{0,N}(X))^2}{e_{\lambda,N}(X)^3} - \frac{(m_0(1,X)-m(1,X))(e(X)-e_{0,N}(X))}{e_{\lambda,N}(X)^2}\right) \\
&\quad + 2(1-Z)\left(\frac{m_0(0,X)-m(0,X)(e_{0,N}(X)-e(X))}{(1-e_{\lambda,N}(X))^2} - \frac{(Y-m_\lambda(0,X))(e_{0,N}(X)-e(X))^2}{(1-e_{\lambda,N}(X))^3}\right).
\end{aligned}$$

We now show that $f_N(\cdot)$, $f_N'(\cdot)$, and $f_N''(\cdot)$ are upper bounded by an integrable random variable on an open interval containing $[0,1]$, so that by the Leibniz rule, we can swap both first and second derivatives with expectations to conclude the function $\lambda \mapsto \mathbb{E}_{0,N}[s(W;\theta_0,\nu_0+\lambda(\nu-\nu_0), e_{0,N}+\lambda(e-e_{0,N}))]$ has second derivative $\mathbb{E}_{0,N}[f_N''(\lambda)]$ for $\lambda \in [0,1]$. Note $e_{\lambda,N} \in \mathcal{F}_\gamma$. Then by repeated application of the triangle inequality, we see that for sufficiently small $\epsilon > 0$, we must have

$$\sup_{\lambda \in (-\epsilon, 1+\epsilon)} |f_N(\lambda)| \leqslant 2(|m(1,X)| + |m_0(1,X)| + |m(0,X)| + |m_0(0,X)|) \\ + 2\gamma^{-1}(|Y(1)| + |Y(0)| + |m(1,X)| + |m_0(1,X)| + |m(0,X)| + |m_0(0,X)|)$$
$$\sup_{\lambda \in (-\epsilon, 1+\epsilon)} |f_N'(\lambda)| \leqslant 2(|m(1,X)| + |m_0(1,X)| + |m(0,X)| + |m_0(0,X)|) \\ + 2\gamma^{-2}(|Y(1)| + |Y(0)| + |m(1,X)| + |m_0(1,X)| + |m(0,X)| + |m_0(0,X)|)$$
$$\sup_{\lambda \in (-\epsilon, 1+\epsilon)} |f_N''(\lambda)| \leqslant 2\gamma^{-3}(|Y(1)| + |Y(0)| + |m(1,X)| + |m_0(1,X)| + |m(0,X)| + |m_0(0,X)|).$$

The right-hand sides above are clearly integrable under $P_{0,N}$. Hence, the mapping $\lambda \mapsto \mathbb{E}_{0,N}[s(W;\theta_0,\nu_0+\lambda(\nu-\nu_0), e_{0,N}+\lambda(e-e_{0,N}))]$ is indeed twice differentiable with second derivative $\mathbb{E}_{0,N}[f_N''(\lambda)]$ for $\lambda \in [0,1]$. This second derivative is continuous for such λ by continuity of $f_N''(\cdot)$ and dominated convergence.

Condition (c) Fix $e \in \mathcal{F}_\gamma$. By stationarity of the mean functions $m_0(z,\cdot)$, $z = 0,1$, for each $t = 1,\dots,T$, $i = 1,\dots,N_t$ we have (letting $\nu_0 = \nu_{0,\mathrm{AIPW}}$ for brevity)

$$\begin{aligned}
\mathbb{E}[s(W_{ti}(1);\theta_0,\nu_0,e) \mid X_{ti}] &= m_0(1,X) - m_0(0,X) - \theta_0 + \mathbb{E}_t\left[\frac{Y(1)-m_0(1,X)}{e(X)} \;\Big|\; X\right] \\
&= m_0(1,X) - m_0(0,X) - \theta_0 + (e(X))^{-1}\mathbb{E}_t[Y(1) - m_0(1,X) \,|\, X] \\
&= m_0(1,X) - m_0(0,X) - \theta_0
\end{aligned}$$

and similarly

$$\begin{aligned}
\mathbb{E}[s(W_{ti}(0);\theta_0,\nu_0,e) \mid X_{ti}] &= m_0(1,X) - m_0(0,X) - \theta_0 - \mathbb{E}_t\left[\frac{Y(0)-m_0(0,X)}{1-e(X)} \;\Big|\; X\right] \\
&= m_0(1,X) - m_0(0,X) - \theta_0 - (1-e(X))^{-1}\mathbb{E}_t[Y(0) - m_0(0,X) \,|\, X] \\
&= m_0(1,X) - m_0(0,X) - \theta_0.
\end{aligned}$$

Subtracting shows (20).

It remains to show that the out-of-fold estimators $(\hat{m}^{(-k)}(0,\cdot), \hat{m}^{(-k)}(1,\cdot), \hat{e}^{(-k)})$ lie in some set $\mathcal{T}_N$ with high probability for all sufficiently large N, where this set $\mathcal{T}_N$ satisfies equations (21) through (26). To construct this $\mathcal{T}_N$, we see that by the rate conditions on the nuisance estimators along with equation (S8), there exists a sequence $\tilde{\delta}_N \downarrow 0$ so that $\|e_{0,N} - e_0\|_{2,P_0^X} \leqslant \tilde{\delta}_N$ for all N. Furthermore, with probability approaching 1 as $N \to \infty$, these four conditions hold for $z = 0,1$ and all folds $k = 1,\dots,K$:

$$\begin{aligned}
\|\hat{m}^{(-k)}(z,\cdot) - m_0(z,\cdot)\|_{2,P_0^X} + \|\hat{e}^{(-k)} - e_{0,N}\|_{2,P_0^X} &\leqslant \tilde{\delta}_N, \\
\|\hat{m}^{(-k)}(z,\cdot)\|_{2,P_0^X} \times \|\hat{e}^{(-k)} - e_{0,N}\|_{2,P_0^X} &\leqslant N^{-1/2}\tilde{\delta}_N, \\
\|\hat{m}^{(-k)}(z,\cdot) - m_0(z,\cdot)\|_{q,P_0^X} &\leqslant C, \\
\hat{e}^{(-k)}(\cdot) &\in \mathcal{F}_\gamma.
\end{aligned}$$

We then define $\mathcal{T}_N$ be the set of functions $(m(0,\cdot), m(1,\cdot), e(\cdot))$ in $\mathcal{T} = \mathcal{N}_{\mathrm{AIPW}} \times \mathcal{E}$ obeying these conditions:

$$\begin{aligned}
\|m(z,\cdot) - m_0(z,\cdot)\|_{2,P_0^X} + \|e - e_{0,N}\|_{2,P_0^X} &\leqslant \tilde{\delta}_N, \\
\|m(z,\cdot)\|_{2,P_0^X} \times \|e - e_{0,N}\|_{2,P_0^X} &\leqslant N^{-1/2}\tilde{\delta}_N, \\
\|m(z,\cdot) - m_0(z,\cdot)\|_{q,P_0^X} &\leqslant C, \\
e(\cdot) &\in \mathcal{F}_\gamma.
\end{aligned}$$

By construction, for all $k = 1, \dots, K$ we have $\Pr((\hat{m}^{(-k)}, \hat{e}^{(-k)}) \in \mathcal{T}_N) \to 1$ as $N \to \infty$.

For the remainder of the proof, we take N large enough so that $1/2 \leqslant \mathrm{d}P_{0,N}^X/\mathrm{d}P_0^X \leqslant 2$. For such N, $\|f\|_{2,P_{0,N}^X} \leqslant \sqrt{2}\|f\|_{2,P_0^X}$ holds for all $f \in L^2(P_0^X)$. We will show equations (21) through (24) hold for a sequence δ_N that is some constant multiple of $\tilde{\delta}_N$.

Equation (21) Fix $(\nu, e) \in \mathcal{T}_N$. By the calculations and notation above in the proof of condition (b) and interchanging differentiation with expectation, we can verify using unconfoundedness that

$$\frac{\partial}{\partial\lambda}\mathbb{E}_{0,N}[s(W;\theta_0,\nu_0,\lambda(\nu-\nu_0),e_{0,N}+\lambda(e-e_{0,N})]\Big|_{\lambda=0} = \mathbb{E}_{0,N}[f_N'(0)] = 0$$

so the left-hand side of the Neyman orthogonality condition (21) is 0. The full calculation is shown in the proof of Theorem 5.1 in [4].

Equation (22) Once again, we fix $(\nu, e) \in \mathcal{T}_N$ and recall the calculations and notations in the proof of condition (b) above. We see that

$$\Big|\frac{\partial^2}{\partial\lambda^2}\mathbb{E}_{0,N}[s(W;\theta_0,\nu_0,\lambda(\nu-\nu_0),e_{0,N}+\lambda(e-e_{0,N})]\Big| = \big|\mathbb{E}_{0,N}[f_N''(\lambda)]\big|$$

and hence

$$\begin{aligned}
\big|\mathbb{E}_{0,N}[f_N''(\lambda)]\big| \leqslant \frac{2}{\gamma^3}\Big(&\big|\mathbb{E}_{0,N}[(Y(1)-m_\lambda(1,X))(e(X)-e_{0,N}(X))^2]\big| \\
&+ \big|\mathbb{E}_{0,N}[(Y(0)-m_\lambda(0,X))(e_{0,N}(X)-e(X))^2]\big|\Big) \\
&+ \frac{2}{\gamma^2}\Big(|\mathbb{E}_{0,N}[(m_0(1,X)-m(1,X))(e(X)-e_0(X))]| \\
&+ \big|\mathbb{E}_{0,N}[(m_0(0,X)-m(0,X))(e(X)-e_0(X))]\big|\Big).
\end{aligned}$$

By Cauchy-Schwarz and the definition of $\mathcal{T}_N$ we have for $z = 0, 1$ that

$$\begin{aligned}
&\mathbb{E}_{0,N}\big[|m_0(z,X) - m(z,X)| \times |e(X) - e_{0,N}(X)|\big] \\
&\leqslant \|m_0(z,\cdot) - m(z,\cdot)\|_{2,P_{0,N}^X} \times \|e - e_{0,N}\|_{2,P_{0,N}^X} \\
&\leqslant 2N^{-1/2}\tilde{\delta}_N.
\end{aligned}$$

Furthermore for $z = 0, 1$ we have $\mathbb{E}_{0,N}\big[(Y(z) - m_0(z, X))(e(X) - e_{0,N}(X))^2\big] = 0$ by conditioning on X. Hence, for all $\lambda \in [0, 1]$,

$$
\begin{aligned}
&\big|\mathbb{E}_{0,N}[(Y(z) - m_\lambda(z, X))(e(X) - e_{0,N}(X))^2]\big| \\
&= \big|\mathbb{E}_{0,N}[\lambda(m(z, X) - m_0(z, X))(e(X) - e_{0,N}(X))^2]\big| \\
&\leqslant \lambda \mathbb{E}_{0,N}\big[|m(z, X) - m_0(z, X)||e(X) - e_{0,N}(X)|\big] \\
&\leqslant 2N^{-1/2}\tilde{\delta}_N
\end{aligned}
$$

where the first inequality uses the fact that $|e(X) - e_{0,N}(X)| \leqslant 1$. Taking suprema over $(\nu, e) \in \mathcal{T}_N$ and $\lambda \in [0, 1]$, we get

$$
\sup_{(\nu,e)\in\mathcal{T}_N} \sup_{\lambda\in[0,1]} \left|\frac{\partial^2}{\partial\lambda^2}\mathbb{E}_{0,N}[s(W; \theta_0, \nu_0, \lambda(\nu-\nu_0), e_{0,N}+\lambda(e-e_{0,N}))]\right| \leqslant N^{-1/2}\left(\frac{8}{\gamma^3} + \frac{8}{\gamma^2}\right)\tilde{\delta}_N
$$

which shows equation (22).

Equation (23) For any $(\nu, e) \in \mathcal{T}_N$, trivially

$$
\mathbb{E}_0\big[|s_a(W; \nu, e) - s_a(W; \nu_0, e_0)|^2\big] = \mathbb{E}_0\big[| - 1 - (-1)|^2\big] = 0.
$$

Equation (24) We fix $(\nu, e) \in \mathcal{T}_N$ and write

$$
\begin{aligned}
s(W; \theta_0, \nu, e) - s(W; \theta_0, \nu_0, e_0) &= \left(1 - \frac{Z}{e_0(X)}\right)(m(1, X) - m_0(1, X)) \\
&\quad + \left(1 - \frac{1 - Z}{1 - e_0(X)}\right)(m_0(0, X) - m(0, X)) \\
&\quad + Z(Y - m(1, X))(e(X)^{-1} - e_0(X)^{-1}) \\
&\quad - (1 - Z)(Y - m(0, X))((1 - e(X))^{-1} - (1 - e_0(X))^{-1}).
\end{aligned}
$$

It now suffices to show that each of the summands has asymptotically vanishing second moment. The definition of $\mathcal{T}_N$ ensures that

$$
\begin{aligned}
\mathbb{E}_0\left[\left(1 - \frac{Z}{e_0(X)}\right)^2(m(1, X) - m_0(1, X))^2\right] &\leqslant (1 + \gamma^{-1})^2\tilde{\delta}_N^2 \quad \text{and} \\
\mathbb{E}_0\left[\left(1 - \frac{1 - Z}{1 - e_0(X)}\right)^2(m_0(0, X) - m(0, X))^2\right] &\leqslant (1 + \gamma^{-1})^2\tilde{\delta}_N^2.
\end{aligned}
$$

Next

$$
\begin{aligned}
&\mathbb{E}_0\big[Z^2(Y(1) - m(1, X))^2(e(X)^{-1} - e_0(X)^{-1})^2\big] \\
&\leqslant \gamma^{-4}\mathbb{E}_0\big[(Y(1) - m(1, X))^2(e(X) - e_0(X))^2\big] \\
&= \gamma^{-4}\mathbb{E}_0\big[(e(X) - e_0(X))^2 v_0(1, X)\big] \\
&\quad + \gamma^{-4}\mathbb{E}_0\big[(e(X) - e_0(X))^2(m(1, X) - m_0(1, X))^2\big] \\
&\leqslant C\gamma^{-4}\|e - e_0\|^2_{2,P_0^X} + \gamma^{-4}\|m(1, \cdot) - m_0(1, \cdot)\|^2_{2,P_0^X} \\
&\leqslant (2C + 1)\gamma^{-4}\tilde{\delta}_N^2.
\end{aligned}
$$

By a similar computation

$$\mathbb{E}_0[(1-Z)^2(Y(0)-m(0,X))^2((1-e(X))^{-1}-(1-e_0(X))^{-1})^2] \leqslant (2C+1)\gamma^{-4}\tilde{\delta}_N^2.$$

This completes the proof of (24).

Equation (25) For any $(\nu, e) \in \mathcal{T}_N$, trivially $\mathbb{E}\big[|s_a(W(z);\nu,e)|^q\big] = 1$ for $z = 0, 1$.

Equation (26) Fix $(\nu, e) \in \mathcal{T}_N$. We write

$$s(W(1);\theta_0,\nu,e) = m(1,X) - m(0,X) - \theta_0 + \frac{Y(1)-m(1,X)}{e(X)}. \tag{S24}$$

Since $(\mathbb{E}_0[|Y(z)|^q])^{1/q} \leqslant C$ for $z = 0, 1$,

$$\mathbb{E}_0\big[|m_0(z,X)|^q\big] = \mathbb{E}\big[|\mathbb{E}_0[Y(z) \mid X]|^q\big] \leqslant \mathbb{E}_0\big[|Y(z)|^q\big] \leqslant C^q$$

and then by the definition of θ_0

$$\begin{aligned}
|\theta_0| &= \big(|\mathbb{E}_0[m_0(1,X) - m_0(0,X)]|^q\big)^{1/q} \\
&\leqslant \big(\mathbb{E}_0\big[|m_0(1,X) - m_0(0,X)|^q\big]\big)^{1/q} \\
&\leqslant 2C.
\end{aligned}$$

With $\|m(z,\cdot) - m_0(z,\cdot)\|_{q,P_0^X} \leqslant C$ by definition of $\mathcal{T}_N$, we have

$$\|m(z,\cdot)\|_{q,P_0^X} \leqslant \|m(z,\cdot) - m_0(z,\cdot)\|_{q,P_0^X} + \|m_0(z,\cdot)\|_{q,P_0^X} \leqslant 2C$$

for $z = 0, 1$. Therefore

$$\begin{aligned}
\sup_{(\nu,e)\in\mathcal{T}_N} \big(\mathbb{E}_0\big[|s(W(1);\theta_0,\nu,e)|^q\big]\big)^{1/q} &\leqslant \sup_{(\nu,e)\in\mathcal{T}_N} \|m(1,\cdot)\|_{q,P_0^X} + \|m(0,\cdot)\|_{q,P_0^X} + |\theta_0| \\
&\quad + \sup_{(\nu,e)\in\mathcal{T}_N} \gamma^{-1}\big((\mathbb{E}_0[|Y(1)|^q])^{1/q} + \|m(1,\cdot)\|_{q,P_0^X}\big) \\
&\leqslant 6C + \frac{3C}{\gamma}.
\end{aligned}$$

Similarly, $\sup_{(\nu,e)\in\mathcal{T}_N}\big(\mathbb{E}_0\big[|s(W(0);\theta_0,\nu,e)|^q\big]\big)^{1/q} \leqslant 6C + 3C/\gamma$ as well.

Equation (27) We begin by considering $S_t^{(-k)}(z)$ for $z = 1$. For each $k = 1, \ldots, K$ we have

$$\begin{aligned}
s(W_{ti}(1);\hat{\nu}^{(-k)},\hat{e}^{(-k)}) - s(W_{ti}(1);\nu_0,e_{0,N}) &= \big(1 - \hat{e}^{(-k)}(X)^{-1}\big)\big(\hat{m}^{(-k)}(1,X_{ti}) - m_0(1,X_{ti})\big) \\
&\quad + \big(Y_{ti}(1) - m_0(1,X_{ti})\big)\big(\hat{e}^{(-k)}(X_{ti})^{-1} - e_0(X_{ti})^{-1}\big) \\
&\quad - \big(\hat{m}^{(-k)}(0,X_{ti}) - m_0(0,X_{ti})\big).
\end{aligned}$$

Taking the conditional expectation given $\mathcal{S}^{(-k)}$ and X_{ti} yields

$$\begin{aligned}
&\mathbb{E}\big[s(W_{ti}(1);\hat{\nu}^{(-k)},\hat{e}^{(-k)}) - s(W_{ti}(1);\nu_0,e_{0,N}) \mid \mathcal{S}^{(-k)}, X_{ti}\big] \\
&= \big(1 - \hat{e}^{(-k)}(X)^{-1}\big)(\hat{m}^{(-k)}(1,X_{ti}) - m_0(1,X_{ti})) - (\hat{m}^{(-k)}(0,X_{ti}) - m_0(0,X_{ti})).
\end{aligned}$$

Then for $z = 1$ we get

$$
\begin{aligned}
S_t^{(-k)}(1) &\leqslant \sqrt{\frac{1}{n_{t,k}} \sum_{i:(t,i)\in\mathcal{I}_k} \left(1 - \hat{e}^{(-k)}(X_{ti})^{-1}\right)^2 \left(\hat{m}^{(-k)}(1, X_{ti}) - m_0(1, X_{ti})\right)^2} \\
&\quad + \sqrt{\frac{1}{n_{t,k}} \sum_{i:(t,i)\in\mathcal{I}_k} (\hat{m}^{(-k)}(0, X_{ti}) - m_0(0, X_{ti}))^2} \\
&= O_p\left(\|\hat{m}^{(-k)}(1,\cdot) - m(1,\cdot)\|_{2,P_0^X} + \|\hat{m}^{(-k)}(0,\cdot) - m(0,\cdot)\|_{2,P_0^X}\right) \\
&= o_p(N^{-1/4}).
\end{aligned}
$$

The first equality above follows from $\left(1 - \hat{e}^{(-k)}(X_{ti})\right)^2 \leqslant (1+\gamma^{-1})^2 < \infty$, followed by an application of the conditional Markov inequality using

$$
\begin{aligned}
\mathbb{E}\left[\frac{1}{n_{t,k}} \sum_{i:(t,i)\in\mathcal{I}_k} (\hat{m}^{(-k)}(z, X_{ti}) - m_0(z, X_{ti}))^2 \;\middle|\; \mathcal{S}^{(-k)}\right] &= \|\hat{m}^{(-k)}(z,\cdot) - m_0(z,\cdot)\|_{2,P_t^X}^2 \\
&\leqslant \kappa_t^{-1}\|\hat{m}^{(-k)}(z,\cdot) - m_0(z,\cdot)\|_{2,P_0^X}^2
\end{aligned}
$$

for $z = 0, 1$. By an identical argument, $S_t^{(-k)}(0) = o_p(N^{-1/4})$, establishing (27).

Having shown all conditions of Assumption 4.1, the conclusion of Corollary 4.1 follows by Theorem 4.1.

C.7 Proof of Corollary 4.2: CLT for feasible $\hat{\theta}_{\mathrm{EPL}}$ in a CSBAE

Corollary 4.2, which shows the feasible estimator $\hat{\theta}_{\mathrm{EPL}}$ satisfies a CLT for estimating $\theta_{0,\mathrm{PL}}$ in a CSBAE, holds under the numbered generalizations at the end of Appendix A, subject to the additional requirement that the mean and variance functions are stationary across batches, as in Appendix C.3. We prove this more general result using the same structure as the proof of Corollary 4.1 in Appendix C.6.

The proof of Corollary 3.2 in Appendix C.3 shows that Assumption A.2 is satisfied with $\theta_0 = \theta_{0,\mathrm{PL}}$, $s(\cdot) = s(\cdot)_{\mathrm{EPL}}$, $\nu_0 = \nu_{0,\mathrm{EPL}} \in \mathcal{N} = \mathcal{N}_{\mathrm{EPL}}$, and $\gamma = 0$. Then it suffices to show the remaining conditions of Assumption 4.1 to complete the proof, in view of Theorem 4.1. Throughout the remainder of this proof we let C_0 be a generic positive finite constant; possibly depending on c and C in Assumption 3.2; different appearances of C_0 may correspond to different constants.

Condition (a) Invertibility of $\mathbb{E}_0[s_a(W;\nu_0,e_0)]$ and existence of $\mathbb{E}_0[\|s(W;\theta_0,\nu_0,e_0)\|^2]$ follow from Assumption 3.2, as shown in the proof of Corollary 3.2.

Condition (b) Fix $(\nu, e) \in \mathcal{T}$. Recall that the weight function in our estimator $\hat{\theta}_{\mathrm{EPL}}$ is $w(X,\nu,e) = (v(0,x)e(x) + v(1,x)(1-e(x)))^{-1}$. We compute

$$
\begin{aligned}
&\mathbb{E}_{0,N}[s(W;\theta_0,\nu_0 + \lambda(\nu-\nu_0), e_{0,N} + \lambda(e - e_{0,N}))] \\
&\qquad = \mathbb{E}_{0,N}[w(X;\nu_\lambda,e_{\lambda,N})(Z - e_{\lambda,N}(X))(Y - m_\lambda(0,X) - Z\psi(X)^\top\theta_0)\psi(X)] \\
&\qquad = \lambda\mathbb{E}_{0,N}[w(X;\nu_\lambda,e_{\lambda,N})(Z - e_{\lambda,N}(X))(m_0(0,X) - m(0,X))\psi(X)] \\
&\qquad = \mathbb{E}_{0,N}[f_N(\lambda)]
\end{aligned}
$$

where the last two equalities follow by conditioning on (X, Z) then just X, and we have defined

$$\begin{aligned}
\nu_\lambda(\cdot) &= \nu_0(\cdot) + \lambda(\nu(\cdot) - \nu_0(\cdot)), \\
e_{\lambda,N}(\cdot) &= e_{0,N}(\cdot) + \lambda(e(\cdot) - e_{0,N}(\cdot)) \quad \text{and} \\
f_N(\lambda) &= \lambda^2 w(X; \nu_\lambda, e_{\lambda,N})(e_{0,N}(X) - e(X))(m_0(0, X) - m(0, X))\psi(X) \\
&= \lambda^2 w(X; \nu_\lambda, e_{\lambda,N}) g_N(X)
\end{aligned}$$

for $g_N(X) = (e_{0,N}(X) - e(X))(m_0(0, X) - m(0, X))\psi(X)$. We compute

$$\begin{aligned}
\frac{\partial}{\partial\lambda} w(X; \nu_\lambda, e_{\lambda,N}) &= -w^2(X; \nu_\lambda, e_{\lambda,N})\Delta_N(\lambda, X) \\
\frac{\partial}{\partial\lambda} w^2(X; \nu_\lambda, e_{\lambda,N}) &= -2w^3(X; \nu_\lambda, e_{\lambda,N})\Delta_N(\lambda, X)
\end{aligned}$$

where

$$\begin{aligned}
\Delta_N(\lambda, X) &= \frac{\partial}{\partial\lambda}\big(v_\lambda(0, X)e_{\lambda,N}(X) + v_\lambda(1, X)(1 - e_{\lambda,N}(X))\big) \\
&= (1 - e_\lambda(X))(v(1, X) - v_0(1, X)) + e_\lambda(X)(v(0, X) - v_0(0, X)) \quad \text{and} \\
&\quad + (e(X) - e_{0,N}(X))(v_\lambda(0, X) - v_\lambda(1, X))
\end{aligned}$$

Then

$$\begin{aligned}
f_N'(\lambda) &= (-\lambda^2\Delta_N(\lambda, X)w^2(X; \nu_\lambda, e_{\lambda,N}) + 2\lambda w(X; \nu_\lambda, e_{\lambda,N}))g_N(X) \\
f_N''(\lambda) &= (2\lambda^2(\Delta_N(\lambda, X))^2 w^3(X; \nu_\lambda, e_{\lambda,N}) - \lambda^2\Delta_N^{(2)}(\lambda, X)w^2(X; \nu_\lambda, e_{\lambda,N}))g_N(X) \\
&\quad + (2w(X; \nu_\lambda, e_{\lambda,N}) - 4\lambda\Delta_N(\lambda, X)w^2(X; \nu_\lambda, e_{\lambda,N}))g_N(X)
\end{aligned}$$

where

$$\begin{aligned}
\Delta_N^{(2)}(\lambda, X) &= \frac{\partial}{\partial\lambda}\Delta_N(\lambda, X) \\
&= 2(e(X) - e_{0,N}(X))(v(0, X) - v_0(0, X) + v_0(1, X) - v(1, X))
\end{aligned}$$

We conclude that for sufficiently small $\epsilon > 0$,

$$\begin{aligned}
\sup_{\lambda\in(-\epsilon,1+\epsilon)} |\Delta_N(\lambda, X)| &\leqslant \sup_{\lambda\in(-\epsilon,1+\epsilon)} |1 - e_\lambda(X))||v(1, X) - v_0(1, X)| + |e_\lambda(X)||v(0, X) - v_0(0, X)| \\
&\quad + \sup_{\lambda\in(-\epsilon,1+\epsilon)} |e(X) - e_{0,N}(X)||v_\lambda(0, X) - v_\lambda(1, X)| \\
&\leqslant C_0 \quad \text{and} \\
\sup_{\lambda\in(-\epsilon,1+\epsilon)} |\Delta_N^{(2)}(\lambda, X)| &\leqslant C_0.
\end{aligned}$$

Taking ϵ smaller if necessary, we can ensure

$$\sup_{\lambda\in(-\epsilon,1+\epsilon)} w(X; \nu_\lambda, e_{\lambda,N}) = \sup_{\lambda\in(-\epsilon,1+\epsilon)} (v_\lambda(0, X)e_{\lambda,N}(X) + v_\lambda(1, X)(1 - e_{\lambda,N}(X)))^{-1} < 2c^{-1}$$

and then

$$\sup_{\lambda\in(-\epsilon,1+\epsilon)} \|f_N(\lambda)\| \leqslant C_0\|g_n(X)\| \leqslant C_0C|e_{0,N}(X)-e(X)||m_0(0,X)-m(0,X)|$$
$$\sup_{\lambda\in(-\epsilon,1+\epsilon)} \|f'_N(\lambda)\| \leqslant C_0\|g_n(X)\| \leqslant C_0C|e_{0,N}(X)-e(X)||m_0(0,X)-m(0,X)|$$
$$\sup_{\lambda\in(-\epsilon,1+\epsilon)} \|f''_N(\lambda)\| \leqslant K\|g_n(X)\| \leqslant C_0C|e_{0,N}(X)-e(X)||m_0(0,X)-m(0,X)|$$

By the Leibniz integral rule, the mapping $\lambda \mapsto \mathbb{E}_{0,N}[s(W;\theta_0,\nu_0+\lambda(\nu-\nu_0),e_{0,N}+\lambda(e-e_{0,N}))]$ then has second derivative $\mathbb{E}_{0,N}[f''_N(\lambda)]$ on $[0,1]$. With $f''_N(\lambda)$ continuous on $[0,1]$, we conclude by dominated convergence that $\mathbb{E}_{0,N}[f''_N(\lambda)]$ is continuous as well.

Condition (c) For $z=0,1$,

$$\mathbb{E}_0[Y(z)\,|\,X] = \mathbb{E}_0[Y\,|\,X, Z=z] = m_0(0,X) + z\psi(X)^\top\theta_0.$$

Fix $e(\cdot)\in\mathcal{F}_0$. For each $t=1,\dots,T$, $i=1,\dots,N_t$ we have

$$\begin{aligned}
&\mathbb{E}[s(W_{ti}(1);\theta_0,\nu_0,e)\mid X_{ti}]\\
&\quad= w(X_{ti};\nu_0,e)(1-e(X))\big(\mathbb{E}_0[Y(1)\,|\,X=X_{ti}] - m_0(0,X_{ti}) - \psi(X_{ti})^\top\theta_0\big)\psi(X_{ti})\\
&\quad= 0
\end{aligned}$$

and similarly

$$\begin{aligned}
&\mathbb{E}[s(W_{ti}(0);\theta_0,\nu_0,e)\mid X_{ti}]\\
&\quad= -w(X_{ti};\nu_0,e)e(X)\big(\mathbb{E}_0[Y(0)\,|\,X=X_{ti}] - m_0(0,X_{ti})\big)\psi(X_{ti})\\
&\quad= 0.
\end{aligned}$$

These are equal so their difference is 0 as required for (20).

Now we show the out-of-fold estimates $(\hat m^{(-k)}(0,\cdot),\hat v^{(-k)}(0,\cdot),\hat v^{(-k)}(1,\cdot),\hat e^{(-k)}(\cdot))$ lie in a set $\mathcal{T}_N$ with high probability for all sufficiently large N, where this $\mathcal{T}_N$ satisfies equations (21) through (26). By the rate and regularity conditions on the nuisance estimators in Corollary 4.2, there exists a sequence $\tilde\delta_N \downarrow 0$ and constants $0<c<C<\infty$ so that with probability approaching 1 as $N\to\infty$, we have

$$\begin{aligned}
\|\hat m^{(-k)}(0,\cdot)-m_0(0,\cdot)\|_{2,P_0^X} + \|\hat v(z,\cdot)-v_0(z,\cdot)\|_{2,P_0^X} + \|\hat e^{(-k)}-e_{0,N}\|_{2,P_0^X} &\leqslant \tilde\delta_N\\
\|\hat m^{(-k))}(0,\cdot)-m_0(0,\cdot\|_{2,P_0^X}\|\hat e^{(-k)}-e_{0,N}\|_{2,P_0^X} &\leqslant N^{-1/2}\tilde\delta_N\\
\|\hat m^{(-k)}(0,\cdot)-m_0(0,\cdot)\|_{q,p_0^X} &\leqslant C\\
\hat v^{(-k)}(z,x) &\geqslant c,\quad z=0,1
\end{aligned}$$

for all folds $k=1,\dots,K$. Then let $\mathcal{T}_N$ be the set of functions $(m(0,\cdot),v(0,\cdot),v(1,\cdot),e(\cdot))$ in $\mathcal{N}_{\mathrm{EPL}}\times\mathcal{F}_0$ for which

$$\begin{aligned}
\|m(0,\cdot)-m_0(0,\cdot)\|_{2,P_0^X} + \|v(z,\cdot)-v_0(z,\cdot)\|_{2,P_0^X} + \|e-e_{0,N}\|_{2,P_0^X} &\leqslant \tilde\delta_N\\
\|m(0,\cdot)-m_0(0,\cdot\|_{2,P_0^X}\|e-e_{0,N}\|_{2,P_0^X} &\leqslant N^{-1/2}\tilde\delta_N\\
\|m(0,\cdot)-m_0(0,\cdot)\|_{q,p_0^X} &\leqslant C\\
v(z,x) &\geqslant c,\quad z=0,1.
\end{aligned}$$

By construction, $\Pr((\hat{m}^{(-k)}(0,\cdot), \hat{v}^{(-k)}(0,\cdot), \hat{v}^{(-k)}(1,\cdot), \hat{e}^{(-k)}(\cdot)) \in \mathcal{T}_N) \to 1$ as $N \to \infty$ for all $k = 1, \ldots, K$. Now we show that $\mathcal{T}_N$ satisfies equations (21) through (26) for all N large enough to ensure that $1/2 \leqslant \mathrm{d}P_{0,N}^X/\mathrm{d}P_0^X \leqslant 2$. For such N, $\|f\|_{2,P_{0,N}^X} \leqslant \sqrt{2}\|f\|_{2,P_0^X}$ holds for all $f \in L^2(P_0^X)$. As in the proof of Corollary 4.1, we will show equations (21) through (24) hold for a sequence δ_N that is some constant multiple of $\tilde{\delta}_N$.

Equation (21) Using the notation from our proof of condition (b) above, we see $f_N'(0) = 0$ with probability 1, so for any $(\nu, e) \in \mathcal{T}_N$ we have

$$\frac{\partial}{\partial \lambda}\mathbb{E}_{0,N}[s(W;\theta_0,\nu_0+\lambda(\nu-\nu_0),e_{0,N}+\lambda(e-e_{0,N}))]\Big|_{\lambda=0} = \mathbb{E}_{0,N}[f_N'(0)] = 0$$

which shows the left-hand side of (21) is identically zero.

Equation (22) Once again we recall the notation and calculations from the proof of condition (b) above. For each $\lambda \in [0,1]$ and $(\nu, e) \in \mathcal{T}_N$ we see

$$\left\|\frac{\partial^2}{\partial \lambda^2}\mathbb{E}_{0,N}\left[s\big(W;\theta_0,\nu_0+\lambda(\nu-\nu_0),e_{0,N}+\lambda(e-e_{0,N})\big)\right]\right\| = \left\|\mathbb{E}_{0,N}[f_N''(\lambda)]\right\|$$

which is no larger than

$$\mathbb{E}_{0,N}\left[\sup_{\lambda\in[0,1]}\|f_N''(\lambda)\|\right] \leqslant C_0\|e-e_{0,N}\|_{2,P_0^X}\|m(0,\cdot)-m_0(0,\cdot)\|_{2,P_0^X} \leqslant C_0 N^{-1/2}\tilde{\delta}_N$$

by the definition of $\mathcal{T}_N$.

Equation (23) Recall that $s_{\mathrm{EPL},a}(W;\nu,e) = -w(X;\nu,e(X))Z(Z-e)\psi(X)\psi(X)^\top$. Fix $(\nu, e) \in \mathcal{T}_N$. We compute

$$\begin{aligned} s_a(W;\nu,e) - s_a(W;\nu_0,e_0) &= \big(w(X;\nu_0,e_0)(Z-e_0(X)) - w(X;\nu,e)(Z-e(X))\big)Z\psi(X)\psi(X)^\top \\ &= w(X;\nu_0,e_0)(e(X)-e_0(X))Z\psi(X)\psi(X)^\top \\ &\quad + (w(X;\nu_0,e_0)-w(X;\nu,e))(Z-e(X))Z\psi(X)\psi(X)^\top. \end{aligned}$$

Hence, using $\|\psi(X)\psi(X)^\top\| \leqslant C^2$,

$$\begin{aligned} &\mathbb{E}_0\big[\|s_a(W;\nu,e)-s_a(W;\nu_0,e_0)\|^2\big]^{1/2} \\ &\leqslant C^2\mathbb{E}_0\big[(w(X;\nu_0,e_0)Z(e(X)-e_0(X)))^2\big]^{1/2} \\ &\quad + C^2\mathbb{E}_0\big[(w(X;\nu_0,e_0)-w(X;\nu,e))^2(Z-e(X))^2Z^2\big]^{1/2}. \end{aligned}$$

To bound the right-hand side, first note that

$$\Big(\mathbb{E}_0\big[(w(X;\nu_0,e_0)Z(e(X)-e_0(X)))^2\big]\Big)^{1/2} \leqslant c^{-1}\|e-e_0\|_{2,P_0^X} \leqslant c^{-1}\tilde{\delta}_N.$$

Next

$$
\begin{aligned}
&|w(X;\nu,e) - w(X;\nu_0,e_0)| \\
&\leqslant c^{-2}\big|v(0,X)e(X) + v(1,X)(1-e(X)) - v_0(0,X)e_0(X) - v_0(1,X)(1-e_0(X))\big| \\
&\leqslant c^{-2}\big(|v(0,X)-v_0(0,X)|\cdot e(X) + v_0(0,X)\cdot|e(X)-e_0(X)|\big) \\
&\quad + c^{-2}\big(|v(1,X)-v_0(1,X)|\cdot(1-e(X)) + v_0(1,X)\cdot|e(X)-e_0(X)|\big) \\
&\leqslant c^{-2}\big(|v(0,X)-v_0(0,X)| + |v(1,X)-v_0(1,X)| + 2C|e(X)-e_0(X)|\big)
\end{aligned}
$$

and so

$$
\begin{aligned}
&\mathbb{E}_0\big[(w(X;\nu_0,e_0) - w(X;\nu,e))^2(Z-e(X))^2Z^2\big]^{1/2} \\
&\leqslant c^{-2}\big(\|v(0,\cdot)-v_0(0,\cdot)\|_{2,P_0^X} + \|v(1,\cdot)-v_0(1,\cdot)\|_{2,P_0^X} + 2C\|e-e_0\|_{2,P_0^X}\big) \\
&\leqslant C_0\tilde{\delta}_N,
\end{aligned}
$$

which shows (23).

Equation (24) Once again, fix $(\nu,e)\in\mathcal{T}_N$. Then

$$
\begin{aligned}
&s(W;\theta_0,\nu,e) - s(W;\theta_0,\nu_0,e_0) \\
&= [w(X;\nu,e)(Z-e(X)) - w(X;\nu_0,e_0)(Z-e_0(X))][Y - m_0(0,X) - Z\psi(X)^\top\theta_0]\psi(X) \\
&\quad + w(X;\nu,e)(Z-e(X))(m_0(0,X)-m(0,X))\psi(X)
\end{aligned}
$$

so that

$$
\begin{aligned}
&\mathbb{E}_0\big[\|s(W;\theta_0,\nu,e) - s(W;\theta_0,\nu_0,e_0)\|^2\big]^{1/2} \\
&\leqslant C\Big(\mathbb{E}_0\big[\big(w(X;\nu,e)(Z-e(X)) - w(X;\nu_0,e_0)(Z-e_0(X))\big)^2\big(Y - m_0(0,X) - Z\psi(X)^\top\theta_0\big)^2\big]\Big)^{1/2} \\
&\quad + C\big(\mathbb{E}_0\big[(w(X;\nu,e)(Z-e(X)))^2(m_0(0,X)-m(0,X))^2\big]\big)^{1/2}.
\end{aligned}
$$

By conditioning on (X,Z) we see that

$$
\begin{aligned}
&\mathbb{E}_0\big[|w(X;\nu,e)(Z-e(X)) - w(X;\nu_0,e_0)(Z-e_0(X))|^2|Y - m_0(0,X) - Z\psi(X)^\top\theta_0|^2\big]^{1/2} \\
&\quad = \mathbb{E}_0\big[v_0(Z,X)|w(X;\nu,e)(Z-e(X)) - w(X;\nu_0,e_0)(Z-e_0(X))|^2\big]^{1/2} \\
&\quad \leqslant C^{1/2}\mathbb{E}_0\big[|w(X;\nu,e)(Z-e(X)) - w(X;\nu_0,e_0)(Z-e_0(X))|^2\big]^{1/2}.
\end{aligned}
$$

Now we write

$$
\begin{aligned}
&\big|w(X;\nu,e)(Z-e(X)) - w(X;\nu_0,e_0)(Z-e_0(X))\big| \\
&\leqslant |w(X;\nu,e) - w(X;\nu_0,e_0)||Z-e(X)| + |w(X;\nu_0,e_0)||e_0(X)-e(X)| \\
&\leqslant c^{-2}(|v(0,X)-v_0(0,X)| + |v(1,X)-v_0(1,X)| + 2C|e(X)-e_0(X)|) + c^{-1}|e_0(X)-e(X)|
\end{aligned}
$$

so that

$$
\begin{aligned}
&\mathbb{E}_0\big[|w(X;\nu,e)(Z-e(X)) - w(X;\nu_0,e_0)(Z-e_0(X))|^2|Y - m_0(0,X) - Z\psi(X)^\top\theta_0|^2\big]^{1/2} \\
&\quad \leqslant C^{1/2}\big[c^{-2}(\|v(0,\cdot)-v_0(0,\cdot)\|_{2,P_0^X} + \|v(1,\cdot)-v_0(1,\cdot)\|_{2,P_0^X} + 2C\|e-e_0\|_{2,P_0^X}) + c^{-1}\|e-e_0\|_{2,P_0^X}\big] \\
&\quad \leqslant C_0\tilde{\delta}_N
\end{aligned}
$$

With $(\nu, e) \in \mathcal{T}_N$ arbitrary and

$$\mathbb{E}_0\big[|w(X;\nu,e)(Z-e(X))|^2|m_0(0,X)-m(0,X)|^2\big]^{1/2} \leqslant c^{-1}\|m(0,\cdot)-m_0(0,\cdot)\|_{2,P_0^X} \leqslant C_0\tilde{\delta}_N$$

we have shown (24).

Equation (25) Fix $(\nu, e) \in \mathcal{T}_N$. For $z = 0, 1$ we have

$$\sup_{z\in\{0,1\}} \|s_a(W(z);\nu,e)\|^q = \sup_{z\in\{0,1\}} \|w(X;\nu,e)z(z-e(X))\psi(X)\psi(X)^\top\|^q \leqslant \left(\frac{C^2}{c}\right)^q$$

which immediately shows that

$$\mathbb{E}_0\big[\|s_a(W(z);\nu,e)\|^q\big]^{1/q} \leqslant C_0.$$

Equation (26) For any $(\nu, e) \in \mathcal{T}_N$ we have

$$\mathbb{E}_0\big[\|s(W(z);\theta_0,\nu,e)\|^q\big]^{1/q} \leqslant \frac{C}{c}\mathbb{E}\big[|Y(z)-m(0,X)-z\psi(X)^\top\theta_0|^q\big]^{1/q} \leqslant C_0$$

since

$$\begin{aligned}
&\mathbb{E}_0\big[|Y(z)-m(0,X)-z\psi(X)^\top\theta_0|^q\big]^{1/q} \\
&\quad\leqslant \mathbb{E}_0\big[|Y(z)|^q\big]^{1/q} + \mathbb{E}_0\big[|m(0,X)|^q\big]^{1/q} + \mathbb{E}_0\big[|m(1,X)|^q\big]^{1/q}.
\end{aligned}$$

We have $(\mathbb{E}_0[|Y(z)|^q])^{1/q} \leqslant C_0$ by Assumption 3.2, hence for $z = 0, 1$

$$\begin{aligned}
(\mathbb{E}_0[|m(z,X)|^q])^{1/q} &\leqslant (\mathbb{E}_0[|m_0(z,X)|^q])^{1/q} + \|m(z,\cdot)-m_0(z,\cdot)\|_{q,P_0^X} \\
&\leqslant (\mathbb{E}_0[|Y(z)|^q])^{1/q} + \|m(z,\cdot)-m_0(z,\cdot)\|_{q,P_0^X} \\
&\leqslant C_0
\end{aligned}$$

which shows (26).

Equation (27) Fix a fold $k \in \{1, \ldots, K\}$. For each $z = 0, 1$ and $(t, i) \in \mathcal{I}_k$ we have

$$\begin{aligned}
&s(W_{ti}(z);\hat{\nu}^{(-k)},\hat{e}^{(-k)}) - s(W_{ti}(z);\nu_0,e_{0,N}) \\
&\quad= \big[w(X_{ti};\hat{\nu}^{(-k)},\hat{e}^{(-k)})(z-\hat{e}^{(-k)}(X_{ti})) - w(X_{ti};\nu_0,e_{0,N})(z-e_{0,N}(X_{ti}))\big] \\
&\qquad\times [Y_{ti}(z) - m_0(0,X_{ti}) - z\psi(X_{ti})^\top\theta_0]\psi(X_{ti}) \\
&\qquad+ w(X_{ti};\hat{\nu}^{(-k)},\hat{e}^{(-k)})(z-\hat{e}^{(-k)}(X_{ti}))(m_0(0,X_{ti}) - \hat{m}^{(-k)}(0,X_{ti}))\psi(X_{ti}).
\end{aligned}$$

Taking the conditional expectation given $\mathcal{S}^{(-k)}, X_{ti}$ gives

$$\begin{aligned}
&\Big\|\mathbb{E}\big[s(W_{ti}(z);\hat{\nu}^{(-k)},\hat{e}^{(-k)}) - s(W_{ti}(z);\nu_0,e_{0,N})\,|\,\mathcal{S}^{(-k)},X_{ti}\big]\Big\|^2 \\
&\quad= w^2(X_{ti};\hat{\nu}^{(-k)},\hat{e}^{(-k)})(z-\hat{e}^{(-k)}(X_{ti}))^2(\hat{m}^{(-k)}(0,X_{ti}) - m_0(0,X_{ti}))^2\|\psi(X_{ti})\|^2 \\
&\quad\leqslant \frac{C^2}{c^2}(\hat{m}^{(-k)}(0,X_{ti}) - m_0(0,X_{ti}))^2
\end{aligned}$$

and so

$$S_t^{(-k)}(z) \leqslant C_0 \sqrt{\frac{1}{n_{t,k}} \sum_{(t,i)\in\mathcal{I}_k} (\hat{m}^{(-k)}(0, X_{ti}) - m_0(0, X_{ti}))^2} = o_p(N^{-1/4})$$

in view of conditional Markov's inequality, as

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{n_{t,k}} \sum_{(t,i)\in\mathcal{I}_k} (\hat{m}^{(-k)}(0, X_{ti}) - m_0(0, X_{ti}))^2 \;\middle|\; \mathcal{S}^{(-k)}\right] &= \|\hat{m}^{(-k)}(0,\cdot) - m_0(0,\cdot)\|_{2,P_t^X} \\
&\leqslant \kappa_t^{-1} \|\hat{m}^{(-k)}(0,\cdot) - m_0(0,\cdot)\|_{2,P_0^X} \\
&= o_p(N^{-1/4})
\end{aligned}$$

by assumption.

Having shown all conditions of Assumption 4.1, we can apply Theorem 4.1 to complete the proof of Corollary 4.2.

C.8 Proof of Lemma 5.1

Here we show that the information functions $\Psi_d(\cdot)$ and $\Psi_a(\cdot)$ for D-optimality and A-optimality, respectively, both satisfy conditions (a) through (d) of Assumption 5.2. That both $\Psi_d(\cdot)$ and $\Psi_a(\cdot)$ are continuous, concave, and non-decreasing on $\mathbb{S}_+^p$ is well known.

If M is singular then $\Psi_d(M) = \Psi_a(M) = -\infty$; however if $M \in \mathbb{S}_{++}^p$ then $\Psi_d(M)$ and $\Psi_a(M)$ are finite. Thus condition (a) holds with $\Psi_0 = -\infty$.

Next, we recall that for $M \in \mathbb{S}_{++}^p$, we have $\nabla\Psi_d(M) = M^{-1}$ and $\nabla\Psi_a(M) = M^{-2}$. Now we fix $0 < k < K$ and $A, B \in \mathbb{S}_{++}^p$ such that $KI \succcurlyeq A \succcurlyeq kI$ and $KI \succcurlyeq B \succcurlyeq kI$. Then

$$\begin{aligned}
\|\nabla\Psi_d(A) - \nabla\Psi_d(B)\| &= \|A^{-1} - B^{-1}\| = \|A^{-1}(B-A)B^{-1}\| \\
&\leqslant \|A^{-1}\|\|B^{-1}\|\|B-A\| \leqslant k^{-2}\|A-B\| \quad \text{and} \\
\|\nabla\Psi_a(A) - \nabla\Psi_a(B)\| &= \|A^{-2} - B^{-2}\| = \|A^{-2}(B^2-A^2)B^{-2}\| \\
&\leqslant \|A^{-2}\|\|B^{-2}\|(\|B\|\|B-A\| + \|B-A\|\|A\|) \\
&\leqslant Kk^{-4}\|A-B\|
\end{aligned}$$

which shows condition (b).

We also have $K^{-1}I \preccurlyeq \nabla\Psi_d(A) \preccurlyeq k^{-1}I$ and $K^{-2}I \preccurlyeq \nabla\Psi_a(A) \preccurlyeq k^{-2}I$. Therefore condition (c) holds.

Finally fix $\tilde{\Psi}_0 > -\infty$, and suppose $0 \preccurlyeq A \preccurlyeq KI$ with $\Psi_d(A) \geqslant \tilde{\Psi}_0$. Letting $\lambda_1 \geqslant \ldots \geqslant \lambda_p$ be the eigenvalues of A, we have $\lambda_j \leqslant K$ for all j and so

$$(p-1)\log(K) + \lambda_p \geqslant \Psi_d(A) = \sum_{j=1}^p \log\lambda_j \geqslant \tilde{\Psi}_0$$

so that $\lambda_p \geqslant \exp(\tilde{\Psi}_0 - (p-1)\log K) > 0$, showing condition (d) for $\Psi_d(\cdot)$. Similarly if $\Psi_a(A) \geqslant \tilde{\Psi}_0$ then

$$-\frac{p-1}{K} - \lambda_p^{-1} \geqslant \Psi_a(A) = -\sum_{j=1}^p \lambda_j^{-1} \geqslant \tilde{\Psi}_0$$

which implies $-\tilde{\Psi}_0 - (p-1)/K \geqslant \lambda_p^{-1} > 0$. Therefore $\lambda_p \geqslant (-\tilde{\Psi}_0 - (p-1)/K)^{-1} > 0$, showing condition (d) for $\Psi_a(\cdot)$ as well.

C.9 Proof of Lemma 5.2: Convergence of generic concave maximization routine

Here we prove Lemma 5.2 about the convergence rates of our generic concave maximization routine.

Proof. Let $\tilde{\delta} = \delta/2$, with $\delta < 0$ as in Assumption 5.1. We will repeatedly use the fact that since $[\tilde{\delta}, 1-\tilde{\delta}] \times \mathcal{W} \subseteq \mathbb{R}^{r+1}$ is compact, Assumption 5.1 implies that $f = f(\cdot,\cdot)$ and all of its partial derivatives up to second order are uniformly bounded above in norm on that set. Then WLOG we can make C from Assumption 5.1 larger so that

$$\sup_{(k,w)\in[\tilde{\delta},1-\tilde{\delta}]\times\mathcal{W}} \|h(k,w)\| \leqslant C, \quad \forall h = h(\cdot,\cdot) \in \{f, f', f'', f_w, f_{ww}, f'_w\} \tag{S25}$$

where $f_w = f_w(\cdot,\cdot)$, $f_{ww} = f_{ww}(\cdot,\cdot)$, and $f'_w = f'_w(\cdot,\cdot)$ are tensors with f_w the partial derivative of f with respect to the second argument, f_{ww} the second partial derivative of f with respect to the second argument, and f'_w the partial derivative of f' with respect to the second argument.

First, we show the existence of $e^*(\cdot)$ satisfying (28). For each propensity $e = e(\cdot) \in \mathcal{E}$ define $\phi(e) = \Psi(M(e))$ and $\hat{\phi}_n(e) = \Psi(\hat{M}_n(e))$ where

$$M(e) = \int_{\mathcal{X}} f(e(x), \eta(x))\mathrm{d}P(x) \quad \text{and} \quad \hat{M}_n(e) = \frac{1}{n}\sum_{i=1}^{n} f(e(X_i), \hat{\eta}(X_i)).$$

For any $e_1 = e_1(\cdot), e_2 = e_2(\cdot) \in \mathcal{E}$ and $\lambda \in [0,1]$ we have

$$\begin{aligned} M(\lambda e_1 + (1-\lambda)e_2) &= \int_{\mathcal{X}} f\big(\lambda e_1(x) + (1-\lambda)e_2(x), \eta(x)\big)\,\mathrm{d}P(x) \\ &\succcurlyeq \int_{\mathcal{X}} \big(\lambda f(e_1(x), \eta(x)) + (1-\lambda) f(e_2(x), \eta(x))\big)\,\mathrm{d}P(x) \\ &= \lambda M(e_1) + (1-\lambda)M(e_2) \end{aligned}$$

where the matrix inequality follows from the fact that the function $u \mapsto f(u, \eta(x))$ is a concave matrix-valued function on $[0,1]$ since its second derivative is globally negative semidefinite by (30). Thus $M = M(\cdot)$ is also a concave matrix-valued function on $\mathcal{E}$. In fact, M is also Lipschitz continuous in the sense that

$$\begin{aligned} \|M(e_1) - M(e_2)\| &\leqslant \int_{\mathcal{X}} \|f(e_1(x), \eta(x)) - f(e_2(x), \eta(x))\|\,\mathrm{d}P(x) \\ &\leqslant C\int_{\mathcal{X}} |e_1(x) - e_2(x)|\,\mathrm{d}P(x) \\ &\leqslant C\|e_1 - e_2\|_{2,P}, \quad \forall e_1, e_2 \in \mathcal{E} \end{aligned} \tag{S26}$$

where the second inequality uses (S25) with $h = f'$ and Taylor's theorem with the Lagrange form of the remainder. With the information function $\Psi = \Psi(\cdot)$ continuous, concave, and increasing in the semidefinite ordering by Assumption 5.2, we conclude that $\phi = \phi(\cdot)$ is continuous and concave on $\mathcal{E}$ (e.g., by Section 3.6 in [5]).

We now consider an extension $\bar{\phi} = \bar{\phi}(\cdot)$ of ϕ to $L^2(P)$,

$$\bar{\phi}(e) = \begin{cases} \phi(e), & e \in \mathcal{E} \\ -\infty, & \text{otherwise.} \end{cases}$$

Since $\phi(\cdot)$ is continuous and concave on $\mathcal{E}$, it is straightforward to show that the extension $\bar{\phi}$ is concave and upper semicontinuous on $L^2(P)$; the latter means that $\bar{\phi}(e_0) = \limsup_{e \to e_0} \bar{\phi}(e)$ for all $e_0 = e_0(\cdot) \in L^2(P)$. The function $\bar{\phi}$ is also "proper" in that it never equals $+\infty$. Since $\mathcal{F}_*$ is a closed, bounded, and convex subset of the Hilbert space $L^2(P)$, by Proposition 1.88 and Theorem 2.11 of [6] we conclude $\bar{\phi}(\cdot)$, and hence $\phi(\cdot)$, attains its maximum on $\mathcal{F}_*$. This shows the existence of $e^* = \arg\max_{e \in \mathcal{F}} \phi(e)$. We will show uniqueness (P-almost surely) later.

Next, note that compactness of $F_n \subseteq \mathbb{R}^n$ and continuity of the map $(e_1, \ldots, e_n) \mapsto \Psi\left(n^{-1} \sum_{i=1}^n f(e_i, \hat{\eta}(X_i))\right)$, which follows from continuity of the information function Ψ and of $e \mapsto f(e, \hat{\eta}(x))$ on $[0,1]$ for each $x \in \mathcal{X}$, ensure the existence of a vector $(\hat{e}_1, \ldots, \hat{e}_n)$ satisfying (32). This shows the second claim of Lemma 5.2. Then condition 2 of Assumption 5.1 ensures that there exists a propensity $\hat{e} = \hat{e}(\cdot) \in \mathcal{E}$ with $\hat{e}(X_i) = \hat{e}_i$ for all $i = 1, \ldots, n$. So all that remains is to show that any such propensity $\hat{e}(\cdot)$ satisfies the rate conditions $\|\hat{e} - e^*\|_{2,P_n} + \|\hat{e} - e^*\|_{2,P} = O_p(n^{-1/4} + \alpha_n)$, along with uniqueness of $e^*(\cdot)$ satisfying (28) (P-almost surely).

We have now established existence of $e^*(\cdot)$, $\hat{e}_1, \ldots, \hat{e}_n$ and $\hat{e}(\cdot)$. It remains to show the desired convergence rates of $\hat{e}(\cdot)$ to $e^*(\cdot)$ in L^2 under both P and $\hat{P}_n$. Before proceeding further, we list a few useful facts. By (S25) with $h = f$ and recalling that the Frobenius norm of a matrix upper bounds its spectral norm, we have

$$M(e) \preccurlyeq CI \quad \text{and} \quad \hat{M}_n(e) \preccurlyeq CI, \quad \forall e \in \mathcal{E}. \tag{S27}$$

Next, we claim that

$$M(e^*) \succcurlyeq k^* I \tag{S28}$$

for some $k^* > 0$. To see this, note that $\Psi(M(e^*)) = \phi(e^*) \geqslant \phi(e_0) = \Psi(M(e_0)) > \Psi_0$ by Assumption 5.2(a) and condition 3 of Assumption 5.1. Then Assumption 5.2(d) shows that the $k_* > 0$ we need exists.

Now, we define the function class

$$\mathcal{F}_n = \left\{ e \in \mathcal{E} \;\middle|\; m_L \leqslant \frac{1}{n} \sum_{i=1}^n e(X_i) \leqslant m_H \right\}.$$

This class contains any propensity $\hat{e}(\cdot)$ derived as above by solving (32) and interpolating within the base propensity class $\mathcal{E}$.

We complete the proof by the following 4 steps:

1. Show that ϕ is strongly concave on $\mathcal{F}_*$ with respect to the $\|\cdot\|_{2,P}$ norm and that $\hat{\phi}_n$ is strongly concave on $\mathcal{F}_n$ with respect to the $\|\cdot\|_{2,P_n}$ norm. This means there exist nonrandom positive constants c_0, r_0, and k_0 such that

$$c_0 \min\bigl(r_0, \|e^* - e\|_{2,P}\bigr) \times \|e^* - e\|_{2,P} \leqslant \phi(e^*) - \phi(e), \quad \forall e \in \mathcal{F}_* \tag{S29}$$

and

$$c_0 \min\bigl(r_0, \|\hat{e} - e\|_{2,P_n}\bigr) \times \|\hat{e} - e\|_{2,P_n} \mathbf{1}(A_n) \leqslant (\hat{\phi}_n(\hat{e}) - \hat{\phi}_n(e))\mathbf{1}(A_n), \quad \forall e \in \mathcal{F}_n \tag{S30}$$

where A_n is the event that $\hat{M}_n(\hat{e}) \succcurlyeq k_0 I$. Equation (S29) shows that $e^*(\cdot)$ is unique P-a.s., because any e that maximizes (28) makes the right hand size of (S29) equal 0 which then makes $\|e - e^*\|_{2,P} = 0$.

2. Conclude by the previous step that with probability approaching 1,

$$c_0 \min(r_0, \|e^* - \hat{e}_{\mathcal{F}}\|_{2,P}) \times \|e^* - \hat{e}_{\mathcal{F}}\|_{2,P} \leqslant \phi(e^*) - \phi(\hat{e}_{\mathcal{F}}) + \hat{\phi}_n(\hat{e}) - \hat{\phi}_n(e_n^*),$$

and

$$c_0 \min(r_0, \|\hat{e} - e_n^*\|_{2,P_n}) \times \|\hat{e} - e_n^*\|_{2,P_n} \mathbf{1}(A_n) \leqslant (\phi(e^*) - \phi(\hat{e}_{\mathcal{F}}) + \hat{\phi}_n(\hat{e}) - \hat{\phi}_n(e_n^*))\mathbf{1}(A_n)$$

for all $\hat{e}_{\mathcal{F}} \in \mathcal{F}$, $e_n^* \in \mathcal{F}_n$.

3. Show that with probability approaching 1, there exist $\hat{e}_{\mathcal{F}} \in \mathcal{F}$ and $e_n^* \in \mathcal{F}_n$ converging at the rate $O_p(n^{-1/2})$ in sup-norm to $\hat{e}$ and e^*, respectively, so that by empirical process arguments we can argue that

$$\phi(e^*) - \phi(\hat{e}_{\mathcal{F}}) + \hat{\phi}_n(\hat{e}) - \hat{\phi}_n(e_n^*) = O_p(n^{-1/2}) + O_p(\alpha_n). \tag{S31}$$

Show that $\Pr(A_n) \to 1$ as $n \to \infty$ and conclude by the previous step that $\|e^* - \hat{e}_{\mathcal{F}}\|_{2,P} + \|\hat{e} - e_n^*\|_{2,P_n} = O_p(n^{-1/4}) + O_p(\alpha_n^{-1/2})$. Then by the definitions of $\hat{e}_{\mathcal{F}}$ and e_n^* we can conclude that

$$\|\hat{e} - e^*\|_{2,P} + \|\hat{e} - e^*\|_{2,P_n} = O_p(n^{-1/4}) + O_p(\alpha_n^{-1/2})$$

as well. In particular, $\hat{e}$ is mean square consistent for e^* both in-sample and out-of-sample.

4. Apply a "peeling" argument, similar to Theorem 3.2.5 in [3], to show that $\|\hat{e} - \tilde{e}\|_{2,P_n} = O_p(\alpha_n)$, where

$$\tilde{e} \in \operatorname*{arg\,max}_{e \in \mathcal{F}_n} \Psi\left(n^{-1} \sum_{i=1}^{n} f(e(X_i), \eta(X_i))\right)$$

is the propensity score we'd estimate with knowledge of η, i.e., by taking $\hat{\eta} = \eta$. Conclude by the previous step that $\|\hat{e} - e^*\|_{2,P_n} = O_p(n^{-1/4}) + O_p(\alpha_n)$, and show the same convergence rate holds for $\|\hat{e} - e^*\|_{2,P}$ by empirical process arguments.

Step 1

Strong concavity will be proven using calculus along with Assumptions 5.1 and 5.2. First, we notice that by Assumption 5.2(c) and (S28), we know $\nabla\Psi(A)|_{A=M(e^*)} \succcurlyeq m^* I$ for some $m^* > 0$. Then by continuity of the smallest eigenvalue function $\lambda_{\min}(\cdot)$ and of $\nabla\Psi(\cdot)$, there exists $r_0 > 0$ such that if $e \in \mathcal{E}$ satisfies $\|e - e^*\|_{2,P} \leqslant r_0$ (which implies $\|M(e) - M(e^*)\| \leqslant Cr_0$ by (S26)), then $M(e) \succcurlyeq (k^*/2)I$ and $\nabla\Psi(A)|_{A=M(e)} \succcurlyeq (m^*/2)I$. We now extend this argument to provide a high probability eigenvalue lower bound on $\hat{M}_n(e)$ for e sufficiently close to $\hat{e}$ in $L^2(\mathcal{X}, P_n)$:

Lemma C.1. *Suppose that all conditions of Lemma 5.2 hold. Fix any $k_0 > 0$ and define A_n to be the event that $\hat{M}_n(\hat{e}) \succcurlyeq k_0 I$. Then there exist $\tilde{r} > 0$ and $0 < \tilde{k} < \tilde{K}$ such that whenever A_n holds, for all $e \in \mathcal{E}$ with $\|e - \hat{e}\|_{2,P_n} \leqslant \tilde{r}$ we have $\hat{M}_n(e) \succcurlyeq (k_0/2)I$ and $\tilde{K}I \succcurlyeq \nabla\Psi(\hat{M}_n(e)) \succcurlyeq \tilde{k}I$.*

Proof of Lemma C.1. The function $\lambda_{\min}(\cdot)$ is uniformly continuous on the compact subset $\mathcal{M} = \{A : 0 \preccurlyeq A \preccurlyeq CI\}$ of $\mathbb{R}^{p\times p}$. Hence, when the event A_n holds, we know that there exists (nonrandom) $\tilde{\delta} > 0$ such that $A \succcurlyeq (k_0/2)I$ for all $A \in \mathcal{M}$ with $\|A - \hat{M}_n(\hat{e})\| \leqslant \tilde{\delta}$. By (S27), $\mathcal{M}$ contains both $\{M(e) : e \in \mathcal{E}\}$ and $\{\hat{M}_n(e) : e \in \mathcal{E}\}$. Noting that

$$\begin{aligned}
\|\hat{M}_n(e_1) - \hat{M}_n(e_2)\| &\leqslant \frac{1}{n}\sum_{i=1}^{n} \|f(e_1(X_i), \hat{\eta}(X_i)) - f(e_2(X_i), \hat{\eta}(X_i)\| \\
&\leqslant \frac{C}{n}\sum_{i=1}^{n} |e_1(X_i) - e_2(X_i)| \\
&\leqslant C\|e_1 - e_2\|_{2,P_n}. \qquad \text{(S32)}
\end{aligned}$$

we see that whenever $e \in \mathcal{E}$ with $\|e - \hat{e}\|_{2,P_n} \leqslant \tilde{r} := \tilde{\delta}/C$, we have $\|\hat{M}_n(e) - \hat{M}_n(\hat{e})\| \leqslant \tilde{\delta}$ and hence $\hat{M}_n(e) \succcurlyeq (k_0/2)I$. The conclusion of Lemma C.1 follows immediately by Assumption 5.2(c). $\square$

Next, we bound directional derivatives of ϕ and $\hat{\phi}_n$.

Lemma C.2. *For any $e_1, e_2 \in \mathcal{F}$ with $M(e_1)$ and $M(e_2)$ nonsingular, the inequality*

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2}\phi(e_1 + t(e_2 - e_1)) \leqslant \operatorname{tr}\left[\nabla\Psi(M(e_1 + t(e_2 - e_1)))^\top \left(\frac{\mathrm{d}^2}{\mathrm{d}t^2}M(e_1 + t(e_2 - e_1))\right)\right] \qquad \text{(S33)}$$

holds for each $t \in (0, 1)$. Similarly, for any $e_1, e_2 \in \mathcal{F}_n$ with $\hat{M}_n(e_1)$ and $\hat{M}_n(e_2)$ nonsingular, we have the inequality

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2}\hat{\phi}_n(e_1 + t(e_2 - e_1)) \leqslant \operatorname{tr}\left[\nabla\Psi(\hat{M}_n(e_1 + t(e_2 - e_1)))^\top \left(\frac{\mathrm{d}^2}{\mathrm{d}t^2}\hat{M}_n(e_1 + t(e_2 - e_1))\right)\right]. \qquad \text{(S34)}$$

Proof of Lemma C.2. Fix $e_1 = e_1(\cdot)$ and $e_2 = e_2(\cdot) \in \mathcal{F}$ with $M(e_1)$ and $M(e_2)$ invertible. Define

$$\tilde{M}(t) = M(e_{(t)})$$

where $e_{(t)} = e_1 + t(e_2 - e_1)$. We first show that

$$\tilde{M}'(t) = \int_{\mathcal{X}} (e_2(x) - e_1(x)) f'(e_{(t)}(x), \eta(x)) \, \mathrm{d}P(x), \quad \forall t \in [0,1], \quad \text{and} \tag{S35}$$

$$\tilde{M}''(t) = \int_{\mathcal{X}} (e_2(x) - e_1(x))^2 f''(e_{(t)}(x), \eta(x)) \, \mathrm{d}P(x), \quad \forall t \in (0,1). \tag{S36}$$

We include the endpoints $t = 0, 1$ in (S35) so that we can apply Taylor's theorem with the Lagrange form of the second order remainder to complete the proof of Lemma C.2. We could also strengthen (S36) to include those endpoints, but this will not be needed.

Consider the difference quotient

$$D_1(t, h; X) = \frac{f(e_{(t+h)}(X), \eta(X)) - f(e_{(t)}(X), \eta(X))}{h}$$

By the chain rule we know that

$$\lim_{h \to 0} D_1(t, h; X) = \frac{\mathrm{d}}{\mathrm{d}t} f(e_{(t)}(X), \eta(X)) = (e_2(X) - e_1(X)) f'(e_{(t)}(X), \eta(X)).$$

Furthermore the fact that $e_1(x), e_2(x) \in [0,1]$ for all $x \in \mathcal{X}$ indicates

$$\tilde{\delta} \leqslant -h \leqslant e_{(t)}(x) \wedge e_{(t+h)}(x) \leqslant e_{(t)}(x) \vee e_{(t+h)}(x) \leqslant 1 + h \leqslant 1 - \tilde{\delta}$$

for all $t \in [0,1]$ and $|h| \leqslant -\tilde{\delta}$. By uniform boundedness of f' on $[\tilde{\delta}, 1 - \tilde{\delta}] \times \mathcal{W}$ and Taylor's theorem, and noting $e_{(t+h)}(x) - e_{(t)}(x) = h(e_2(x) - e_1(x))$ we conclude that

$$\sup_{0 < |h| \leqslant -\tilde{\delta}} \|D_1(t, h; X)\| \leqslant |e_2(X) - e_1(X)| \times \sup_{e \in [\tilde{\delta}, 1 - \tilde{\delta}]} \|f'(e, \eta(X))\| \leqslant C$$

so by dominated convergence

$$\begin{aligned} \lim_{h \to 0} \frac{\tilde{M}(t+h) - \tilde{M}(t)}{h} &= \lim_{h \to 0} \int_{\mathcal{X}} D_1(t, h; x) \, \mathrm{d}P(x) \\ &= \int_{\mathcal{X}} (e_2(x) - e_1(x)) f'(e_{(t)}(x), \eta(x)) \, \mathrm{d}P(x) \end{aligned}$$

which establishes (S35).

Similarly we define the second difference quotient

$$D_2(t, h; X) = \frac{f'(e_{(t+h)}(X), \eta(X)) - f'(e_{(t)}(X), \eta(X))}{h}.$$

By the chain rule we once again have

$$\lim_{h \to 0} D_2(t, h; X) = \frac{\mathrm{d}}{\mathrm{d}t} f'(e_{(t)}(X), \eta(X)) = (e_2(X) - e_1(X)) f''(e_{(t)}(X), \eta(X)).$$

By uniform boundedness of f'' on $[\tilde{\delta}, 1 - \tilde{\delta}] \times \mathcal{W}$ we get

$$\sup_{0 < |h| \leqslant -\tilde{\delta}} |D_2(t, h; X)| \leqslant |e_2(X) - e_1(X)| \times \sup_{e \in [\tilde{\delta}, 1 - \tilde{\delta}]} \|f''(e, \eta(X))\| \leqslant C.$$

Then in view of (S35) we can apply dominated convergence to conclude that

$$\lim_{h\to 0}\frac{\tilde{M}'(t+h)-\tilde{M}'(t)}{h} = \lim_{h\to 0}\int_{\mathcal{X}} D_2(t,h;x)(e_2(x)-e_1(x))\,\mathrm{d}P(x)$$
$$= \int_{\mathcal{X}} (e_2(x)-e_1(x))^2 f''(e_{(t)}(x),\eta(x))\,\mathrm{d}P(x)$$

establishing (S36).

Now we differentiate $\tilde{\phi}(t) := \phi(e_{(t)})$. Note $M(e_{(t)}) \succ 0$ for all $t \in [0,1]$ by concavity of $M(\cdot)$, shown previously. Then using (S35) and (S36), we apply the chain rule to get

$$\tilde{\phi}'(t) = \mathrm{tr}\big[\nabla\Psi(\tilde{M}(t))^\top \tilde{M}'(t)\big], \quad t \in [0,1].$$

Similarly for all $t \in (0,1)$

$$\tilde{\phi}''(t) = D^2\Psi(\tilde{M}(t))(\tilde{M}'(t),\tilde{M}'(t)) + \mathrm{tr}[\nabla\Psi(\tilde{M}(t))^\top \tilde{M}''(t)]$$
$$\leqslant \mathrm{tr}\big[\nabla\Psi(\tilde{M}(t))^\top \tilde{M}''(t)\big],$$

which shows (S33). Here $D^2\Psi(\tilde{M}(t))$ is the second derivative mapping of Ψ evaluated at $\tilde{M}(t)$, viewed as a bilinear function from $\mathbb{R}^{p\times p}\times\mathbb{R}^{p\times p}$ to $\mathbb{R}$; the inequality in the preceding display follows from concavity of Ψ.

Equation (S34) follows by a very similar calculation, though the argument is simplified, since dominated convergence is no longer needed as we are dealing with finite sums instead of integrals. Instead we immediately perform term-by-term differentiation to conclude

$$\tilde{M}_n'(t) = \frac{1}{n}\sum_{i=1}^n (e_2(X_i)-e_1(X_i)) f'(e_{(t)}(X_i),\hat{\eta}(X_i)), \quad \forall t\in[0,1], \quad \text{and} \tag{S37}$$

$$\tilde{M}_n''(t) = \frac{1}{n}\sum_{i=1}^n (e_2(X_i)-e_1(X_i))^2 f''(e_{(t)}(X_i),\hat{\eta}(X_i)), \quad \forall t\in(0,1) \tag{S38}$$

where $\tilde{M}_n(t) := \hat{M}_n(e_{(t)})$, and then use the chain rule as above. □

We are now ready to prove (S29). We apply Lemma C.2 with $e_1 = e^*$ and any $e_2 \in \mathcal{F}$ with $\|e_2 - e^*\|_{2,P} \leqslant r_0$. Note that our definition of r_0 (in the paragraph before the statement of Lemma C.1) along with (S28) ensures that $M(e_1)$ and $M(e_2)$ are nonsingular. Also, note that $\tilde{\phi}'(0) \leqslant 0$ by optimality of e^*. Lemma C.2 along with Taylor's theorem with the Lagrange form of the remainder then enables us to conclude

$$\phi(e_2) = \tilde{\phi}(1) = \tilde{\phi}(0) + \tilde{\phi}'(0) + \frac{1}{2}\tilde{\phi}''(t) \leqslant \phi(e^*) + \frac{1}{2}\mathrm{tr}\big[\nabla\Psi(\tilde{M}(t))^\top\tilde{M}''(t)\big]$$

for some $t \in (0,1)$, where $\tilde{M}(t) = M(e_1 + t(e_2 - e_1))$ as in the proof of Lemma C.2. By (30) and (S36), we know that $\tilde{M}''(t) \preccurlyeq 0$. Recalling that the trace of the product of two symmetric positive semidefinite matrices is nonnegative, we have

$$0 \geqslant \mathrm{tr}[(\nabla\Psi(\tilde{M}(t)) - (m^*/2)I)^\top\tilde{M}''(t)] = \mathrm{tr}[\nabla\Psi(\tilde{M}(t))^\top\tilde{M}''(t)] - \frac{m^*}{2}\mathrm{tr}(\tilde{M}''(t))$$

since $\nabla\Psi(\tilde{M}(t)) \succcurlyeq (m^*/2)I$. Then

$$\begin{aligned}\operatorname{tr}[\nabla\Psi(\tilde{M}(t))^\top \tilde{M}''(t)] &\leqslant \frac{m^*}{2}\operatorname{tr}(\tilde{M}''(t)) \\ &\leqslant -\frac{cm^*}{2}\int_{\mathcal{X}}(e^*(x)-e_2(x))^2\,\mathrm{d}P(x) \\ &= -\frac{cm^*}{2}\|e^*-e_2\|_{2,P}^2\end{aligned}$$

where the second inequality follows by (30) and (S36) once again. We conclude that whenever $e_2 \in \mathcal{F}$ with $\|e_2 - e^*\|_{2,P} \leqslant r_0$ we have

$$\phi(e^*) \geqslant \phi(e_2) + \frac{cm^*}{4}\|e_2 - e^*\|_{2,P}^2.$$

Now take any $e_2 \in \mathcal{F}$ with $\|e_2 - e^*\|_{2,P} > r_0$. Define $t = 1 - r_0/\|e_2 - e^*\|_{2,P} \in (0,1)$ and consider $\tilde{e}_2 = te^* + (1-t)e_2$ so that $\|\tilde{e}_2 - e^*\|_{2,P} = r_0$. Note $\tilde{e}_2 \in \mathcal{F}$ by convexity of $\mathcal{F}_*$, so by the preceding display

$$\phi(e^*) \geqslant \phi(\tilde{e}_2) + \frac{cm^*r_0^2}{4} \geqslant t\phi(e^*) + (1-t)\phi(e_2) + \frac{cm^*r_0^2}{4}$$

where the second inequality is by concavity of ϕ. Rearranging we have

$$\phi(e^*) \geqslant \phi(e_2) + \frac{cm^*r_0^2}{4(1-t)} = \phi(e_2) + \frac{cm^*r_0}{4}\|e^* - e_2\|_{2,P}.$$

Letting $c_0 = cm^*/4$, we conclude that for all $e \in \mathcal{F}$ we have

$$\phi(e^*) \geqslant \phi(e) + c_0\min(r_0, \|e^* - e\|_{2,P})\|e^* - e\|_{2,P}$$

which shows (S29).

The proof of (S30) is quite similar. Take $k_0 > 0$ such that with k^* as in (S28), whenever $0 \preccurlyeq A \preccurlyeq CI$ with $\Psi(A) \geqslant \inf_{B \succcurlyeq (k^*/4)I}\Psi(B)$, we have $A \succcurlyeq k_0 I$. Such a k_0 exists by Assumptions 5.2(a) and 5.2(d). By Lemma C.1, on the truncation event A_n that $\hat{M}_n(\hat{e}) \succcurlyeq k_0 I$, we have $\tilde{K}I \geqslant \nabla\Psi(\hat{M}_n(e)) \geqslant \tilde{k}I$ whenever $\|e - \hat{e}\|_{2,P_n} \leqslant \tilde{r}$, for some $\tilde{r} > 0$ and $0 < \tilde{k} < \tilde{K}$. Now we apply (S34) with $e_1 = \hat{e}$ and any $e_2 \in \mathcal{F}_n$ with $\|e_2 - \hat{e}\|_{2,P_n} \leqslant \tilde{r}$ (note we must have $\hat{M}_n(e_1)$ and $\hat{M}_n(e_2)$ nonsingular). Defining $\tilde{\phi}_n(t) := \hat{\phi}_n(e_{(t)})$ for $e_{(t)} = e_1 + t(e_2 - e_1)$, by optimality of $\hat{e}$ we must have $\tilde{\phi}_n'(0) \leqslant 0$. Taylor's theorem and the second part of Lemma C.2 then allow us to conclude that

$$\begin{aligned}\hat{\phi}_n(e_2)\mathbf{1}(A_n) = \tilde{\phi}_n(1)\mathbf{1}(A_n) &= \Big(\tilde{\phi}_n(0) + \tilde{\phi}_n'(0) + \frac{1}{2}\tilde{\phi}_n''(t)\Big)\mathbf{1}(A_n) \\ &\leqslant \Big(\hat{\phi}_n(\hat{e}) + \frac{1}{2}\operatorname{tr}\big[\nabla\Psi(\tilde{M}_n(t))^\top \tilde{M}_n''(t)\big]\Big)\mathbf{1}(A_n) \qquad \text{(S39)}\end{aligned}$$

for some $t \in (0,1)$, where $\tilde{M}_n(t) = \hat{M}_n(e_{(t)})$ as in the proof of Lemma C.2. With $\nabla\Psi(\tilde{M}_n(t)) \succcurlyeq \tilde{k}I$ for all $t \in (0,1)$ whenever A_n holds, we have by (S38) that

$$0 \geqslant \operatorname{tr}\big[(\nabla\Psi(\tilde{M}_n(t)) - \tilde{k}I)^\top \tilde{M}_n''(t)\big]\mathbf{1}(A_n) = \big(\operatorname{tr}[(\nabla\Psi(\tilde{M}_n(t))^\top \tilde{M}_n''(t)] - \tilde{k}\operatorname{tr}[\tilde{M}_n''(t)]\big)\mathbf{1}(A_n).$$

Hence by (30)

$$\mathrm{tr}[(\nabla\Psi(\tilde{M}_n(t))^\top \tilde{M}_n''(t)]\mathbf{1}(A_n) \leqslant \tilde{k}\mathrm{tr}[\tilde{M}_n''(t)]\mathbf{1}(A_n) \leqslant -\tilde{k}c\|\hat{e}-e\|_{2,P_n}^2\mathbf{1}(A_n).$$

Then by (S39) we conclude that whenever $\|e-\hat{e}\|_{2,P_n} \leqslant \tilde{r}$ we have

$$(\hat{\phi}_n(\hat{e}) - \hat{\phi}_n(e))\mathbf{1}(A_n) \geqslant c\frac{\tilde{k}}{2}\|\hat{e}-e\|_{2,P_n}^2\mathbf{1}(A_n).$$

Redefining r_0 to be the minimum of the r_0 appearing in the proof of (S29) and $\tilde{r}$, since $\hat{\phi}_n$ is always concave on $\mathcal{F}_n$. we can repeat the argument at the end of the proof of (S29) to conclude that

$$(\hat{\phi}_n(\hat{e}) - \hat{\phi}_n(e))\mathbf{1}(A_n) \geqslant (c_0 \min(r_0, \|\hat{e}-e\|_{2,P_n})\|\hat{e}-e\|_{2,P_n})\,\mathbf{1}(A_n)$$

for all $e \in \mathcal{F}_n$; here $c_0 := c\tilde{k}/2 > 0$.

Step 2

Fix $\hat{e}_{\mathcal{F}} \in \mathcal{F}_*$ and $e_n^* \in \mathcal{F}_n$. The result of Step 1 shows that for some positive constants c_0 and r_0, we have

$$\begin{aligned} c_0 \min(r_0, \|e^*-\hat{e}_{\mathcal{F}}\|_{2,P})\|e^*-\hat{e}_{\mathcal{F}}\|_{2,P} &\leqslant \phi(e^*) - \phi(\hat{e}_{\mathcal{F}}), \quad \text{and} \\ c_0 \min(r_0, \|\hat{e}-e_n^*\|_{2,P_n})\|\hat{e}-e_n^*\|_{2,P_n}\mathbf{1}(A_n) &\leqslant (\hat{\phi}_n(\hat{e}) - \hat{\phi}_n(e_n^*))\mathbf{1}(A_n). \end{aligned}$$

With $\phi(e^*) - \phi(\hat{e}_{\mathcal{F}}) \geqslant 0$ and $\hat{\phi}_n(\hat{e}) - \hat{\phi}_n(e_n^*) \geqslant 0$ by the definitions of e^* and $\hat{e}$, we can further upper bound the right-hand sides by

$$\begin{aligned} \phi(e^*) - \phi(\hat{e}_{\mathcal{F}}) &\leqslant \phi(e^*) - \phi(\hat{e}_{\mathcal{F}}) + \hat{\phi}_n(\hat{e}) - \hat{\phi}_n(e_n^*), \quad \text{and} \\ (\hat{\phi}_n(\hat{e}) - \hat{\phi}_n(e_n^*))\mathbf{1}(A_n) &\leqslant \phi(e^*) - \phi(\hat{e}_{\mathcal{F}}) + \hat{\phi}_n(\hat{e}) - \hat{\phi}_n(e_n^*). \end{aligned}$$

Step 3

For brevity, in this section we introduce the empirical process notation

$$Pe = \int_{\mathcal{X}} e(x)\,\mathrm{d}P(x)$$

for all $e \in \mathcal{E}$. For instance, with P_n the empirical measure induced by $X_1, \ldots, X_n$, we have $P_n e = n^{-1}\sum_{i=1}^n X_i$ for all $e \in \mathcal{E}$.

We first show that we can choose particular $\hat{e}_{\mathcal{F}} = \hat{e}_{\mathcal{F}}(\cdot) \in \mathcal{F}$ and $e_n^*(\cdot) \in \mathcal{F}_n$ that are very close in sup norm to $\hat{e} \in \mathcal{F}_n$ and $e^* \in \mathcal{F}$, respectively, with high probability.

Lemma C.3. *Under the conditions of Lemma 5.2, there exist $\hat{e}_{\mathcal{F}} = \hat{e}_{\mathcal{F}}(\cdot) \in \mathcal{F}$ and $e_n^* = e_n^*(\cdot) \in \mathcal{F}_n$ such that* $\sup_{x\in\mathcal{X}}|\hat{e}(x) - \hat{e}_{\mathcal{F}}(x)| + \sup_{x\in\mathcal{X}}|e^*(x) - e_n^*(x)| = O_p(n^{-1/2})$.

Proof of Lemma C.3. With the functions in $\mathcal{E}$ uniformly bounded by 1, by Lemma B.3 we conclude

$$\mathbb{E}_P\left[\sup_{e\in\mathcal{E}}\Big|P_n e - Pe\Big|\right] \leqslant KCn^{-1/2}$$

so that $\sup_{e\in\mathcal{E}}\big|P_n e - Pe\big| = O_p(n^{-1/2})$ by Markov's inequality. In view of the fact that $m_L \leqslant P_n\hat{e} \leqslant m_H$ and $m_L \leqslant Pe^* \leqslant m_H$ since $\hat{e} \in \mathcal{F}_n$ and $e^* \in \mathcal{F}$, we have

$$m_L - \sup_{e\in\mathcal{E}}\big|P_n e - Pe\big| \leqslant P_n e^* \leqslant m_H + \sup_{e\in\mathcal{E}}\big|P_n e - Pe\big|, \quad \text{and} \tag{S40}$$

$$m_L - \sup_{e\in\mathcal{E}}\big|P_n e - Pe\big| \leqslant P\hat{e} \leqslant m_H + \sup_{e\in\mathcal{E}}\big|P_n e - Pe\big|. \tag{S41}$$

Next, with $e_L = e_L(\cdot)$ and $e_H = e_H(\cdot)$ as in the assumptions of Lemma 5.2, define

$$e_n^*(x) = \begin{cases} e^*(x), & m_L \leqslant P_n e^* \leqslant m_H \\ e^*(x) + \lambda_n(e_L(x) - e^*(x)), & P_n e^* < m_L \\ e^*(x) + \lambda_n(e_H(x) - e^*(x)), & P_n e^* > m_H \end{cases}$$

where

$$\lambda_n = \begin{cases} \dfrac{m_L - P_n e^*}{P_n[e_L - e^*]}, & P_n e^* < m_L \\ \dfrac{P_n e^* - m_H}{P_n[e^* - e_H]}, & P_n e^* > m_H \\ 0 & \text{otherwise.} \end{cases}$$

On the event

$$A_n = \left\{P_n e_L \geqslant \frac{Pe_L + m_L}{2}, P_n e_H \leqslant \frac{Pe_H + m_H}{2}\right\}$$

we must have $0 \leqslant \lambda_n \leqslant 1$ since $Pe_L > m_L$ and $Pe_H < m_H$, and so when A_n holds we know $e_n^* \in \mathcal{E}$ by convexity of $\mathcal{E}$. Furthermore we have $P_n e_n^* = \max(m_L, \min(m_H, \mathbb{E}_{P_n}[e^*(X)]))$ so that in fact $e_n^* \in \mathcal{F}_n$. But with $M = 2\max\left((Pe_L - m_L)^{-1}, (m_H - Pe_H)^{-1}\right)$ we have

$$\begin{aligned} 0 \leqslant \lambda_n \mathbf{1}(A_n) &\leqslant M\mathbf{1}(A_n)[(m_L - P_n e^*)\mathbf{1}(P_n e^* < m_L) + (P_n e^* - m_H)\mathbf{1}(P_n e^* > m_H)] \\ &\overset{\text{(S40)}}{\leqslant} 2M\sup_{e\in\mathcal{E}}|P_n e - P_e|\mathbf{1}(A_n). \end{aligned}$$

With $\Pr(A_n) \to 1$ as $n \to \infty$ by the law of large numbers, we get $\lambda_n = O_p(n^{-1/2})$, and hence for each $x \in \mathcal{X}$ we have

$$|e_n^*(x) - e^*(x)| \leqslant \lambda_n(|e_L(x) - e^*(x)| \vee |e_H(x) - e^*(x)|) \leqslant \lambda_n = O_p(n^{-1/2}).$$

Next, define

$$\hat{e}_{\mathcal{F}}(x) = \begin{cases} \hat{e}(x), & m_L \leqslant P\hat{e} \leqslant m_H \\ \hat{e}(x) + \tilde{\lambda}_n(e_L(x) - \hat{e}(x)), & P\hat{e} < m_L \\ \hat{e}(x) + \tilde{\lambda}_n(e_H(x) - \hat{e}(x)), & P\hat{e} > m_H \end{cases}$$

with

$$\tilde{\lambda}_n = \begin{cases} \dfrac{m_L - P\hat{e}}{m_H - \hat{e}}, & P\hat{e} < m_L \\ \dfrac{P\hat{e} - m_H}{P\hat{e} - m_L}, & P\hat{e} > m_H \\ 0, & \text{otherwise} \end{cases}$$

so that $0 \leqslant \tilde{\lambda}_n \leqslant 1$ and $\hat{e}_{\mathcal{F}} \in \mathcal{F}$ always. In fact

$$\begin{aligned} \tilde{\lambda}_n &\leqslant \frac{M}{2}\big[(m_L - P\hat{e})\mathbf{1}(P\hat{e} < m_L) + (P\hat{e} - m_H)\mathbf{1}(P\hat{e} > m_H)\big] \\ &\leqslant M \sup_{e\in\mathcal{E}} |P_n e - Pe| \end{aligned}$$

so that $\tilde{\lambda}_n = O_p(n^{-1/2})$ as well and the lemma follows since for all $x \in \mathcal{X}$

$$|\hat{e}(x) - \hat{e}_{\mathcal{F}}(x)| \leqslant \tilde{\lambda}_n(|e_L(x) - \hat{e}(x)| \vee |e_H(x) - \hat{e}(x)|) \leqslant \tilde{\lambda}_n.$$

□

We are now ready to prove consistency. Taking $\hat{e}_{\mathcal{F}}$ as in Lemma C.3, we upper bound the right-hand side of the inequalities in step 2:

$$\begin{aligned} \phi(e^*) - \phi(\hat{e}_{\mathcal{F}}) + \hat{\phi}_n(\hat{e}) - \hat{\phi}_n(e_n^*) \leqslant{}& |\phi(e^*) - \phi(e_n^*)| + |\phi(e_n^*) - \hat{\phi}_n(e_n^*)| \\ &+ |\hat{\phi}_n(\hat{e}) - \hat{\phi}_n(\hat{e}_{\mathcal{F}})| + |\hat{\phi}_n(\hat{e}_{\mathcal{F}}) - \phi(\hat{e}_{\mathcal{F}})|. \end{aligned} \tag{S42}$$

In view of (S26) and (S32), Lemma C.3 shows that

$$\|M(e^*) - M(e_n^*)\| = O_p(n^{-1/2}), \quad \text{and} \tag{S43}$$

$$\|\hat{M}_n(\hat{e}) - \hat{M}_n(\hat{e}_{\mathcal{F}})\| = O_p(n^{-1/2}). \tag{S44}$$

We now use (S43) to show that

$$|\phi(e^*) - \phi(e_n^*)| = O_p(n^{-1/2}). \tag{S45}$$

As shown at the start of step 1, whenever $\|e_n^* - e^*\|_{2,P} \leqslant r_0$ we have $M(e_n^*) \succcurlyeq (k^*/2)I$, so that

$$tM(e^*) + (1-t)M(e_n^*) \succcurlyeq (k^*/2)I \quad \forall t \in [0,1].$$

Applying Taylor's theorem to $\Psi(\cdot)$ we have for some $K < \infty$ that

$$\begin{aligned} |\phi(e^*) - \phi(e_n^*)|\mathbf{1}(\tilde{A}_n) &= |\Psi(M(e^*)) - \Psi(M(e_n^*))|\mathbf{1}(\tilde{A}_n) \\ &\leqslant \sup_{t\in[0,1]} \Big|\mathrm{tr}[\nabla\Psi(tM(e^*) + (1-t)M(e_n^*))]^\top [M(e^*) - M(e_n^*)]\Big|\mathbf{1}(\tilde{A}_n) \\ &\leqslant \sup_{t\in[0,1]} \|\nabla\Psi(tM(e^*) + (1-t)M(e_n^*))\| \cdot \|M(e^*) - M(e_n^*)\|\mathbf{1}(\tilde{A}_n) \\ &\leqslant K\sqrt{p} \cdot \|M(e^*) - M(e_n^*)\|\mathbf{1}(\tilde{A}_n) \end{aligned} \tag{S46}$$

where $\tilde{A}_n$ is the event $\|e_n^* - e^*\|_{2,P} \leqslant r_0$ and the last inequality follows from Assumption 5.2(c) and the fact that $\|A\| \leqslant \sqrt{p}\lambda_{\max}(A)$ for any $A \in \mathbb{S}_+^p$. Here $\lambda_{\max}(A)$ denotes the largest eigenvalue of A. Hence $|\phi(e^*) - \phi(e_n^*)|\mathbf{1}(A_n) = O_p(n^{-1/2})$ by (S43). But $\Pr(\tilde{A}_n) \to 1$ by (S43), so indeed $|\phi(e^*) - \phi(e_n^*)| = O_p(n^{-1/2})$.

Convergence of the remaining three terms in (S42) depends on the following result:

$$\sup_{e\in\mathcal{E}} \|\hat{M}_n(e) - M(e)\| = O_p(n^{-1/2}) + O_p(\alpha_n). \tag{S47}$$

To show this, define

$$M_n(e) = \frac{1}{n}\sum_{i=1}^n f(e(X_i), \eta(X_i))$$

which replaces the estimated nuisance function $\hat{\eta}$ in the definition of $\hat{M}_n$ with the true η. First note that by uniform boundedness of $\|f_w\|_2$,

$$\frac{1}{n}\sum_{i=1}^n \sup_{e\in[0,1]} \|f(e, \hat{\eta}(X_i)) - f(e, \eta(X_i))\| \leqslant Cn^{-1}\sum_{i=1}^n \|\hat{\eta}(X_i) - \eta(X_i)\|_2 \leqslant C\|\hat{\eta} - \eta\|_{2,P_n}.$$

Then by (35) we have

$$\begin{aligned}\sup_{e\in\mathcal{E}} \|\hat{M}_n(e) - M_n(e)\| &= \sup_{e\in\mathcal{E}} \left\| \frac{1}{n}\sum_{i=1}^n f(e(X_i), \hat{\eta}(X_i)) - f(e(X_i), \eta(X_i)) \right\| \\ &\leqslant \frac{1}{n}\sum_{i=1}^n \sup_{e\in[0,1]} \|f(e, \hat{\eta}(X_i)) - f(e, \eta(X_i))\| \\ &= O_p(\alpha_n). \end{aligned} \tag{S48}$$

Now for $i, j \in \{1, \dots, p\}$ define the class of functions

$$\mathcal{G}_{ij} = \{x \mapsto f_{ij}(e(x), \eta(x)) \mid e \in \mathcal{E}\}$$

where f_{ij} denotes the (i,j)-th entry of the function f. By Lemma B.4

$$\begin{aligned} n^{-1/2}\int_0^1 \sqrt{\log \mathcal{N}(r, \mathcal{G}_{ij}, L^2(P_n))}\,\mathrm{d}r &= Cn^{-1/2}\int_0^{C^{-1}} \sqrt{\log \mathcal{N}(C\epsilon, \mathcal{G}_{ij}, L^2(P_n))}\,\mathrm{d}\epsilon \\ &\leqslant Cn^{-1/2}\int_0^{C^{-1}} \sqrt{\log \mathcal{N}(\epsilon, \mathcal{E}, L^2(P_n))}\,\mathrm{d}\epsilon. \end{aligned}$$

Let

$$D(e) = \frac{1}{n}\sum_{i=1}^n f(e(X_i), \eta(X_i)) - \int_{\mathcal{X}} f(e(x), \eta(x))\,\mathrm{d}P(x) = M_n(e) - M(e).$$

Then Lemma B.3 and Assumption 5.1 indicate that $\sup_{e\in\mathcal{E}} |D_{ij}(e)| = O_p(n^{-1/2})$ and so

$$\sup_{e\in\mathcal{E}} \|D(e)\| = \sup_{e\in\mathcal{E}} \left(\sum_{i=1}^p\sum_{j=1}^p D_{ij}(e)^2\right)^{1/2} \leqslant \left(\sum_{i=1}^p\sum_{j=1}^p \sup_{e\in\mathcal{E}} |D_{ij}(e)|^2\right)^{1/2} = O_p(n^{-1/2}). \tag{S49}$$

The result (S47) follows by the triangle inequality in view of (S48) and (S49).

We are finally ready to bound the remaining terms on the right-hand side of (S42), and show that $\Pr(A_n) \to 1$ as $n \to \infty$ where A_n is the event $\hat{M}_n(\hat{e}) \succcurlyeq k_0 I$ for $k_0 > 0$ defined in Step 1. Choose $\delta > 0$ so that for any A_1, A_2 in $\mathcal{M} = \{A \in \mathbb{S}_+^p : 0 \preccurlyeq A \preccurlyeq CI\}$, with $\|A_1 - A_2\| \leqslant \delta$, we have $|\lambda_{\min}(A_1) - \lambda_{\min}(A_2)| \leqslant \min(k_0/2, k^*/4)$, where k^* satisfies (S28). Such δ exists by uniform continuity of $\lambda_{\min}(\cdot)$ on the compact subset $\mathcal{M}$ of $\mathbb{R}^{p\times p}$ (cf. the proof of Lemma C.1). Also define the event B_n that all of the following are true:

$$\|e_n^* - e^*\|_{2,P} \leqslant r_0, \tag{S50}$$

$$\sup_{e\in\mathcal{E}} \|\hat{M}_n(e) - M(e)\| \leqslant \delta \quad \text{and} \tag{S51}$$

$$\|\hat{M}_n(\hat{e}) - \hat{M}_n(\hat{e}_{\mathcal{F}})\| \leqslant \delta. \tag{S52}$$

We claim that B_n implies the following conditions:

$$M(e_n^*) \succcurlyeq \frac{k^*}{2} I, \quad \hat{M}_n(e_n^*) \succcurlyeq \frac{k^*}{4} I, \quad \hat{M}_n(\hat{e}) \succcurlyeq k_0 I \quad \text{and} \quad \hat{M}_n(\hat{e}_{\mathcal{F}}) \succcurlyeq \frac{k_0}{2} I.$$

We prove these statements briefly. Assume B_n holds. First, note that $M(e_n^*) \succcurlyeq (k^*/2)I$ holds by definition of r_0 and (S50). Next, the definition of δ immediately ensures by (S51) that $\hat{M}_n(e_n^*) \succcurlyeq (k^*/4)I$. But then

$$\Psi(\hat{M}_n(\hat{e})) \geqslant \Psi(\hat{M}_n(e_n^*)) \geqslant \inf_{B \succcurlyeq (k^*/4)I} \Psi(B)$$

so that $\hat{M}_n(\hat{e}) \succcurlyeq k_0 I$ by the definition of k_0, and in particular we have shown $B_n \subseteq A_n$. Finally (S52) shows $\hat{M}_n(\hat{e}_{\mathcal{F}}) \succcurlyeq (k_0/2)I$.

Now take $\tilde{K} < \infty$ to be as derived from Assumption 5.2(c) with $k = \min(k^*/4, k_0/2)$ and $K = C$. Then repeated applications of arguments analogous to (S46) show that

$$\begin{aligned}
|\phi(e_n^*) - \hat{\phi}_n(e_n^*)|\mathbf{1}(B_n) &= |\Psi(M(e_n^*)) - \Psi(\hat{M}_n(e_n^*))|\mathbf{1}(B_n) \\
&\leqslant \sup_{0 \preccurlyeq A \preccurlyeq \tilde{K} I} |\mathrm{tr}(A^\top [M(e_n^*) - \hat{M}_n(e_n^*)])| \\
&\leqslant \tilde{K}\sqrt{p} \cdot \|M(e_n^*) - \hat{M}_n(e_n^*)\| \\
&= O_p(n^{-1/2}) + O_p(\alpha_n)
\end{aligned}$$

by (S47) and similarly

$$|\hat{\phi}_n(\hat{e}) - \hat{\phi}_n(\hat{e}_{\mathcal{F}})|\mathbf{1}(B_n) \leqslant \tilde{K}\sqrt{p} \cdot \|\hat{M}_n(\hat{e}) - \hat{M}_n(\hat{e}_{\mathcal{F}})\| = O_p(n^{-1/2})$$

by (S44). Also

$$|\hat{\phi}_n(\hat{e}_{\mathcal{F}}) - \phi(\hat{e}_{\mathcal{F}})|\mathbf{1}(B_n) \leqslant \tilde{K}\sqrt{p} \cdot \|\hat{M}_n(\hat{e}_{\mathcal{F}}) - M(\hat{e}_{\mathcal{F}})\| = O_p(n^{-1/2}) + O_p(\alpha_n)$$

by (S47).

However, by (S43), (S26), (S44), and (S47) we know that $\Pr(B_n) \to 1$ as $n \to \infty$. Since $B_n \subseteq A_n$ we also have $\Pr(A_n) \to 1$. We conclude from the preceding displays that

$$|\phi(e_n^*) - \hat{\phi}_n(e_n^*)| + |\hat{\phi}_n(\hat{e}) - \hat{\phi}_n(\hat{e}_{\mathcal{F}})| + |\hat{\phi}_n(\hat{e}_{\mathcal{F}}) - \phi(\hat{e}_{\mathcal{F}})| = O_p(n^{-1/2}) + O_p(\alpha_n).$$

Then by step 2, (S42), and (S45), we conclude that

$$\begin{aligned}\|e^* - \hat{e}_{\mathcal{F}}\|_{2,P} &= O_p(n^{-1/4}) + O_p(\alpha_n^{-1/2}) \quad \text{and} \\ \|\hat{e} - e_n^*\|_{2,P_n} &= O_p(n^{-1/4}) + O_p(\alpha_n^{-1/2}).\end{aligned}$$

But by Lemma C.3 we know that $\|\hat{e} - \hat{e}_{\mathcal{F}}\|_{2,P} + \|e^* - e_n^*\|_{2,P_n} = O_p(n^{-1/2})$. So by the triangle inequality we conclude $\|e^* - \hat{e}\|_{2,P} + \|e^* - \hat{e}\|_{2,P_n} = O_p(n^{-1/4}) + O_p(\alpha_n^{-1/2})$ as well.

Step 4

The final step in the argument to derive our best convergence rates is a variation of a standard "peeling" argument used in deriving convergence rates of M-estimators. The main argument requires deriving a bound on the "locally centered empirical process" as in our next result.

Lemma C.4. *For each $e \in \mathcal{E}$, let $\phi_n(e) = \Psi(M_n(e))$ and take $\tilde{e} \in \arg\max_{e \in \mathcal{F}_n} \phi_n(e)$. Then there exists $\beta > 0$ and a universal constant $C_0 < \infty$ such that for all $u \leqslant \beta$*

$$\sup_{e \in \mathcal{E}: \|e - \tilde{e}\|_{2,P_n} \leqslant u} \big[(\hat{\phi}_n(e) - \phi_n(e)) - (\hat{\phi}_n(\tilde{e}) - \phi_n(\tilde{e}))\big] \mathbf{1}(B_n) \leqslant C_0\big(u\|\hat{\eta} - \eta\|_{2,P_n} + \|\hat{\eta} - \eta\|_{2,P_n}^2\big) \tag{S53}$$

for some sequence of events B_n with $\Pr(B_n) \to 1$ as $n \to \infty$.

Proof of Lemma C.4. Let $U_n(e) = \hat{\phi}_n(e) - \phi_n(e)$ for each $e \in \mathcal{E}$. By Taylor's theorem, for each $e \in \mathcal{E}$

$$U_n(e) = \Psi(\hat{M}_n(e)) - \Psi(M_n(e)) = \nabla\Psi(R_n(e))^\top(\hat{M}_n(e) - M_n(e))$$

for some $R_n(e)$ lying on the line segment between $M_n(e)$ and $\hat{M}_n(e)$. Write

$$\begin{aligned}U_n(e) - U_n(\tilde{e}) = {} & (\nabla\Psi(R_n(e)) - \nabla\Psi(R_n(\tilde{e})))^\top(\hat{M}_n(e) - M_n(e)) \\ & + \nabla\Psi(R_n(\tilde{e}))^\top[(\hat{M}_n(e) - M_n(e)) - (\hat{M}_n(\tilde{e}) - M_n(\tilde{e}))].\end{aligned} \tag{S54}$$

Because $\hat{\eta} = \eta$ trivially satisfies the conditions of Lemma 5.2, all of our results in steps 1–3 apply if we replace $\hat{M}_n$ with M_n. In particular with $k_0 > 0$ as chosen in step 1, we have $\Pr(M_n(\tilde{e}) \succcurlyeq k_0 I) \to 1$ as $n \to \infty$ by the argument at the end of step 3. Take $\tilde{r}$, $\tilde{k}$, and $\tilde{K}$ derived from Lemma C.1 with this choice of k_0. By the proof of Lemma C.1, we know that for all $0 \preccurlyeq A \preccurlyeq CI$ with $\|A - M_n(\tilde{e})\| \leqslant C\tilde{r}$, we must have $A \succcurlyeq (k_0/2)I$ and $\tilde{K}I \succcurlyeq \nabla\Psi(A) \succcurlyeq \tilde{k}I$. Let B_n be the intersection of the events $M_n(\tilde{e}) \succcurlyeq k_0 I$ and $\sup_{e \in \mathcal{E}} \|\hat{M}_n(e) - M_n(e)\| \leqslant C\tilde{r}/3$. Now $\Pr(B_n) \to 1$ as $n \to \infty$ by (S48). Then for any $e \in \mathcal{E}$ with $\|e - \tilde{e}\|_{2,P_n} \leqslant \tilde{r}/3$, in view of (S32) we have

$$\begin{aligned}\|\hat{M}_n(e) - M_n(\tilde{e})\|\mathbf{1}(B_n) &\leqslant \|\hat{M}_n(e) - \hat{M}_n(\tilde{e})\| + \|\hat{M}_n(\tilde{e}) - M_n(\tilde{e})\|\mathbf{1}(B_n) \\ &\leqslant \frac{2C\tilde{r}}{3} \quad \text{and} \\ \|M_n(e) - M_n(\tilde{e})\|\mathbf{1}(B_n) &\leqslant \|M_n(e) - \hat{M}_n(e)\|\mathbf{1}(B_n) + \|\hat{M}_n(e) - M_n(\tilde{e})\|\mathbf{1}(B_n) \\ &\leqslant C\tilde{r}.\end{aligned}$$

We conclude by the preceding display that whenever B_n holds, for all $e \in \mathcal{E}$ with $\|e - \tilde{e}\|_{2,P_n} \leqslant \tilde{r}/3$ we have $(k_0/2)I \preccurlyeq \hat{M}_n(e) \preccurlyeq CI$ and $(k_0/2)I \preccurlyeq M_n(e) \preccurlyeq CI$, and thus $(k_0/2)I \preccurlyeq R_n(e) \preccurlyeq CI$ along with $\tilde{K}I \succcurlyeq \nabla\Psi(R_n(e)) \succcurlyeq \tilde{k}I$. Then by Assumption 5.2(b), for all $u \leqslant \tilde{r}/3$ there exists a constant $K_0 < \infty$ (independent of u) for which

$$\begin{aligned}
\|\nabla\Psi(R_n(e)) - \nabla\Psi(R_n(\tilde{e}))\|\mathbf{1}(B_n) &\leqslant K_0\|R_n(e) - R_n(\tilde{e})\|\mathbf{1}(B_n) \\
&\leqslant K_0(\|R_n(e) - \hat{M}_n(e)\| + \|\hat{M}_n(e) - \hat{M}_n(\tilde{e})\| + \|\hat{M}_n(\tilde{e}) - R_n(\tilde{e})\|)
\end{aligned}$$

The preceding inequality holds for all $e \in \mathcal{E}$ with $\|e - \tilde{e}\|_{2,P_n} \leqslant u$. Taking a supremum over such $e(\cdot)$, another application of (S32) and the fact that $\|R_n(e) - \hat{M}_n(e)\| \leqslant \|M_n(e) - \hat{M}_n(e)\|$ for all $e \in \mathcal{E}$ show that

$$\sup_{e\in\mathcal{E}:\|e-\tilde{e}\|_{2,P_n}\leqslant u} \|\nabla\Psi(R_n(e)) - \nabla\Psi(R_n(\tilde{e}))\|\mathbf{1}(B_n) \leqslant K_0\big(Cu + 2S_n(\mathcal{E})\big)$$

where $S_n(\mathcal{E}) = \sup_{e\in\mathcal{E}} \|\hat{M}_n(e) - M_n(e)\|$. Then by Cauchy-Schwarz

$$\begin{aligned}
&\sup_{e\in\mathcal{E}:\|e-\tilde{e}\|_{2,P_n}\leqslant u} \big|(\nabla\Psi(R_n(e)) - \nabla\Psi(R_n(\tilde{e})))^\top(\hat{M}_n(e) - M_n(e))\big|\mathbf{1}(B_n) \\
&\leqslant K_0\big(Cu + 2S_n(\mathcal{E})\big)S_n(\mathcal{E})
\end{aligned} \tag{S55}$$

for all $u \leqslant \tilde{r}/3$.

Next, define

$$c(x; e, \tilde{e}, \eta, \hat{\eta}) = \big[f(e(X_i), \hat{\eta}(X_i)) - f(e(X_i), \eta(X_i))\big] - \big[f(\tilde{e}(X_i), \hat{\eta}(X_i)) - f(\tilde{e}(X_i), \eta(X_i))\big]$$

for all $x \in \mathcal{X}$ so that

$$\big(\hat{M}_n(e) - M_n(e)\big) - \big(\hat{M}_n(\tilde{e}) - M_n(\tilde{e})\big) = \frac{1}{n}\sum_{i=1}^n c(X_i; e, \tilde{e}, \eta, \hat{\eta}).$$

Fix $k, \ell \in \{1, \dots, p\}$. By Taylor's theorem with the Lagrange form of the remainder,

$$\begin{aligned}
|c(X_i; e, \tilde{e}, \eta, \hat{\eta})| &= |[f(e(X_i), \hat{\eta}(X_i)) - f(e(X_i), \eta(X_i))] - [f(\tilde{e}(X_i), \hat{\eta}(X_i)) - f(\tilde{e}(X_i), \eta(X_i))]| \\
&= |f_w(e(X_i), \eta_1(X_i))^\top(\hat{\eta}(X_i) - \eta(X_i)) - f_w(\tilde{e}(X_i), \eta_2(X_i))^\top(\hat{\eta}(X_i) - \eta(X_i))| \\
&\leqslant \|\hat{\eta}(X_i) - \eta(X_i)\|_2\|f_w(e(X_i), \eta_1(X_i)) - f_w(e(X_i), \eta(X_i))\|_2 \\
&\quad + \|\hat{\eta}(X_i) - \eta(X_i)\|_2\|f_w(e(X_i), \eta(X_i)) - f_w(\tilde{e}(X_i), \eta(X_i))\|_2 \\
&\quad + \|\hat{\eta}(X_i) - \eta(X_i)\|_2\|f_w(\tilde{e}(X_i), \eta(X_i)) - f_w(\tilde{e}(X_i), \eta_2(X_i))\|_2 \\
&\leqslant 2C\|\hat{\eta}(X_i) - \eta(X_i)\|_2^2 + C\|\hat{\eta}(X_i) - \eta(X_i)\|_2|\tilde{e}(X_i) - e(X_i)|.
\end{aligned}$$

We have omitted subscripts $k\ell$ on c and f everywhere in the preceding display for brevity (i.e., c above denotes $c_{k\ell}$ and f above denotes $f_{k\ell}$). The functions $e_1(x)$ and $e_2(x)$ are somewhere on the line segment between $e(x)$ and $\tilde{e}(x)$, and he functions $\eta_1(x)$ and $\eta_2(x)$ are somewhere on the line segment between $\eta(x)$ and $\hat{\eta}(x)$, and the final inequality follows from uniform boundedness of f_{ww} and f_w'. We conclude

$$\begin{aligned}
\sup_{e\in\mathcal{E}:\|e-\tilde{e}\|_{2,P_n}\leqslant u} \|(\hat{M}_n(e) - M_n(e)) - (\hat{M}_n(\tilde{e}) - M_n(\tilde{e}))\| &\leqslant \frac{1}{n}\sum_{i=1}^n \|c(X_i; e, \tilde{e}, \eta, \hat{\eta})\| \\
&\leqslant Cp^2\big(2\|\hat{\eta} - \eta\|_{2,P_n}^2 + u\|\hat{\eta} - \eta\|_{2,P_n}\big)
\end{aligned}$$

where the last inequality follows by Cauchy-Schwarz. Recalling $\nabla\Psi(R_n(\tilde{e})) \preccurlyeq \tilde{K}I$ whenever B_n holds, we conclude

$$\begin{aligned}&\sup_{e\in\mathcal{E}:\|e-\tilde{e}\|_{2,P_n}\leqslant u} \nabla\Psi(R_n(\tilde{e}))^\top[(\hat{M}_n(e)-M_n(e))-(\hat{M}_n(\tilde{e})-M_n(\tilde{e}))]\mathbf{1}(B_n)\\ &\leqslant \tilde{K}Cp^2(2\|\hat{\eta}-\eta\|_{2,P_n}^2+u\|\hat{\eta}-\eta\|_{2,P_n})\end{aligned}$$

for all $u \leqslant r/3$. The preceding display and (S55) imply Lemma C.4 in view of the decomposition (S54). □

Continuing with the proof of Step 4, let $C_n = A_n \cap B_n$ and $r = r_0 \wedge \beta > 0$, where A_n and r_0 are as in (S30) with $\hat{e} = \tilde{e}$ and B_n and β are as in (S53). Fix $M > -\infty$ and an arbitrary sequence $a_n \uparrow \infty$. For each $j > M$ with $2^j a_n\alpha_n \leqslant r$, define the "shell" $S_j = \{e \in \mathcal{F}_n : 2^{j-1}a_n\alpha_n < \|e-\tilde{e}\|_{2,P_n} \leqslant 2^j a_n\alpha_n\}$. It follows that for each such j, whenever $e \in S_j$ we have $r_0 \geqslant r \geqslant \|e-\tilde{e}\|_{2,P_n} \geqslant 2^{j-1}a_n\alpha_n$, and so by (S30), we have

$$(\phi_n(e)-\phi_n(\tilde{e}))\mathbf{1}(C_n) \leqslant -c_0\|e-\tilde{e}\|_{2,P_n}^2\mathbf{1}(C_n) \leqslant -c_0 2^{2j-2}a_n^2\alpha_n^2\mathbf{1}(C_n)$$

for all $e \in S_j$. Hence using the definition of $\hat{e}$

$$\begin{aligned}\mathbf{1}(\hat{e}\in S_j)\mathbf{1}(C_n) &\leqslant \mathbf{1}\left(\sup_{e\in S_j}\hat{\phi}_n(e)-\hat{\phi}_n(\tilde{e})\geqslant 0\right)\mathbf{1}(C_n)\\ &\leqslant \mathbf{1}\left(\sup_{e:\|e-\tilde{e}\|_{2,P_n}\leqslant 2^j a_n\alpha_n}(\hat{\phi}_n(e)-\hat{\phi}_n(\tilde{e}))-(\phi_n(e)-\phi_n(\tilde{e}))\geqslant c_0 2^{2j-2}a_n^2\alpha_n^2\right)\mathbf{1}(C_n)\\ &\leqslant \mathbf{1}\big(2^j a_n\alpha_n\|\hat{\eta}-\eta\|_{2,P_n}+\|\hat{\eta}-\eta\|_{2,P_n}^2\geqslant c_0C_0^{-1}2^{2j-2}a_n^2\alpha_n^2\big)\mathbf{1}(C_n)\\ &\leqslant \frac{2^j a_n\alpha_n\|\hat{\eta}-\eta\|_{2,P_n}+\|\hat{\eta}-\eta\|_{2,P_n}^2}{c_0C_0^{-1}2^{2j-2}a_n^2\alpha_n^2}\mathbf{1}(C_n)\\ &= \frac{C_0}{c_0}\left(\frac{\|\hat{\eta}-\eta\|_{2,P_n}}{2^{j-2}a_n\alpha_n}+\frac{\|\hat{\eta}-\eta\|_{2,P_n}^2}{2^{2j-2}a_n^2\alpha_n^2}\right)\mathbf{1}(C_n)\end{aligned}$$

where the last inequality follows from Lemma C.4. Then

$$\begin{aligned}\mathbf{1}\left(\frac{r}{2}\geqslant\|\hat{e}-\tilde{e}\|_{2,P_n}>2^M a_n\alpha_n\right)\mathbf{1}(C_n) &\leqslant \sum_{j>M,2^j a_n\alpha_n\leqslant r}\mathbf{1}(\hat{e}\in S_j)\mathbf{1}(C_n)\\ &\leqslant \frac{C_0}{c_0}\sum_{j=M+1}^{\infty}\left(\frac{\|\hat{\eta}-\eta\|_{2,P_n}}{2^{j-2}a_n\alpha_n}+\frac{\|\hat{\eta}-\eta\|_{2,P_n}^2}{2^{2j-2}a_n^2\alpha_n^2}\right)\\ &= \frac{C_0}{c_0}\left(\frac{\|\hat{\eta}-\eta\|_{2,P_n}}{2^{M-2}a_n\alpha_n}+\frac{4}{3}\frac{\|\hat{\eta}-\eta\|_{2,P_n}^2}{2^{2M}a_n^2\alpha_n^2}\right).\end{aligned}$$

Since $\|\hat{\eta}-\eta\|_{2,P} = o_p(a_n\alpha_n)$, we conclude

$$\Pr(r/2 \geqslant \|\hat{e}-\tilde{e}\|_{2,P_n} > 2^M a_n\alpha_n, C_n) = o(1)$$

for each M. Now, by step 3 we know that $\|\hat{e}-e^*\|_{2,P_n} = o_p(1)$. Applying step 3 again but with $\hat{\eta} = \eta$ shows $\|\tilde{e}-e^*\|_{2,P_n} = O_p(n^{-1/4}) = o_p(1)$. Thus

$$\|\hat{e}-\tilde{e}\|_{2,P_n} \leqslant \|\hat{e}-e^*\|_{2,P_n} + \|e^*-\tilde{e}\|_{2,P_n} = o_p(1)$$

so that $\Pr(\|\hat{e}-\tilde{e}\|_{2,P_n} > r/2) = o(1)$. Since $\Pr(C_n) \to 1$ as $n \to \infty$, we can conclude that

$$\Pr(\|\hat{e}-\tilde{e}\|_{2,P_n} > 2^M a_n \alpha_n) = o(1).$$

With $a_n \uparrow \infty$ arbitrary, by the preceding display and Lemma B.2 we have $\|\hat{e}-\tilde{e}\|_{2,P_n} = O_p(\alpha_n)$ and

$$\|\hat{e}-e^*\|_{2,P_n} \leqslant \|\hat{e}-\tilde{e}\|_{2,P_n} + \|\tilde{e}-e^*\|_{2,P_n} = O_p(n^{-1/4}) + O_p(\alpha_n)$$

as desired.

It remains to show the same convergence rate holds out of sample, i.e., $\|\hat{e}-e^*\|_{2,P} = O_p(n^{-1/4}) + O_p(\alpha_n)$. We do this by showing that $\big|\|\hat{e}-e^*\|_{2,P_n} - \|\hat{e}-e^*\|_{2,P}\big| = O_p(n^{-1/4})$. With $\mathcal{E}_2^-$ as defined in Lemma B.5 in terms of the collection $\mathcal{E}$ of Assumption 5.1, we know that

$$\begin{aligned}\big|\|\hat{e}-e^*\|_{2,P_n}^2 - \|\hat{e}-e^*\|_{2,P}^2\big| &= \Big|\frac{1}{n}\sum_{i=1}^n [(\hat{e}(X_i)-e^*(X_i))^2 - \int_{\mathcal{X}} (\hat{e}(x)-e^*(x))^2 \,\mathrm{d}P(x)]\Big| \\ &\leqslant \sup_{e\in\mathcal{E}_2^-} |(P_n-P)e|.\end{aligned}$$

Furthermore, by Lemma B.3 and Lemma B.5 we know that for some $K_0 < \infty$ we have

$$\begin{aligned}\sup_{e\in\mathcal{E}_2^-} |(P_n-P)f| &\leqslant K_0 n^{-1/2} \int_0^1 \sqrt{\log \mathcal{N}(\epsilon, \mathcal{E}_2^-, L^2(P_n))}\,\mathrm{d}\epsilon \\ &\leqslant K_0\sqrt{2} n^{-1/2} \int_0^1 \sqrt{\log \mathcal{N}(\epsilon/4, \mathcal{E}, L^2(P_n))}\,\mathrm{d}\epsilon \\ &= 4K_0\sqrt{2} n^{-1/2} \int_0^{1/4} \sqrt{\log \mathcal{N}(\delta, \mathcal{E}, L^2(P_n))}\,\mathrm{d}\delta.\end{aligned}$$

Then by (31) we conclude $\sup_{e\in\mathcal{E}_2^-} |(P_n-P)e| = O_p(n^{-1/2})$. Since

$$|a-b| \leqslant a+b \implies \sqrt{|a-b|} \leqslant \sqrt{a+b} \implies |a-b| \leqslant \sqrt{a+b}\sqrt{|a-b|} = \sqrt{|a^2-b^2|}$$

for any $a, b \geqslant 0$, we have

$$\big|\|\hat{e}-e^*\|_{2,P_n} - \|\hat{e}-e^*\|_{2,P}\big| = O_p(n^{-1/4})$$

as desired. □

C.10 Proof of Theorem 5.1

Here we prove convergence of Algorithm 1, our concave maximization procedure for designing an optimal CSBAE. We begin by proving we can design for $\hat{\theta}_{\mathrm{AIPW}}$, as stated in the first numbered condition of Theorem 5.1. The proof proceeds by showing that the objective defining $e^*_{t,\mathrm{AIPW}}(\cdot)$ in (36) can be written in a form so that Assumption 5.1 and (35) are satisfied, the latter with $\alpha_N = N^{-1/4}$. Then we conclude by applying Lemma 5.2.

Many of our expressions will include the cumulative sum of batch frequencies $\sum_{u=1}^t \kappa_u$. We use $\kappa_{1:t}$ to denote this quantity below. We similarly abbreviate $\sum_{u=1}^t N_u$ to $N_{1:t}$.

With $V_{0:t,\text{AIPW}}$ scalar, the information function $\Psi = \Psi(\cdot)$ is simply an increasing scalar-valued function by Assumption 5.2, and hence we have

$$
\begin{aligned}
e^*_{t,\text{AIPW}}(\cdot) = \operatorname*{arg\,max}_{e_t(\cdot)\in\mathcal{F}_{*,t}} (V_{0:t,\text{AIPW}})^{-1} &= \operatorname*{arg\,min}_{e_t(\cdot)\in\mathcal{F}_{*,t}} V_{0:t,\text{AIPW}} \\
&= \operatorname*{arg\,max}_{e_t(\cdot)\in\mathcal{F}_{*,t}} h(V_{0:t,\text{AIPW}}) \\
&= \operatorname*{arg\,max}_{e_t(\cdot)\in\mathcal{F}_{t,*}} \mathbb{E}_{P^X}\left[\frac{2C}{\gamma_0} - \frac{v_0(1,X)}{e_0^{(t)}(X)} - \frac{v_0(0,X)}{1-e_0^{(t)}(X)}\right]
\end{aligned}
$$

where in the final equality we dropped the additive term $\mathbb{E}[(\tau_0(X)-\theta_0)^2]$ which is independent of $e_t(\cdot)$, and defined $h(x) = 2C/\gamma_0 - x$ for

$$
\gamma_0 := \frac{\epsilon_1}{2}\frac{1}{\kappa_{1:t}}\min(\kappa_1,\kappa_t) > 0.
$$

Evidently $h(\cdot)$ is decreasing. We can now define the information matrix

$$
\mathcal{I} = \mathcal{I}(e_t,\eta_0) = \mathbb{E}_{P^X}\left[\frac{2C}{\gamma_0} - \frac{v_0(1,X)}{e_0^{(t)}(X)} - \frac{v_0(0,X)}{1-e_0^{(t)}(X)}\right]
$$

which is of the form (29) with

$$
f(e,w) = \frac{2C}{\gamma_0} - \frac{w_1}{1-w_3-w_4e} - \frac{w_2}{w_3+w_4e}
$$

and

$$
\eta(x) = \left(v_0(0,x),\ v_0(1,x),\ \frac{1}{\kappa_{1:t}}\sum_{u=1}^{t-1}\kappa_u e_u(x),\ \frac{1}{\kappa_{1:t}}\kappa_t\right).
$$

Let $\mathcal{W} = [c,C]^2\times\mathcal{W}_+$ where $\mathcal{W}_+ = \{(x,y)\in\mathbb{R}^2 \mid x\geqslant\gamma_0, y\geqslant\gamma_0, x+y\leqslant 1-\gamma_0\}$. By (3) and the assumptions of the Theorem (specifically the uniform bounds on the variance functions and the assumption that $\epsilon_1 \leqslant e_1(x) = \hat{e}_1^{(k)}(x) \leqslant 1-\epsilon_1$ for all $x\in\mathcal{X}$ and folds $k=1,\dots,K$), we can verify that $\eta(x)\in\mathcal{W}$ and $\hat{\eta}^{(k)}(x)\in\mathcal{W}$ for all $x\in\mathcal{X}$, folds $k=1,\dots,K$, and sufficiently large N, where

$$
\hat{\eta}^{(k)}(x) = \left(\hat{v}^{(k)}(0,x),\ \hat{v}^{(k)}(1,x),\ \frac{1}{N_{1:t}}\sum_{u=1}^{t-1}N_u\hat{e}_u^{(k)}(x),\ \frac{N_t}{N_{1:t}}\right).
$$

Also we have

$$
1-\gamma_0 \geqslant w_3+w_4e \geqslant \gamma_0 \quad \text{and} \quad f(e,w)\geqslant 0, \quad \forall (e,w)\in[0,1]\times\mathcal{W}.
$$

Evidently $\mathcal{W}$ is closed and bounded, hence compact. With w_3+w_4e linear in e, there exists $\delta<0$ such that for all (e,w) in a neighborhood containing $(\delta,1-\delta)\times\mathcal{W}$, w_3+w_4e is uniformly bounded away from 0 and then $f(e,w)$ evidently has continuous second partial derivatives on this neighborhood. Finally, we compute

$$
-f''(e,w) = \frac{2w_2w_4^2}{(w_3+w_4e)^3} + \frac{2w_1w_4^2}{(1-w_3-w_4e)^3} \geqslant \frac{2c\gamma_0^2}{(1-\gamma_0)^3} > 0, \quad \forall (e,w)\in[0,1]\times\mathcal{W}
$$

which shows that all conditions of Assumption 5.1 have been satisfied. Equation (35) holds with $\alpha_N = N^{-1/4}$ by (47) and (18), completing the proof of the first numbered condition of Theorem 5.1, pertaining to design for $\hat{\theta}_{\text{AIPW}}$.

It remains to show the second numbered condition holds. As above, the proof proceeds by showing the objective for $\hat{e}^*_{t,\text{EPL}}(\cdot)$ in (36) can be written in a form so that Assumption 5.1 and (35) are satisfied, the latter with $\alpha_N = N^{-1/4}$. To that end, we take $\mathcal{I} = \mathcal{I}(e_t, \eta_0) = V^{-1}_{0:t,\text{EPL}}$ which takes the form (29) with

$$f(e, w) = f(e, w_1, w_2, w_3, w_4, w_5) = \frac{(w_3 + w_4 e)(1 - w_3 - w_4 e)}{w_1(w_3 + w_4 e) + w_2(1 - w_3 - w_4 e)} w_5 w_5^\top$$

and

$$\eta(x) = \left(v_0(0, x),\, v_0(1, x),\, \frac{1}{\kappa_{1:t}} \sum_{u=1}^{t-1} \kappa_u e_u(x),\, \frac{\kappa_t}{\kappa_{1:t}},\, \psi(x) \right).$$

Let $\mathcal{W} = [c, C]^2 \times \mathcal{W}_+ \times [-C, C]^p \subseteq \mathbb{R}^{4+p}$, where once again $\mathcal{W}_+ = \{(x, y) \in \mathbb{R}^2 \mid x \geqslant \gamma_0, y \geqslant \gamma_0, x + y \leqslant 1 - \gamma_0\}$ for

$$\gamma_0 := \frac{\epsilon_1}{2\kappa_{1:t}} \min(\kappa_1, \kappa_t) > 0.$$

Evidently $\mathcal{W}$ is closed and bounded, hence compact. Then the assumptions of the Theorem (specifically the uniform bounds on the variance functions and the assumption that $\epsilon_1 \leqslant e_1(x) = \hat{e}_1^{(k)}(x) \leqslant 1 - \epsilon_1$ for all $x \in \mathcal{X}$ and folds $k = 1, \ldots, K$) ensure that for each fold $k = 1, \ldots, K$, if we take

$$\hat{\eta}^{(k)}(x) = \left(\hat{v}^{(k)}(0, x),\, \hat{v}^{(k)}(1, x),\, \frac{1}{N_{1:t}} \sum_{u=1}^{t-1} N_u \hat{e}_u^{(k)}(x),\, \frac{N_t}{N_{1:t}},\, \psi(x) \right)$$

then $\eta(x) \in \mathcal{W}$ and $\hat{\eta}^{(k)}(x) \in \mathcal{W}$ for all $x \in \mathcal{X}$ whenever N is sufficiently large. We note that for each $e \in [0, 1]$,

$$f(e, w) = \frac{(w_3 + w_4 e)(1 - w_3 - w_4 e)}{w_1(w_3 + w_4 e) + w_2(1 - w_3 - w_4 e)} w_5 w_5^\top$$

is positive semidefinite because the lead constant above is nonnegative. Furthermore, note that the denominator $w_1(w_3 + w_4 e) + w_2(1 - w_3 - w_4 e)$ is bounded below by c for any $(e, w) \in [0, 1] \times \mathcal{W}$ and continuous on $(e, w_1, w_2, w_3, w_4) \in \mathbb{R}^5$. This denominator is linear in e (for fixed w) and so additionally there exists $\delta < 0$ such that on some open neighborhood containing $(\delta, 1 - \delta) \times \mathcal{W}$, this denominator is strictly positive. Therefore $f(e, w)$ has two continuous partial derivatives with respect to e. Finally, we compute

$$-f''(e, w) = \frac{2w_4^2(w_3 + w_4 e)(1 - w_3 - w_4 e)}{(w_1(w_3 + w_4 e) + w_2(1 - w_3 - w_4 e))^3} w_5 w_5^\top.$$

This is positive semidefinite since as above, $1 - \gamma_0 \geqslant w_3 + w_4 e \geqslant \gamma_0$ for all $(e, w) \in [0, 1] \times \mathcal{W}$, so

$$\frac{2w_4^2(w_3 + w_4 e)(1 - w_3 - w_4 e)}{(w_1(w_3 + w_4 e) + w_2(1 - w_3 - w_4 e))^3} > 0$$

on $[0,1] \times \mathcal{W}$. Furthermore, as all diagonal entries of $w_5 w_5^\top$ are nonnegative, the inclusion of an intercept in $\psi(x)$ ensures that

$$\inf_{(e,w)\in[0,1]\times\mathcal{W}} \operatorname{tr}(-f''(e,w)) \geqslant \inf_{(e,w)\in[0,1]\times\mathcal{W}} \frac{2w_4^2(w_3+w_4e)(1-w_3-w_4e)}{[w_1(w_3+w_4e)+w_2(1-w_3-w_4e)]^3}$$
$$\geqslant \frac{2\gamma_0^3(1-\gamma_0)}{C^3}.$$

As before, equation (35) holds with $\alpha_N = N^{-1/4}$ by (47) and (18), enabling us to apply Lemma 5.2 and completing the proof of the Theorem.

D Additional simulations

We present results from some additional numerical simulations in the framework of Section 6.

D.1 Unequal budget constraints

Tables 1 and 2 reproduce Tables 1 and 2 using simulations with budget constraints $m_{L,2} = m_{H,2} = 0.4$. The results are qualitatively similar to those in the main text. One notable difference is that there seem to be some additional gains to pooling in ATE estimation, both asymptotically and in finite samples. For example, in the homoskedastic DGPs, we see about a 5% asymptotic efficiency gain from pooling in Table 1 when either $d = 1$ or $d = 10$, which translates well to finite sample gains, particularly for the $d = 10$ DGP. By contrast there is no asymptotic gain for the homoskedastic DGPs in Table 1.

D.2 Perfect nuisance estimation

To better isolate the performance effects of our specific choices of nuisance estimation methods in the numerical study of Section 6, in Tables 3 and 4 we reproduce Tables 1 and 2, respectively, but assume all nuisance functions are known exactly at both the design and estimation stages. For ATE estimation (Table 3), the flexible designs are aware of a perfectly constant variance function in the homoskedastic DGPs, which induces them to always learn the (optimal) simple RCT in every simulation. However, for the binned designs in the homoskedastic DGPs and both the binned and flexible designs in the heteroskedastic DGPs, there is some cross-simulation variability in the propensity learned. This stems from variation in the parts of the variance function being sampled due to variation in the covariates across simulations. Consequently, the simulated finite sample efficiency gain ends up being somewhat lower than the asymptotic gain. This suggests that the finite sample efficiency gains from pooling observed in Table 1 are due to improved use of nuisance function estimates by the pooled estimator $\hat{\theta}_{\mathrm{AIPW}}$. One reason we might expect this is that the pooled estimator uses nuisance estimates from observations pooled across both batches of the experiment, while each component of the linearly aggregated estimator only uses nuisance estimates from a single batch.

In Table 4, however, we still see some finite sample efficiency gains from pooling, though the effect is not as large as in Table 2 in the main text. We attribute this to the fact that

Table 1: Same as Table 1, but for simulations with $m_{L,2} = m_{H,2} = 0.4$

DGP	Estimator	Design	Sim. rel. eff. (90% CI)	Asymp. rel. eff.
$d = 1$, Homoskedastic	$\hat{\theta}_{\text{AIPW}}$	Flexible	1.051 (1.015, 1.088)	1.050
	$\hat{\theta}_{\text{AIPW}}$	Binned	1.034 (0.995, 1.073)	1.050
	$\hat{\theta}_{\text{AIPW}}$	Simple RCT	1.050 (1.024, 1.077)	1.050
	$\hat{\theta}^{(\text{LA})}_{\text{AIPW}}$	Flexible	0.993 (0.977, 1.008)	1.000
	$\hat{\theta}^{(\text{bin})}_{\text{AIPW}}$	Binned	0.916 (0.871, 0.962)	0.919
$d = 1$, Heteroskedastic	$\hat{\theta}_{\text{AIPW}}$	Flexible	1.024 (0.978, 1.071)	1.073
	$\hat{\theta}_{\text{AIPW}}$	Binned	1.099 (1.046, 1.154)	1.067
	$\hat{\theta}_{\text{AIPW}}$	Simple RCT	1.043 (1.019, 1.067)	1.034
	$\hat{\theta}^{(\text{LA})}_{\text{AIPW}}$	Flexible	1.032 (1.003, 1.063)	1.012
	$\hat{\theta}^{(\text{bin})}_{\text{AIPW}}$	Binned	1.016 (0.961, 1.071)	0.993
$d = 10$, Homoskedastic	$\hat{\theta}_{\text{AIPW}}$	Flexible	1.084 (1.031, 1.139)	1.050
	$\hat{\theta}_{\text{AIPW}}$	Binned	1.081 (1.030, 1.133)	1.029
	$\hat{\theta}_{\text{AIPW}}$	Simple RCT	1.128 (1.096, 1.161)	1.050
	$\hat{\theta}^{(\text{LA})}_{\text{AIPW}}$	Flexible	0.994 (0.970, 1.018)	1.000
	$\hat{\theta}^{(\text{bin})}_{\text{AIPW}}$	Binned	0.520 (0.478, 0.565)	0.448
$d = 10$, Heteroskedastic	$\hat{\theta}_{\text{AIPW}}$	Flexible	1.051 (0.999, 1.107)	1.064
	$\hat{\theta}_{\text{AIPW}}$	Binned	1.001 (0.952, 1.051)	1.045
	$\hat{\theta}_{\text{AIPW}}$	Simple RCT	1.062 (1.031, 1.094)	1.035
	$\hat{\theta}^{(\text{LA})}_{\text{AIPW}}$	Flexible	0.981 (0.948,1.015)	1.012
	$\hat{\theta}^{(\text{bin})}_{\text{AIPW}}$	Binned	0.626 (0.582, 0.671)	0.611

for estimating $\theta_{0,\text{PL}}$, pooling allows the asymptotic variance to be approached more quickly as a function of the total sample size N. Such an effect does not show up in ATE estimation with AIPW, since for $\hat{\theta}^*_{\text{AIPW}}$ and $\hat{\theta}^*_{\text{EPL}}$ the oracle estimators of Section 3, we can see that $N\text{Var}(\hat{\theta}^*_{\text{AIPW}}) = V_{0,\text{AIPW}}$ exactly for all N while $N\text{Var}(\hat{\theta}^*_{\text{EPL}})$ only approaches $V_{0,\text{EPL}}$ asymptotically as $N \to \infty$. So letting A^*_N be the (finite sample) AMSE of $\hat{\theta}^*_{\text{EPL}}$ computed on a sample of size N, we'd expect $NA^*_N > 2NA^*_{2N}$. Then averaging two independent copies of $\hat{\theta}^*_{\text{EPL}}$ on N observations yields an estimator with AMSE $A^*_N/2$, while pooling would yield an estimator with AMSE A_{2N^*}.

Table 2: Same as Table 2, but for simulations with $m_{L,2} = m_{H,2} = 0.4$

DGP	Estimator	Design	Sim. rel. eff. (90% CI)	Asymp. rel. eff.
$d = 1$, Homoskedastic	$\hat{\theta}_{\text{EPL}}$	Flexible	1.118 (1.076, 1.162)	1.092
	$\hat{\theta}_{\text{EPL}}$	Binned	1.084 (1.038, 1.131)	1.051
	$\hat{\theta}_{\text{EPL}}$	Simple RCT	1.099 (1.072, 1.126)	1.050
	$\hat{\theta}_{\text{EPL}}^{(\text{LA})}$	Flexible	0.995 (0.976, 1.015)	1.006
$d = 1$, Heteroskedastic	$\hat{\theta}_{\text{EPL}}$	Flexible	1.139 (1.035, 1.249)	1.090
	$\hat{\theta}_{\text{EPL}}$	Binned	1.050 (0.905, 1.188)	1.013
	$\hat{\theta}_{\text{EPL}}$	Simple RCT	1.074 (0.980, 1.172)	1.034
	$\hat{\theta}_{\text{EPL}}^{(\text{LA})}$	Flexible	1.028 (0.970, 1.090)	1.017
$d = 10$, Homoskedastic	$\hat{\theta}_{\text{EPL}}$	Flexible	1.234 (1.203, 1.266)	1.056
	$\hat{\theta}_{\text{EPL}}$	Binned	1.204 (1.172, 1.237)	0.994
	$\hat{\theta}_{\text{EPL}}$	Simple RCT	1.262 (1.234, 1.292)	1.050
	$\hat{\theta}_{\text{EPL}}^{(\text{LA})}$	Flexible	0.997 (0.982, 1.013)	1.002
$d = 10$, Heteroskedastic	$\hat{\theta}_{\text{EPL}}$	Flexible	1.104 (1.065, 1.144)	1.069
	$\hat{\theta}_{\text{EPL}}$	Binned	1.055 (1.016, 1.094)	1.043
	$\hat{\theta}_{\text{EPL}}$	Simple RCT	1.078 (1.042, 1.116)	1.035
	$\hat{\theta}_{\text{EPL}}^{(\text{LA})}$	Flexible	0.993 (0.967,1.019)	1.013

Table 3: Same as Table 1, but for oracle simulations that assume knowledge of the true nuisance functions

DGP	Estimator	Design	Sim. rel. eff. (90% CI)	Asymp. rel. eff.
$d = 1$, Homoskedastic	$\hat{\theta}_{\text{AIPW}}$	Flexible	1.000 (1.000, 1.000)	1.000
	$\hat{\theta}_{\text{AIPW}}$	Binned	0.978 (0.960, 0.997)	0.999
	$\hat{\theta}_{\text{AIPW}}$	Simple RCT	1.000 (1.000, 1.000)	1.000
	$\hat{\theta}_{\text{AIPW}}^{(\text{LA})}$	Flexible	1.000 (1.000, 1.000)	1.000
	$\hat{\theta}_{\text{AIPW}}^{(\text{bin})}$	Binned	0.828 (0.794, 0.863)	0.875
$d = 1$, Heteroskedastic	$\hat{\theta}_{\text{AIPW}}$	Flexible	1.005 (0.960, 1.051)	1.051
	$\hat{\theta}_{\text{AIPW}}$	Binned	0.977 (0.935, 1.020)	1.044
	$\hat{\theta}_{\text{AIPW}}$	Simple RCT	1.000 (1.000, 1.000)	1.000
	$\hat{\theta}_{\text{AIPW}}^{(\text{LA})}$	Flexible	1.008 (0.977, 1.039)	1.026
	$\hat{\theta}_{\text{AIPW}}^{(\text{bin})}$	Binned	0.905 (0.859, 0.952)	0.976
$d = 10$, Homoskedastic	$\hat{\theta}_{\text{AIPW}}$	Flexible	1.000 (1.000, 1.000)	1.000
	$\hat{\theta}_{\text{AIPW}}$	Binned	0.968 (0.923, 1.013)	0.965
	$\hat{\theta}_{\text{AIPW}}$	Simple RCT	1.000 (1.000, 1.000)	1.000
	$\hat{\theta}_{\text{AIPW}}^{(\text{LA})}$	Flexible	1.000 (1.000,1.000)	1.000
	$\hat{\theta}_{\text{AIPW}}^{(\text{bin})}$	Binned	0.402 (0.368, 0.439)	0.432
$d = 10$, Heteroskedastic	$\hat{\theta}_{\text{AIPW}}$	Flexible	1.023 (0.986, 1.060)	1.043
	$\hat{\theta}_{\text{AIPW}}$	Binned	1.010 (0.967, 1.054)	1.019
	$\hat{\theta}_{\text{AIPW}}$	Simple RCT	1.000 (1.000, 1.000)	1.000
	$\hat{\theta}_{\text{AIPW}}^{(\text{LA})}$	Flexible	1.020 (0.990,1.051)	1.026
	$\hat{\theta}_{\text{AIPW}}^{(\text{bin})}$	Binned	0.618 (0.575, 0.663)	0.612

Table 4: Same as Table 2, but for oracle simulations that assume knowledge of the true nuisance functions

DGP	Estimator	Design	Sim. rel. eff. (90% CI)	Asymp. rel. eff.
$d=1$, Homoskedastic	$\hat{\theta}_{\mathrm{EPL}}$	Flexible	1.129 (1.080, 1.179)	1.100
	$\hat{\theta}_{\mathrm{EPL}}$	Binned	1.026 (0.984, 1.069)	1.021
	$\hat{\theta}_{\mathrm{EPL}}$	Simple RCT	1.009 (1.003, 1.016)	1.000
	$\hat{\theta}_{\mathrm{EPL}}^{(\mathrm{LA})}$	Flexible	1.073 (1.041, 1.107)	1.056
$d=1$, Heteroskedastic	$\hat{\theta}_{\mathrm{EPL}}$	Flexible	1.126 (1.070, 1.184)	1.128
	$\hat{\theta}_{\mathrm{EPL}}$	Binned	0.966 (0.916, 1.017)	0.969
	$\hat{\theta}_{\mathrm{EPL}}$	Simple RCT	1.008 (1.001, 1.015)	1.000
	$\hat{\theta}_{\mathrm{EPL}}^{(\mathrm{LA})}$	Flexible	1.082 (1.043, 1.123)	1.079
$d=10$, Homoskedastic	$\hat{\theta}_{\mathrm{EPL}}$	Flexible	1.079 (1.068, 1.090)	1.025
	$\hat{\theta}_{\mathrm{EPL}}$	Binned	0.997 (0.979, 1.015)	0.966
	$\hat{\theta}_{\mathrm{EPL}}$	Simple RCT	1.047 (1.041, 1.053)	1.000
	$\hat{\theta}_{\mathrm{EPL}}^{(\mathrm{LA})}$	Flexible	1.021 (1.013, 1.028)	1.019
$d=10$, Heteroskedastic	$\hat{\theta}_{\mathrm{EPL}}$	Flexible	1.132 (1.115, 1.148)	1.075
	$\hat{\theta}_{\mathrm{EPL}}$	Binned	1.082 (1.058, 1.106)	1.054
	$\hat{\theta}_{\mathrm{EPL}}$	Simple RCT	1.059 (1.052, 1.066)	1.000
	$\hat{\theta}_{\mathrm{EPL}}^{(\mathrm{LA})}$	Flexible	1.055 (1.041,1.068)	1.062

References

[1] Dudley RM. The sizes of compact subsets of Hilbert space and continuity of Gaussian processes. Journal of Functional Analysis. 1967;1(3):290-330.

[2] Kitagawa T, Tetenov A. Who should be treated? Empirical welfare maximization methods for treatment choice. Econometrica. 2018;86(2):591-616.

[3] van der Vaart AW, Wellner JA. Weak Convergence and Empirical Processes. Springer Series in Statistics. New York: Springer; 1996.

[4] Chernozhukov V, Chetverikov D, Demirer M, Duflo E, Hansen C, Newey W, et al. Double/debiased machine learning for treatment and structural parameters. The Econometrics Journal. 2018 01;21(1):C1-C68. Available from: `https://doi.org/10.1111/ectj.12097`.

[5] Boyd S, Vandenberghe L. Convex optimization. Cambridge: Cambridge University Press; 2004.

[6] Barbu V, Precupanu T. Convexity and optimization in Banach spaces. 4th ed. Dordrecht: Springer Science & Business Media; 2012.