Electronic Companion for *Heterogeneous Treatment Effects in Panel Data: Insights into the Healthy Incentives Program*

In this electronic companion, we provide detailed proofs of the results from the main paper, along with additional computational findings. We begin by bounding the approximation error in Section EC.1 with the proof of Theorem 1. In Section EC.2, we prove the error decomposition given in Lemma 1. Next, Section EC.3 outlines the steps of the proof of the estimation error bound presented in Theorem 2. Finally, Section EC.4 provides further details and results from our computational experiments.

EC.1. Proofs for the approximation error

Proof of Theorem 1. We define the diameter of a given set C of points as the maximum distance between two points in the set C:

$$\text{Diam}(C) = \sup_{(z_1,t_1),(z_2,t_2)\in C} \|X^{z_1t_1} - X^{z_2t_2}\|.$$

Because the treatment effect function $\mathcal{T}_i^\star$ is Lipschitz continuous for every $i \in [q]$, with Lipschitz constant L_i, it suffices to show that, for any leaf x in $\mathbb{T}$, the set C_x of points within that leaf has a diameter bounded above as follows: $\text{Diam}(C_x) \le \frac{2\sqrt{p}}{\ell^s}$.

In this proof we assume, without loss of generality, that all leaves are at depth exactly $\lfloor c \log \ell \rfloor$. If a leaf were instead even deeper in the tree, the diameter of the associated cluster would be even smaller. We begin by establishing a lower bound for the number of observations n_a in any node a of the tree $\mathbb{T}$. This is done by leveraging the fact that every node is at depth at most $c \log \ell$, the upper bound on the number of leaves ℓ, and the property of α-regularity. Specifically, due to α-regularity, we have $n_a \ge nT \cdot \alpha^{c \log \ell}$. Hence,

$$n_a \ge nT\alpha^{c\log\ell} = nT\ell^{c\log\alpha} \le nT\left(\frac{\alpha}{2c_mM}\right)^{\frac{-1}{c\log\alpha}\cdot c\log\alpha} = \frac{2nT}{\alpha}c_mM.$$

Rearranging, we have the following for any node a of the tree $\mathbb{T}$:

$$c_mM \le \frac{\alpha n_a}{2nT}. \tag{EC.1}$$

Next, for every covariate $j \in [p]$ and any leaf x, we aim to establish an upper bound for the length of C_x along the j-th coordinate, denoted $\text{Diam}_j(C_x)$. This is defined as follows:

$$\text{Diam}_j(C_x) := \sup_{(z_1,t_1),(z_2,t_2)\in C_x} \left|X_j^{z_1t_1} - X_j^{z_2t_2}\right|.$$

Utilizing the lower bound in Assumption 1, the volume V_a of the rectangular hull enclosing the points in a given node a can be upper bounded as follows:

$$V_a \leq \frac{1}{\underline{c}}\left(\frac{n_a}{nT} + c_m M\right) \leq \frac{2}{\underline{c}}\left(\frac{n_a}{nT}\right), \tag{EC.2}$$

where the last inequality made use of (EC.1). Given the α-regularity of splits, we know that a child of node a has at most $n_a(1-\alpha)$ observations. This implies that the volume V_r of the rectangular hull enclosing the points removed by the split on node a satisfies:

$$V_r \geq \frac{1}{\bar{c}}\left(\frac{\alpha n_a}{nT} - c_m M\right) \geq \frac{1}{2\bar{c}}\left(\frac{\alpha n_a}{nT}\right). \tag{EC.3}$$

Putting (EC.2) and (EC.3) together, we have $V_r \geq \left(\frac{\alpha \underline{c}}{4\bar{c}}\right) V_a$, which means that a split along a given covariate removes at least $\frac{\alpha \underline{c}}{4\bar{c}}$ fraction of the length along that coordinate.

Because the tree is a fair-split tree, for every $(\pi+1)p$ splits, at least one split is made on every covariate. Hence, the number of splits leading to node x along each coordinate is at least $\left\lfloor \frac{\lfloor c \log \ell \rfloor}{(\pi+1)p} \right\rfloor \geq \frac{c \log \ell}{(\pi+1)p} - 2$.

As a result, for every $j \in [p]$, we have $\mathrm{Diam}_j(C_x) \leq \left(1 - \frac{\alpha \underline{c}}{4\bar{c}}\right)^{\frac{c \log \ell}{(\pi+1)p} - 2}$, which implies

$$\mathrm{Diam}(C_x) \leq \sqrt{p}\left(1 - \frac{\alpha \underline{c}}{4\bar{c}}\right)^{\frac{c}{(\pi+1)p}\log \ell - 2} \leq \sqrt{p}\left(1 - \frac{\alpha \underline{c}}{4\bar{c}}\right)^{-2} (\ell)^{\frac{c}{(\pi+1)p}\log\left(1 - \frac{\alpha \underline{c}}{4\bar{c}}\right)} \leq \sqrt{p}\left(\frac{3}{4}\right)^{-2} (\ell)^{-\frac{c}{(\pi+1)p}\frac{\alpha \underline{c}}{4\bar{c}}},$$

which proves the desired bound. □

The following lemma is a generalization of the multivariate case of the Dvoretzky–Kiefer–Wolfowitz inequality. Specifically, it provides a bound on the maximum deviation of the empirical fraction of observations within a hyper-rectangle from its expected value. It will be used in the proof of Lemma EC.2, and its proof is provided in the accompanying technical report.

LEMMA EC.1. *Let X be a d-dimensional random vector with pdf f_X. Let $X_1, X_2, \ldots, X_n$ be an i.i.d. sequence of d-dimensional vectors drawn from this distribution. For any $x_1, x_2 \in \mathbb{R}^d$, define*

$$G(x_1, x_2) := \Pr[x_1 \leq X \leq x_2] \qquad \textit{and} \qquad G_n(x_1, x_2) := \frac{1}{n}\sum_{i=1}^{n} I_{\{X_i \in [x_1, x_2]\}}.$$

Then, for any $\epsilon > 0$, we have both of the following bounds:

$$\Pr\left[\sup_{x_1,x_2\in\mathbb{R}^d} |G_n(x_1,x_2) - G(x_1,x_2)| > \epsilon\right] \le 2n^{2d}\exp\left(-2n\epsilon^2 + 16\epsilon d\right),$$
$$\Pr\left[\sup_{x_1,x_2\in\mathbb{R}^d} |G_n(x_1,x_2) - G(x_1,x_2)| > \frac{4d}{n} + \epsilon\right] \le 2n^{2d}\exp\left(-2n\epsilon^2\right).$$

LEMMA EC.2. *For large enough n and T, the covariates $X^{zt} \in \mathbb{R}^p$ for $z \in [n]$ and $t \in [T]$ satisfy Assumption 1 with probability at least $1 - \frac{6}{(nT)^2}$ if, for any nonnegative integer x, y, and z that add up to p, every X^{zt} is the concatenation of vectors $w^{zt} \in \mathbb{R}^x, u^{zt} \in \mathbb{R}^y, v^{zt} \in \mathbb{R}^z$, which are independently generated as follows:*

1. ***Covariates varying by unit and time:*** *Each w^{zt} is i.i.d., sampled from a distribution W on $[0,1]^x$ with density uniformly lower bounded by $\underline{c}^{1/3}$ and upper bounded by $\bar{c}^{1/3}$.*
2. ***Covariates varying only across units:*** *Draw n i.i.d. samples from a distribution U on $[0,1]^y$, bounded in density between $\underline{c}^{1/3}$ and $\bar{c}^{1/3}$. These make up u^{z1} for $z \in [n]$, with arbitrary ordering. For every z and $t \in [T]$, we have $u^{zt} = u^{z1}$.*
3. ***Covariates varying only over time:*** *Draw T i.i.d. samples from a distribution V on $[0,1]^z$ with the same density bounds. These form v^{1t} for $t \in [T]$, with arbitrary ordering. For every t and $z \in [n]$, we have $v^{zt} = v^{1t}$.*

Proof. The proof of the lemma proceeds as follows. We first formally define an event $\mathcal{E}$, which intuitively states that the maximum deviation between the fraction of samples falling within a specified interval and its expected value is bounded. Our second step is to show that the event occurs with high probability, and our final step is to show that under the event $\mathcal{E}$, Assumption 1 is satisfied.

We begin by defining our notation. For any random vector D, we define $\hat{G}^D(a,b)$ as the empirical fraction of observations that fall within a specified interval (a,b), with $a \le b$ having the same dimension as D. The expected value of this quantity is denoted $G^D(a,b) := \Pr[a \le D \le b]$. Note that we have

$$\hat{G}^U(a,b) = \frac{1}{n}\sum_z I_{\{u^{z1}\in[a,b]\}} \quad \hat{G}^V(a,b) = \frac{1}{T}\sum_t I_{\{v^{1t}\in[a,b]\}} \quad \hat{G}^X(a,b) = \frac{1}{nT}\sum_{z,t} I_{\{X^{zt}\in[a,b]\}}.$$

Defining event $\mathcal{E}$. Now we will formally define $\mathcal{E}$, which is the intersection of three events $\mathcal{E} = \mathcal{E}_{\mathcal{A}} \cap \mathcal{E}_{\mathcal{B}} \cap \mathcal{E}_{\mathcal{C}}$. The events $\mathcal{E}_{\mathcal{A}}$ (resp. $\mathcal{E}_{\mathcal{B}}$) is the event that maximum deviation between the fraction of

samples from random variable U (resp. V) falling within an interval and its expected value is bounded. To simplify our notation, we let $\epsilon := \sqrt{\frac{(p+1)\ln(nT)}{\min(n,T)}}$. Mathematically,

$$\mathcal{E}_{\mathcal{A}} := \left\{ \sup_{u^{(1)} \le u^{(2)}} \left| \hat{G}^U(u^{(1)}, u^{(2)}) - G^U(u^{(1)}, u^{(2)}) \right| \le \epsilon + \frac{4y}{n} \right\},$$
$$\mathcal{E}_{\mathcal{B}} := \left\{ \sup_{v^{(1)} \le v^{(2)}} \left| \hat{G}^V(v^{(1)}, v^{(2)}) - G^V(v^{(1)}, v^{(2)}) \right| \le \epsilon + \frac{4z}{T} \right\}.$$

We will now turn to defining event $\mathcal{E}_C$, which intuitively bounds the deviation for the W samples. In the event $\mathcal{E}_C$, rather than considering all of the samples of W, we are only considering the samples of W for which the corresponding U and V samples satisfy given constraints. That is, $\mathcal{E}_C$ is the event that for *any* constraints on U and V, the maximum deviation between the following is bounded: 1) the proportion of samples with U and V satisfying the constraint, that have W fall within a specified interval and 2) the expected fraction of W samples falling within a specified interval. We will define the event $\mathcal{E}_C$ formally below.

To define event $\mathcal{E}_C$, we begin by defining S_U (resp. S_V), which is the set of all possible subsets of units (resp. time periods) selected by some constraint on U (resp. V). Given the random variables u^{z1} for $z \in [n]$, we denote by $\tilde{u}_1, \ldots, \tilde{u}_{n^y}$ all n^y vectors formed by selecting its i-th coordinate (for each $i \in [y]$) from this set of observed values: $\{u_i^{z1}, z \in [n]\}$. We similarly define $\tilde{v}_1, \ldots, \tilde{v}_{T^z}$ using the variables v^{1t} for $t \in [T]$. Next, for any choice of indices $i_1, i_2 \in [n^y]$ such that $\tilde{u}_{i_1} \le \tilde{u}_{i_2}$, we denote by $\mathcal{Z}(i_1, i_2)$ the subset of units that lie in the hyper-rectangle defined by $\tilde{u}_{i_1}$ and $\tilde{u}_{i_2}$, that is

$$\mathcal{Z}(i_1, i_2) := \left\{ z \in [n], \tilde{u}_{i_1} \le u^{z1} \le \tilde{u}_{i_2} \right\}.$$

We define $\mathcal{T}(j_1, j_2)$ similarly, for any $j_1, j_2 \in [T^z]$ with $\tilde{v}_{j_1} \le \tilde{v}_{j_2}$. Finally, $S_U := \{\mathcal{Z}(i_1, i_2) : i_1, i_2 \in [n^y], \tilde{u}_{i_1} \le \tilde{u}_{i_2}\}$ and S_V is similarly defined. We are now ready to define the last event:

$$\mathcal{E}_C := \left\{ \sup_{w^{(1)} \le w^{(2)}} \left| \frac{1}{|\mathcal{Z}||\mathcal{T}|} \sum_{z \in \mathcal{Z}, t \in \mathcal{T}} I_{w^{zt} \in [w^{(1)}, w^{(2)}]} - G^W(w^{(1)}, w^{(2)}) \right| \le \frac{nT\epsilon + 4x}{|\mathcal{Z}||\mathcal{T}|}, \quad \forall \mathcal{Z} \in S_U, \mathcal{T} \in S_V \right\}.$$

Bounding the probability of event $\mathcal{E}$. By Lemma EC.1, we directly have $\Pr(\mathcal{E}_{\mathcal{A}}) \ge 1 - 2n^{2y} e^{-2n\epsilon^2}$ and $\Pr(\mathcal{E}_{\mathcal{B}}) \ge 1 - 2T^{2z} e^{-2T\epsilon^2}$. We now show that $\mathcal{E}_C$ occurs with high probability. Note that conditioned on the realizations of u^{z1} and v^{1t} for $z \in [n]$ and $t \in [T]$, and for any choice of unit and time period subsets $\mathcal{Z} \in S_U$ and $\mathcal{T} \in S_V$, all the random variables w^{zt} for $z \in \mathcal{Z}, t \in \mathcal{T}$ are still i.i.d. according to the distribution W since they are sampled independently from samples of U and V. As a result, we can apply Lemma EC.1 to obtain

$$\sup_{w^{(1)} \le w^{(2)}} \left| \frac{1}{|\mathcal{Z}||\mathcal{T}|} \sum_{z \in \mathcal{Z}, t \in \mathcal{T}} I_{w^{zt} \in [w^{(1)}, w^{(2)}]} - G^W(w^{(1)}, w^{(2)}) \right| \le \frac{nT\epsilon + 4x}{|\mathcal{Z}||\mathcal{T}|},$$

with probability at least $1-2(|\mathcal{Z}||\mathcal{T}|)^{2x}e^{-2(nT\epsilon)^2/(|\mathcal{Z}||\mathcal{T}|)} \geq 1-2(nT)^{2x}e^{-2nT\epsilon^2}$ for any given $\mathcal{Z}$ and $\mathcal{T}$ and for any realization of the random variables u^{z1} for $z \in [n]$ and v^{1t} for $t \in [T]$. Taking the union bound, over all choices of indices $i_1, i_2 \in [n^y]$ and $j_1, j_2 \in [T^z]$ then shows that

$$\Pr(\mathcal{E}_C \mid u^{zt}, v^{zt}; \forall z \in [n], t \in [T]) \geq 1-2n^{2y}T^{2z}(nT)^{2x}e^{-2nT\epsilon^2}.$$

In particular, the same lower bound still holds for $\Pr(\mathcal{E}_C)$ after taking the expectation over the samples from U and V. Putting all the probabilistic bounds together, we obtain

$$\Pr[\mathcal{E}_{\mathcal{A}} \cap \mathcal{E}_{\mathcal{B}} \cap \mathcal{E}_C] \geq 1-6n^{2(x+y)}T^{2(x+z)}e^{-2\min(n,T)\epsilon^2}.$$

Plugging in $\epsilon = \sqrt{\frac{(p+1)\ln(nT)}{\min(n,T)}}$, this event has probability at least $1-\frac{6}{(nT)^2}$.

Assumption 1 is satisfied under event $\mathcal{E}$. We suppose that event $\mathcal{E}$ holds for the rest of this proof. Consider an arbitrary hyper-rectangle in $[0,1]^p$ with lower and upper corners $x^{(1)}$ and $x^{(2)}$, where $x^{(1)} \leq x^{(2)}$. For convenience, we decompose $x^{(1)}$ into three distinct vectors: $w^{(1)} \in \mathbb{R}^x$, $u^{(1)} \in \mathbb{R}^y$, and $v^{(1)} \in \mathbb{R}^z$, such that the concatenation of these three vectors forms $x^{(1)}$. Similarly, $x^{(2)}$ is decomposed into vectors $w^{(2)}$, $u^{(2)}$, and $v^{(2)}$. Our goal is to bound the number of observations that simultaneously satisfy the constraints:

$$w^{(1)} \leq w^{zt} \leq w^{(2)} \qquad u^{(1)} \leq u^{zt} \leq u^{(2)} \qquad v^{(1)} \leq v^{zt} \leq v^{(2)}.$$

First, the set $\mathcal{Z}$ of units that satisfies the constraint $u^{(1)} \leq u^{z1} \leq u^{(2)}$ is included within the considered sets S_U. Similarly, the set $\mathcal{T}$ of time periods satisfying the constraint $v^{(1)} \leq v^{1t} \leq v^{(2)}$ is also in S_V. As a result, because $\mathcal{E}_C$ is satisfied, we have

$$\left|\frac{nT}{|\mathcal{Z}||\mathcal{T}|}\hat{G}^X(x^{(1)}x^{(2)}) - G^W(w^{(1)}, w^{(2)})\right| \leq \frac{nT\epsilon + 4x}{|\mathcal{Z}||\mathcal{T}|}.$$

Multiplying both sides by $\frac{|\mathcal{Z}||\mathcal{T}|}{nT}$, we have

$$\left|\hat{G}^X(x^{(1)}, x^{(2)}) - \frac{|\mathcal{Z}||\mathcal{T}|}{nT}G^W(w^{(1)}, w^{(2)})\right| \leq \epsilon + \frac{4x}{nT}. \tag{EC.4}$$

We then use the events $\mathcal{E}_{\mathcal{A}}$ and $\mathcal{E}_{\mathcal{B}}$ to bound the number of units and times satisfying their respective constraints:

$$\left|\frac{|\mathcal{Z}|}{n} - G^U(u^{(1)}, u^{(2)})\right| \leq \epsilon + \frac{4y}{n}, \qquad \left|\frac{|\mathcal{T}|}{T} - G^V(v^{(1)}, v^{(2)})\right| \leq \epsilon + \frac{4z}{T}. \tag{EC.5}$$

Now let $P := G^U(u^{(1)}, u^{(2)})G^V(v^{(1)}, v^{(2)})G^W(w^{(1)}, w^{(2)})$. Provided that n and T are large enough such that $\frac{4x}{nT}, \frac{4y}{n}, \frac{4z}{T} \leq \epsilon \leq \frac{1}{4}$, inequalities (EC.4) and (EC.5) imply

$$\left|\hat{G}^X(x^{(1)}, x^{(2)}) - P\right| \leq 7\epsilon. \tag{EC.6}$$

Because of the assumed density bounds on U, V, and W, we have

$$\underline{c}\prod_{i=1}^{p}\left(x_i^{(2)}-x_i^{(1)}\right)\le P\le \bar{c}\prod_{i=1}^{p}\left(x_i^{(2)}-x_i^{(1)}\right).$$

Combining this with (EC.6) completes the proof. □

EC.2. Proof of Lemma 1

LEMMA EC.3. *Define the matrix subspace* $\mathbf{T}=\left\{UA^\top+BV^\top+a\mathbf{1}^\top \mid A\in\mathbb{R}^{T\times r}, B\in\mathbb{R}^{n\times r}, a\in\mathbb{R}^n\right\}$.

The projection onto its orthogonal space is

$$P_{\mathbf{T}^\perp}(X)=(I-UU^\top)X\left(I-\frac{rr^\top}{\|r\|^2}\right)(I-VV^\top),\quad \textit{where } r=(I-VV^\top)\mathbf{1}$$

Proof. For a given matrix X, we define $P:=(I-UU^\top)X\left(I-\frac{rr^\top}{\|r\|^2}\right)(I-VV^\top)$. To show that $P=P_{\mathbf{T}^\perp}(X)$, it suffices to show that $P\in\mathbf{T}^\perp$ and $X-P\in\mathbf{T}$.

First, note that $U^\top P=0$ and $PV=0$ because of the construction of P. This implies that for any matrices $A\in\mathbb{R}^{T\times r}$ and $B\in\mathbb{R}^{n\times r}$, we have $\langle P,UA^\top\rangle=\langle P,BV^\top\rangle=0$. Next, we note that

$$P\mathbf{1}=(I-UU^\top)X\left(I-\frac{rr^\top}{\|r\|^2}\right)r=0.$$

As a result, we also have $\langle P,a\mathbf{1}^\top\rangle=0$ for all $a\in\mathbb{R}^n$. This ends the proof that $P\in\mathbf{T}^\perp$.

Next, we compute

$$\begin{aligned}X-P&=\underbrace{U\,U^\top X\left(I-\frac{rr^\top}{\|r\|^2}\right)(I-VV^\top)}_{A'^\top}+\underbrace{X\left(I-\frac{rr^\top}{\|r\|^2}\right)V}_{B'}V^\top+X\frac{rr^\top}{\|r\|^2}\\&=UA'^\top+\left(B'-X\frac{r\mathbf{1}^\top V}{\|r\|^2}\right)V^\top+X\frac{r}{\|r\|^2}\mathbf{1}^\top\end{aligned}$$

Hence, $X-P\in\mathbf{T}$, which concludes the proof of the lemma. □

Proof of Lemma 1.. It is known Cai et al. (2010) that the set of subgradients of the nuclear norm of an arbitrary matrix $X\in\mathbb{R}^{n_1\times n_2}$ with SVD $X=U\Sigma V^\top$ is $\partial\|X\|_\star=\{UV^\top+W: W\in\mathbb{R}^{n_1\times n_2}, U^\top W=0, WV=0, \|W\|\le 1\}$.

Thus, the first-order optimality conditions of (3) are

$$\left\langle Z_j, O - \hat{M} - \hat{m}\mathbf{1}^\top - \sum_{i=1}^{k} \hat{\tau}_i Z_i \right\rangle = 0 \qquad \text{for } j = 1, 2, \ldots, k \tag{EC.7a}$$

$$O - \hat{M} - \hat{m}\mathbf{1}^\top - \sum_{i=1}^{k} \hat{\tau}_i Z_i = \lambda \left(\hat{U}\hat{V}^\top + W \right) \tag{EC.7b}$$

$$\hat{U}^\top W = 0 \tag{EC.7c}$$

$$W\hat{V} = 0 \tag{EC.7d}$$

$$\|W\| \leq 1 \tag{EC.7e}$$

$$\hat{m} = \frac{1}{T}\left(O - \hat{M} - \sum_{i=1}^{k} \hat{\tau}_i Z_i \right)\mathbf{1}. \tag{EC.7f}$$

Combining equations (EC.7a) and (EC.7b), we have

$$\left\langle Z_j, \lambda \left(\hat{U}\hat{V}^\top + W \right) \right\rangle = 0 \qquad \text{for } j = 1, 2, \ldots, k \tag{EC.8}$$

By the definition of our model, we have $O = M^\star + \sum_{i=1}^{k} \tau_i^\star Z_i + \hat{E}$. We plug this into equation (EC.7b):

$$M^\star - \hat{M} - \hat{m}\mathbf{1}^\top + \sum_{i=1}^{k} \left(\tau_i^\star - \hat{\tau}_i \right) Z_i + \hat{E} = \lambda \left(\hat{U}\hat{V}^\top + W \right).$$

We apply $P_{\hat{\mathbf{1}}^\perp}(\cdot)$ to both sides, where the formula for the projection is given in Lemma EC.3:

$$\underbrace{P_{\hat{\mathbf{1}}^\perp}(M^\star + \hat{E}) + \sum_{i=1}^{k} \left(\tau_i^\star - \hat{\tau}_i \right) P_{\hat{\mathbf{1}}^\perp}(Z_i)}_{B} = \lambda P_{\hat{\mathbf{1}}^\perp}(W)$$

$$\begin{aligned}
&= \lambda (I - \hat{U}\hat{U}^\top) W \left(I - \frac{(I - \hat{V}\hat{V}^\top)\mathbf{1}\mathbf{1}^\top (I - \hat{V}\hat{V}^\top)}{\left\| (I - \hat{V}\hat{V}^\top)\mathbf{1} \right\|^2} \right) (I - \hat{V}\hat{V}^\top) \\
&= \lambda W \left(I - \frac{\mathbf{1}\mathbf{1}^\top (I - \hat{V}\hat{V}^\top)}{\left\| (I - \hat{V}\hat{V}^\top)\mathbf{1} \right\|^2} \right) \\
&= \lambda W - \lambda W \left(\frac{\mathbf{1}\mathbf{1}^\top}{T} \right).
\end{aligned}$$

The last line makes use of the following claim:

CLAIM EC.1. $\hat{V}^\top \mathbf{1} = 0$.

To compute $\lambda W\left(\frac{\mathbf{1}\mathbf{1}^\top}{T}\right)$, we right-multiply (EC.7b) by $\left(\frac{\mathbf{1}\mathbf{1}^\top}{T}\right)$. We note that by (EC.7f), we have $\left(O - \hat{M} - \hat{m}\mathbf{1}^\top - \sum_{i=1}^k \hat{\tau}_i Z_i\right)\mathbf{1} = 0$. Hence, we find $\lambda W\left(\frac{\mathbf{1}\mathbf{1}^\top}{T}\right) = -\lambda \hat{U}\hat{V}^\top\left(\frac{\mathbf{1}\mathbf{1}^\top}{T}\right) \overset{(i)}{=} 0$, where (i) is due to Claim EC.1. As a result, we can conclude that $B = \lambda W$.

Substituting $\lambda W = B$ into (EC.8), we obtain for all $j = 1, 2, \ldots, k$:

$$\left\langle Z_j, \lambda \hat{U}\hat{V}^\top + P_{\hat{\mathbf{1}}^\perp}(M^\star + \hat{E}) + \sum_{i=1}^k \left(\tau_i^\star - \hat{\tau}_i\right) P_{\hat{\mathbf{1}}^\perp}(Z_i) \right\rangle = 0 \tag{EC.9}$$

Rearranging, we have

$$\left[\left\langle P_{\hat{\mathbf{1}}^\perp}(Z_i), P_{\hat{\mathbf{1}}^\perp}(Z_j)\right\rangle\right]_{i,j\in[k]} \left(\hat{\tau} - \tau^\star\right) = \left[\left\langle Z_i, \lambda \hat{U}\hat{V}^\top + P_{\hat{\mathbf{1}}^\perp}(M^\star + \hat{E})\right\rangle\right]_{i\in[k]}$$

This is equivalent to the error decomposition in the statement of the lemma, completing the proof. □

Proof of Claim EC.1. Suppose for contradiction that $\hat{V}^\top \mathbf{1} \neq 0$. Then, there is a solution that for which the objective is lower, contradicting the optimality of $\hat{M}, \hat{\tau}, \hat{m}$. That solution can be constructed as follows. Define V' to be the matrix resulting from projecting each column of $\hat{V}$ onto the space orthogonal to the vector $\mathbf{1}$. That is, the i-th column of V' is $V'_i = \hat{V}_i - \mathbf{1}\left(\frac{\hat{V}_i^\top \mathbf{1}}{\mathbf{1}^\top \mathbf{1}}\right)$. Let m' be the vector such that $m'\mathbf{1}^\top = \hat{U}\hat{\Sigma}\left(\hat{V} - V'\right)^\top + \hat{m}\mathbf{1}^\top$. The solution $\left(\hat{U}\hat{\Sigma}V'^\top, \hat{\tau}, m'\right)$ achieves lower objective value because the value of the expression in the Frobenius norm remains unchanged, but $\left\|\hat{U}\hat{\Sigma}V'^\top\right\|_\star < \left\|\hat{U}\hat{\Sigma}\hat{V}^\top\right\|_\star = \left\|\hat{M}\right\|_\star$. The inequality is due to the fact that matrix V' is still an orthogonal matrix, but its columns have norm that is at most 1, and at least one of its columns has norm strictly less than 1. □

EC.3. Proof of Theorem 2

In this section, we outline the key steps of the proof for Theorem 2. Due to space constraints, the complete proof is detailed in a separate technical report.

We aim to bound $\left|\tau_i^d - \tilde{\tau}_i^\star\right|$. Throughout this section, we assume that the assumptions made in the statement of Theorem 2 are satisfied. To simplify notation, we define $\eta_i := \|\tilde{Z}_i\|_{\mathrm{F}}$. We have

$$\left|\tau_i^d - \tilde{\tau}_i^\star\right| \leq \left\|\tau^d - \tilde{\tau}^\star\right\| = \frac{1}{\eta_i}\left\|\eta_i \tau^d - \tau^\star\right\| \overset{(i)}{=} \frac{1}{\eta_i}\left\|D^{-1}(\Delta^2 + \Delta^3)\right\| \leq \frac{1}{\sigma_{\min}(D)\eta_i}\left(\left\|\Delta^2\right\| + \left\|\Delta^3\right\|\right), \tag{EC.10}$$

where (i) is due to the error decomposition in Lemma 1 and the definition of τ^d.

Hence, we aim to upper bound $\left\|\Delta^2\right\|$ and $\left\|\Delta^3\right\|$ and lower bound $\sigma_{\min}(D)$. The main challenge lies in upper bounding $\left\|\Delta^3\right\|$. In fact, the desired bounds for $\sigma_{\min}(D)$ and $\left\|\Delta^2\right\|$ are derived as part of the process for bounding $\left\|\Delta^3\right\|$. We now discuss the strategy for bounding $\left\|\Delta^3\right\|$. In order to control $P_{\hat{\mathbf{T}}^\perp}(M^\star)$, we aim to show that the true counterfactual matrix $M^\star$ has tangent space close to that of $\hat{M}$. Because we introduced $\hat{m}$ in the convex formulation, which centers the rows of $\hat{M}$ without penalization from the nuclear norm term, we will focus on the projection of $M^\star$ that has rows with mean zero, $P_{\mathbf{1}^\perp}(M^\star)$. Recall that $P_{\mathbf{1}^\perp}(M^\star) = U^\star\Sigma^\star V^{\star\top}$ is its SVD, and define $X^\star := U^\star\Sigma^{\star 1/2}$ and $Y^\star := V^\star\Sigma^{\star 1/2}$. We similarly define these quantities for $\hat{M}$: $\hat{X} := \hat{U}\hat{\Sigma}^{1/2}$ and $\hat{Y} := \hat{V}\hat{\Sigma}^{1/2}$. We aim to show that we have $(\hat{X},\hat{Y}) \approx (X^\star, Y^\star)$. This is stated formally in Lemma EC.4, which requires introducing a few additional notations first.

Let $g(M,\tau,m)$ denote the convex function that we are optimizing in (3). Recall that $(\hat{M},\hat{\tau},\hat{m})$ is a global optimum of function g. We define a non-convex proxy function f, which is similar to g, except M is split into two variables and expressed as $XY^\top$.

$$f(X,Y) = \min_{m\in\mathbb{R}^n,\tau\in\mathbb{R}^k} \frac{1}{2}\left\|O - XY^\top - m\mathbf{1}^\top - \sum_{i=1}^k \tau_i Z_i\right\|_{\mathrm{F}}^2 + \frac{1}{2}\lambda\langle X,X\rangle + \frac{1}{2}\lambda\langle Y,Y\rangle.$$

Our analysis will relate $(\hat{X},\hat{Y})$ to a specific local optimum of f, which we will show is close to $(X^\star,Y^\star)$. The local optimum of f considered is the limit of the gradient flow of f, initiated at $(X^\star,Y^\star)$, formally defined by the following differential equation:

$$\begin{cases} (\dot{X}^t,\dot{Y}^t) = -\nabla f(X^t,Y^t) \\ (X^0,Y^0) = (X^\star,Y^\star). \end{cases} \tag{EC.11}$$

We define $H_{X,Y}$ as the rotation matrix that optimally aligns (X,Y) with $(X^\star,Y^\star)$. That is, letting $O^{r\times r}$ denote the set of $r\times r$ orthogonal matrices, $H_{X,Y}$ is formally defined as follows:

$$H_{X,Y} := \operatorname*{arg\,min}_{R\in O^{r\times r}} \left\|XR - X^\star\right\|_{\mathrm{F}}^2 + \left\|YR - Y^\star\right\|_{\mathrm{F}}^2.$$

Define $F^\star$ to be the vertical concatenation of matrices $X^\star$ and $Y^\star$. That is, $F^\star = [(X^\star)^\top, (Y^\star)^\top]^\top$. We are now ready to present Lemma EC.4.

LEMMA EC.4. *For sufficiently large n, the gradient flow of f, starting from* $(X^\star,Y^\star)$*, converges to* (X,Y) *such that* $X = \hat{X}R$ *and* $Y = \hat{Y}R$ *for some rotation matrix* $R\in O^{r\times r}$*. Furthermore,* $(X,Y)\in\mathcal{B}$*, where*

$$\mathcal{B} := \left\{(X,Y) \mid \left\|XH_{X,Y} - X^\star\right\|_{\mathrm{F}}^2 + \left\|YH_{X,Y} - Y^\star\right\|_{\mathrm{F}}^2 \le \rho^2\right\}, \quad \rho := \frac{\sigma\sqrt{nr}\log^6(n)}{\sigma_{\min}}\left\|F^\star\right\|_{\mathrm{F}}.$$

Let us derive an upper bound on ρ. Because $\|X^\star\|_{\mathrm{F}} \le \|X^\star\|\sqrt{r} = \sqrt{\sigma_{\max} r}$ and $\|Y^\star\|_{\mathrm{F}} \le \sqrt{\sigma_{\max} r}$, we can conclude that $\|F^\star\|_{\mathrm{F}} \lesssim \sqrt{\sigma_{\max} r}$. Using this fact and the assumption $\frac{\sigma\sqrt{n}}{\sigma_{\min}} \lesssim \frac{1}{\kappa^2 r^2 \log^{12.5}(n)}$ provided in the theorem, we can upper bound ρ as follows: $\rho \lesssim \frac{\sqrt{\sigma_{\min}}}{\log^{6.5}(n)\kappa}$.

With Lemma EC.4 and the bound on ρ, we can complete the bound on $\left\|\Delta^3\right\|$ as follows.

$$\left\|\Delta^3\right\| \le \sqrt{k}\left\|\Delta^3\right\|_\infty \le \sqrt{k}\left\|P_{\hat{\mathbf{T}}^\perp}(M^\star)\right\|_{\mathrm{F}} = \sqrt{k}\left\|(I-\hat{U}\hat{U}^\top)(X^\star Y^{\star\top} + m^\star \mathbf{1}^\top)\left(I - \frac{\hat{r}\hat{r}^\top}{\hat{r}^\top\hat{r}}\right)(I-\hat{V}\hat{V}^\top)\right\|_{\mathrm{F}},$$

where $\hat{r} = (I-\hat{V}\hat{V}^\top)\mathbf{1}$, and the last line comes from the closed form of the projection derived in Lemma EC.3. We can use the fact that $\hat{V}^\top\mathbf{1} = 0$ from Claim EC.1 to simplify the expression. Note that with $\hat{V}^\top\mathbf{1} = 0$, $\hat{r}$ simply evaluates to $\mathbf{1}$.

$$\sqrt{k}\left\|P_{\hat{\mathbf{T}}^\perp}(M^\star)\right\|_{\mathrm{F}} = \sqrt{k}\left\|(I-\hat{U}\hat{U}^\top)X^\star Y^{\star\top}\left(I - \frac{\mathbf{1}\mathbf{1}^\top}{T}\right)(I-\hat{V}\hat{V}^\top)\right\|_{\mathrm{F}} = \sqrt{k}\left\|(I-\hat{U}\hat{U}^\top)X^\star Y^{\star\top}(I-\hat{V}\hat{V}^\top)\right\|_{\mathrm{F}},$$

where the last step is due to the fact that $X^\star Y^{\star\top}\mathbf{1} = 0$ because $X^\star Y^{\star\top} = P_{\mathbf{1}^\perp}(M^\star)$.

Finally, let (X, Y) be the limit of the gradient flow of f starting from $(X^\star, Y^\star)$. By Lemma EC.4, we have $X = \hat{X}R$ and $Y = \hat{Y}R$ for some rotation matrix $R \in O^{r\times r}$. This, combined with the definition of $\hat{\mathbf{T}}$, implies $P_{\hat{\mathbf{T}}^\perp}(XA^\top) = P_{\hat{\mathbf{T}}^\perp}(BY^\top) = 0$ for any $A \in \mathbb{R}^{T\times r}$ and $B \in \mathbb{R}^{n\times r}$. Hence,

$$\begin{aligned}
\left\|\Delta^3\right\| &\le \sqrt{k}\left\|(I-\hat{U}\hat{U}^\top)(XH_{X,Y} - X^\star)(YH_{X,Y} - Y^\star)^\top(I-\hat{V}\hat{V}^\top)\right\|_{\mathrm{F}} \\
&\le \sqrt{k}\left\|XH_{X,Y} - X^\star\right\|_{\mathrm{F}}\left\|YH_{X,Y} - Y^\star\right\|_{\mathrm{F}} \\
&\lesssim \sqrt{k}\rho^2 \\
&\lesssim \frac{\sigma^2 r^2 \kappa n \log^{12.5}(n)}{\sigma_{\min}},
\end{aligned} \tag{EC.12}$$

where the last step is due to the fact that $k = O(\log n)$.

EC.3.1. Discussion of Error Assumption

We now argue that the assumptions on the noise matrix E from Assumption 3 are mild. The following lemma shows that under standard sub-Gaussianity assumptions, these are satisfied with very high probability.

LEMMA EC.5. *Suppose that the entries of E are independent, zero-mean, sub-Gaussian random variables, and the sub-Gaussian norm of each entry is bounded by σ, and E is independent from Z_i for $i \in [k]$. Then, for n sufficiently large, with probability at least $1 - e^{-n}$, we have*

$$\|E\| \lesssim \sigma\sqrt{n} \quad \textit{and} \quad |\langle Z_i, E\rangle| \lesssim \sigma\sqrt{n}, \quad \forall i \in [k].$$

Proof of Lemma EC.5. We start by proving that with probability at least $1-2e^{-n}$, we have $\|E\| \lesssim \sigma\sqrt{n}$. We use the following result bounding the norm of matrices with sub-Gaussian entries.

THEOREM EC.1 **(Theorem 4.4.5, Vershynin (2018))**. *Let A be an $m \times n$ random matrix whose entries A_{ij} are independent, mean zero, sub-Gaussian random variables with sub-Gaussian norm bounded by σ. Then, for any $t > 0$, we have $\|A\| \lesssim \sigma(\sqrt{m}+\sqrt{n}+t)$ with probability at least $1-2\exp(-t^2)$.*

As a result of the above theorem, recalling that E is a $n \times T$ matrix with $T \lesssim n$ and using $t = \sqrt{2n}$, we have $\|E\| \lesssim \sigma\sqrt{n}$ with probability at least $1-2\exp(-2n)$.

We next turn to the second claim. The general version of Hoeffding's inequality states that: for zero-mean independent sub-Gaussian random variables $X_1, \ldots, X_n$, we have

$$\Pr\left[\left|\sum_{i=1}^{n} X_i\right| \geq t\right] \leq 2\exp\left(-\frac{ct^2}{\sum_{i=1}^{n}\|X_i\|_{\psi_2}^2}\right),$$

for some absolute constant $c > 0$. Hence, for any $i \in [k]$, we have

$$\Pr\left[|\langle Z_i, E\rangle| \geq \sigma\sqrt{2n/c}\right] \leq 2\exp\left(-\frac{2\sigma^2 n}{\sigma^2\|Z_i\|_{\mathrm{F}}^2}\right) \leq 2e^{-2n}.$$

Applying the Union Bound over all $i \in [k]$:

$$\Pr\left[\max_{i\in[k]}|\langle Z_i, E\rangle| \geq \sigma\sqrt{n/c}\right] \leq 2ke^{-2n} \lesssim 2\log(n)e^{-2n}.$$

Applying the Union Bound shows that the two desired claims hold with probability at least $1-2e^{-2n}(1+\log n) \geq 1-e^{-n}$ for sufficiently large n. □

EC.4. Details of Computational Experiments

In this section, we present more details and results for Section 4.

Compute resources: Our experiments utilized a high-performance computing cluster, leveraging CPU-based nodes for all experimental runs. Our home directory within the cluster provides 10 TB of storage. Experiments were conducted on nodes equipped with Intel Xeon CPUs, configured with 5 CPUs and 20 GB of RAM. We utilized approximately 40 nodes operating in parallel to conduct 200 iterations of experiments for each parameter set. While a single experiment completes within minutes on a personal laptop, executing the full suite of experiments on the cluster required approximately one day.

Method implementations: Similar to the methodology outlined in the supplemental materials of Farias et al. (2021), we implemented PaCE using an alternating minimization technique to solve the convex optimization problem. The hyperparameter λ was initialized to a large value and gradually reduced until $\hat{M}$ reached the fixed rank $r = 6$. In all of our experiments, we fixed r to 6. We did not tune r, and performance may improve with a different choice.

Notes about the data: We imputed missing census data for 2020 by averaging values from 2019 and 2021.

EC.4.1. Additional computational results

Table EC.1 and Table EC.2 present the complete results for the average nMAE of each method, along with the corresponding standard deviations. Table EC.1 reports the nMAE values for estimating HTEs across all observations. In contrast, Table EC.2 focuses on the treated observations only. Note that many methods like MC-NNM only appear in Table EC.2, as they do not directly give estimates of HTEs for untreated observations.

In the following tables, the "All" column presents the result for all instances across all parameter settings. These settings include treatment proportions (0.05, 0.25, 0.5, 0.75, and 1.0), adaptive or non-adaptive approaches, and additive or multiplicative treatment effects. Each parameter set is tested in 200 instances, totaling 4000 instances reflected in each value within the "All" column. Subsequent columns detail outcomes when one parameter is held constant. Specifically, the "Adaptive" and "Effect" columns each summarize results from 2000 instances, while each result under the "Proportion treated" column corresponds to 800 instances. The standard deviations, shown in parentheses within Table EC.1 and Table EC.2, are calculated for each result by analyzing the variability among the instances that contribute to that particular data point.

Table EC.1: Average nMAE across methods. Standard deviations in parentheses.

		Adaptive		Proportion treated					Effect	
	All	Y	N	0.05	0.25	0.50	0.75	1.0	Add.	Mult.
SNAP										
PaCE	0.13	0.14	0.12	0.32	0.18	0.13	0.1	0.09	0.08	0.18
	(0.1)	(0.07)	(0.11)	(0.2)	(0.11)	(0.07)	(0.06)	(0.05)	(0.05)	(0.11)
Causal Forest	0.17	0.19	0.14	0.33	0.21	0.18	0.14	0.12	0.15	0.19
	(0.08)	(0.04)	(0.1)	(0.15)	(0.06)	(0.04)	(0.05)	(0.07)	(0.06)	(0.09)
LinearDML	0.25	0.26	0.23	0.34	0.27	0.27	0.22	0.21	0.14	0.36
	(0.14)	(0.13)	(0.15)	(0.2)	(0.13)	(0.13)	(0.14)	(0.15)	(0.06)	(0.11)
DML	0.33	0.28	0.37	0.57	0.39	0.34	0.29	0.26	0.2	0.45
	(0.18)	(0.18)	(0.16)	(0.25)	(0.17)	(0.15)	(0.16)	(0.17)	(0.12)	(0.13)
XLearner	0.44	0.43	0.45	0.62	0.58	0.49	0.38	0.28	0.43	0.45
	(0.13)	(0.13)	(0.13)	(0.11)	(0.07)	(0.05)	(0.02)	(0.02)	(0.12)	(0.13)
SLearner	0.5	0.56	0.44	0.87	0.72	0.54	0.41	0.28	0.49	0.52
	(0.2)	(0.14)	(0.24)	(0.06)	(0.09)	(0.09)	(0.11)	(0.12)	(0.2)	(0.2)
ForestDRLearner	0.71	0.81	0.61	1.06	0.96	0.74	0.6	0.49	0.7	0.72
	(0.25)	(0.12)	(0.3)	(0.15)	(0.1)	(0.12)	(0.18)	(0.22)	(0.24)	(0.25)
TLearner	1.1	1.11	1.09	3.01	1.72	1.24	0.78	0.41	1.08	1.13
	(0.64)	(0.47)	(0.77)	(0.99)	(0.13)	(0.09)	(0.11)	(0.14)	(0.63)	(0.65)
DomainAdaptationLearner	1.11	1.1	1.11	2.95	1.71	1.26	0.79	0.4	1.09	1.13
	(0.63)	(0.46)	(0.77)	(1.06)	(0.13)	(0.1)	(0.09)	(0.14)	(0.63)	(0.64)
CausalForestDML	5.76	0.67	10.85	95.2	9.27	1.05	0.29	0.18	5.78	5.75
	(23.98)	(1.03)	(33.14)	(91.04)	(8.91)	(0.63)	(0.05)	(0.06)	(24.38)	(23.6)
State										
PaCE	0.27	0.28	0.25	0.53	0.37	0.24	0.16	0.13	0.21	0.33
	(0.19)	(0.19)	(0.19)	(0.27)	(0.13)	(0.08)	(0.1)	(0.05)	(0.13)	(0.22)
Causal Forest	0.32	0.34	0.29	0.56	0.37	0.31	0.23	0.2	0.27	0.36
	(0.16)	(0.15)	(0.16)	(0.22)	(0.07)	(0.09)	(0.05)	(0.07)	(0.12)	(0.18)
XLearner	0.32	0.41	0.23	0.8	0.3	0.25	0.21	0.2	0.29	0.36
	(0.25)	(0.29)	(0.16)	(0.34)	(0.08)	(0.11)	(0.08)	(0.06)	(0.2)	(0.29)
DML	0.33	0.42	0.23	0.53	0.29	0.29	0.37	0.21	0.28	0.37
	(0.23)	(0.24)	(0.16)	(0.18)	(0.23)	(0.18)	(0.29)	(0.06)	(0.23)	(0.21)

Continued on next page

Table EC.1: Average nMAE across methods. Standard deviations in parentheses (continued).

		Adaptive		Proportion treated					Effect	
	All	Y	N	0.05	0.25	0.50	0.75	1.0	Add.	Mult.
CausalForestDML	0.33	0.36	0.3	0.68	0.41	0.28	0.21	0.19	0.27	0.39
	(0.22)	(0.18)	(0.24)	(0.23)	(0.15)	(0.16)	(0.08)	(0.07)	(0.15)	(0.25)
ForestDRLearner	0.49	0.47	0.51	1.12	0.51	0.38	0.38	0.25	0.46	0.52
	(0.33)	(0.2)	(0.42)	(0.42)	(0.14)	(0.13)	(0.1)	(0.07)	(0.32)	(0.34)
LinearDML	0.49	0.67	0.32	1.26	0.36	0.5	0.35	0.24	0.45	0.54
	(0.45)	(0.56)	(0.14)	(0.74)	(0.16)	(0.17)	(0.04)	(0.08)	(0.43)	(0.46)
SLearner	0.62	0.67	0.56	0.83	0.75	0.65	0.48	0.44	0.6	0.64
	(0.18)	(0.13)	(0.2)	(0.11)	(0.04)	(0.09)	(0.11)	(0.1)	(0.18)	(0.17)
NNEstimator	0.63	0.55	0.71	0.84	0.88	0.4	0.5	0.61	0.62	0.65
	(0.33)	(0.2)	(0.41)	(0.49)	(0.35)	(0.12)	(0.15)	(0.17)	(0.39)	(0.26)
DomainAdaptationLearner	0.88	1.12	0.65	2.5	0.87	0.66	0.5	0.43	0.88	0.88
	(0.83)	(0.97)	(0.56)	(1.15)	(0.29)	(0.2)	(0.17)	(0.13)	(0.85)	(0.81)
TLearner	0.89	1.09	0.69	2.37	0.9	0.69	0.52	0.46	0.9	0.88
	(0.73)	(0.83)	(0.55)	(0.89)	(0.24)	(0.19)	(0.16)	(0.12)	(0.81)	(0.65)
County										
Causal Forest	0.27	0.32	0.22	0.31	0.35	0.24	0.24	0.22	0.21	0.34
	(0.16)	(0.15)	(0.16)	(0.23)	(0.17)	(0.08)	(0.15)	(0.08)	(0.09)	(0.19)
PaCE	0.29	0.33	0.25	0.33	0.39	0.24	0.24	0.24	0.23	0.35
	(0.16)	(0.15)	(0.16)	(0.21)	(0.16)	(0.1)	(0.17)	(0.08)	(0.11)	(0.18)
DML	0.31	0.42	0.19	0.34	0.32	0.21	0.27	0.39	0.18	0.43
	(0.26)	(0.29)	(0.17)	(0.15)	(0.34)	(0.11)	(0.2)	(0.37)	(0.12)	(0.31)
XLearner	0.33	0.37	0.29	0.46	0.4	0.27	0.27	0.23	0.28	0.37
	(0.15)	(0.17)	(0.1)	(0.17)	(0.14)	(0.05)	(0.12)	(0.03)	(0.09)	(0.18)
ForestDRLearner	0.37	0.43	0.31	0.53	0.46	0.29	0.3	0.26	0.31	0.43
	(0.19)	(0.2)	(0.15)	(0.22)	(0.18)	(0.06)	(0.16)	(0.07)	(0.1)	(0.23)
LinearDML	0.37	0.52	0.22	0.4	0.51	0.22	0.26	0.46	0.23	0.51
	(0.35)	(0.43)	(0.15)	(0.16)	(0.54)	(0.11)	(0.2)	(0.42)	(0.13)	(0.44)
SLearner	0.67	0.74	0.59	0.78	0.78	0.63	0.59	0.56	0.64	0.69
	(0.16)	(0.12)	(0.17)	(0.19)	(0.12)	(0.09)	(0.16)	(0.08)	(0.15)	(0.17)
CausalForestDML	1.06	0.41	1.72	3.73	0.86	0.28	0.26	0.24	0.99	1.14
	(1.92)	(0.24)	(2.54)	(3.07)	(0.37)	(0.06)	(0.15)	(0.1)	(1.86)	(1.98)
DomainAdaptationLearner	1.08	0.89	1.26	1.79	1.33	0.91	0.76	0.59	1.09	1.06
	(0.51)	(0.25)	(0.63)	(0.54)	(0.27)	(0.14)	(0.08)	(0.02)	(0.53)	(0.5)
TLearner	1.1	0.92	1.27	1.82	1.35	0.94	0.79	0.6	1.12	1.08
	(0.51)	(0.25)	(0.63)	(0.54)	(0.25)	(0.12)	(0.08)	(0.03)	(0.52)	(0.5)

Table EC.2: Average nMAE among treated observations. Standard deviations in parentheses.

		Adaptive		Proportion treated					Effect	
	All	Y	N	0.05	0.25	0.50	0.75	1.0	Add.	Mult.
SNAP										
PaCE	0.1	0.11	0.08	0.12	0.11	0.09	0.09	0.09	0.06	0.13
	(0.06)	(0.03)	(0.08)	(0.09)	(0.08)	(0.05)	(0.05)	(0.05)	(0.03)	(0.07)
Causal Forest	0.12	0.18	0.06	0.14	0.14	0.13	0.11	0.11	0.11	0.14
	(0.07)	(0.02)	(0.04)	(0.06)	(0.08)	(0.07)	(0.07)	(0.07)	(0.07)	(0.07)
MC-NNM	0.18	0.24	0.12	0.22	0.18	0.17	0.16	0.16	0.16	0.2
	(0.12)	(0.04)	(0.14)	(0.2)	(0.1)	(0.09)	(0.08)	(0.08)	(0.09)	(0.14)
CausalForestDML	0.18	0.24	0.12	0.27	0.21	0.18	0.16	0.15	0.15	0.21
	(0.09)	(0.03)	(0.1)	(0.19)	(0.1)	(0.08)	(0.07)	(0.07)	(0.08)	(0.1)
SLearner	0.23	0.35	0.1	0.27	0.25	0.23	0.21	0.21	0.22	0.24
	(0.14)	(0.04)	(0.06)	(0.14)	(0.15)	(0.13)	(0.13)	(0.13)	(0.14)	(0.13)
DomainAdaptationLearner	0.24	0.38	0.09	0.24	0.25	0.24	0.23	0.24	0.23	0.25
	(0.16)	(0.05)	(0.07)	(0.15)	(0.15)	(0.15)	(0.16)	(0.16)	(0.16)	(0.15)
TLearner	0.26	0.38	0.14	0.3	0.29	0.26	0.24	0.24	0.24	0.28
	(0.15)	(0.05)	(0.12)	(0.22)	(0.16)	(0.14)	(0.15)	(0.15)	(0.15)	(0.16)

Continued on next page

Table EC.2: Average nMAE among treated observations. Standard deviations in parentheses (continued).

		Adaptive		Proportion treated					Effect	
	All	Y	N	0.05	0.25	0.50	0.75	1.0	Add.	Mult.
LinearDML	0.28	0.25	0.31	0.35	0.33	0.3	0.24	0.24	0.11	0.44
	(0.28)	(0.12)	(0.38)	(0.28)	(0.39)	(0.27)	(0.21)	(0.21)	(0.06)	(0.32)
DML	0.37	0.27	0.46	0.56	0.48	0.37	0.31	0.28	0.22	0.51
	(0.31)	(0.17)	(0.38)	(0.25)	(0.48)	(0.24)	(0.18)	(0.19)	(0.15)	(0.36)
XLearner	0.57	0.4	0.74	1.89	0.86	0.57	0.39	0.29	0.45	0.69
	(0.75)	(0.13)	(1.02)	(2.94)	(0.77)	(0.32)	(0.14)	(0.07)	(0.2)	(1.03)
DC-PR	0.59	0.55	0.63	1.88	0.47	0.53	0.57	0.64	0.47	0.72
	(0.66)	(0.2)	(0.91)	(3.22)	(0.21)	(0.19)	(0.18)	(0.22)	(0.16)	(0.9)
ForestDRLearner	1.01	1.18	0.83	2.89	1.63	0.95	0.66	0.52	0.86	1.15
	(1.05)	(0.38)	(1.42)	(3.77)	(1.04)	(0.42)	(0.34)	(0.28)	(0.5)	(1.39)
Median Effect	1.01	0.78	1.24	2.18	1.26	0.97	0.84	0.8	0.92	1.09
	(0.58)	(0.15)	(0.73)	(2)	(0.62)	(0.28)	(0.17)	(0.16)	(0.42)	(0.69)
DID	1.32	0.69	1.94	4.1	1.87	1.24	0.93	0.85	1	1.63
	(1.78)	(0.06)	(2.36)	(6.71)	(2.11)	(0.85)	(0.41)	(0.26)	(0.54)	(2.42)
Mean Effect	1.93	0.88	2.98	6.7	2.9	1.83	1.26	1.08	1.27	2.59
	(3.24)	(0.25)	(4.33)	(11.89)	(3.94)	(1.67)	(0.83)	(0.52)	(0.9)	(4.39)
State										
PaCE	0.21	0.21	0.21	0.35	0.3	0.18	0.15	0.12	0.18	0.25
	(0.14)	(0.13)	(0.15)	(0.23)	(0.11)	(0.04)	(0.09)	(0.04)	(0.11)	(0.16)
MC-NNM	0.23	0.28	0.17	0.23	0.21	0.21	0.24	0.25	0.23	0.23
	(0.12)	(0.14)	(0.07)	(0.17)	(0.11)	(0.1)	(0.09)	(0.11)	(0.11)	(0.13)
XLearner	0.27	0.3	0.23	0.51	0.28	0.22	0.2	0.2	0.25	0.29
	(0.17)	(0.09)	(0.21)	(0.3)	(0.07)	(0.05)	(0.05)	(0.07)	(0.11)	(0.21)
CausalForestDML	0.27	0.31	0.23	0.47	0.29	0.25	0.22	0.19	0.24	0.3
	(0.15)	(0.13)	(0.16)	(0.22)	(0.08)	(0.14)	(0.08)	(0.07)	(0.11)	(0.18)
Causal Forest	0.31	0.32	0.3	0.48	0.38	0.32	0.22	0.2	0.26	0.36
	(0.16)	(0.14)	(0.17)	(0.26)	(0.11)	(0.07)	(0.05)	(0.07)	(0.12)	(0.17)
DomainAdaptationLearner	0.34	0.43	0.24	0.42	0.33	0.3	0.29	0.36	0.33	0.34
	(0.14)	(0.06)	(0.14)	(0.22)	(0.08)	(0.12)	(0.11)	(0.14)	(0.12)	(0.16)
DML	0.34	0.45	0.23	0.66	0.27	0.28	0.37	0.22	0.3	0.38
	(0.3)	(0.36)	(0.14)	(0.47)	(0.21)	(0.17)	(0.3)	(0.06)	(0.29)	(0.29)
LinearDML	0.35	0.38	0.31	0.48	0.31	0.43	0.3	0.25	0.31	0.38
	(0.14)	(0.12)	(0.15)	(0.22)	(0.05)	(0.12)	(0.04)	(0.08)	(0.11)	(0.15)
TLearner	0.38	0.44	0.32	0.47	0.39	0.33	0.33	0.4	0.37	0.39
	(0.15)	(0.07)	(0.18)	(0.28)	(0.03)	(0.08)	(0.11)	(0.13)	(0.11)	(0.17)
SLearner	0.5	0.55	0.44	0.61	0.55	0.52	0.43	0.41	0.47	0.52
	(0.14)	(0.11)	(0.15)	(0.2)	(0.1)	(0.09)	(0.12)	(0.09)	(0.14)	(0.14)
DID	0.59	0.55	0.63	0.8	0.57	0.56	0.52	0.57	0.54	0.64
	(0.24)	(0.11)	(0.31)	(0.54)	(0.11)	(0.08)	(0.08)	(0.09)	(0.12)	(0.31)
NNEstimator	0.61	0.52	0.7	0.85	0.78	0.39	0.5	0.61	0.57	0.65
	(0.36)	(0.23)	(0.45)	(0.71)	(0.31)	(0.08)	(0.16)	(0.17)	(0.27)	(0.43)
DC-PR	0.7	0.53	0.87	1.69	0.37	0.62	0.57	0.57	0.46	0.94
	(1.43)	(0.23)	(1.99)	(3.57)	(0.15)	(0.18)	(0.17)	(0.23)	(0.19)	(1.98)
Median Effect	0.75	0.69	0.8	0.86	0.73	0.84	0.7	0.64	0.69	0.8
	(0.24)	(0.23)	(0.23)	(0.34)	(0.19)	(0.12)	(0.17)	(0.27)	(0.23)	(0.23)
ForestDRLearner	0.88	0.64	1.12	3.39	0.77	0.39	0.4	0.25	0.77	0.99
	(1.43)	(0.52)	(1.93)	(2.56)	(0.24)	(0.06)	(0.13)	(0.07)	(1.07)	(1.72)
Mean Effect	0.97	0.74	1.19	1.55	0.96	1.03	0.77	0.72	0.8	1.14
	(0.75)	(0.27)	(0.98)	(1.71)	(0.32)	(0.27)	(0.23)	(0.32)	(0.34)	(0.98)
County										
Causal Forest	0.22	0.31	0.12	0.24	0.24	0.22	0.19	0.19	0.17	0.26
	(0.15)	(0.13)	(0.08)	(0.22)	(0.16)	(0.11)	(0.1)	(0.09)	(0.1)	(0.17)
PaCE	0.23	0.27	0.18	0.24	0.25	0.22	0.21	0.22	0.18	0.27
	(0.13)	(0.08)	(0.15)	(0.11)	(0.18)	(0.11)	(0.13)	(0.08)	(0.08)	(0.15)
CausalForestDML	0.26	0.36	0.17	0.36	0.31	0.21	0.21	0.22	0.21	0.31
	(0.18)	(0.2)	(0.1)	(0.3)	(0.17)	(0.07)	(0.11)	(0.11)	(0.1)	(0.23)
DML	0.3	0.42	0.18	0.27	0.31	0.22	0.3	0.4	0.16	0.44
	(0.29)	(0.28)	(0.24)	(0.15)	(0.35)	(0.13)	(0.3)	(0.37)	(0.12)	(0.33)
XLearner	0.3	0.35	0.25	0.41	0.36	0.26	0.26	0.21	0.26	0.34
	(0.18)	(0.18)	(0.16)	(0.27)	(0.17)	(0.08)	(0.16)	(0.04)	(0.11)	(0.22)

Continued on next page

Table EC.2: Average nMAE among treated observations. Standard deviations in parentheses (continued).

		Adaptive		Proportion treated					Effect	
	All	Y	N	0.05	0.25	0.50	0.75	1.0	Add.	Mult.
LinearDML	0.33	0.47	0.18	0.29	0.39	0.2	0.28	0.47	0.19	0.47
	(0.32)	(0.34)	(0.21)	(0.24)	(0.36)	(0.13)	(0.28)	(0.44)	(0.12)	(0.39)
DomainAdaptationLearner	0.46	0.59	0.33	0.49	0.48	0.44	0.47	0.43	0.4	0.52
	(0.26)	(0.21)	(0.23)	(0.39)	(0.25)	(0.16)	(0.27)	(0.1)	(0.15)	(0.32)
TLearner	0.48	0.6	0.35	0.5	0.47	0.47	0.49	0.44	0.42	0.53
	(0.26)	(0.21)	(0.24)	(0.41)	(0.24)	(0.17)	(0.28)	(0.1)	(0.16)	(0.32)
SLearner	0.5	0.65	0.36	0.5	0.54	0.49	0.49	0.49	0.48	0.52
	(0.18)	(0.08)	(0.14)	(0.27)	(0.17)	(0.15)	(0.17)	(0.11)	(0.19)	(0.18)
MC-NNM	0.53	0.49	0.57	0.62	0.58	0.51	0.49	0.48	0.45	0.62
	(0.47)	(0.09)	(0.66)	(0.8)	(0.56)	(0.32)	(0.2)	(0.14)	(0.1)	(0.65)
ForestDRLearner	0.53	0.73	0.34	1.28	0.65	0.29	0.25	0.22	0.42	0.65
	(0.62)	(0.76)	(0.34)	(0.96)	(0.41)	(0.12)	(0.14)	(0.1)	(0.36)	(0.78)
Median Effect	0.62	0.77	0.48	0.58	0.61	0.56	0.66	0.71	0.52	0.73
	(0.23)	(0.15)	(0.2)	(0.24)	(0.26)	(0.18)	(0.23)	(0.19)	(0.18)	(0.22)
DC-PR	0.66	0.85	0.48	0.56	0.68	0.57	0.74	0.76	0.53	0.8
	(0.33)	(0.25)	(0.3)	(0.37)	(0.37)	(0.21)	(0.35)	(0.27)	(0.2)	(0.38)
DID	0.74	0.76	0.72	0.67	0.79	0.67	0.82	0.74	0.61	0.87
	(0.36)	(0.2)	(0.47)	(0.33)	(0.4)	(0.23)	(0.53)	(0.14)	(0.13)	(0.45)
Mean Effect	1.1	1.12	1.08	1	1.26	0.89	1.29	1.05	0.73	1.47
	(0.96)	(0.58)	(1.22)	(0.84)	(1.1)	(0.66)	(1.41)	(0.4)	(0.31)	(1.21)

EC.4.2. Parameter selection

In this subsection, we detail how practitioners can select the key parameters: the rank r, the regularization parameter λ, and the number of leaves ℓ.

Choosing the rank r. To determine the appropriate rank for applying PaCE, we adopt a data-driven approach based on the SVD of the observed matrix of outcomes O. Let $\{s_i\}_{i=1}^{d}$ denote the singular values of the data matrix O, sorted in descending order. We select the smallest rank r such that the cumulative variance captured by the top r singular values accounts for at least 99% of the total variance. Formally, we choose r to be the smallest integer satisfying

$$\frac{\sum_{i=1}^{r} s_i^2}{\sum_{i=1}^{d} s_i^2} \geq 0.99.$$

Choosing λ. Theorem 2 suggests setting $\lambda = \Theta(\sigma\sqrt{nr}\log^{4.5}(n))$, which depends on knowledge of both the rank r and the noise level σ up to constant factors. However, this guidance is asymptotic and does not specify an exact value of λ for practical implementation.

In practice, we adopt the following heuristic. We initialize λ as the largest singular value of the observed outcome matrix O, then iteratively reduce it by 10% until the estimated counterfactual matrix has rank r, where r is the estimated rank of O obtained via the procedure described above.

We emphasize that the procedures for selecting both λ and r are practical heuristics. While empirically effective, they do not guarantee the theoretical guarantees of Theorem 2 in the presence of parameter misspecification.

Choosing the number of leaves ℓ. The number of leaves is primarily determined by the dataset size and the practitioner's goals. As stated in Theorem 1, the number of leaves ℓ is upper bounded: $\ell \leq \left(\frac{\alpha}{2c_m M}\right)^{\frac{1}{c \log 1/\alpha}}$, where the main term is $M := \sqrt{\frac{\ln(nT)(p+1)}{\min(n,T)}}$ showcasing that the bound increases in n and T, and decreases in p. The rest of the constants are as follows: α is the proportion of data retained on each side of the split (e.g., $\alpha = 0.05$), c is such that each leaf of the tree has depth at least $c \log n$, and c_m depends on the data (see Assumption 1).

For greater interpretability, a practitioner may choose to use fewer leaves than the maximum allowed. As shown in Figure 2, the initial splits provide the largest improvements in accuracy, suggesting that the marginal loss in precision from using fewer leaves may be offset by gains in simplicity and interpretability.

EC.4.3. Runtime analysis

In this subsection we analyze the runtime of our implementation of Algorithm 1. The runtime scales with the number of leaves ℓ, covariates p, interventions q, units n, and time periods T as follows.

When the tree is split into ℓ leaves, Algorithm 1 runs for ℓ iterations. We analyze the runtime per iteration by separately considering **Step 1** (solving the convex program in (1)) and **Step 2** (selecting a split).

Step 1 is solved via alternating minimization, an iterative procedure in which each variable—M, m, and τ—is updated in turn while holding the others fixed. When all but one of the variables are fixed, the optimization with respect to the remaining variable admits a closed-form solution, making each update step computationally efficient. This process is repeated until convergence: when the objective function stabilizes or a predefined maximum number of iterations is reached. Updating τ requires solving a linear regression problem with nT observations and $q\ell$ predictors, resulting in a runtime of $O(nT(q\ell)^2 + (q\ell)^3)$. The update for m is comparatively efficient, taking $O(q\ell nT)$ time. Updating M is dominated by computing a singular value decomposition (SVD) of the estimated counterfactual matrix, with runtime $O(nT \min(n,T))$.

Step 2 searches for the best split across all p covariates in each of the ℓ leaves, for each of the q interventions. Using binary search to find the optimal threshold for each covariate means that we may evaluate $O(q\ell p \log(nT))$ potential splits. For each given potential split, we must compute the objective value for the potential split, which involves solving a linear regression problem with nT observations and $q\ell$ predictors—this requires a runtime of $O(nT(q\ell)^2 + (q\ell)^3)$.

Combining all components, the total runtime of our implementation of the algorithm is

$$O\left(\ell nT \min(n,T) + \ell(q\ell)^3 p \log(nT)\,(nT + q\ell)\right).$$

References

Cai, J.-F., Candès, E. J., and Shen, Z. (2010). A singular value thresholding algorithm for matrix completion. *SIAM Journal on optimization*, 20(4):1956–1982.

Farias, V., Li, A., and Peng, T. (2021). Learning treatment effects in panels with general intervention patterns. *Advances in Neural Information Processing Systems*, 34:14001–14013.

Vershynin, R. (2018). *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press.