Supplementary Material for Synthetic Learner: Model-Free Inference on Treatments over Time.

Davide Viviano[1*] Jelena Bradic[2 †]

Stanford GSB[1] UC San Diego[1,2]

This version: July 14, 2022

*Present address: Stanford Graduate School of Business, 655 Knight Way, Stanford, CA 94305. Email: dviviano@stanford.edu. This work was mostly conducted while at the Department of Economics, University of California at San Diego, La Jolla, CA, 92093.

†Department of Mathematics and Halicioğlu Data Science Institute, University of California at San Diego, La Jolla, CA, 92093. Email: jbradic@ucsd.edu.

A Proof of the Lemmas in the Main Text

A.1 Proof of Lemma 2.1 and 4.3

We prove Lemma 4.3, while Lemma 2.1 directly follows from stationarity. Note first that under the null hypothesis $Y_{0t}^o = Y_{0t}^0$. As a result, we can write (under the null) $(Y_{0t}^o - \bar{Y}_t, Y_{1t}^1 - \bar{Y}_t, \cdots, Y_{nt} - \bar{Y}_t, Z_{0:n,t}) = (Y_{0t}^0 - \bar{Y}_t, Y_{1t}^1 - \bar{Y}_t, \cdots, Y_{nt} - \bar{Y}_t, Z_{0:n,t})$. Under Assumption 1, we have

$$(Y_{0t}^0 - \bar{Y}_t, Y_{1t}^1 - \bar{Y}_t, \cdots, Y_{nt} - \bar{Y}_t, Z_{0:n,t}) = (\varepsilon_{0t}^0 - \bar{\varepsilon}_t + \iota_0^0 - \bar{\iota}^0, \cdots, \varepsilon_{nt}^0 - \bar{\varepsilon}_t + \iota_n^0 - \bar{\iota}^0, Z_{0:n,t})$$

where $\bar{\varepsilon}_t = \frac{1}{n}\sum_{j=1}^J \varepsilon_{jt}, \bar{\iota}^0 = \frac{1}{n}\sum_{j=1}^J \iota_j^0$. Stationarity directly follows from Assumption 6.

A.2 Proof of Lemma 2.2 and 4.2

We prove Lemma 4.2. Lemma 2.2 follows directly from Assumption 1 and exogeneity of T_0 (Assumption 2). First, note that under Assumption 2 the distribution of potential outcomes and covariates remains invariant conditionally or unconditionally on T_0. Observe that under Assumption 6, we have

$$\tau = \frac{1}{T - T_0}\sum_{t > T_0} \mathbb{E}\left[Y_{0t}^1\right] - \kappa_t - \iota_j^0 - \mathbb{E}\left[\varepsilon_{jt}^0\right] = \frac{1}{T - T_0}\sum_{t > T_0} \mathbb{E}\left[Y_{0t}^1\right] - \kappa_t - \iota_j^0.$$

where, by stationarity and the presence of the unit fixed effect ι_j^0, $\mathbb{E}\Big[\varepsilon_{jt}^0\Big] = 0$, since the expectation is incorporated in the component ι_j^0. Observe that we can write

$$\frac{1}{n}\sum_{j>0}\mathbb{E}\Big[Y_{jt}^0\Big|T_0\Big] = \frac{1}{n}\sum_{j>0}\mathbb{E}\Big[Y_{jt}^0\Big] = \kappa_t^0 + \frac{1}{n}\sum_{j>0}\iota_j^0,$$

where, again, $\mathbb{E}[\varepsilon_{jt}^0] = 0$. Under Assumption 6

$$\frac{1}{T_0}\sum_{s=1}^{T_0}\Big\{\mathbb{E}\Big[Y_{0s}^0|T_0\Big] - \frac{1}{n}\sum_{j>0}\mathbb{E}\Big[Y_{js}^0\Big|T_0\Big]\Big\} = \iota_0^0 - \frac{1}{n}\sum_{j>0}\iota_j^0.$$

Therefore,

$$-\frac{1}{T-T_0}\sum_{t>T_0}\frac{1}{n}\sum_{j>0}\mathbb{E}\Big[Y_{jt}^0\Big|T_0\Big] - \frac{1}{T_0}\sum_{s=1}^{T_0}\Big\{\mathbb{E}\Big[Y_{0s}^0|T_0\Big] - \frac{1}{n}\sum_{j>0}\mathbb{E}\Big[Y_{js}^0\Big|T_0\Big]\Big\} = -\frac{1}{T-T_0}\sum_{t>T_0}\kappa_t^0 - \iota_0^0$$

which completes the proof of Lemma 4.2.

B Theorem 3.1

B.1 Definitions

Definition B.1. (β-mixing)*Let Y be a stochastic process and $(\Omega, \mathcal{F}, Y_\infty)$ be the probability space. The β-mixing coefficient $\beta_Y(h)$ is given by*

$$\beta_Y(h) = \sup_t ||\mathcal{P}_{-\infty:t}\otimes\mathcal{P}_{(t+h):\infty} - \mathcal{P}_{-\infty:t}\mathcal{P}_{(t+h):\infty}||_{TV}$$

where $||.||_{TV}$ is the total variation norm, $\mathcal{P}_{-\infty:t}\otimes\mathcal{P}_{(t+h):\infty}$ is the joint distribution and $\mathcal{P}_{-\infty:t}\mathcal{P}_{(t+h):\infty}$ is the product measure. The process is β-mixing if $\beta_Y(h)\to 0$ as $h\to\infty$

In this section we provide a set of definitions before discussing the main theorem. For a set $\mathbb{A}$ we denote the space of bounded functions on $\mathbb{A}$ by

$$l^\infty(\mathbb{A},\mathbb{B}) = \{f:\mathbb{A}\to\mathbb{B},\ \text{such that } \|f\|_\infty < \infty\},$$

where $\|f\|_\infty = \sup_{a\in\mathbb{A}}|f(a)|$. We let $\mathcal{C}[0,1]$ to be the space of cad-lag functions, i.e. right-continuous with left-hand limits equipped with Skorokhod metric. We define the parameter of interest as a function of the joint law of the data. More precisely, for any law $\mathcal{P}$, for some parameter of interest θ, we can define $\mathcal{P} \mapsto \theta(\mathcal{P})$ to be a measurable map from a domain $\mathcal{C}[0,1]$ to Θ. For a metric space $\mathbb{A}$ with norm $||.||_{\mathbb{A}}$ we denote the set of Lipschitz functionals whose level and Lipschitz constant are bounded by one by

$$BL_1(\mathbb{A}) = \{f : \mathbb{A} \to \mathbb{R} : |f(a)| \leq 1 \text{ and } |f(a) - f(a')| \leq ||a - a'||_{\mathbb{A}} \text{ for all } a, a' \in \mathbb{A}\}.$$

The definition above helps us discussing the definition of weak convergence, provided below.

Definition B.2. (Weak Convergence) We say that $\mathbb{X}_n$ converges weakly in probability conditional on the data to $\mathbb{X}$, or $\mathbb{X}_n \rightsquigarrow \mathbb{X}$ if

$$\sup_{f \in BL_1} \Big| \mathbb{E}[f(\mathbb{X}_n)] - \mathbb{E}[f(\mathbb{X})] \Big| \to 0.$$

Consider the generic problem of studying the limiting distribution of

$$r_n(\phi(\mathbb{X}_n) - \phi(\mathbb{X}))$$

for some $\phi : \mathbb{D}_\phi \subset \mathbb{D} \to \mathbb{E}$. The asymptotic distribution of interest can be derived whenever ϕ satisfies some differentiability requirements such that

$$r_n\{\phi(\mathbb{X}_n) - \phi(\mathbb{X})\} = \phi'_{\theta_0}(r_n(\mathbb{X}_n - \mathbb{X})) + o_{\mathbb{P}}(1).$$

The main condition on ϕ is that is it satisfies a notion of differentiability denoted as Hadamard differentiability. The definition is provided below.

Definition B.3. (Hadamard Differentiable Map) *Let $\mathbb{D}$ and $\mathbb{E}$ be Banach spaces with norms $||.||_{\mathbb{D}}$ and $||.||_{\mathbb{E}}$ respectively, and $\phi : \mathbb{D}_\phi \subseteq \mathbb{D} \to \mathbb{E}$. The map ϕ, is Hadamard differentiable at $\theta \in \mathbb{D}_\phi$ tangentially to a set $\mathbb{D}_0 \subset \mathbb{D}$ if there exist a continuous linear map $\phi'_\theta : \mathbb{D}_0 \to \mathbb{E}$ such that*

$$\lim_{n \to \infty} \left\| \frac{\phi(\theta + t_n k_n) - \phi(\theta)}{t_n} - \phi'_\theta(k) \right\|_{\mathbb{E}} = 0,$$

$\forall$ converging sequences $t_n \to 0$, $\{t_n\} \subset \mathbb{R}$ and $k_n \in \mathbb{D}$, $k_n \to k \in \mathbb{D}_0$ as $n \to \infty$ and $\theta + t_n k_n \in \mathbb{D}_\phi$ for all $n \geq 1$ sufficiently large.

It can be shown that Hadamard differentiability is equivalent to the difference in the previous expression in tending to zero uniformly on k in compact subsets of $\mathbb{D}$ (Van der Vaart, 2000). We move to define the parameters of interest as functionals that map a space of bounded functions to a Banach space. We do this in the following definition.

Definition B.4. Let the weights be $w(\cdot) : \mathcal{A} \subset l^\infty(\mathbb{R}^p, \mathbb{R}) \to \mathcal{W}$, where $\mathcal{W} = [0,1]^p$.

As a second step we need to define the tests as a function of the parameter of interest and of a vector $z \in \mathbb{R}^{p+1}$.

Definition B.5. We define $S_0(z, w) \in l^\infty(\mathbb{R}^{p+1} \times \mathcal{W}, \mathbb{R})$ where $z \in \mathbb{R}^{p+1}$ and $w \in \mathcal{W}$, with $S_0(z, w) = |z_1 - z_2 w|^2$ where z_1 is the first entry of z and z_2 all the remaining entries.

With an abuse of notation, we define

$$Z_t = (Y_{0t}^o, g(X_t)),$$

the vector of observations after imposing the null hypothesis and of transformations of the learners, and refer to Y_t as Y_{0t} throughout our discussion. We will consider $g(\cdot)$ to be a fixed function independent of Z_t for the moment (hence Z_t stationary under Assumption 1), and return to $g(\cdot)$ being estimated with past data at the end of the proof (Appendix B.3).

We denote the empirical measures for control and treatment period, respectively,

$$\mathcal{P}_{T_0} = \frac{1}{T_0} \sum_{t=1}^{T_0} \delta_{Z_t}, \qquad \mathcal{P}_{T_1} = \frac{1}{T - T_0} \sum_{t > T_0} \delta_{Z_t}, \quad \mathcal{P}_T = \frac{1}{T} \sum_{t=1}^{T} \delta_{Z_t}.$$

Similarly $\mathcal{P}_T^*, \mathcal{P}_{T_0}^*$ denote the bootstrapped counterparts constructed based on the entire sample $\mathcal{P}_T$.

For $z \in \mathbb{R}^{p+1}$ let the operator $\leq$ be the component wise operator. We now let $\mathcal{F}$ be the function class:

$$\mathcal{F} = \{f_s : s \in \mathbb{R}^{p+1}, f_s(z) = 1_{z \leq s}\}.$$

It follows that the empirical distribution function for the control and treatment period can be expressed point wise as

$$F_{T_0}(s) = \mathcal{P}_{T_0} f_s, \quad F_{T_1}(s) = \mathcal{P}_{T_1} f_s$$

respectively and similarly this hold for bootstrap measures. Notice that we can see $F_{T_0}(\cdot)$ and $F_{T_1}(\cdot)$ as elements of $l^\infty(\Omega \times \mathcal{F}, \mathbb{R})$ and similarly $F_{T_0}^*$, F_T^* as element of $l^\infty(\Omega \times \bar{\Omega} \times \mathcal{F}, \mathbb{R})$ where $\bar{\Omega}$ is the probability space associated with bootstrap weights. For a fixed sample path we can view this mappings as belonging to $l^\infty(\mathcal{F}, \mathbb{R})$. We define

$$\mathbb{H}_T(f_s) = \sqrt{T_0}\begin{bmatrix} \mathcal{P}_{T_0} f_s - \mathcal{P} f_s \\ \mathcal{P}_{T_1} f_s - \mathcal{P} f_s \end{bmatrix}, \quad \mathbb{H}^*_T(f_s) = \sqrt{T_0}\begin{bmatrix} \mathcal{P}^*_{T_0} f_s - \mathcal{P}_T f_s \\ \mathcal{P}^*_{T_1} f_s - \mathcal{P}_T f_s \end{bmatrix}. \tag{B.1}$$

Here $\mathcal{P}$ denotes the distribution of $(Y_t^o, g(X_t))$ for fixed $g(\cdot)$ (i.e., the distribution induced by applying a function $[i(\cdot), g(\cdot)]$ to the vector (Y_t^o, X_t), with $g(\cdot)$ evaluated at T_- independent data points from $(Y_t^o, X_t)_{t=1}^T$, and conditional on such data points, and $i(\cdot)$ being the identity function).

We express the test statistics of interest as functionals of $\mathbb{H}_T$ and a fixed null trajectory $\{a_t^o\}$. That is, for Equation (7) (divided by $(T_0 - T)^{1/2}$) we can define $(A, B) \to T(A, B)$ with

$$T(A, B) = \int S_0(z, w(A))dB.$$

B.2 Auxiliary Lemmas

Lemma B.1 (Functional Delta Method, Van Der Vaart and Wellner (1996) Theorem 3.9.4)**.** *Let $\mathbb{D}$ and $\mathbb{E}$ be metrizable topological vector spaces. Let $\phi : \mathbb{D}_\phi \subset \mathbb{D} \mapsto \mathbb{E}$ be Hadamard-differentiable at θ tangentially to $\mathbb{D}_0$. Let $X_n : \Omega_n \mapsto \mathbb{D}_\phi$ be maps with $r_n(X_n - \theta) \rightsquigarrow X$ for some sequence of constants $r_n \to \infty$, where X is separable and takes its values in $\mathbb{D}_0$. Then $r_n(\phi(X_n) - \phi(\theta)) \rightsquigarrow \phi'_\theta(X)$. If ϕ'_θ is defined and continuous on the whole of $\mathbb{D}$, then the sequence $r_n(\phi(X_n) - \phi(\theta)) - \phi'_\theta(r_n(X_n - \theta))$ converges to zero in outer probability.*

We will use the above lemma and modifications of the arguments from the functional delta method for the bootstrap (Theorem 3.9.11 in Van Der Vaart and Wellner, 1996), to derive the validity of the bootstrap procedure.

Lemma B.2. *(Lemma 3.8 in Lunde and Shalizi, 2017) Let Z_t be a β-mixing (stationary) process with mixing rate that decays at least at a cubic rate. Consider $\mathbb{H}_t$ defined in (B.1). Then*

$$\mathbb{H}_t \rightsquigarrow \mathbb{H} = \mathbb{G} \times \mathbb{G}$$

where $\mathbb{H}$ is a bivariate Gaussian process with $\times$ symbol denoting independence. Furthermore, is a mean *zero Gaussian Process with covariance structure given by*

$$\Gamma(f,g) = \lim_{k\to\infty} \sum_{i=1}^{\infty} \{\mathbb{E}[f(Z_k)g(Z_i)] - \mathbb{E}[f(Z_k)]\mathbb{E}[g(Z_i)]\}, \quad \forall f,g \in \mathcal{F}. \tag{B.2}$$

Lemma B.3. *(Kosorok, 2008, Theorem 11.26) Let Y be a stationary sequence in $\mathbb{R}^d$ with marginal distribution P and let $\mathcal{F}$ be a class of functions in $L_2(P)$. Let $\mathbb{G}_n^*(f) = Y_n^* f - Y_n f$. Also assume that $Y_1^*, Y_2^*, \ldots, Y_n^*$ are generated by the circular block bootstrap procedure with $b(n) \to \infty$ as $n \to \infty$ and that there exist a $2 < v < \infty$, $q > v/(v-2)$ and $0 < \rho < (v-2)/[2v-2]$ such that:*

1. $\limsup_{k\to\infty} k^q \beta(k) < \infty$;
2. *$\mathcal{F}$ is permissible, VC and has envelope F satisfying $PF^v < \infty$;*
3. $\limsup_{n\to\infty} n^{-\rho} b(n) < \infty$.

Then $\mathbb{G}_n^* \rightsquigarrow \mathbb{G} \in l^\infty(\mathcal{F})$ where $\mathbb{G}$ is a mean 0 Gaussian Process with covariance structure $\Gamma(f,g) = \lim_{k\to\infty} \sum_{i=1}^{\infty} \{\mathbb{E}[f(Y_i)g(Y_k)] - \mathbb{E}[f(Y_i)]\mathbb{E}[g(Y_k)]\}, \quad \forall f, g, \in \mathcal{F}$.

Lemma B.4. *(Lunde and Shalizi, 2017, Lemma A.3.3) Given a continuous function A and a function of bounded variation B in the sense of Hardy-Krause in the hyper-rectangol $\mathcal{R} = \prod_{i=1}^{d} [a_i, b_i]$ define*

$$\phi(A,B) = \int_{[a,b]} A dB.$$

Then, $\phi : C(\mathcal{R}) \times BV_M(\mathcal{R}) \to \mathbb{R}$ is Hadamard differentiable at each $(A,B) \in \mathbb{D}_\phi$ such that $\int |dA| < \infty$. The derivative is given by

$$\phi'_{A,B}(a,\beta) = \int_{[a,b]} A d\beta + \int_{[a,b]} a dB.$$

Lemma B.5. *Consider two multivariate processes $\{X_t, Y_t\}$ and $\{f(X_t), Y_t\}$ for some measurable function f. Let $\beta_1(h)$ and $\beta_2(h)$ be respectively the beta-mixing coefficient of $\{X_t, Y_t\}$ and $\{f(X_t), Y_t\}$. Then*

$$\beta_2(h) \leq \beta_1(h).$$

Proof of Lemma B.5. For two random variables (X,Y) and a measurable function f, $\sigma(f(X)) \subseteq \sigma(X)$, since the pullback $f \circ X(A)^{-1} = X^{-1}(f^{-1}(A)) \in \sigma(X)$ by measurability of f for a given event A. Henceforth for any t,

$$\sigma((f(X_t), Y_t)) \subseteq \sigma((X_t, Y_t)).$$

This implies that

$$\sup_{A\in\sigma((f(X_t),Y_t)),B\in\sigma((f(X_{t+h}),Y_{t+h}))}\sum_i\sum_j|P(A_i\cap B_j)-P(A_i)P(B_j)|$$
$$\leq \sup_{A\in\sigma((X_t,Y_t)),B\in\sigma((X_{t+h},Y_{t+h})}\sum_i\sum_j|P(A_i\cap B_j)-P(A_i)P(B_j)|.$$

where the supremum is over all paris of finite partitions $\{A_i\},\{B_j\}$ such that $A_i\in\mathcal{A}$ and $B_j\in\mathcal{B}$ where $\mathcal{A}$ and $\mathcal{B}$ are the sigma algebras generated by the random variables of interest. Henceforth, exploting the definition of β-mixing given in Bradley et al. (2005) we have

$$\beta_2(h)=\sup_t\sup_{A\in\sigma((f(X_t),Y_t)),B\in\sigma((f(X_{t+h}),Y_{t+h}))}\sum_i\sum_j|P(A_i\cap B_j)-P(A_i)P(B_j)|$$
$$\leq \sup_t\sup_{A\in\sigma((X_t,Y_t)),B\in\sigma((X_{t+h},Y_{t+h})}\sum_i\sum_j|P(A_i\cap B_j)-P(A_i)P(B_j)|=\beta_1(h).$$

□

Lemma B.6. *(Lunde and Shalizi, 2017, Lemma 3.7) Let Q and R be probability distributions on $(\Omega^j,\mathcal{A}^j)$ where j is allowed to be ∞. Define the space of histories $(H_t,\mathcal{H}_t)$, where $H_t=\Omega^t$ and $\mathcal{H}_t=\mathcal{A}^t$ and let $h_t\in\mathcal{H}_t$ be an history. Suppose Q, R permit regular conditional probabilities, denoted $Q(\cdot|h_t)$, and $R(\cdot|h_t)$, respectively. Then, for $||\cdot||_{TV}$ denoting the total variation distance,*

$$||Q-R||_{TV}<ab\Rightarrow Q(||Q(\cdot|h_t)-R(\cdot|h_t)||_{TV}>a)<b.$$

B.3 Proof of Theorem 3.1

Recall that the theorem is derived under the null hypothesis H_0. Consider the case where $g(\cdot)$ are fixed functions. We then return to the case where these are estimated based on F_- (the hold-out sample) at the end of the proof. We first prove that $\mathbb{H}_T^*$ and $\mathbb{H}_T$ which we define as the empirical and bootstrapped measures as in (B.1), converge to the same process $\mathbb{H}$. We let $2T_0=T$ for notational convenience, but the results directly hold as $T_0\propto T$ after rescaling by an appropriate constant.[1] By Lemma B.5 beta-mixing conditions on (Y_t^o,X_t) imply the same conditions on the beta-mixing coefficients of $(Y_t^o,g(X_t))$. Similarly, stationarity of (Y_t^o,X_t) also implies stationarity of $(Y_t^o,g(X_t))$.

[1]In that case, we would also rescale appropriately each component on $\mathbb{H}_T$ by its corresponding constant term.

To apply the functional delta method we first need to show that $\mathbb{H}_T^*$ and $\mathbb{H}_T$ converge to the same process up to a multiplicative constant. By Lemma B.2 $\mathbb{H}_T \to_d \mathbb{H}$. Therefore, by Lemma B.1, we have that for an Hadamard differentiable map as defined in Lemma B.1,

$$\sqrt{T_0}\Big(\phi([\mathcal{P}_{T_0}, \mathcal{P}_{T_1}]) - \phi([\mathcal{P}, \mathcal{P}])\Big) \to_d \phi'_{\mathcal{P}}(\mathbb{H}).$$

By Lemma B.3, $\mathcal{P}_{T_0}^* - \mathcal{P}_T$ and $\mathcal{P}_T^* - \mathcal{P}_T$ (appropriately rescaled) converge marginally to a Gaussian Process with covariance matrix described in Lemma B.3.[2] Since $\mathcal{P}_{T_0}^*, \mathcal{P}_T^*$ are drawn independently conditional on $\mathcal{P}_T$, we can conclude that $\mathbb{H}_T^* \to_d \mathbb{H}$ conditionally on $\mathcal{P}_T$.

We now follow the proof of Theorem 3.9.11 in Van Der Vaart and Wellner (1996) with some minor modifications given the presence of $(\mathcal{P}_{T_0}, \mathcal{P}_{T_1})$ to show consistency of the bootstrap for a genering Hadamard differentiable functional ϕ as defined in Lemma B.1. Define $\mathbb{P}_T^* = [\mathcal{P}_{T_0}^*, \mathcal{P}_{T_1}^*]$. Note that $\mathbb{H}_T^* = \sqrt{T_0}(\mathbb{P}_T^* - [\mathcal{P}_T, \mathcal{P}_T])$. Let $\mathbb{P}_T = [\mathcal{P}_T, \mathcal{P}_T]$. Note that by Theorem 11.25 in Kosorok (2008), $\sqrt{2T_0}(\mathcal{P}_T - \mathcal{P}) \to_d \mathbb{G}$ where $\mathbb{G}$ is as defined in Lemma B.2. Since $\mathbb{H}_T \to_d \mathbb{H}$, by Lemma B.1,

$$\begin{aligned} &\sup_{h \in \mathrm{BL}_1(\mathbb{E})} \Big| \mathbb{E}h\Big(\sqrt{T_0}(\phi(\mathbb{P}_T^*) - \phi(\mathbb{P}_T)\Big) - \mathbb{E}h\Big(\sqrt{T_0}(\phi([\mathcal{P}_{T_0}, \mathcal{P}_{T_1}]) - \phi([\mathcal{P}, \mathcal{P}])\Big) \Big| \\ &\le \underbrace{\sup_{h \in \mathrm{BL}_1(\mathbb{E})} \Big| \mathbb{E}h\Big(\sqrt{T_0}(\phi(\mathbb{P}_T^*) - \phi(\mathbb{P}_T)\Big) - \mathbb{E}h(\phi'_{\mathcal{P}}(\mathbb{H}))\Big|}_{(i)} + o(1). \end{aligned}$$

We want to show that (i) converges to zero.

Following Van Der Vaart and Wellner (1996) (Page 380), without loss of generality, we can assume that $\phi'_P : \mathbb{D} \mapsto \mathbb{E}$ is defined and continuous on the whole space. For every $h \in \mathrm{BL}_1(\mathbb{E})$, the function $h \circ \phi'_P$ is contained in $\mathrm{BL}_{||\phi'_P||}(\mathbb{D})$. Therefore, since $\mathbb{H}_T^* \to_d \mathbb{H}$, we can write

$$\sup_{h \in \mathrm{BL}_1(\mathbb{E})} \Big| \mathbb{E}h(\phi'_{\mathcal{P}}(\mathbb{H}_T^*)) - \mathbb{E}h(\phi'_{\mathcal{P}}(\mathbb{H}))\Big| = o(1).$$

[2]Conditions in Lemma B.3 are justified for the following reasons. As discussed in Section 9.1.1, Kosorok (2008) the class $\mathcal{F}$ has bounded VC dimension. In addition, the class is also permissible since it satisfies the two requirements of permissibility: we can index the class by a set $T = \mathbb{R}^p$ that is a valid Polish space equipped with Borel sigma field.

Next,

$$\sup_{h\in \mathrm{BL}_1(\mathbb{E})}\left|\mathbb{E}h\Big(\sqrt{T_0}(\phi(\mathbb{P}_T^*)-\phi(\mathbb{P}_T)\Big)-\mathbb{E}h(\phi_P'(\mathbb{H}_T^*))\right| = \sup_{h\in \mathrm{BL}_1(\mathbb{E})}\left|\mathbb{E}h\Big(\sqrt{T_0}(\phi(\mathbb{P}_T^*)-\phi(\mathbb{P}_T)\Big)-\mathbb{E}h(\phi_P'(\sqrt{T_0}(\mathbb{P}_T^*-\mathbb{P}_T)))\right|$$
$$\leq \varepsilon + 2P\Big(\Big|\Big|\sqrt{T_0}(\phi(\mathbb{P}_T^*)-\phi(\mathbb{P}_T)\Big)-\phi_P'(\sqrt{T_0}(\mathbb{P}_T^*-\mathbb{P}_T)))\Big|\Big|>\varepsilon\Big),$$

since we took without loss of generality the supremum over a ball with radius one. Notice now that $\sqrt{T_0}(\mathbb{P}_T^*-[\mathcal{P},\mathcal{P}])$ and $\sqrt{T_0}(\mathbb{P}_T-[\mathcal{P},\mathcal{P}])$ they both converge (unconditionally) in distribution to a separable random element that concentrate on $\mathbb{D}_0$. The first case follows since $\mathbb{H}_T^*$ converges in distribution to $\mathbb{H}$ and $\sqrt{2T_0}(\mathcal{P}_T-\mathcal{P})$ converges in distribution to $\mathbb{G}$. The second case follows directly from the fact that $\sqrt{2T_0}(\mathcal{P}_T-\mathcal{P})$ converges in distribution to $\mathbb{G}$. Using the functional delta method, we can write

$$\sqrt{T_0}(\phi(\mathbb{P}_T^*)-\phi(\mathcal{P},\mathcal{P}))=\phi_{\mathcal{P}}'(\sqrt{T_0}(\mathbb{P}_T^*-(\mathcal{P},\mathcal{P})))+o_p(1),\quad \sqrt{T_0}(\phi(\mathbb{P}_T)-\phi(\mathcal{P},\mathcal{P}))=\phi_{\mathcal{P}}'(\sqrt{T_0}(\mathbb{P}_T-(\mathcal{P},\mathcal{P})))+o_p(1).$$

Subctracting the two equations, we obtain the desired result.

We now want to show Hadamard differentiability. We now show Hadamard differentiability for $T(\cdot,\cdot)$ for $S_0(z,w)=|z(-1,w)|^2$, with $T(\cdot,\cdot)$ being the test statistic of interest. The proof invokes the chain rule for Hadamard differentiable maps (Van der Vaart, 2000). We can see $T(\cdot,\cdot)$ as the composition of maps

$$T:(A,B)\overset{(a)}{\to}(B,w(A))\overset{(b)}{\to}(B,S_0(z,w(A)))\overset{(c)}{\to}\int S_0(z,w(A))dB.$$

The map (c) is Hadamard differentiable by Lemma B.4. We have to show that $S_0(z,w)$ is itself Hadamard differentiable in w at $w(\mathcal{P})$ that we write as $\bar{w}$ for short. We start by proving Hadamard differentiability of

$$S_0(z,w)=[z'(-1,w)]^2.$$

We omit the S_0 notation for sake of simplicity in the next few lines. To show Hadamard differentiability we need to show that

$$\left|\left|\frac{S(.,\bar{w}+t_nh_n)-S(.,\bar{w})}{t_n}-S_{\bar{w}}'(.,h)\right|\right|_\infty\to 0 \tag{B.3}$$

where

$$S_{\bar{w}}'((z_1,z_2),h)=2(z_1-z_2\bar{w})h'z_2.$$

We can rewrite the LHS above as

$$||2(z_1-z_2'\bar{w})z_2'(h_n-h)||_\infty+|t_n|||(z_2'h_n)^2||_\infty. \tag{B.4}$$

For the first term in (B.4) by the assumption of compact support, there must exist $c < \infty$ such that

$$\sup_{z_1, z_2} |2(z_1 - z_2'\bar{w}) z_2'(h_n - h)| < c||h_n - h||_1$$

and since $||h_n - h|| \to 0$ the term goes to zero. Equivalently for the second term, since $(z_2' h_n)^2 < \infty$ by the compactness assumption, and $t_n \to 0$ also the second term converges to zero. We are left to show that $A \mapsto w(A)$ is Hadamard differentiable, for w. This directly follows by assumption.

Finally, consider the case where $g(\cdot)$ are estimated using information from F_- (i.e., using the hold-out sample indexed by $t < 0$), where T_- is kept fixed with T (recall the definition of $\mathcal{P}$ at the end of Appendix B.1). To invoke the functional delta method in this case, we need to show that the asymptotic statement in Lemma B.2 holds conditionally on the initial sample $(Z_{T_-}, \cdots, Z_0)$, letting $T \to \infty$. Using the definition of weak convergence, we need to show that

$$\sup_{h \in \mathrm{BL}_1} \Big| \mathbb{E}\Big[h\Big(\sqrt{T_0}[\mathcal{P}_{T_0} - \mathcal{P}, \mathcal{P}_{T_1} - \mathcal{P}] \Big) \Big| Z_{T_-}, \cdots, Z_0 \Big] - \mathbb{E}[h(c\mathbb{H})] \Big| \to 0,$$

up to a rescaling constant c. To show the claim, we exploit the β-mixing conditions. First, we observe that we can write

$$\sqrt{T_0}[\mathcal{P}_{T_0} - \mathcal{P}, \mathcal{P}_{T_1} - \mathcal{P}] = \sqrt{T_0}\Big[\underbrace{\frac{1}{T_0}\sum_{t=1}^{m} \delta_{Z_t} - \frac{m}{T_0}\mathcal{P}}_{(A)} + \underbrace{\frac{1}{T_0}\sum_{T_0 \geq t > m} \delta_{Z_t} - (1 - \frac{m}{T_0})\mathcal{P}}_{:=\mathcal{P}_{m, T_0}}, \mathcal{P}_{T_1} - \mathcal{P} \Big],$$

for any m. We take $m \to \infty$ and $m/\sqrt{T_0} \to 0$, which implies $(A) = o(1)$ almost surely. Since $(A) = o(1)$ we can write

$$\begin{aligned}
&\sup_{h \in \mathrm{BL}_1} \Big| \mathbb{E}\Big[h\Big(\sqrt{T_0}[\mathcal{P}_{T_0} - \mathcal{P}, \mathcal{P}_{T_1} - \mathcal{P}] \Big) \Big| Z_{T_-}, \cdots, Z_0 \Big] - \mathbb{E}[h(\mathbb{H})] \Big| \\
&= \sup_{h \in \mathrm{BL}_1} \Big| \mathbb{E}\Big[h\Big(\sqrt{T_0}[\mathcal{P}_{m, T_0}, \mathcal{P}_{T_1} - \mathcal{P}] \Big) \Big| Z_{T_-}, \cdots, Z_0 \Big] - \mathbb{E}[h(\mathbb{H})] \Big| + \varepsilon + 2P\Big(||\sqrt{T_0}(A)|| > \varepsilon | Z_{T_-}, \cdots, Z_0 \Big) \\
&= \sup_{h \in \mathrm{BL}_1} \Big| \mathbb{E}\Big[h\Big(\sqrt{T_0}[\mathcal{P}_{m, T_0}, \mathcal{P}_{T_1} - \mathcal{P}] \Big) \Big| Z_{T_-}, \cdots, Z_0 \Big] - \mathbb{E}[h(\mathbb{H})] \Big| + o(1),
\end{aligned}$$

for a suitable choice of $\varepsilon = o(1)$. Also, note that $\sqrt{T_0}[\mathcal{P}_{m, T_0}, \mathcal{P}_{T_1} - \mathcal{P}]$ satisfies the mixing conditions in Lemma B.2.

Define μ_m a sequence of probability distributions of $(Z_{T_-}, \cdots, Z_0, Z_m, \cdots)$, and μ_∞ the product measure of $(Z_{T_-}, \cdots, Z_0)$ and $(Z_m, \cdots)$. Then under the assumed mixed conditions, by Lemma B.6

$$\sum_{m=1}^{\infty} P(||\mu_m(\cdot|h_0) - \mu_\infty(\cdot|h_0)||_{TV} > a_m) < \infty,$$

for $a_m \to 0$, which implies $||\mu_m(\cdot|h_0) - \mu_\infty(\cdot|h_0)||_{TV} \to \infty$ by the Borel-Cantelli Lemma. This implies that for any bounded function f

$$\int f d\mu_m(\cdot|h_0) \to \int f d\mu_\infty(\cdot|h_0).$$

Hence, we can write

$$\sup_{h \in \mathrm{BL}_1} \Big|\mathbb{E}\Big[h\Big(\sqrt{T_0}[\mathcal{P}_{m,T_0}, \mathcal{P}_{T_1} - \mathcal{P}]\Big)\Big| Z_{T_-}, \cdots, Z_0\Big] - \mathbb{E}[h(\mathbb{H})]\Big| = \sup_{h \in \mathrm{BL}_1} \Big| \int h\Big(\sqrt{T_0}[\mathcal{P}_{m,T_0}, \mathcal{P}_{T_1} - \mathcal{P}]\Big) d\mu_\infty - \mathbb{E}[h(\mathbb{H})]\Big| + o(1).$$

The same reasoning applies to the bootstrapped process $\mathbb{H}_T^*$. Since μ_∞ is the product measure of $(Z_{T_-}, \cdots, Z_0)$ and $(Z_m, \cdots)$ the rest of the proof follows directly from the unconditional case discussed at the beginning of the proof.

B.4 Proof of Theorem 4.1

Note that under Assumption 2 we can fix T_0 as non-random (condition on T_0) without affecting the distribution of observables and unobservables. To prove the claim we can write the estimator as follows

$$\hat{\tau} = \underbrace{\frac{1}{T - T_0} \sum_{t > T_0} Y_t - \frac{1}{T_0/2} \sum_{t = T_0/2+1}^{T_0} Y_t}_{(A)} + \underbrace{\sum_{t = T_0/2+1}^{T_0} w(F_{0,-1/2}) g(X_t) - \sum_{t > T_0}^{T_0} w(F_0) g(X_t)}_{(B)}.$$

Consider first (A). By beta-mixing (and hence alpha-mixing), by the Ergodic Theorem[3]

$$(A) - \tau \to_p 0.$$

We are left to show that (B) converges to zero. We define $\tilde{F}_1$ the empirical distribution of X_t for $t > T_0$ and $\tilde{F}_{0,1/2}$ the empirical distribution of X_t for $T_0 \geq t > T_0/2 + 1$. Note that by Assumption

[3]See Theorem 3.34 in White (2014).

2 and stationarity (Assumption 1), we obtain $\mathbb{E}[\tilde{F}_1] = \mathbb{E}[\tilde{F}_{0,1/2}] = \mathcal{P}_X$ with $\mathcal{P}_X$ denoting the marginal distribution of X_t. Here, $\mathcal{P}_X$ is time-invariant by Assumption 1 and the fact that we subctract the fixed effects components. We also define $\mathcal{P} = \mathbb{E}[F_0]$ denoting the distribution of (X_t, Y_t) (time invariant by Assumption 1). Finally, recall the definition of F_- denoting the empirical distribution over the sample $t < 0$.

We now write, in compact form, with a change of notation, and making explicit the dependence of $g(X_t, F_-)$ with F_-,

$$(B) = \int w(F_{0,-1/2}) g(X_t; F_-) d\tilde{F}_{0,1/2} - \int w(F_0) g(X_t; F_-) d\tilde{F}_1.$$

We now discuss each component after *conditioning* on F_-. In particular, we note that we can interpret each component as map

$$T(A, B; F_-) = \int w(A) g(x, F_-) dB,$$

which is Hadamard differentiable in A and B by Lemma B.4 (but not necessarily in F_-). Following the argument of the proof of Theorem 3.1 verbatim and invoking Lemma B.2 and beta-mixing as in the proof of Theorem 3.1 (see Appendix B.3), we obtain

$$\sqrt{T_0}(T(F_{0,-1/2}, \tilde{F}_{0,1/2}; F_-) - T(\mathcal{P}, \mathcal{P}_X; F_-)) \to_d T'_{\mathcal{P},\mathcal{P}_X}(\mathbb{Y}; F_-),$$
$$\sqrt{T_0}(T(F_0, \tilde{F}_1; F_-) - T(\mathcal{P}, \mathcal{P}_X; F_-)) \to_d T'_{\mathcal{P},\mathcal{P}_X}(\tilde{\mathbb{Y}}; F_-)$$

for two tight processes $(\mathbb{Y}, \tilde{\mathbb{Y}})$. This implies that $T(F_{0,-1/2}, \tilde{F}_{0,1/2}; \mathcal{P}), T(F_0, \tilde{F}_1)$ converges in probability to the same asymptotic limit. This implies that

$$\int w(F_{0,-1/2}) g(X_t; F_-) d\tilde{F}_{0,1/2} - \int w(F_0) g(X_t; F_-) d\tilde{F}_1 = o_P(1),$$

completing the proof.

B.5 Additional Results

It is interesting to study Hadamard differentiability of the weights. Hadamard differentiability for Least Squares has been shown in other papers, such as in Lunde and Shalizi (2017). Hence, we only need to show Hadamard differentiability for exponential weights.

Theorem B.7. *Consider exponential weights as in Equation* (11) *with* $\eta \propto 1/T_0$*, and* $(Y_t, g(X_t))$ *uniformly bounded. Then such weights are Hadamard differentiable.*

Proof of Theorem B.7. Since we consider a finite dimensional parameter space, we can prove Hadamard differentiability by showing Hadamard differentiability for each coordinate. We will assume that

$$\eta_t = \frac{\eta}{t}.$$

We let $l : \mathbb{R}^2 \to \mathbb{R}_+$ the loss function and we assume the loss can be at most $C < \infty$ and $\eta > 0$(which follows for quadratic loss functions and bounded random variables). The aim is to show that

$$\hat{w}_0^{(i)}(F) = \frac{\exp(-\eta \int l(z_1, z_i)dF)}{\sum_{j=2}^{p+1} \exp(-\eta \int l(z_1, z_j)dF)} \tag{B.5}$$

is Hadamard differentiable with $Z = (z_1, \ldots, z_{p+1}) \in [-M, M]^{p+1}$. We will use the chain rule for the Hadamard derivative. We can see $w^{(i)}(F)$ is the composition of mappings:

$$\begin{aligned}
A &\overset{(a)}{\mapsto} \begin{bmatrix} \int l(z_1, z_2)dA \\ \ldots \\ \int l(z_1, z_{p+1})dA \end{bmatrix} \\
&\overset{(b)}{\mapsto} \begin{bmatrix} \exp(-\eta \int l(z_1, z_2)dA) \\ \ldots \\ \exp(-\eta \int l(z_1, z_{p+1})dA) \end{bmatrix} \\
&\overset{(c)}{\mapsto} \begin{bmatrix} \exp(-\eta \int l(z_1, z_i)dA) \\ \sum_{j=1}^{p} \exp(-\eta \int l(z_1, z_{j+1})dA) \end{bmatrix} \\
&\overset{(d)}{\mapsto} \frac{\exp(-\eta \int l(z_1, z_i)dA)}{\sum_{j=1}^{p} \exp(-\eta \int l(z_1, z_{j+1})dA)}
\end{aligned}.$$

We will prove that each map is Hadamard differentiable under the conditions stated component wise. We start by proving that (a) is Hadamard differentiable component wise. For $||h_n - h||_\infty \to 0$, $t_n \to 0$

$$\begin{aligned}
\frac{\int l(z_1, z_j)d(F + t_n h_n) - \int l(z_1, z_j)dF}{t_n} &= \frac{t_n \int l(z_1, z_j)dh_n}{t_n} + \frac{\int l(z_1, z_j)d(F - F)}{t_n} \\
&\to \int l(z_1, z_j)dh.
\end{aligned}$$

Since $l(z_1, z_j)$ is bounded, uniform convergence follows. Hence

$$\left\| \frac{\int l(z_1, z_j)d(F + t_n h_n) - \int l(z_1, z_j)dF}{t_n} - \int l(z_1, z_j)dh \right\|_\infty \to 0.$$

We now move to (b). Let x be the argument of the map. Recall that the argument is positive. Using the mean value theorem, for $\bar{h}_n \in [h_n, h]$,

$$\begin{aligned} \frac{\exp(-\eta(x + t_n h_n)) - \exp(-\eta(x))}{t_n} &= \frac{\exp(-\eta x)[\exp(-\eta t_n h_n) - 1]}{t_n} \\ &= \frac{\exp(-\eta x)[-\eta t_n \exp(-\eta t_n \bar{h}_n) h_n]}{t_n} \\ &= -\eta \exp(-\eta x - \eta t_n \bar{h}_n) h_n \to -\eta \exp(-\eta x) h. \end{aligned} \tag{B.6}$$

Hence we have

$$\begin{aligned} &\left\| \frac{\exp(-\eta(x + t_n h_n)) - \exp(-\eta(x))}{t_n} + \eta \exp(-\eta x) h \right\|_\infty \\ &= \sup_{x \in \mathbb{R}_+} |\exp(-\eta x)[\frac{\exp(-\eta t_n h_n) - 1}{t_n} + \eta h]| \\ &\le |\frac{\exp(-\eta t_n h_n) - 1}{t_n} + \eta h| = | - \eta \exp(-\eta t_n \bar{h}_n) h_n + \eta h| \to 0 \end{aligned}$$

since $\exp(-\eta x) \le 1$ for $x \ge 0$. Since $\eta h \exp(-\eta x)$ is linear in h, (b) is Hadamard differentiable. The map (c) is Hadamard differentiable by linearity of the Hadamard derivative. We are left to prove (d). Let

$$t_n \to 0, \quad \left\| h_n - \begin{pmatrix} h_1 \\ h_2 \end{pmatrix} \right\|_\infty \to 0.$$

Take N such that $pe^{-\eta C} > |t_N h_{N,2}|$. Such N exists since $t_n \to 0$ and $h_{n,2} \to h_2 < \infty$. For $n > N$ we have

$$
\begin{aligned}
&\left\| \frac{\frac{x+t_nh_{n,1}}{y+t_nh_{n,2}} - \frac{x}{y}}{t_n} - \frac{h_1}{y} + h_2\frac{x}{y^2} \right\|_\infty \\
&= \sup_{0\le x\le 1, y\ge re^{-\eta C}} \left\| \frac{(h_{n,1}-h_1) - (h_{n,2}-h_2)x/y - h_1t_nh_{n,2}/y + h_2xh_{n,2}t_n/y^2}{y+t_nh_{n,2}} \right\|_\infty \\
&\le \sup_{0\le x\le 1, y\ge re^{-\eta C}} \frac{|h_{n,1}-h_1|}{|y+t_nh_{n,2}|} + \frac{|h_{n,2}-h_2||\frac{x}{y}|}{|y+t_nh_{n,2}|} + |t_n|\frac{|h_1h_{n,2}|/|y|}{|y+t_nh_{n,2}|} + \frac{|h_2h_{n,2}x/y^2|}{|y+t_nh_{n,2}|}|t_n| \qquad \text{(B.7)} \\
&\le \frac{|h_{n,1}-h_1|}{pe^{-\eta C} - |t_nh_{n,2}|} + \frac{|h_{n,2}-h_2|}{pe^{-\eta C}(pe^{-\eta C} - |t_nh_{n,2}|)} + |t_n|\frac{|h_1h_{n,2}|}{pe^{-\eta C}(pe^{-\eta C} - |t_nh_{n,2}|)} \\
&+ \frac{|h_2h_{n,2}|}{pe^{-2\eta C}(pe^{-\eta C} - |t_nh_{n,2}|)}|t_n| \\
&\to 0.
\end{aligned}
$$

since each term is going to zero and the denominator of each term is bounded away from zero. Hence the Hadamard derivative is $-\frac{h_1}{y} + h_2\frac{x}{y^2}$, which is linear in h. □

We discuss the implication of our results for the asymptotic distribution of the test statistic in the following theorem.

Theorem B.8. *Under the conditions in Theorem 3.1 the test statistics $\mathcal{T}(A,B)$, are Hadamard differentiable tangentially at $(\mathcal{P}_0,\mathcal{P}_0)$ with derivative $\mathcal{T}'_{\mathcal{S},\mathcal{P}_0,\mathcal{P}_0}$. Therefore,*

$$\mathcal{T}_{\mathcal{S}}(\mathcal{P}_{T_0},\mathcal{P}_T) - \mathcal{T}_{\mathcal{S}}(\mathcal{P}_0,\mathcal{P}_0) \to_d \mathcal{T}'_{\mathcal{S},\mathcal{P}_0,\mathcal{P}_0}(\mathbb{H})$$

where $\mathcal{P}_0,\mathbb{H}$ are as defined respectively as the true CDF of (Y_t^0, X_t) and as defined in Lemma B.2.

The proof follows directly from the differentiability properties discussed in the derivation of Theorem 3.1, and by the Functional Delta Method. Observe that the asymptotic distribution depends on the Hadamard derivative of the test statistics which ultimately depends on the weighting mechanism. Its expression is obtained through the chain rule in the proof of Theorem 3.1 .

C Proofs of the Results in Section 5

With an abuse of notation we define the prediction of the potential outcome at time t as

$$\hat{Y}_t^0(\mathcal{F}_{t-1}) := \hat{m}_t(X_t, w^{t-1})$$

where w^{t-1} are exponential weights as in (11) computed using information available only until time $t-1$.

Lemma C.1. *Consider the Synthetic Learner with the exponential weighting scheme with* $\hat{g}_i : \mathcal{X} \to [-\frac{M}{2}, \frac{M}{2}]$. *Then*

$$\frac{1}{T_0}\sum_{t=1}^{T_0}(Y_t - \hat{m}_t(X_t, w^{t-1}))^2 - \min_{i\in\{1,\dots,p\}}\frac{1}{T_0}\sum_{t=1}^{T_0}(Y_t - \hat{g}_i(X_t))^2 \leq \log(p)/\eta + \sum_{t=1}^{T_0} C_t^2\eta/8 \tag{C.1}$$

where $C_t = Y_t^2 + M^2$.

Proof of Lemma C.1 . The proof follows similarly to Cesa-Bianchi et al. (1999) with appropriate modifications for the context under consideration. Let

$$w_{i,t} = \sum_{i=1}^{p}\exp(-\eta\sum_{s=1}^{t} l(Y_s, \hat{g}_i(X_s))), \quad W_t = \sum_{i=1}^{p} w_{i,t}.$$

We denote $\hat{g}_i(X_t) := \hat{g}_{i,t}$ for expositional convenience. We recall that the weights in the first period are uniform across the learners.

We start by deriving a lower bound:

$$\begin{aligned}\log(W_{T_0}/w) = \log(W_{T_0}) - \log(p) &= \log\Big(\sum_{i=1}^{p}\exp(-\eta\sum_{t=1}^{T_0}(Y_t - \hat{g}_{i,t})^2)\Big) - \log(p) \\ &\geq \log(\max_{i\in\{1,\dots,p\}}\exp(-\eta\sum_{t=1}^{T_0}(Y_t - \hat{g}_{i,t})^2) - \log(p) \\ &= -\eta\min_{i\in\{1,\dots,p\}}\sum_{t=1}^{T}(Y_t - \hat{g}_{i,t})^2 - \log(p).\end{aligned}$$

Next, we derive an upper bound on the same quantity of interest.

$$\begin{aligned}\log(W_{T_0}/w) = \log(\prod_{t=1}^{T_0} W_t/W_{t-1}) &= \sum_{t=1}^{T_0}\log\Big(\sum_{i=1}^{p}\frac{w_{i,t-1}}{W_{t-1}}\exp(-\eta(Y_t - \hat{g}_{i,t})^2)\Big) \\ &= \sum_{t=1}^{T_0}\log\Big(\mathbb{E}_{\hat{\mathbf{g}}\sim Q_t}[\exp(-\eta(Y_t - \hat{g}_{i,t})^2)]\Big)\end{aligned}$$

where $\mathbb{E}_{\hat{\mathbf{g}}\sim Q_t}$ denotes the expectation conditional on the data taken with respect to a distribution Q_t on base-learners which assigns a probability proportional to $\exp(-\eta\sum_{s=1}^{t-1}(Y_t - \hat{g}_{i,s})^2)$ to each base algorithm.

Recalling Hoeffding bound on the moment generating function of a bounded random variable we observe that

$$\begin{aligned}\log(\mathbb{E}_{\hat{\mathbf{g}}\sim Q_t}[\exp(-\eta(Y_t-\hat{g}_{i,t})^2)]) &\leq -\eta\mathbb{E}_{\hat{\mathbf{g}}\sim Q_t}[(Y_t-\hat{g}_{i,t})^2]+\frac{\eta^2C_t^2}{8}\\ &\leq -\eta(Y_t-\mathbb{E}_{\hat{\mathbf{g}}\sim Q_t}[\hat{g}_{i,t}])^2+\frac{\eta^2C_t^2}{8}\\ &= -\eta(Y_t-\sum_{i=1}^p\frac{w_{i,t}}{W_t}\hat{g}_{i,t})^2+\frac{\eta^2C_t^2}{8}\\ &= -\eta(Y_t-\hat{m}_t(X_t))^2+\frac{\eta^2C_t^2}{8}\end{aligned}$$

where $C_t = Y_t^2 + M^2$. Hence we have

$$\begin{aligned}&-\eta\min_{i\in\{1,\dots,p\}}\sum_{t=1}^T(Y_t-\hat{m}_t(X_t,w^{t-1}))^2-\log(p) &\text{(C.2)}\\ &\leq \log(W_{T_0}/w)\leq\sum_{t=1}^{T_0}-\eta(Y_t-\sum_{i=1}^p\frac{w_{i,t}}{W_t}\hat{g}_{i,t})^2+\sum_{t=1}^{T_0}\frac{\eta^2C_t^2}{8}. &\text{(C.3)}\end{aligned}$$

Rearranging terms we get

$$\sum_{t=1}^{T_0}(Y_t-\hat{m}_t(X_t,w^{t-1}))^2-\min_{i\in\{1,\dots,p\}}\sum_{t=1}^{T_0}(Y_t-\hat{g}_{i,t})^2\leq\frac{\log(p)}{\eta}+\sum_{t=1}^{T_0}\frac{C_t^2\eta}{8}. \tag{C.4}$$

□

Corollary 1 (Theorem 5.1)**.** *Suppose that $Y_t \leq M < \infty$ almost surely. Consider $\eta = \sqrt{\frac{8\log(p)}{M^2T_0}}$ and the above conditions hold. Then we obtain that*

$$\frac{1}{T_0}\sum_{t=1}^{T_0}(Y_t-\hat{m}_t(X_t,w^{t-1}))^2-\min_{i\in\{1,\dots,p\}}\frac{1}{T_0}\sum_{t=1}^{T_0}(Y_t-\hat{g}_{i,t})^2\leq C\sqrt{\frac{\log(p)}{T_0}}. \tag{C.5}$$

for a constant $C < \infty$. Therefore, Theorem 5.1 holds.

C.1 Proof of Theorem 5.2

We now discuss the main proof of the theorem.

We let w^{t-1} denote the exponential weights computed using information only from time 1 up to time $t-1$ for estimating $\hat{Y}_t^0$.

We start by decomposing the average loss as follows.

$$\begin{aligned}\frac{1}{T_0}\sum_{t=1}^{T_0}(Y_t - \hat{m}_t(X_t, w^{t-1}))^2 &= \frac{1}{T_0}\sum_{t=1}^{T_0}(\hat{m}_t(X_t, w^{t-1}) - \mu_t(X_t))^2 + \varepsilon_{0t}^2 + 2(\hat{m}_t(X_t, w^{t-1}) - \mu_t(X_t))\varepsilon_{0t} \\ \frac{1}{T_0}\sum_{t=1}^{T_0}(Y_t - \hat{g}_i(X_t))^2 &= \frac{1}{T_0}\sum_{t=1}^{T_0}(\hat{g}_i(X_t) - \mu_t(X_t))^2 + \varepsilon_{0t}^2 + 2(\hat{g}_i(X_t) - \mu_t(X_t))\varepsilon_{0t}\end{aligned}. \tag{C.6}$$

Let $\hat{g}_{i,t} = \hat{g}_i(X_t)$. Note that

$$\begin{aligned}&\frac{1}{T_0}\Big\{\sum_{t=1}^{T_0}(Y_t - \hat{m}_t(X_t, w^{t-1}))^2 - \min_{i\in\{1,\dots,p\}}\sum_{t=1}^{T_0}(\hat{g}_{i,t} - Y_t)^2\Big\} \\ &= \frac{1}{T_0}\sum_{t=1}^{T_0}(\mu_t(X_t) - \hat{m}_t(X_t, w^{t-1}))^2 + 2\min_{i\in\{1,\dots,p\}}\frac{1}{T_0}\sum_{t=1}^{T_0}(\hat{m}_t(X_t, w^{t-1}) - \hat{g}_i(X_t))\varepsilon_{0t} - \frac{1}{T_0}\sum_{t=1}^{T_0}(\hat{g}_{i,t} - \mu_t(X_t))^2.\end{aligned} \tag{C.7}$$

Next, we provide a bound on the cumulative one step ahead prediction error.

Using Lemma C.1 and the decomposition in (C.7), for a finite constant C independent of p, T_0, we write

$$\begin{aligned}&\frac{1}{T_0}\{\sum_{t=1}^{T_0}(\mu_t(X_t) - \hat{m}_t(X_t, w^{t-1}))^2 \\ &\leq \min_{i\in\{1,\dots,p\}}\frac{1}{T_0}\sum_{t=1}^{T_0}\{(\hat{g}_{i,t} - \hat{m}_t(X_t, w^{t-1}))\varepsilon_{0t} + (\hat{g}_{i,t} - \mu_t(X_t))^2\} + C\sqrt{\frac{\log(p)}{T_0}} \\ &\leq \underbrace{\max_{j\in\{1,\dots,p\}}\frac{1}{T_0}\sum_{t=1}^{T_0}(\hat{g}_{j,t} - \hat{m}_t(X_t, w^{t-1}))\varepsilon_{0t}}_{(I)} + \underbrace{\min_{i\in\{1,\dots,p\}}\frac{1}{T_0}\sum_{t=1}^{T_0}(\hat{g}_{i,t} - \mu_t(X_t))^2 + C\sqrt{\frac{\log(p)}{T_0}}}_{(II)}.\end{aligned} \tag{C.8}$$

We discuss the term (I). Define

$$V_t = \varepsilon_t(\hat{m}_t(X_t, w^{t-1}) - \hat{g}_i(X_t)) - \mathbb{E}[\varepsilon_t(\hat{m}_t(X_t, w^{t-1}) - \hat{g}_i(X_t))|\mathcal{F}_{t-1}].$$

We have

$$(I) = \max_{i\in\{1,\dots,p\}}\frac{1}{T_0}\sum_{t=1}^{T_0}\Big[V_t + \mathbb{E}[\varepsilon_t(\hat{m}_t(X_t, w^{t-1}) - \hat{g}_i(X_t))|\mathcal{F}_{t-1}]\Big].$$

Using the triangular inequality, we have

$$
\begin{aligned}
|(I)| \leq & \max_{i\in\{1,\cdots,p\}} \Big|\frac{1}{T_0}\sum_{t=1}^{T_0} \varepsilon_t(\hat{m}_t(X_t, w^{t-1}) - \hat{g}_i(X_t)) - \mathbb{E}\Big[\varepsilon_t(\hat{m}_t(X_t, w^{t-1}) - \hat{g}_i(X_t))\Big|\mathcal{F}_{t-1}\Big]\Big| \\
& + \Big|\frac{1}{T_0}\sum_{t=1}^{T_0} \mathbb{E}\Big[\varepsilon_t(\hat{m}_t(X_t, w^{t-1}) - \hat{g}_i(X_t))\Big|\mathcal{F}_{t-1}\Big]\Big|.
\end{aligned} \tag{C.9}
$$

Observe that V_t is a martingale difference sequence. As a result, we can use Azuma inequality (Boucheron et al., 2013), and the union bound over p, and obtain with probability at least $1-\delta$,

$$
\max_{i\in\{1,\cdots,p\}} \Big|\frac{1}{T_0}\sum_{t=1}^{T_0} \varepsilon_t(\hat{m}_t(X_t, w^{t-1}) - \hat{g}_i(X_t)) - \mathbb{E}\Big[\varepsilon_t(\hat{m}_t(X_t, w^{t-1}) - \hat{g}_i(X_t))\Big|\mathcal{F}_{t-1}\Big]\Big| \leq C_0\sqrt{\frac{\log(p/\delta)}{T_0}},
$$

where C_0 is a finite constant independent of T_0, p. Since $\mathbb{E}\Big[\varepsilon_t(\hat{m}_t(X_t, w^{t-1}) - \hat{g}_i(X_t))\Big|\mathcal{F}_{t-1}\Big] = 0$, we have with probability at least $1-\delta$,

$$
|(I)| \leq C_0\sqrt{\frac{\log(p/\delta)}{T_0}},
$$

for a constant $C_0 < \infty$.

D Examples

D.1 Failure of Stationarity with Spillovers

Although our results remain correct in the presence of carry-over effects we do not expect them to hold for the case of spillover over covariates. Observe that spillovers effects on X_t violate stationarity of X_t and hence the validity of Theorem 3.1.

We provide a counterexample where the bootstrap is likely to fail under such circumstances. We consider a case in which covariates are a function of the treatment assignment indicator - i.e., there are spillovers from Y_t to X_t while keeping $m = 0$. A factor model, $Y_t = a_t D_t + F_t + \mathcal{N}(0, \sigma^2)$ and 10-covariates that possibly depend on the treatment groups d, $X_{j,t}(d) = F_t + \mathcal{N}(0, 1+d)$ are considered with $F_t \sim \mathcal{N}(0,1)$. Moreover, $a_t = 0$. The best estimator then is a simple average $\hat{Y}_t^0 = \bar{X}_t$ where $\bar{X}_t$ is the sample average of $X_{t,j}$'s. Then, $\mathcal{T}_S = \tau_1\chi^2_{T-T_0}$ for $\tau_1 = (\sigma^2 + 2J^{-1})/\sqrt{T-T_0}$, while $\mathcal{T}_S^* = \tau_2\chi^2_{T-T_0}$ with $\tau_2 = (\sigma^2 + J^{-1})/\sqrt{T-T_0}$. Figure D.1 illustrates over-rejection of the nulls whenever the spillover effects are strong (right panel). For larger values of σ^2, spillover effects become negligible and the density of the

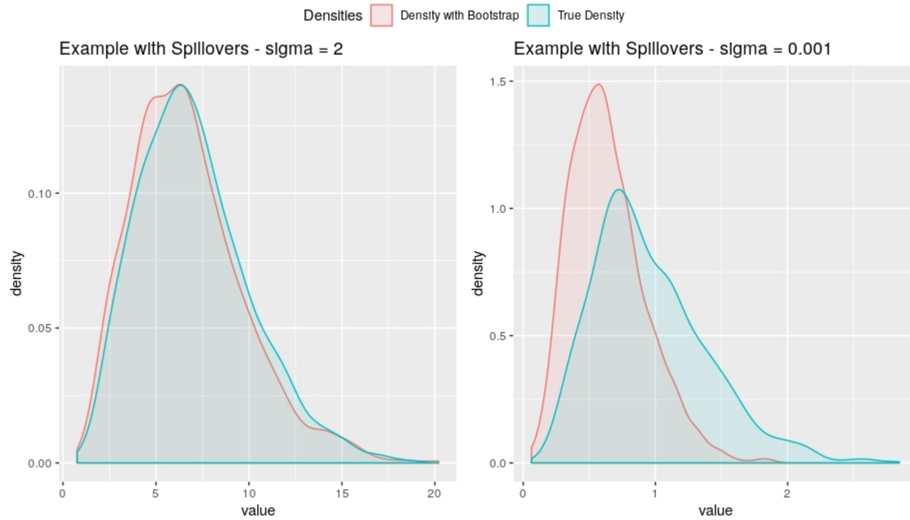

Figure D.1: Explanatory example on failure of the bootstrap in the presence of spillover effects on X_t. Blue histogram indicates $\mathcal{T}_S = \tau_1\chi^2(T-T_0)$ whereas the red corresponds to its bootstrapped counterpart $\mathcal{T}_S^* = \tau_2\chi^2(T-T_0)$ where $\tau_1 = (\sigma^2+2J^{-1})/\sqrt{T-T_0}$, $\tau_2 = (\sigma^2+J^{-1})/\sqrt{T-T_0}$. Large values of σ correspond to diminishing effects of the spillovers in which case we see good approximation properties; although in this case stationarity is still violated. For small values of σ we see clear departures of the two distributions.

bootstrapped test statistic approximately agrees with the true density in this also, non-stationary case (left panel).

E Additional Numerical Experiments

E.1 Results for Other DGPs

In this section we illustrate results for other DGPs considered. Results are robust across DGPs and illustrate substantial improvements of the proposed methods. These can be observed in Figure E.1.

E.2 Varying $T-T_0$

In this section we vary $T-T_0$, i.e., the post treatment period, and show that results remain robust for different choices of the post-treatment period, both longer and shorter than ten. This is illustrated in Figure E.2, E.3. In the figures, we report the power for different values of T and $T^* = T - T_0$.

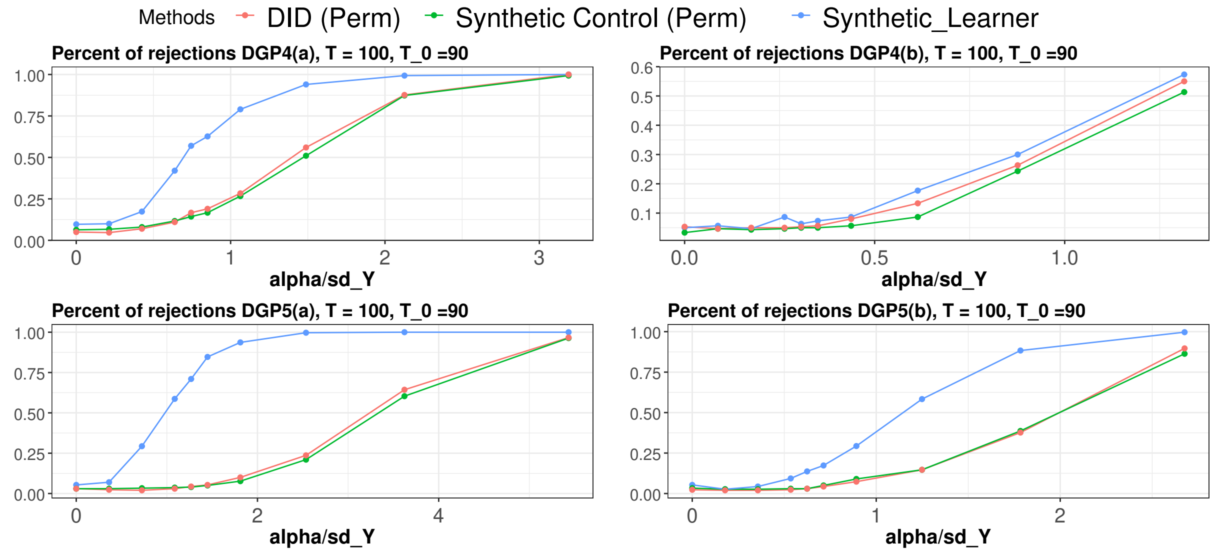

Figure E.1: $T = 100, p = 10$ Percentage of rejections of the null hypothesis of no treatment effects over 300 repetitions. The x-axis reports the policy's effect rescaled by the outcome's standard deviation. Synthetic learner has XGboost,Support Vector Regression and ARIMA(0,1,1) and 50 additional non informative predictions. The blue line denotes the proposed method; red line denotes the difference-in-differences and green line denotes the Synthetic Control method.

E.3 Oracle Study: Known Critical Value

In order to better understand the drivers of the power performance, we study the case where the critical value of the test is known to the researcher; we estimated it by Monte Carlo simulation since no closed-form expression for its density is available. Figure E.4 collects our results. Synthetic Learner does not have uninformative learners, and the class consists of XGboost, Support Vector Regression, and ARIMA(0,1,1). We take $T = 300$, $T_0 = 280$, and we let $T_{-} = 140$.

Since the critical quantile are assumed to be known, SC and DiD are estimated using information until time T_0 (Abadie et al., 2010), and not on the full sample as imposed by permutation methods. The proposed method outperforms uniformly DiD, and SC in almost all **DGP**s considered.

Additional results on the oracle study are included in Figure E.5.

E.4 Endogenous Time of the Treatment

Additional results for the case of endogenous time of treatment are discussed in Figure E.6, which are consistent with results in the main text.

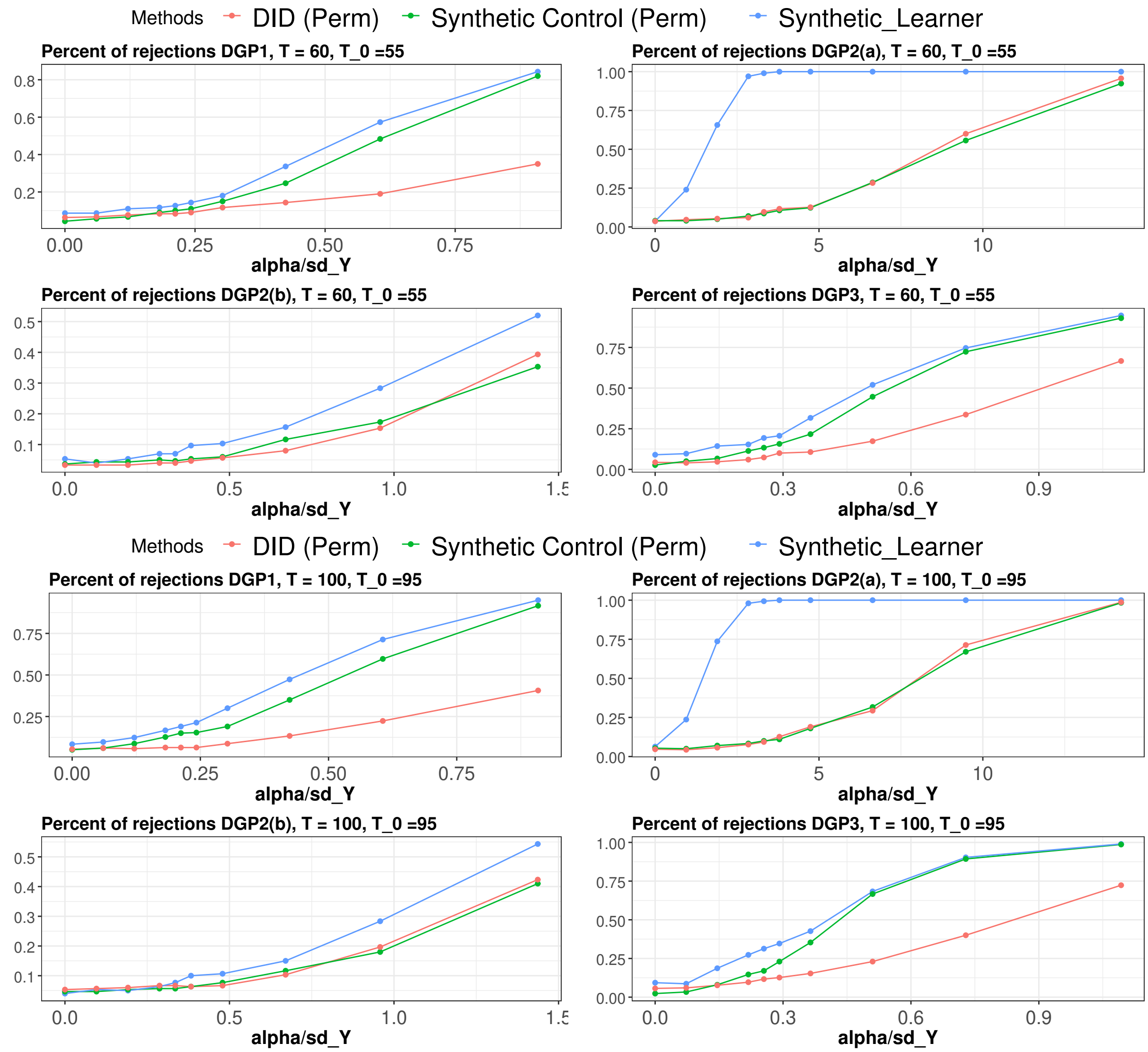

Figure E.2: $T \in \{60, 100\}, T^* = 5, p = 10$ Percentage of rejections of the null hypothesis of no treatment effects over 300 repetitions. The x-axis reports the policy's effect rescaled by the outcome's standard deviation. Synthetic learner has XGboost,Support Vector Regression and ARIMA(0,1,1) and 50 additional non informative predictions. The blue line denotes the proposed method; red line denotes the difference-in-differences and green line denotes the Synthetic Control method.

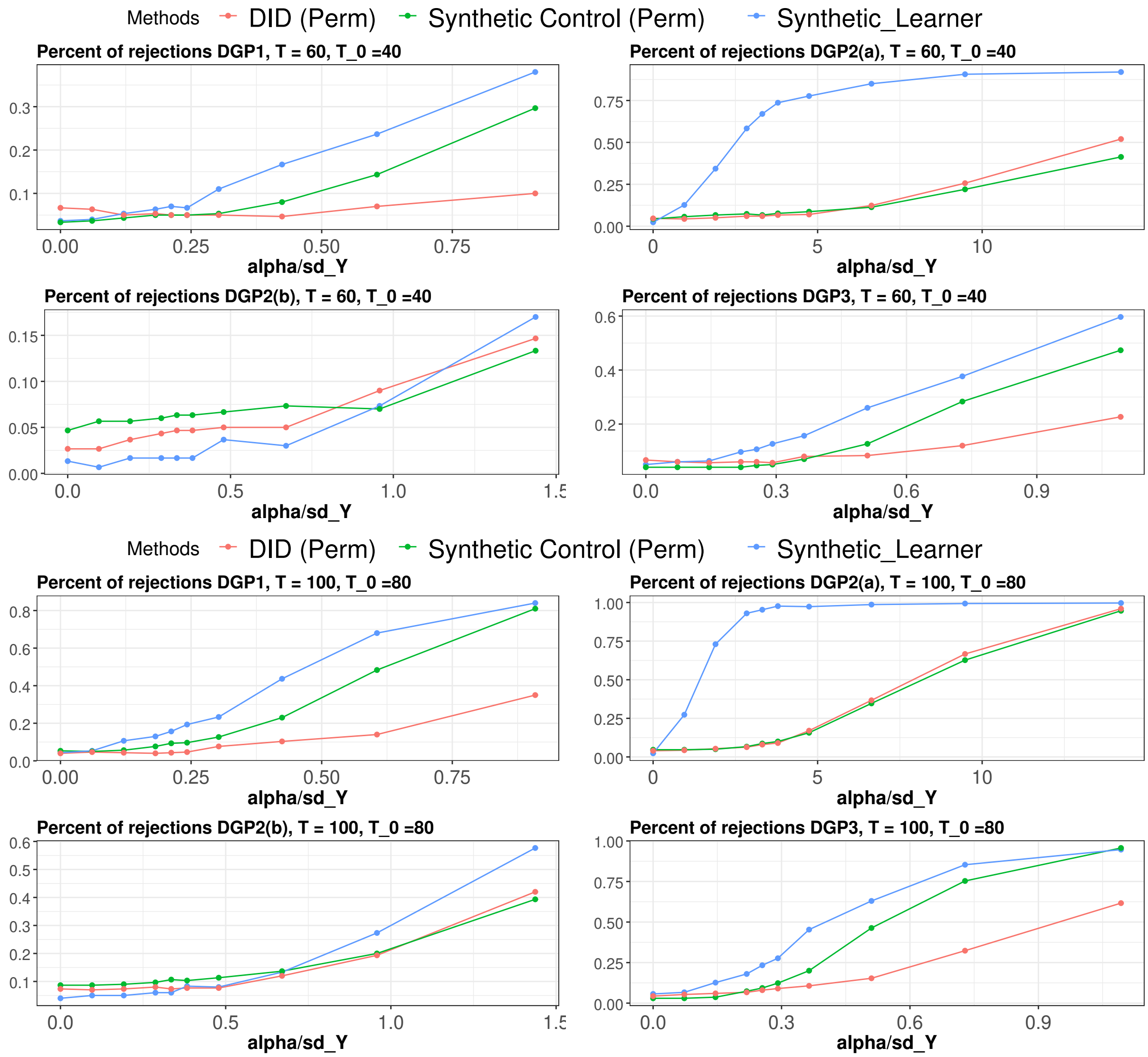

Figure E.3: $T \in \{60, 100\}, T^* = 20, p = 10$ Percentage of rejections of the null hypothesis of no treatment effects over 300 repetitions. The x-axis reports the policy's effect rescaled by the outcome's standard deviation. Synthetic learner has XGboost,Support Vector Regression and ARIMA(0,1,1) and 50 additional non informative predictions. The blue line denotes the proposed method; red line denotes the difference-in-differences and green line denotes the Synthetic Control method.

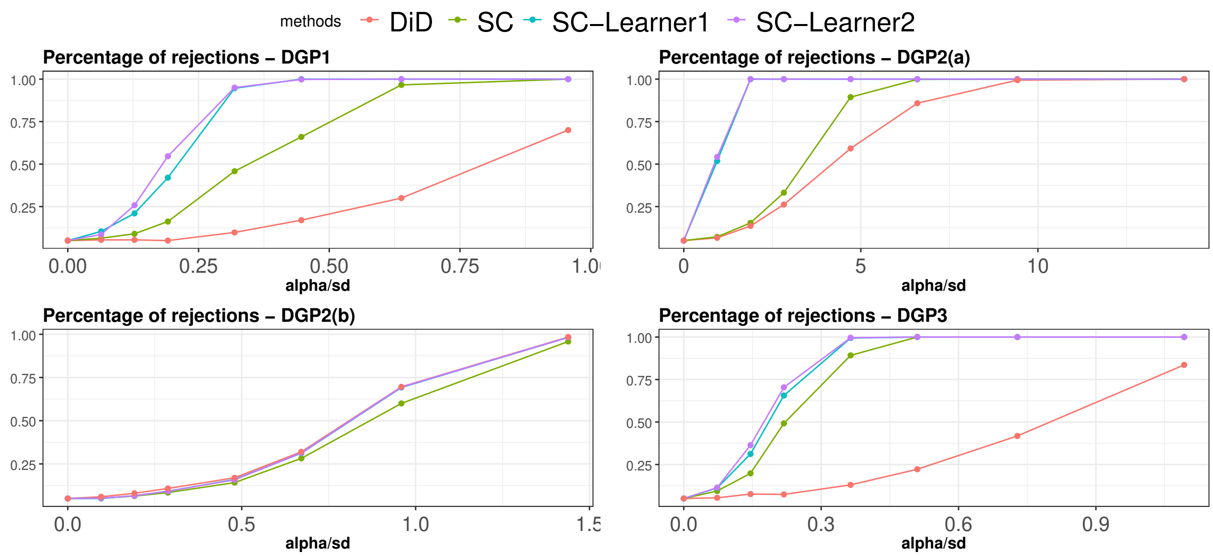

Figure E.4: Percentage of rejections over 500 repetitions with $T = 300$, $T_0 = 280, p = 50$, and $T_- = 140$ when the critical quantile is *known*(oracle case). The x-axis reports the policy's effect rescaled by the outcome's standard deviation. SC and DiD are estimated using all information until time T_0(no permutation test required). SC-Learner1 is the Synthetic Learner trained with XGboost, Support Vector Regression, ARIMA(0,1,1) and 50 additional non informative predictions. SC-Learner2 is the Synthetic Learner which also includes classical SC and it does not include non-informative predictions. The red line corresponds to the DiD method; the yellow line corresponds to the SC method; the blue line is SC-Learner 1 method and the purple line in SC-Learner 2.

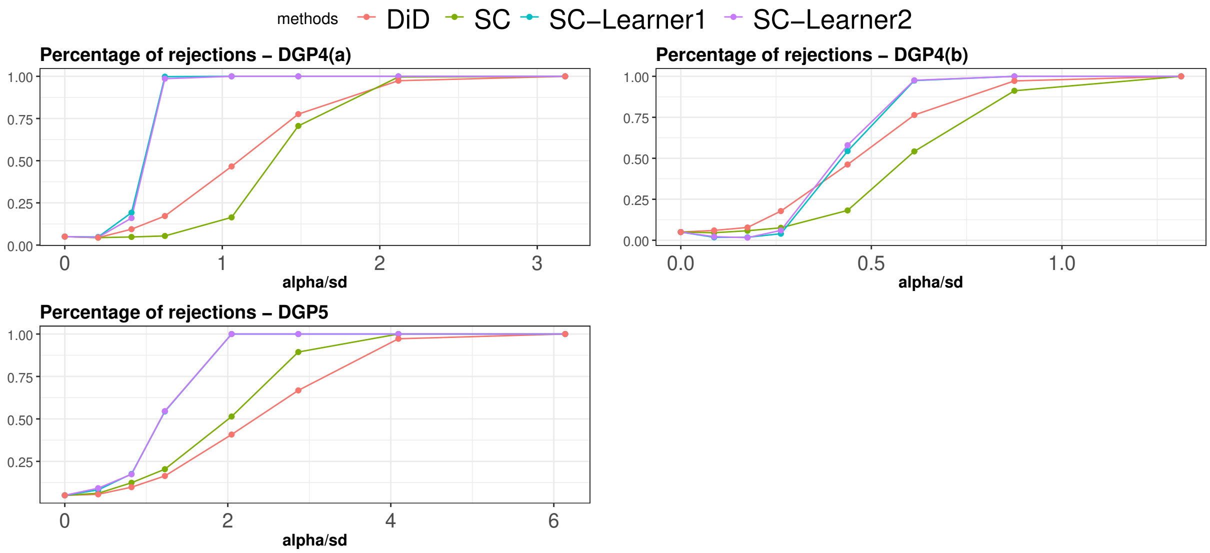

Figure E.5: Percentage of rejections over 500 repetitions with $T = 300$, $T_0 = 280$, $p = 50$ and $T_{-} = 140$ when the critical quantile is *known*(oracle case). The x-axis reports the policy's effect rescaled by the outcome's standard deviation. SC and DiD are estimated using all information until time T_0(no permutation test required). SC-Learner1 is the Synthetic Learner trained with XGboost, Support Vector Regression, ARIMA(0,1,1) and 50 additional non informative predictions. SC-Learner2 is the Synthetic Learner which also includes classical SC and it does not include non-informative predictions. The red line corresponds to the DiD method; the yellow line corresponds to the SC method; the blue line is SC-Learner 1 method and the purple line in SC-Learner 2.

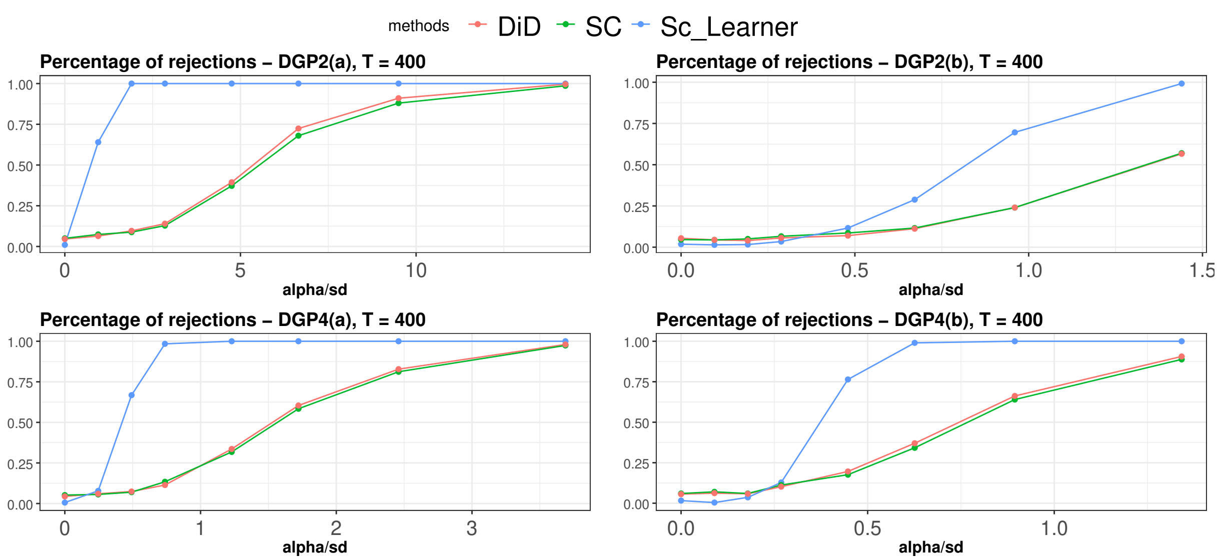

Figure E.6: Percentage of rejections over 500 repetitions with $T = 400$, $T_0 = 280$, $J = 50$ and $T_{-} = 140$ when the critical quantile is estimated via resampling and the time of treatment is *endogenous*..

F Empirical Analysis: Additional Results

In Table F.1 we report results when all states with the exception of Oregon are considered, since Oregon also implemented a Medicaid reform. Table F.2 F.3 collect results for different choices of the tuning parameter η and show robustness of the results.

Table F.1: 90% and 80% critical values, t-statistic and ATT using all the states as controls. The effect estimated is over the time window 2010-2014 (first row), and consecutive windows 2011-2014 ($m = 1yr$), 2012-2014 ($m = 2yr$), 2013-2014 ($m = 3yr$). Period 1 collects results when learners are estimated using the window between 1998-2006 and weights are estimated over the period 1993-1997. Period 2 corresponds to the opposite scenario. Demeaned SL denotes the SL with time-varying fixed effects.

Period 1	SL				Demeaned SL			
	CV90	CV80	t stat	ATT	CV90	CV80	t stat	ATT
m = 0	1.723	1.598	1.883	3.525	1.671	1.548	1.820	3.503
m = 1yr	1.664	1.525	1.830	3.754	1.583	1.455	1.754	3.701
m = 2yr	1.614	1.477	1.829	3.973	1.529	1.401	1.741	3.892
m = 3yr	1.535	1.389	1.776	4.098	1.450	1.313	1.679	3.991
Period 2	SL				Demeaned SL			
	CV90	CV80	t stat	ATT	CV90	CV80	t stat	ATT
m = 0	1.707	1.574	1.279	5.840	1.086	1.007	1.049	5.520
m = 1yr	1.642	1.514	1.221	6.040	1.027	0.948	0.990	5.708
m = 2yr	1.609	1.481	1.226	6.243	1.001	0.921	0.995	5.901
m = 3yr	1.555	1.420	1.194	6.351	0.964	0.883	0.972	6.001

In Figure F.1, we plot the observed outcome and the estimated counterfactual for the percentage of individuals not having economic access to health-care.

Table F.2: 90% and 80% critical values, t-statistic and ATT using the southern states as controls and $\eta = 1/(T\text{Var}(Y_t))$. The effect estimated is over the time window 2010-2014 (first row), and consecutive windows 2011-2014 ($m = 1yr$), 2012-2014 ($m = 2yr$), 2013-2014 ($m = 3yr$). Period 1 collects results when learners are estimated using the window between 1998-2006 and weights are estimated over the period 1993-1997. Period 2 corresponds to the opposite scenario. Demeaned SL denotes the SL with time-varying fixed effects.

Period 1	SL				Demeaned SL			
	CV90	CV80	t stat	ATT	CV90	CV80	t stat	ATT
m = 0	1.321	1.237	1.328	1.786	0.901	0.844	0.834	0.969
m = 1yr	1.237	1.155	1.253	1.932	0.787	0.743	0.728	1.034
m = 2yr	1.200	1.114	1.244	2.096	0.751	0.706	0.695	1.077
m = 3yr	1.151	1.057	1.200	2.178	0.709	0.659	0.648	1.063
Period 2	**SL**				**Demeaned SL**			
	CV90	CV80	t stat	ATT	CV90	CV80	t stat	ATT
m = 0	0.713	0.665	0.321	3.280	0.521	0.473	0.246	3.137
m = 1yr	0.642	0.597	0.231	3.297	0.462	0.416	0.166	3.142
m = 2yr	0.631	0.586	0.220	3.365	0.453	0.408	0.157	3.204
m = 3yr	0.618	0.571	0.214	3.392	0.452	0.405	0.152	3.198

Table F.3: 90% and 80% critical values, t-statistic and ATT using the southern states as controls and $\eta = 1\%$. The effect estimated is over the time window 2010-2014 (first row), and consecutive windows 2011-2014 ($m = 1yr$), 2012-2014 ($m = 2yr$), 2013-2014 ($m = 3yr$). Period 1 collects results when learners are estimated using the window between 1998-2006 and weights are estimated over the period 1993-1997. Period 2 corresponds to the opposite scenario. Demeaned SL denotes the SL with time-varying fixed effects.

Period 1	SL				Demeaned SL			
	CV90	CV80	t stat	ATT	CV90	CV80	t stat	ATT
m = 0	1.336	1.254	1.362	1.865	0.895	0.841	0.827	0.988
m = 1yr	1.256	1.170	1.292	2.023	0.792	0.744	0.720	1.054
m = 2yr	1.221	1.131	1.286	2.197	0.755	0.709	0.686	1.094
m = 3yr	1.155	1.062	1.245	2.289	0.713	0.659	0.641	1.082
Period 2	**SL**				**Demeaned SL**			
	CV90	CV80	t stat	ATT	CV90	CV80	t stat	ATT
m = 0	1.281	1.186	0.513	4.690	0.547	0.496	0.236	3.143
m = 1yr	1.219	1.127	0.441	4.805	0.486	0.437	0.157	3.161
m = 2yr	1.197	1.108	0.439	4.928	0.485	0.431	0.149	3.225
m = 3yr	1.164	1.074	0.427	4.970	0.480	0.427	0.144	3.211

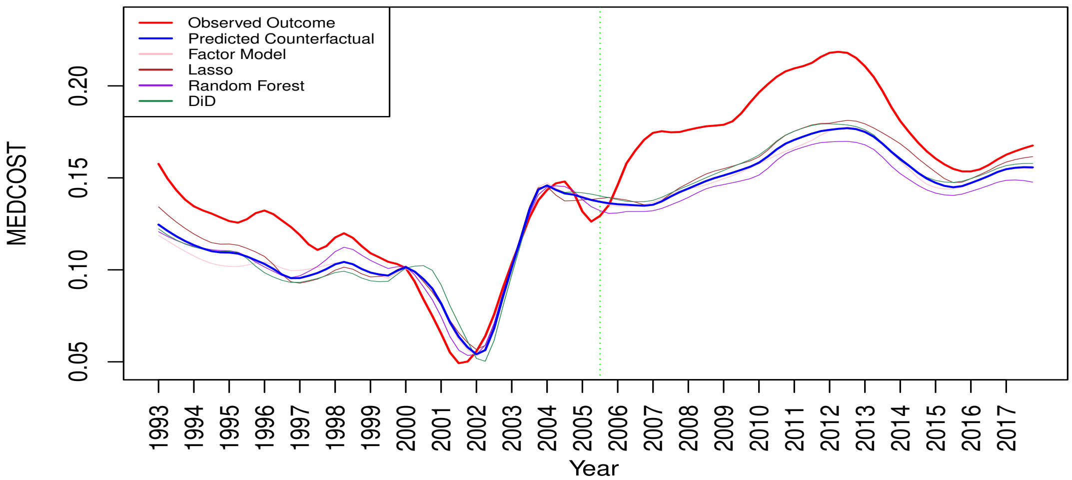

Figure F.1: Observed and predicted counterfactual of the percentage of childless adults who are not able to afford health-care expenses with demeaned SL and southern states as controls. The red line is the observed time series in Tennessee, and the blue line is the estimated counterfactual under no dis-enrollment. The other lines are predictions of the counterfactual for each base-learner used. In the plot, the time series are smoothed using a local polynomial.

G Additional Algorithms

Algorithm G.1 Synthetic Control Bootstrap

Require: Observations $\{Y_t, X_t\}_{t=T_-}^{T}$, time of the treatment-T_0, carryover effect size-m, tuning parameter $\eta > 0$, learners $f_1, \ldots, f_p$, null hypothesis values $\{a_t^o\}_{t>T_0}$

Split the pre-treatment period into two parts: $t \in [T_-, 0]$ and $t \in [1, T_0]$;

Form predictions $\hat{g}_j = f_j(\{Y_t, X_t\}_{t<1})$, $j \in 1, \ldots, p$;

Compute Y_{0t}^o for all t

for $b = 1, \ldots, B$ **do**

  Return a sample $\{\tilde{Y}_t^*, X_t^*\}$ of size T by performing circular block bootstrap on $\{Y_{0t}^o, X_t\}$ for $t \in \{1, \ldots, T\}$

  Compute $\widehat{\mathbf{w}}_0^*$ on $\{\tilde{Y}_t^*, \hat{\mathbf{g}}(X_t^*)\}_{1 \le t \le T_0}$ according to (11)

  Compute the predicted counterfactual $\hat{Y}_{0t}^{0^*} = \sum_{i=1}^{p} \widehat{\mathbf{w}}_0^{*(i)} \hat{g}_i^0(X_t^*)$, $t > T_0$.

  Compute the test statistics of interest for sharp and average null

$$\mathcal{T}_S^b = (T - T_0)^{-1/2} \sum_{t>T_0} \left(\tilde{Y}_t^* - \hat{Y}_{0t}^{0^*}\right)^2 \qquad \text{or} \qquad \mathcal{T}_A^b = (T - T_0)^{-1} \left(\sum_{t>T_0} \tilde{Y}_t^* - \hat{Y}_{0t}^{0^*}\right)^2;$$

end for

  return $q_{1-\alpha}^*$ as $(1-\alpha)$-th quantile of the sample

$$\mathcal{T}_S^1, \mathcal{T}_S^2, \ldots, \mathcal{T}_S^B \qquad \text{or} \qquad \mathcal{T}_A^1, \mathcal{T}_A^2, \ldots, \mathcal{T}_A^B.$$

References

Abadie, A., A. Diamond, and J. Hainmueller (2010). Synthetic control methods for comparative case studies: Estimating the effect of california's tobacco control program. *Journal of the American statistical Association 105*(490), 493–505.

Boucheron, S., G. Lugosi, and P. Massart (2013). *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press.

Bradley, R. C. et al. (2005). Basic properties of strong mixing conditions. a survey and some open questions. *Probability surveys 2*, 107–144.

Cesa-Bianchi, N., G. Lugosi, et al. (1999). On prediction of individual sequences. *The Annals of Statistics 27*(6), 1865–1895.

Kosorok, M. R. (2008). *Introduction to empirical processes and semiparametric inference.* Springer.

Lunde, R. and C. R. Shalizi (2017). Bootstrapping generalization error bounds for time series. *arXiv preprint arXiv:1711.02834*.

Van der Vaart, A. W. (2000). *Asymptotic statistics*, Volume 3. Cambridge university press.

Van Der Vaart, A. W. and J. A. Wellner (1996). Weak convergence. In *Weak convergence and empirical processes*, pp. 16–28. Springer.

White, H. (2014). *Asymptotic theory for econometricians.* Academic press.