SYNTHETIC LEARNER: MODEL-FREE INFERENCE ON TREATMENTS OVER TIME

DAVIDE VIVIANO AND JELENA BRADIC

Department of Economics and Department of Mathematics, University of California at San Diego

Abstract. Understanding of the effect of a particular treatment or a policy pertains to many areas of interest – ranging from political economics, marketing to health-care and personalized treatment studies. In this paper, we develop a non-parametric, model-free test for detecting the effects of treatment over time that extends widely used Synthetic Control tests. The test is built on counterfactual predictions arising from many learning algorithms. In the Neyman-Rubin potential outcome framework with possible carry-over effects, we show that the proposed test is asymptotically consistent for stationary, beta mixing processes. We do not assume that class of learners captures the correct model necessarily. We also discuss estimates of the average treatment effect, and we provide regret bounds on the predictive performance. To the best of our knowledge, this is the first set of results that allow for example any Random Forest to be useful for provably valid statistical inference in the Synthetic Control setting. In experiments, we show that our Synthetic Learner is substantially more powerful than classical methods based on Synthetic Control or Difference-in-Differences, especially in the presence of non-linear outcome models.

Keywords: Synthetic Control, Difference In Differences, Causal Inference, Random Forests.

E-mail address: jbradic@ucsd.edu.

1. Introduction

In this paper, we consider the problem of estimating and testing the effect of a treatment or policy of interest on a single unit observed over multiple periods. We consider the framework where one single unit, such as an individual, firm or state, observed over multiple periods, is exposed to treatment from one point in time onwards. In economics and more broadly in social sciences, for instance, treatments are often associated with changes in policy regimes. Examples are changes in legislation or implementation of new welfare programs. In marketing, on the other hand, practitioners often run over long periods geo-localized experiments for testing advertising campaigns, which requires precise knowledge for estimating and testing treatment effects with time-dependent data (Vaver and Koehler, 2011; Brodersen et al., 2015; Varian, 2016).

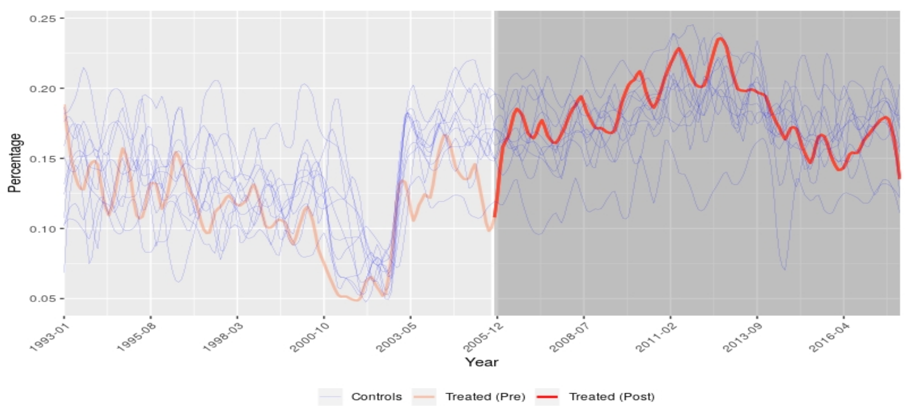

Figure 1. BRFFS data. Percentage of adults in Tennessee (treated unit, red) and southern states (controls, blue) who could not afford health-care from January 1993 to December 2017. Y-axes is the percentage of individuals (without children) who answered yes to the following survey question: "Was there a time in the past 12 months when you needed to see a doctor but could not because of the cost?." Individuals were disenrolled from public health insurance between 2005 and 2006.

In this paper, we advocate for a model-free, flexible approach, where we would be able to detect effects of treatment even when the underlying outcome model is misspecified. To illustrate the benefits of our approach, consider the following BRFFS data about U.S. residents regarding their health-related risk behaviors[1]. The question of interest is whether 2005 disenrollment (TennCare) program from public health insurances in Tennessee had a significant impact on the access to health-care services in Tennessee, after controlling for the access to health care in other states that did not experience changes in the policy. Figure 1 illustrates time-dependent data that exhibits a potentially highly non-linear relationship between health-care affordability in Tennessee and health-care in other southern states. Even if we believe in an economy driven by a factor model, the relation between observations and latent factors might be highly non-linear and unknown to the researcher. Pinpointing such dependences is nearly an impossible task.

Inferential approaches for Synthetic Control include a widely successful placebo testing as introduced in (Abadie et al., 2010, 2015). Firpo and Possebom (2018) interpret such tests as Fisher randomization tests whose validity requires strong symmetry assumptions which in practice are often violated; see, e.g., Hahn and Shi (2017). Although these methods perform well in settings

[1] Behavioral Risk Factor Surveillance System Data can be freely downloaded from https://www.cdc.gov/brfss/annual_data/annual_data.htm.

where the model is correctly specified, they quickly break down when the model is misspecified. In this paper, we explore the use of ideas from the machine learning literature to improve the Synthetic Controlinferential tools whenever the model is not fully captured.

However, there are essential hurdles that need to be cleared before methods like Random Forest can be directly useful for inference in Synthetic Control. Construction of a valid test is particularly difficult as machine-learning methods do not allow for tractable asymptotic theory. Moreover, the dependence structure of the data gives rise to substantial challenges from causal identification perspective. We formalize the problem of inference in quasi-experimental designs under the Neyman-Rubin potential outcome framework. We generalize the notion of treatment effects for time-dependent data by allowing for carry-over effects, i.e., treatment assignment is possibly dependent on past treatment assignments. The framework rephrases causal estimands as a function of the entire treatment history and leans on the recent literature on treatments over time that include, among others, Iavor Bojinov (2019); Robins et al. (2000); Blackwell and Glynn (2018); Hernán and Robins (2015); Abraham and Sun (2018); Athey and Imbens (2018); Imai et al. (2018), while it brings substantial innovation in the definition of the causal estimands of interest with respect to the past literature on the Synthetic Control.

A treatment is implemented at time T_0 and researchers observe $\{Y_t, X_t\}$, the outcome, and the covariates, respectively; X_t includes units that are not exposed to the treatment or further information on the outcome itself. Our approach starts by considering a general, fully non-parametric estimator, of the likes of Random Forest or kernel smoothing or Neural Network. We allow for a construction of many different kinds of estimators, i.e., learners; $\widehat{\mathbf{g}} = (\hat{g}_1, \dots, \hat{g}_p)^\top$. We propose an estimator and a test statistic designed as learning under experts advice, where very many learners are utilized to form the counterfactual predictions. Our construction utilizes weights that are constructed through a potential function satisfying the Blackwell condition (see Chapter 2 in Cesa-Bianchi and Lugosi (2006)). Given the weights, $\widehat{\mathbf{w}}_0 = (\hat{w}_0^{(1)}, \dots, \hat{w}_0^{(p)})^\top$ and the learners $\widehat{\mathbf{g}}_0$, constructed on a pre-treatment period, $t \leq T_0$, we propose an estimate of the average treatment effect (ATE) as

$$\widehat{\text{ATE}} = {T_m}^{-1} \sum_{t \geq T_0+m} (Y_t - \hat{Y}_t^0) - (T_0)^{-1} \sum_{t=1}^{T_0} (Y_t - \hat{Y}_t^0). \tag{1}$$

with counterfactual predictions computed on a post-treatment period

$$\hat{Y}_t^0 = \widehat{\mathbf{w}_0}^\top \widehat{\mathbf{g}_0}(X_t)$$

where $T_m = T - T_0 - m$ and m denotes the length of the carryover effects". The first term in (1) computes the difference between the average outcome and the predicted counterfactual outcome on the post-treatment period. The second term in (1) subtracts any additive bias due to possible model misspecification. Section 2 contains complete framework.

Our main result regarding inference on treatment effects shows that consistency of a test statistic; for the case of testing whether the average treatment effect is zero, it takes the form of

$$T = T_m^{-1/2} \sum_{t=T_0+m}^{T} \left(Y_t - \hat{Y}_t^0 \right)^2.$$

We showcase that the validity of our test does not depend on the underlying model being correctly specified and it does not require strong conditions on the learners (e.g., consistency for example). Even in the simple case of one learner, may it be a Random Forest or a Neural Network, our result considerably broadens those achieved in the existing literature. The main result of Chernozhukov et al. (2018) is that if Lasso is able to estimate well the underlying model, then a variant of the test above leads to a valid test. However, in observational studies, where the true model is rarely known, the results of Chernozhukov et al. (2018) depend on the rate at which we can estimate the model well.

We also provide guarantees on the one-step-ahead regret bound, i.e., the difference between the expected mean squared error and the best mean squared error that could be achieved from deploying any single learner in a given class. Mean squared error over time is defined as an average of one-step-ahead prediction errors. One of our results is that, under mild regularity assumptions, such regret scales logarithmically with the number of learners used and is of the order of $\sqrt{\log(p)/T}$; comparing favorably to *i.i.d.*, linear rates of Künzel et al. (2019). This result does not require correctness of the specified model. We allow possible non-stationary and time-dependent covariates. The reason we obtain strong guarantees is closely tied to the results in the literature on "learning under experts advise" (Cesa-Bianchi et al., 1999) as well as literature on exponential aggregation (Rigollet et al., 2012).

1.1. **Related work.** There has been a longstanding understanding that machine-learning methods, of the likes of Random Forests and Neural Network, are beneficial for predictive purposes. However, utilizing such methods for inference presents with fundamental challenges. The main contribution of this paper is a method enabling hypothesis testing in the context of Synthetic Controls (SC) using predictions arising from many machine-learning like methods. SC has been proven hugely successful in estimating counterfactuals in the presence of time-dependent outcomes.

An increasingly growing literature has considered the construction of valid confidence intervals in the context of SC. A result closes to us is that of Chernozhukov et al. (2017) who provide permutation based construction following constrained estimators of the likes of Abadie et al. (2010) for which they require stability at estimation. Permutation-based approaches are known however to lead to conservative statements when utilized with large post-treatment periods. Inference for SC with linear, high-dimensional outcome model was recently developed by Chernozhukov et al. (2018). Imai et al. (2018) discusses bootstrap approximations following K-nearest neighbors and matching. However, formal inferential justifications were not further discussed therein.

Our contribution can be viewed as complementary to this literature. To the best of our knowledge, we provide the first set of conditions under which predictions made by many-machine learning methods, including Random Forests and Neural Network, can be used to develop valid tests for SC. The difficulty of choosing which method to use often counterbalances the extensive choice of machine-learning methods. Our tools allow for predictions to be combined from many different algorithms simultaneously, therefore offering diversification gains over-relying on predictions arising from a single model only.

Our contribution belongs to the broader literature on machine learning methods for causal inference. Athey et al. (2018) discusses asymptotic properties on balancing estimators combined with regularized regression adjustments. Wager and Athey (2018) and Athey et al. (2019) derive asymptotic properties of Random Forests of Breiman (2001). Farrell et al. (2018) provides the theory for semiparametric inference on treatment effects with deep Neural Network whose validity relies on sufficiently fast rates of the risk bounds of the estimators. Orthogonal scores have been discussed in a serious of papers; see, e.g., Chernozhukov et al. (2018); Belloni et al. (2014, 2017). A specific limitation of this line of work is that it has not addressed, until now, time-dependent observations.

A large number of papers use machine-learning methods for estimating heterogeneous treatment effects. Athey et al. (2018) proposes matrix completion methods for Synthetic Control. Amjad et al. (2018) proposes singular value thresholding, whereas matching has been discussed by Imai and Kim (2019). Difference-in-difference methods were recently revised in Athey and Imbens (2018) and Arkhangelsky et al. (2018). Alternatives for SC include ridge-regression (Ben-Michael et al., 2018), kernel balancing (Hazlett and Xu, 2018), structural models with auto-regressive elements (Blackwell and Glynn, 2018), generalized SC (Xu, 2017; Doudchenko and Imbens, 2016). Outside SC, propensity score estimation via boosting (McCaffrey et al., 2004; Zhu et al., 2015) was proposed, together with deep learning for instrumental variables (Hartford et al., 2017) and covariate balancing (Kallus, 2018), bayesian regression trees (Hahn et al., 2017); support vector machines (Imai et al., 2013), causal boosting (Powers et al., 2017) and Bayesian non parametric methods (Taddy et al., 2016). For further reference, we refer to Knaus et al. (2018); Athey (2018).

Several papers average learners in the context of estimating treatment effects. Künzel et al. (2019) classifies such learners into three class, denoted as S, T, and X learners. While S-learners estimate the conditional expectation of the outcome variable by treating the set of covariates and the treatment assignments as features without giving the treatment a particular role, T-learners estimate the conditional expectation of the outcome, separately for treated and controls. The X-learner uses three steps based on a modification of the T-learner (Athey and Imbens, 2016). S-learners have been studied with BART (Green and Kern, 2010) and trees (Athey and Imbens, 2015). However, they all exclusively focus on *i.i.d.* observations. We pioneer the idea of T-learning and X-learning with time-dependent data, offering an alternative and simple weighting scheme which is inspired by the literature on boosting (Schapire and Freund, 2012) and online learning (Cesa-Bianchi et al., 1997, 1999; Cesa-Bianchi and Lugosi, 2006). We aggregate learners using the exponential weighting scheme, and we derive inferential properties under such a scheme. Exponential weights are well known to have desirable properties for prediction, while little or no attention has been given in deriving inferential properties.

1.2. **Organization.** The paper is organized as follows. In Section 2 we introduce Synthetic Learner. In Section 3 we show that the algorithm guarantees nominal coverage under stationarity conditions and we discuss oracle regret guarantees. Section 4 discusses numerical experiments, showing that the method outperforms Synthetic Control and Difference-In-Difference under a wide variety of data generating processes. Section 5 discusses applications in health-economics.

2. Methodology

Throughout this article for each unit i we observe outcome variable Y_{it} and a binary treatment variable D_{it}. We denote with Y_{0t}, here and onward denoted with Y_t, the treated unit; the remaining $i = 1, \cdots, n$ units, $Y_{1t}, \ldots, Y_{nt}$ are the controls. Additional covariate information for each unit are denoted with Z_{it}. For the sake of simplicity we denote with $X_t = (Y_{1t}, \ldots, Y_{nt}, Z_{0t}^\top, Z_{1t}^\top, \ldots, Z_{nt}^\top) \in \mathbb{R}^J$. The treated unit is observed from T_-, that for notational convenience we let to be negative, it gets treated at time T_0 and remains treated until T_+; D_t denote the treatment assignment indicator for the treated unit, being equal to one if treatment is assigned at time t and zero otherwise.

2.1. **Potential Outcomes.** At each time step, t , we expose a single unit to either treatment, $D_t = 1$, or control, $D_t = 0$ and subsequently measure an outcome. We consider binary treatment and denote the treatment path up to time t as a vector $\mathbf{d}_t \in \{0, 1\}^t$, where the observed history is denoted with $(D_{T-}, \ldots, D_{T_0}, \ldots, D_t)$. In classical causal inference settings the treatment path is of length 1. However, for time-series observations a more general notion is desired. Following the potential outcomes framework we then posit the existence of potential outcomes $Y_t(\mathbf{d}_t)$, corresponding respectively to the response the treated subject would have experienced at time t while being exposed to the treatment assignment contained in the treatment path $\mathbf{d}_t$. Formulating treatments and potential outcomes as paths was introduced initially by Robins (1986).

We impose no leads effect, namely future treatment assignment does not affect the current outcome. We also require that the realizations of potential outcomes do not depend on future treatment assignment as well as treatment assignment of past m lags or more. Using the same notation as in Rambachan and Shephard (2019), we formalize these conditions below.

Assumption 1. For $t \in \{T_-, \ldots, 0, 1, \ldots, T_0, \ldots, T_+\}$ we assume that the following holds

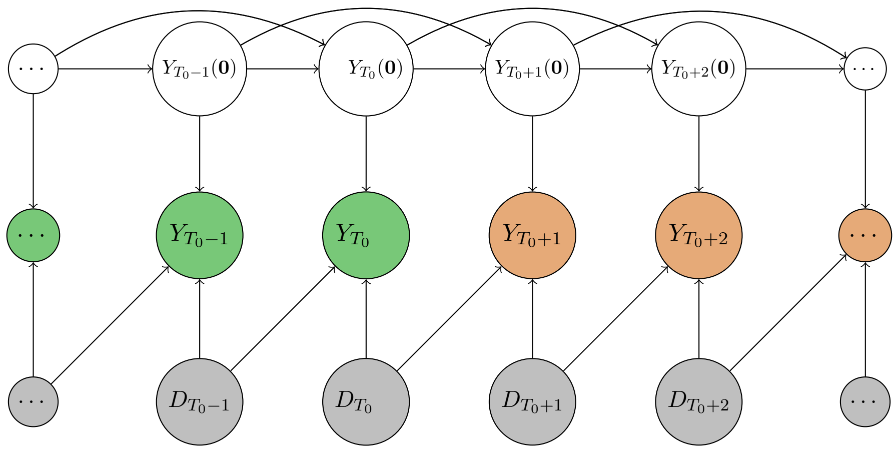

FIGURE 2. Graphical Interpretation of example (2) where $Y_t(\mathbf{0})$ is let to be a *potential-autoregression* of order two, i.e., depending on the past two lags $Y_{t-1}(\mathbf{0})$ and $Y_{t-2}(\mathbf{0})$ for example. We considered $m = 1$. Different color denotes pre and post-treatment periods. The observed outcome Y_t is equal to $Y_t(\mathbf{0})$ for $t \leq T_0$ and to $Y_t(\mathbf{0}) + \alpha_1 D_t + \alpha_2 D_{t-1}$ otherwise.

(a) (No Anticipation) for all $\mathbf{d}_{(t+1):T_+}, \mathbf{d}'_{(t+1):T_+} \in \{0,1\}^{T_+-t-1}$,

$$Y_t(\mathbf{d}_{T_-:t}, \mathbf{d}_{(t+1):T_+}) = Y_t(\mathbf{d}_{T_-:t}, \mathbf{d}'_{(t+1):T_+});$$

(b) (Carryover Effects) for all $\mathbf{d}_{T_-:(t-m)}, \mathbf{d}'_{T_-:(t-m)} \in \{0,1\}^{t-m+|T_-|}$,

$$Y_t(\mathbf{d}_{T_-:(t-m)}, \mathbf{d}_{(t-m+1):T_+}) = Y_t(\mathbf{d}'_{T_-:(t-m)}, \mathbf{d}_{(t-m+1):T_+}).$$

With a slight abuse of notation, we consider potential outcomes of the form $Y_t(\mathbf{d}_{(t-m):t})$. The no-anticipation assumption has been previously discussed in Abbring and Heckman (2007), Athey and Imbens (2018), while the restricted carryover effect is analogous to the identification assumption stated in Imai et al. (2018), Iavor Bojinov (2019), Blackwell and Glynn (2018) among others. Carryover effects have not been considered in previous literature on Synthetic Control.
Wrongly assuming the absence of carry-over effects can in fact lead to misspecified causal estimands and hence possibly biased estimates. For example, consider a simple case

$$Y_t(\mathbf{d}_{(t-m):t}) = Y_t(\mathbf{0}) + \sum_{s=0}^{m} \alpha_{s+1} d_{t-s} \tag{2}$$

for a sequence of constants $\alpha_s \in \mathbb{R}$. We provide a graphical illustration of this example (2) in Figure 2, by letting $Y_t(\mathbf{0})$ be dependent on the past two outcome $Y_{t-1}(\mathbf{0})$ and $Y_{t-2}(\mathbf{0})$ – "potential-autoregression" as in Iavor Bojinov (2019) – and setting $m = 1$ (for simplicity). In Figure 2 we showcase that the observed outcome depends on the potential outcome under no treatment and on present and past treatment assignment indicators. To see that a naive ATE estimate, defined as a difference between pre- and post-treatment averages is possibly biased, it suffices to observe that it's mean $|t : t > T_0|^{-1} \sum_{s=0}^{m-1} (m-s)\alpha_{s+1} + \sum_{s=0}^{m} \alpha_s$ is not necessarily equal to $Y_t(\mathbf{1}) - Y_t(\mathbf{0}) = \sum_{s=0}^{m} \alpha_{s+1}$. Another example illustrates a clear under-estimation of the naive ATE estimate. Consider a case where $Y_t(\mathbf{d}_{(t-m):t}) = Y_t(\mathbf{0}) + \alpha \prod_{s=0}^{m-1} d_{t-s}(1 - d_{t-m}) + \beta \prod_{s=0}^{m} d_{t-s}$ for some constants $\beta, \alpha \in \mathbb{R}$. Then, $Y_t(\mathbf{1}) - Y_t(\mathbf{0}) = \beta$. However, $Y_t(1, 1, \ldots, 0) - Y_t(\mathbf{0}) = \alpha$ and $Y_t(1, 0, \ldots, 0) - Y_t(\mathbf{0}) = 0$, therefore naive ATE estimate will underestimate the true treatment effect for all $\alpha < \beta$.

Throughout the paper we assume $1 < T_0 < T_+$, implying that the treated unit is observed to incur the treatment at some point in time[2]. In Section 4.3.2 we consider extensions where T_0 is random. Therefore, the realized outcome of Y_t, is the realization of $Y_t(\mathbf{0})$ for all $t \leq T_0$ while it is the realization of $Y_t(\mathbf{1})$ for all $t > T_0 + m$. The observations in between, are polluted by the carryover effects and contain a mixture of the two outcomes. Similarly to Abadie et al. (2010), Imai and Kim (2019), Chernozhukov et al. (2018), Athey and Imbens (2018), Doudchenko and Imbens (2016) among others we define point-wise treatment effects and we consider in addition a temporal average m-lag treatment effects

$$\mathcal{A}_t = Y_t(\mathbf{1}) - Y_t(\mathbf{0}), \quad \bar{\mathcal{A}} = T_m^{-1} \sum_{t=T_0+m}^{T_+} \mathcal{A}_t.$$

We consider testing hypothesis of interest related to $\mathcal{A}_t$ and $\bar{\mathcal{A}}$. The average treatment effect (ATE) denotes the difference between the effect of a policy the effect of the policy had it never been implemented. Intuitively, temporal average m-lag treatment effect captures the long-term effect of a policy. These estimands generalize treatment effects as studied in the literature on Synthetic Control, where potential outcome are defined only as a function of the current treatment assignment.

2.2. Estimation and Inference with Synthetic Learner.

2.2.1. *Estimation of ATE.* The basic idea of our synthetic learner can be explained in three steps. For notational convenience, we denote with Y_t^1 and Y_t^0 the potential outcomes $Y_t(\mathbf{1})$ and $Y_t(\mathbf{0})$, respectively. Split the pre-treatment period into three regions, denotes with $[T_-, 1], [1, T_0/2]$ and $[T_0/2 + 1, T_0]$.

1. Estimate the outcome of the treated unit for the duration of the first pre-treated period

$$E[Y_{0t}|Y_{1t}, \dots, Y_{nt}, Z_{0t}^\top, Z_{1t}^\top, \dots, Z_{nt}^\top] := E[Y_{0t}|X_t] = f(X_t), \qquad t \leq 1$$

   using time-dependent control samples and/or additional covariates. For a learning algorithm j, denote such estimate with $\hat{g}_j^0$. Form p such estimates by considering p different learning algorithms.

2. Learn the efficacy of the utilized learning algorithm, j, by constructing weights, $\hat{w}_0^{(j)}$, based on their out-of-sample prediction errors denoted with

$$\sum_{t=1}^{T_0/2} l\Big(Y_t, \hat{g}_j^0(X_t)\Big)$$

   for a chosen loss function l, and evaluated on the second pre-treatment period. Typical example is a quadratic loss, $\sum_{t=1}^{T_0}\Big(Y_t - \hat{g}_j^0(X_t)\Big)^2$. Such weights then mimic the importance of each learning method; larger weight corresponds to better predictions. Ideally, weights should accommodate a possibly large number of learning algorithms. Choice of weights is discussed in details in Section 2.3.

   Formulate the improved estimate of the counterfactual for the post-treatment period, by considering weighted average of the predictions formed in Step 1 and evaluated in the

[2]Most of the theoretical results in the literature on Synthetic Control are derived for T_0 being deterministic and treatment effects being fixed. The reader might refer to Arkhangelsky et al. (2018), Chernozhukov et al. (2018), Abadie et al. (2010) to cite some. Similarly, we consider T_0 as deterministic, fixed treatment effects while we let potential outcomes to be random. We extend the setting to random treatment effects and random T_0 in Section 4.3.2.

post-treatment period,

$$\hat{Y}_t^0 = \sum_{j=1}^{p} \hat{w}_0^{(j)} \hat{g}_j^0(X_t), \qquad t \geq T_0 + m \tag{3}$$

3. Define ATE estimate as the average between the predicted counterfactual outcome and the observed outcome,

$$T_m{}^{-1} \sum_{t \geq T_0+m} (Y_t - \hat{Y}_t^0) = T_m{}^{-1} \sum_{t \geq T_0+m} (Y_t^1 - \hat{Y}_t^0)$$

where $T_m = T_+ - t_0 - m$. The expression above computes the average difference between the observed outcome under the treatment regime and the predicted counterfactual under control. In the presence of model misspecification, this term may not consistently estimate the ATE. We make correction for model-misspecification by further subtracting the out-of-sample predictions of the pre-treatment period,

$$(T_0/2)^{-1} \sum_{t=T_0/2+1}^{T_0} (Y_t - \hat{Y}_t^0) := (T_0/2)^{-1} \sum_{t=T_0/2+1}^{T_0} (Y_t^0 - \hat{Y}_t^0).$$

This leads to a final ATE estimate

$$T_m{}^{-1} \sum_{t \geq T_0+m} (Y_t^1 - \hat{Y}_t^0) - (T_0/2)^{-1} \sum_{t=T_0/2+1}^{T_0} (Y_t^0 - \hat{Y}_t^0). \tag{4}$$

Sometimes it will be beneficial to estimate the learners closer to the time of the implemented treatment, T_0; weights and experts can be estimated in an opposite order without affecting the validity of our results. Moreover, we note that the choice of $T_0/2$ is chosen out of convenience. One can imagine a sample-splitting that takes many such splits; however, we work around this simple case for the duration of the article.

2.2.2. *From estimation to hypothesis testing.* Our main hypothesis of interest includes, sharp nulls,

$$H_0 : \mathcal{A}_t = a_t^o, \qquad a_t^o \in \mathbb{R}, \qquad t > T_0 + m.$$

However, we can incorporate a more general glass of hypothesis $H_0 : Y_t(\mathbf{1}) = f(Y_t(\mathbf{0}), a_t^o),\ a_t^o \in \mathbb{R},\ t > T_0 + m$. For function f is invertible in its first argument. If we are interested in testing the null hypothesis of no effect then $f(x, y) = x$. On the other hand testing the null hypothesis of a linear trend would require $a_t^o = \delta(t - T_0)$ for an arbitrary $\delta \in \mathbb{R}$ as well as $f(x, y) = x + y$. In addition, testing the null hypothesis of a constant multiplicative effect can simply be done with $f(x, y) = x \times y$ and $\alpha_t^o = \alpha$ for some constant α. Throughout our discussion, we denote Y_t^o the outcome after imposing the null hypothesis of interest on the outcome under control. That is, $Y_t^o = Y_t$ for $t \leq T_0$, while $Y_t^o = f^{-1}(Y_t, a_t^o)$ for $t > T_0 + m$. Consider the simple example of testing $Y_t(\mathbf{1}) - Y_t(\mathbf{0}) = a_t^o$. Then $Y_t^o = Y_t - a_t^o$ for $t > T_0 + m$. For a multiplicative case instead, namely for testing $Y_t(\mathbf{1}) = a_t^o Y_t(\mathbf{0})$ for some $a_t^o \neq 0$, $Y_t^o = Y_t / a_t^o$ for $t > T_0 + m$.

We are also interested in the average null hypothesis

$$H_0 : \bar{\mathcal{A}} = a^o, \qquad a^o \in \mathbb{R}.$$

While sharp null hypothesis implicitly assumes that the missing value of the potential outcome is known under the null, the average nulls, do not state conditions on the missing value of the potential outcome at each point in time. Average and sharp nulls as described above are discussed, for example, in Imbens and Rubin (2015).

Algorithm 1 Synthetic ATE Learner

Require: Observations $\{Y_t, X_t\}_{t=T_-}^{T_+}$, time of the treatment-$T_0$, carryover effect size-$m$, tuning parameter $\eta > 0$, learners $f_1, \dots, f_p$

1: Split the pre-treatment period into two parts: $t \in [T_-, 1]$ and $t \in [2, T_0]$

2: Form predictions $\hat{g}_j^0 = f_j(\{Y_t, X_t\}_{t<1})$, $j \in 1, \dots, p$.

3: Use the second pre-treatment period, $\{Y_t, X_t\}_{t=1}^{T_0}$, to estimate the weights of the learners, $\widehat{\mathbf{w}_0} = \hat{\mathbf{w}}_0(1, T_0)$ with jth entry of $\hat{\mathbf{w}}_d(u, v)$, $\hat{\mathbf{w}}_d^{(j)}(u, v)$, defined as follows

$$\hat{\mathbf{w}}_d^{(j)}(u,v) = \frac{\exp\left\{-\eta \sum_{t=u}^{v} l\Big(Y_t, \hat{g}_j^d(X_t)\Big)\right\}}{\sum_{i=1}^{p} \exp\left\{-\eta \sum_{t=u}^{v} l\Big(Y_t, \hat{g}_i^d(X_t)\Big)\right\}}, \quad d \in \{0,1\} \tag{5}$$

where l denotes a loss function that evaluates prediction quality.

4: Compute the predicted counterfactual

$$\hat{Y}_t^0 = \textstyle\sum_{j=1}^{p} \hat{w}_0^{(j)} \hat{g}_j^0(X_t) \text{ for } t \geq T_0 + m$$

5: Compute the out-of-sample pre-treatment prediction

$$\hat{Y}_t^0 = \textstyle\sum_{j=1}^{p} \hat{w}_0^{(j)} \hat{g}_j^0(X_t) \text{ for } t \in [T_0/2 + 1, T_0]$$

return Estimate the average treatment effect

$$\widehat{\text{ATE}} = T_m{}^{-1} \sum_{t \geq T_0 + m} (Y_t^1 - \hat{Y}_t^0) - (T_0/2)^{-1} \sum_{t=T_0/2+1}^{T_0} (Y_t^0 - \hat{Y}_t^0). \tag{6}$$

Given predictions of the counterfactuals, $\hat{Y}_t^0$ is as in (3), we consider the following test statistics regarding the sharp nulls

$$\mathcal{T}_S = T_m^{-1/2} \sum_{t > T_0 + m} \left(Y_t^o - \hat{Y}_t^0\right)^2. \tag{7}$$

High values of statistic $\mathcal{T}_S$ would indicate departures from the null hypothesis. When considering the average null hypothesis, we consider the test statistic $\mathcal{T}_A$ where

$$\mathcal{T}_A^{1/2} = T_m^{-1/2} \sum_{t > T_0 + m} \left(Y_t^o - \hat{Y}_t^0\right). \tag{8}$$

Under Assumption 1 we let potential outcomes depend on the past m treatment assignment indicators. The predictions, $\hat{Y}_t^0$, are from pre- while the statistic is evaluated on the post-treatment period at least m periods ahead of the initial implementation of the policy. Treatment effect over the period $T_0 + 1, \dots, T_0 + m$ denotes "short-run effects", while the treatment effect on the subsequent period quantifies "long-run" effects. This distinction, whereas simple and intuitive, is an important novelty with respect to previous literature on Synthetic Control. We show that the test statistics proposed above are theoretically valid regardless of the misspecification bias of $\hat{Y}_t^0$. Such robustness property justifies why we do not need bias-adjustments to the test statistic or learners themselves.

Building on recent literature on sample splitting for inference recently revisited in the context of $i.i.d$ data in Rinaldo et al. (2016) and Fithian et al. (2014) among others, we construct a consistent method for estimating the critical value of the proposed tests. We use a portion of the data to train learners, while the remaining observations are used for the bootstrap estimate of the critical value. We only estimate learners once and not on each bootstrapped sample. In Figure 3 we provide an intuitive explanation of the testing protocol.

Each bootstrapped sample of size $T_+ - m$ is designed across the time-index after the null hypothesis is imposed and utilizes circular block bootstrap; it "wraps" observations in a circle, creates

Algorithm 2 Synthetic learner for testing sharp nulls

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

1: Split the pre-treatment period into two parts: $t \in [T_-, 1]$ and $t \in [2, T_0]$;
2: Form predictions $\hat{g}_j^0 = f_j(\{Y_t, X_t\}_{t<1})$, $j \in 1, \dots, p$;
3: Compute $\widehat{\mathbf{w}_0}$ on $\{Y_t, \hat{g}_1^0(X_t), \dots, \hat{g}_p^0(X_t)\}_{1<t\leq T_0}$ according to (5);
4: Compute the predicted counterfactual $\hat{Y}_t^0 = \sum_{i=1}^p \hat{\mathbf{w}}_0^{(i)} \hat{g}_i^0(X_t)$, $t > T_0 + m$.
5: Compute the outcomes under the null Y_t^o for $t > T_0 + m$ of the form $Y_t^o = f^{-1}(Y_t, a_t^o)$.
6: Compute the test statistics

$$\mathcal{T}_{\mathcal{S}} = T_m^{-1/2} \sum_{t=T_0+m}^{T_+} \left(Y_t^o - \hat{Y}_t^0\right)^2.$$

7: Run Bootstrap Algorithm 3 to obtain $\hat{q}_{1-\alpha}$;
return Reject the null hypothesis if $\mathcal{T}_{\mathcal{S}} > \hat{q}_{1-\alpha}$.

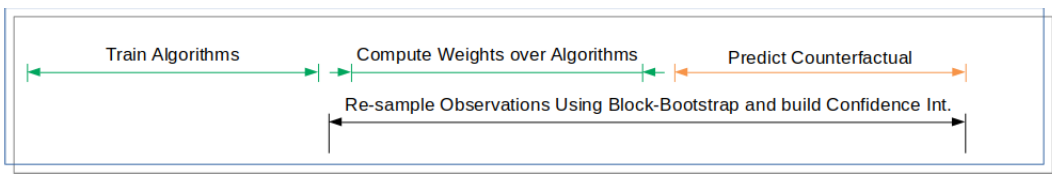

Figure 3. Testing for $m = 0$. Train learners on an initial sample; compute the weights on a consecutive block of observations and then predict the counterfactual. Green color denotes pre-treatment period while orange color denotes post-treatment period; bootstrap observations after imposing the null hypothesis.

blocks of size $b(T_+)$ of consecutive samples and draws blocks with replacements. When creating a circle, we remove part of the post-treatment we suspect has carry-over effects; in our notation $t \in [T_0+1, T_0+m]$. This procedure guarantees that observations, once reshuffled by the bootstrap, are consistent with the realized potential outcomes under control for any t. Details of the bootstrap are presented in Algorithm 3.

The circular block bootstrap was first introduced in Politis and Romano (1992), where the authors show consistency for approximating the distribution of the sample mean. Here we consider a more demanding setting with a test statistic which is a non-linear functional of the data. This creates additional challenges from a theoretical perspective that we discuss in the next Section.

2.3. **Potential weights.** Weights can be computed in several ways. Although least-squares have been considered (see, e.g., Künzel et al. (2019); Polley and Van Der Laan (2010)), it can perform poorly when the number of learners is large compared to the sample size. Equal weighting, on the other hand, can perform poorly in the presence of many uninformative learners. To equip the method to better learn in the presence of a large number of learns, some of which may potentially be ineffective, we discuss an alternative weighting scheme.

Before discussing the weighting scheme in detail, we need to introduce some necessary notation. We denote $\hat{Y}_t(\mathcal{F}_{t-1})$ the one step ahead prediction implying that weights are computed using information available only until time $t-1$. We define the i-th *regret* as the gain that the learner would have incurred if it predicted $\hat{g}_i^0(X_t)$ instead of $\hat{Y}_t$ at time t. We let be the cumulative regret vector, for the time period $\{t_1, \dots, t_2\}$, utilizing loss function l, $\mathbf{R}_{t_1,t_2}^0 = \left(\sum_{t=t_1}^{t_2} l(Y_t, \hat{Y}_t(\mathcal{F}_{t-1})) - l(Y_t, \hat{g}_1^0(X_t)), \dots, \sum_{t=t_1}^{t_2} l(Y_t, \hat{Y}_t(\mathcal{F}_{t-1})) - l(Y_t, \hat{g}_p^0(X_t))\right)^\top$. Potential weights

Algorithm 3 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 $f_1, \dots, f_p$, null hypothesis values $\{a_t^o\}_{t \geq T_0+m}$

Split the pre-treatment period into two parts: $t \in [T_-, 1]$ and $t \in [2, T_0]$;

Form predictions $\hat{g}_j^0 = f_j(\{Y_t, X_t\}_{t<1})$, $j \in 1, \dots, p$;

Compute Y_t^o for all t

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

  Return a sample $\{\tilde{Y}_t^*, X_t^*\}$ of size $T_+ - m$ by performing circular block bootstrap on $\{Y_t^o, X_t\}$ for $t \in \{1, \dots, T_+\} \setminus \{T_0+1, \dots, T_0+m\}$

  Compute $\widehat{\mathbf{w}}_0^*$ on $\{\tilde{Y}_t^*, \hat{\mathbf{g}}(X_t^*)\}_{1 \leq t \leq T_0}$ according to (5)

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

  Compute the test statistics of interest

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

end for

  return $\hat{q}_{1-\alpha}$ as $(1-\alpha)$-th quantile of the sample

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

are computed by evaluating the derivative of the potential function at the given regret vector and rescaling over the sum of the weight, as

$$\hat{\mathbf{w}}_0(t_1, t_2) = \frac{\nabla \Phi'(\mathbf{R}_{(t_1,t_2)}^0)}{\sum_{i=1}^p \nabla \Phi'(\mathbf{R}_{(t_1,t_2)}^0)_i}. \tag{9}$$

Here, the function $\Phi : \mathbb{R}^p \to \mathbb{R}$ is the *potential* function with $\Phi(\mathbf{u}) = \psi\left(\sum_{i=1}^p \phi(\mathbf{u}_i)\right)$ where $\phi : \mathbb{R} \to \mathbb{R}$ is non-negative, strictly increasing and twice differentiable function and $\psi : \mathbb{R} \to \mathbb{R}$ is nonnegative, strictly increasing, concave and twice differentiable *auxiliary* function. We illustrate some examples.

(*Exponential Weights*) Whenever $\phi(x) = \exp(\eta x)$ as well as $\psi(x) = \frac{1}{\eta}\log(x)$ potential weights of (9) match those of (5).

(*Polynomial Weights*) For $\Phi(\mathbf{u}) = ||\mathbf{u}_+||_q^2$ we obtain

$$\hat{\mathbf{w}}_0^{(i)}(1, T_0) = \frac{\left(\sum_{t=1}^{T_0} l(Y_t, \hat{Y}_t(\mathcal{F}_{t-1})) - l(Y_t, \hat{g}_i^0(X_t))\right)_+^{q-1}}{\sum_{j=1}^p \left(\sum_{t=1}^{T_0} l(Y_t, \hat{Y}_t(\mathcal{F}_{t-1})) - l(Y_t, \hat{g}_j^0(X_t))\right)_+}. \tag{10}$$

(*Following the Leader*) With $\Phi(\mathbf{u}) = \max(\mathbf{u}_1, \dots, \mathbf{u}_p)$ and by assuming that there are no ties we obtain

$$\hat{\mathbf{w}}_0^{(i)}(1, T_0) = \begin{cases} 1 \text{ if } \sum_{t=1}^{T_0-1} l(Y_t, \hat{g}_i^0(X_t)) - l(Y_t, \hat{g}_j^0(X_t))) > 0, & \forall j \neq i \\ 0 \text{ otherwise .} \end{cases} \tag{11}$$

As shown in the examples above, potential weights do not require any matrix inversion and can be computed easily. In the discussion that follows we showcase some key desirable properties. Following Theorem 2.1 of Cesa-Bianchi and Lugosi (2006) we can illustrate a simple result that guarantees

$$\frac{1}{T_0} \sum_{t=1}^{T_0} l(Y_t, \hat{Y}_t(\mathcal{F}_{t-1})) - \min_{i \in \{1,\dots,p\}} \frac{1}{T_0} \sum_{t=1}^{T_0} l(Y_t, g_{i,t}) \leq M \sqrt{\frac{2\log(p)}{T_0}}$$

for a choice of bounded and convex loss function l. Since $\hat{Y}_t(\mathcal{F}_{t-1})$ is estimated only on previous data and evaluated at X_t this notion of performance is rooted in out-of-sample performance metric.

Therefore, the cumulative loss incurred by our prediction converges to the cumulative loss incurred by the best learner in the class under consideration at rate $\sqrt{\log(p)/T_0}$. We can achieve similar bounds also with polynomial weights. For a polynomial weighting scheme, by choosing $q = 2\log(p)$ for $p > 2$, rate can be shown to be $\sqrt{2(\log(p) - 1)e/T_0}$. Rates of the order of $\sqrt{p/T_0}$ for the apparently naive weighting of (11) can be guaranteed if we randomize our choice by inflating losses by a given random variable $\xi_{i,t}$.

2.4. **Which Experts? A Practical Overview.** In this section we provide a non-comprehensive discussion on which learners can be used with our formulation of testing.

Synthetic Control. In absence of additional covariates, the Synthetic Controlmethod postulates the model $Y_t = X_t\beta + u_t + a_t D_t$, $||\beta||_1 = 1, \beta \geq 0$, $\mathbb{E}[u_t|X_t] = 0$ where control units serves as covariates (i.e., $X_t = (Y_{1,t}, Y_{2,t}, \ldots, Y_{n,t})$. The model can be extended; see for example Doudchenko and Imbens (2016) which included a time-independent intercept and/or relaxations on the contraints of the parameter space, letting $||\beta||_1 \leq K$ (Chernozhukov et al., 2017, 2018). Abadie et al. (2010) also discusses the case of additional covariates. We denote $\mathbf{Z}_0 = (Z_{0,1}, \ldots, Z_{0,T_+})$ the vector of covariates for the treated unit and $\mathbf{Z} = (\mathbf{Z}_1, \ldots, \mathbf{Z}_n)$ the matrix of covariates for control units. The authors propose a matching-type estimator to estimate β $\arg\min_\beta (\mathbf{Z}_0 - \mathbf{Z}\beta)^\top V(\mathbf{Z}_0 - \mathbf{Z}\beta)$, $\quad ||\beta||_1 = 1, \beta_1 \geq 0$ for a positive definite matrix V.

Interactive Fixed Effect Models. Interactive fixed effects, and factor models for Synthetic Control have been discussed in Hsiao et al. (2012); Xu (2017); Chernozhukov et al. (2017); Athey et al. (2018); Arkhangelsky et al. (2018); Amjad et al. (2018) among others. The model is defined as follows $Y_t = \lambda_y^\top F_t + Z_{0,t}\beta + a_t D_t + u_{y,t}$ with $Y_{j,t} = \lambda_j^\top F_t + Z_{j,t}\beta + u_{j,t}$ where F_t are time varying unobserved factors, λ_j are unit specific loadings and β is a vector of common coefficients; here, $Y_{j,t}$ denote control units. $\hat{Y}_t^0$ is obtained by first estimating $\hat{\lambda}_y, \hat{F}_t, \hat{\beta}$ using alternating Least Squares (Bai, 2009; Chernozhukov et al., 2017) and then predicting $\hat{Y}_t^0 = \hat{\lambda}_y'\hat{F}_t + Z_{0,t}\hat{\beta}$.

Support Vector Regression. The non-linear relation between Y_t^0 and X_t can be accomodated using feature expansions of X_t (Drucker et al., 1997). Hazlett and Xu (2018) discuss feature expansions of X_t, $\phi(X_t)$, in the contest of balancing and Synthetic Control. We consider $\hat{Y}_t^0 = \phi(X_t)\hat{w} + \hat{b}$ where the unknown slope and intercept are computed as

$$(\hat{w}, \hat{b}) = \min_{w,b} \frac{1}{2}||w||_2^2, \quad \text{s.t.} - \epsilon - \xi_t^* \leq Y_t - \phi(X_t)w - b \leq \epsilon + \xi_t, \quad \xi_t, \xi_t^* \geq 0.$$

In the above, ξ_t, ξ_t^* are slack variables used to relax the constraint implying that residuals cannot be larger than a pre-specified (soft) threshold. The predicted counterfactual becomes

$$\hat{Y}_t^0 = \sum_{s=T_-}^{0} (a_s - a_s^*)\phi(X_t)^\top\phi(X_s) + \hat{b} = \sum_{s=T_-}^{0} (a_s - a_s^*)\mathcal{K}(X_t, X_s) + \hat{b}$$

where a_s, a_s^* denote the Lagrange multipliers of the initial optimization. Predictions do not depend directly on the function $\phi(\cdot)$ but only on the kernel $\mathcal{K}(\cdot,\cdot)$. Classical choices include radial and polynomial kernels. For a comprehensive discussion on support vector machine, Burges (1998)

Recurrent Neural Network. Recurrent Neural Network(RNN) predicts future outcome by exhibiting temporal dynamics in its architecture. Formally, RNNs consists in a series of hidden layers $h_t = g_t(X_t, y_{t-1}, h_{t-1})$ and predictions $y_t = l_t(h_t)$ where g_t and l_t are some activation/loss functions. An example is the Jordan Network, where $h_t = \sigma_h(X_t W_h + y_{t-1}U_h + b_h), y_t = \sigma_y(h_t W_y + b_y)$ where W_h, U_h, b_h, W_y, b_y are coefficients to be estimated and $\sigma_h(\cdot), \sigma_y(\cdot)$ are activation functions(e.g. indicator functions or sigmoids). In the context under consideration, we estimate RNN parameters on the sub-sample indexed by $t \leq 1$. Alternative RNN include Hoepfield Network, Elman Network and others. For a comprehensive overview, Bishop et al. (1995).

Random Forest. Random Forest, first discussed in Breiman (2001), is gaining growing popularity for causal inference (Athey et al., 2019). It has been observed that in order to achieve fast rates and consistency, Breiman's forests need to be modified. One approach proposed initially by (Athey and Imbens, 2015) defines honesty as a desirable property. To the best of our knowledge, honest Random Forest has not been discussed yet in the context of panel data and Synthetic Control. Here we propose a modification that takes into account the time structure of the data.

Algorithm 4 Random Forest for Synthetic Control

Require: Observations $\{Y_t, X_t\}_{t=1}^{t_0}$, minimum leaf size k, block size b

1: For each tree in the forest split the observations into two equal parts: $t \in [1, t_0/2]$ and $t \in [t_0/2+1, t_0]$;
2: Draw a bootstrap sample $\mathcal{I}$ from $\{Y_t, X_t\}_1^{t_0/2}$ via the circular block(b) bootstrap;
3: Draw a bootstrap sample $\mathcal{J}$ from $\{Y_t, X_t\}_{t_0/2+1}^{t_0}$ via the circular block(b) bootstrap;
4: Grow a tree via recursive partitioning where a random subset of features as well as the best splits are designed using $\mathcal{I}$-samples only;
5: Estimate leaf-wise responses using $\mathcal{J}$-samples only.

▷ Random Forest for Synthetic Control makes predictions $\hat{Y}_t^0$ using leaf-averages on the leaf containing X_t, only using the $\{X_t\}_{t\geq t_0+m}$-sample observations. The splitting criteria can be a standard one minimizing mean-squared error of out-of sample predictions (computed on $\mathcal{I}$-samples only). Splits are restricted so that each leaf of the tree must contain k or more $\mathcal{J}$-sample observations.

For *i.i.d.* observations a tree is honest if, for each training example it only uses the response to estimate the predictions or to decide where to place the splits, but not both (Wager and Athey, 2018). Honesty relies on the independence of the data. Due to time dependence, trees can be approximately honest if the block of observations used to construct splits is "approximately" independent on the block of observations used to estimate leaf-wise responses. If we consider two consecutive blocks, as the size of these two blocks increases, under mixing conditions, we would expect that the dependence between observations in each block decay with the sample size (see for instance Yu (1994)). A further reason to consider such honesty criterion is related to the interest of researcher to predict *future* observations, hence imposing splitting rules based on past observations but computing leaf-wise responses use observations most close in time to the future ones.

Other methods Further possible base learners of interest are standard Time Series Methods (Hamilton, 1994), K- nearest neighbors, local regressions and many others.

3. Theoretical Analysis

3.1. **Hypothesis Testing.** Theoretical analysis of the Synthetic Learner builds on the literature of empirical processes and block bootstrap (Lunde and Shalizi, 2017). Recall the definition of Y_t^o as in Section 2.2.2. We define $\mathcal{F}_t = \sigma(X_t, Y_t, X_{t-1}, Y_{t-1}, \dots)$ as the natural filtration up to time t. For expositional convenience we consider the case of no-carryover effects, but the same results hold for a fixed level of carryover effects $m = k \ll T_0$. Next, we impose conditions on the dependence structure of the data.

Assumption 2. Conditional on $\mathcal{F}_0$, assume that $\{Y_t^o, X_t\}_{t\geq 1}$ is (a) β-mixing with mixing coefficients $\sum_{k=1}^{\infty}(k+1)^2\beta(k) < \infty$ and (b) strictly stationary. In addition, $\{Y_t^o, X_t\}_t$ is bounded.

Stationarity is fairly standard in the literature on block bootstrap (Politis and Romano, 1992, 1994; Lahiri et al., 1999; Lunde and Shalizi, 2017). For inference via Synthetic Control, similar stationarity conditions have been imposed also in Chernozhukov et al. (2017) to show the validity of permutation tests. Chernozhukov et al. (2018) requires only covariance-stationary conditions of the data, but the validity of the results relies on the constrained Lasso estimator. Stationarity and beta-mixing conditions as stated above cover a large class of ARMA processes (Pham and Tran, 1985), AR-ARCH processes (Lange et al., 2011), Markov Switching Processes (see for example Lee (2005)), GARCH (Carrasco and Chen, 2002), to cite some.

Assumption 3. Weights are either computed via Least Squares or (5) with hyper-parameter $\eta \propto 1/T_+$, or as generic potential weights with Hadamard differentiable potential function (9).

Assumption 4. The reference class of learners contains only bounded functions.

The boundedness condition can be easily guaranteed in practice if predictions are made to be constant if they exceed a given threshold. For large enough thresholds, this requirement has no strong implications in practice. We consider the case of large pre and post-treatment period, namely we consider the asymptotic regime for $T_0, T_+ \to \infty$ and $T_0 = \lambda T_+$ for $\lambda \in (0,1)$. We impose conditions on the block size $b(T_+)$.

Theorem 3.1. *Let* $\limsup_{T_+\to\infty} b(T_+)/\sqrt{T_+} < \infty$ *and* $\lim_{T_+\to\infty} b(T_+) \to \infty$. *Let Assumptions 2, 3 and 4 hold. Then, under the null hypothesis, whenever* $p < \infty$

$$\sup_x \left|\mathbb{P}_0(\mathcal{T}_{\mathcal{S}}^* - \mathcal{T}_{\mathcal{S}} \le x | Y_{1:T_+}, X_{1:T_+}) - \mathbb{P}_0(\mathcal{T}_{\mathcal{S}} - \mathbb{E}[\mathcal{T}_{\mathcal{S}}] \le x)\right| = o_p(1)$$
$$\sup_x \left|\mathbb{P}_0(\mathcal{T}_{\mathcal{A}}^* - \mathcal{T}_{\mathcal{A}} \le x | Y_{1:T_+}, X_{1:T_+}) - \mathbb{P}_0(\mathcal{T}_{\mathcal{A}} - \mathbb{E}[\mathcal{T}_{\mathcal{A}}] \le x)\right| = o_p(1)$$

as $T_+ \to \infty$.

The proof is presented in the Supplement. We observe that Theorem 3.1 does not impose any restrictions on the ambient dimension of the feature space X_t, therefore, allowing for extremely high-dimensional observations. Moreover, class of learners does not need to be only restricted to differentiable estimators or estimators with specific consistency guarantees; for example, stumps or sigmoid link functions are often desirable as is apparent in Random Forests or Neural Network, respectively. We achieve such generality by our sample-splitting procedure as outlined in Algorithm 2.

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 our result. 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 4 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 bootstrapped test statistic approximately agrees with the true density in this also, non-stationary case (left panel).

3.2. **Average treatment effects.** Here, we relax the condition of fixed treatment effects imposed in Section 3.1 and we consider estimands that involve the treatment distribution in the data, similarly to Boruvka et al. (2018). Namely, we consider the average treatment effect $\mathbb{E}[Y_t(\mathbf{1}) - Y_t(\mathbf{0})]$ under stationarity. We let $T_0 = \lambda T_+$ where $T_+ \to \infty$ and λ is potentially a random variable strictly between zero and one.

Theorem 3.2. *Assume that the exponential weights (5) are chosen with* $\eta \propto 1/T_+$, *and that Assumption 4 holds. Conditional on* $\mathcal{F}_0$, $\{(Y_t(\mathbf{0}), X_t)\}_{t\ge1}$ *and* $\{(Y_t(\mathbf{1}), X_t)\}_{t\ge1}$ *are* α*-mixing and stationary, with* Y_t *having finite second moment and* $T_0 \perp (Y_t(\mathbf{d}), X_t)$. *Then, for* $l(x,y)$ *being either*

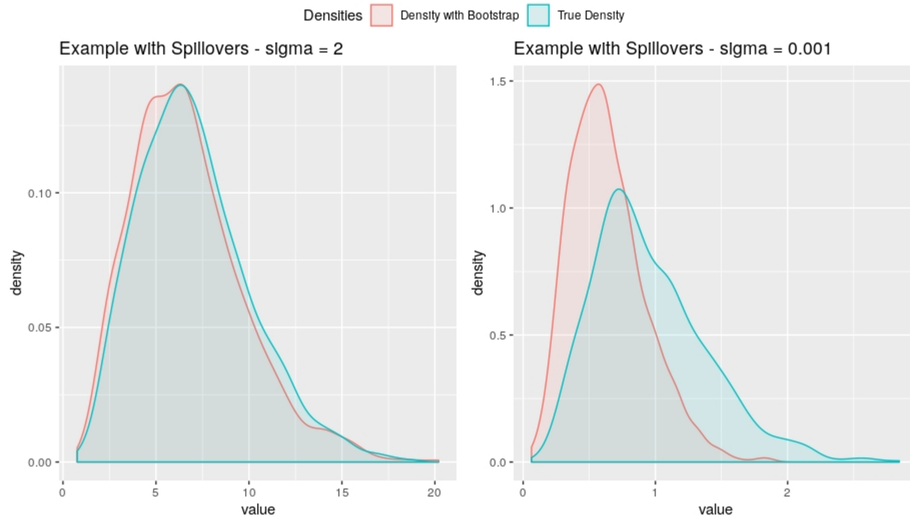

FIGURE 4. Explanatory example on failure of the bootstrap in the presence of spillover effects on X_t. Blue histogram indicates $\mathcal{T}_S = \tau_1\chi^2(T_+ - T_0)$ whereas the red corresponds to its bootstrapped counterpart $\mathcal{T}_S^* = \tau_2\chi^2(T_+ - T_0)$ where $\tau_1 = (\sigma^2 + 2J^{-1})/\sqrt{T_+ - T_0}$, $\tau_2 = (\sigma^2 + J^{-1})/\sqrt{T_+ - T_0}$. 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.

bounded, a quadratic or absolute deviation loss, $\widehat{\text{ATE}}$ *as defined in* (4)*, satisfies*

$$\widehat{\text{ATE}} \to_p \mathbb{E}[Y_t(\mathbf{1}) - Y_t(\mathbf{0})]$$

for $t \geq 1$*, and* $T_+ \to \infty$*.*

The result above shows that our proposed estimator is consistent despite the possible lack of the model specification; the bias correction of (4) guarantees consistency of the final estimator. Recent literature on Synthetic Control (Chernozhukov et al., 2018; Arkhangelsky et al., 2018) treated T_0 as deterministic. Here we consider the more general case of random T_0, but we impose that it is independent of potential outcomes and X_t. In addition, our result also accounts for the presence of carryover effects and random treatment effects.

In our next result, we provide stronger guarantees. We denote $\hat{g}_i^0(X_t) = h_i(X_t, F_-)$ where F_- denotes the empirical distribution of $(X_{T_-}, Y_{T_-}, \dots, X_0, Y_0)$. Similarly, we denote $\widehat{\text{ATE}}(F_-)$ as a function of F_-. In the next theorem we provide uniform guarantees over $F_- \in \mathcal{D}$, where $\mathcal{D}$ denotes the space of compactly supported probability distributions.

Theorem 3.3. *Assume that the exponential weights* (5) *are chosen with* $\eta \propto 1/T_+$*, Assumption 4 holds, and* $\{(Y_t(\mathbf{0}), X_t)\}$ *and* $\{(Y_t(\mathbf{1}), X_t)\}$ *are* α*-mixing and stationary and bounded. Then, for* $T_0 \perp (Y_t(\mathbf{d}), X_t)$ *and* $l(x, y)$ *being either bounded, a quadratic or absolute deviation loss,* $\widehat{\text{ATE}}(F_-)$ *as defined in* (4)*, satisfies*

$$\sup_{F_- \in \mathcal{D}} \left| \widehat{\text{ATE}}(F_-) - \mathbb{E}[Y_t(\mathbf{1}) - Y_t(\mathbf{0})] \right| = o_p(1),$$

as $T_+ \to \infty$*.*

Theorem 3.3 does not involve conditional statements on $\mathcal{F}_0$ and it holds uniformly over the initial sample used to train learners, namely uniformly the part of $(X_{T_-}, Y_{T_-}, \dots, X_0, Y_0)$. Our proof relies on a functional law of large numbers appropriately derived for the context under consideration.

3.3. **Heterogeneous treatment effects.** Assessing the performance of a prediction algorithm can be very challenging. One metric of performance, becoming more popular in the literature on causal inference in recent years, are error regret bounds of the estimator of interest. In this section, we discuss finite-sample regret bounds. We postulate the following data generating process(DGP) for potential outcomes,

$$Y_t(\mathbf{d}) = \mu(X_t, d_{t-m}, \dots, d_t) + \varepsilon_t(\mathbf{d}). \tag{12}$$

where $\mathbb{E}[\varepsilon_t(\mathbf{d})|X_t, \mathcal{F}_{t-1}] = 0$. We assume *sequential ignorability*, i.e., $\varepsilon_t(\mathbf{d}) \perp D_t|X_t, \mathcal{F}_{t-1}$, where $\mathcal{F}_{t-1}$ denotes past filtration. For T_0 being deterministic, this condition trivially holds. In the following we do not treat T_0 as deterministic and we impose only such weaker condition. In this new setting, the *unconfoundeness* assumption can equivalently be stated as in the potential outcomes framework, i.e., $Y_t(\mathbf{d}) \perp T_0|X_t, \mathcal{F}_{t-1}$. We observe that overlap assumption is not needed for our results on regret. Examples that satisfy the above conditions include settings like $Y_t = \alpha_1 D_t + \dots + \alpha_m D_{t-m} + \mu(X_t) + \varepsilon_t$ or $Y_t = X_t\beta \times (D_t + D_{t-1}) + \varepsilon_t$, where the treatment effect is random and heterogenous and whose realizations depends on X_t.

We consider

$$\text{CATE} = \mathbb{E}[Y_t(\mathbf{1}) - Y_t(\mathbf{0})|X_t = x] = \mu(x, \mathbf{1}) - \mu(x, \mathbf{0})$$

where X_t contains current and past values of control units and additional covariates. Our notation deviates from the standard notation of causal framework in longitudinal studies (Robins, 1989), since we consider estimands that involve the treatment distribution in the data, similarly to Boruvka et al. (2018), where our definition of CATE is consistent with what the latter paper defines as moderated proximal effect[3].

We predict $\hat{Y}_t^0$, using the sample splitting rule discussed in Section 2.2.1. Similarly, we predict $\hat{Y}_t^1$ using the same procedure, applied only to post-treatment period, where we split data into $\{T_0 + m + 1, \dots, T\}$, used to estimate weights, and $\{T + 1, \dots, T_+\}$ used for training $\{\hat{g}_i^1\}_{i=1}^p$. The method belongs to a class of T-learning algorithms as defined in (Künzel et al., 2019); extensions to X-learning are presented in Section 4.3.2.

Assumption 5. Regression function $\mu(\cdot, \cdot)$ is bounded.

Ideally, we would like to provide bounds on the mean squared error (MSE) of the CATE estimator, namely $\mathbb{E}[(\mu(X_t, \mathbf{1}) - \mu(X_t, \mathbf{0}) - (\hat{Y}_t^1 - \hat{Y}_t^0))^2]$. On the other hand, the lack of exchangeability creates substantial challenges in deriving theoretical properties of such an object. $\hat{Y}_t^1$ and $\hat{Y}_t^0$ are estimated over two different periods, and it is not clear at which period the MSE should be evaluated. A formal definition of performance metrics for causal estimators with nonexchangeable data goes far beyond the scope of this paper.

We study instead the behavior of our algorithm trained only on $t-1$ observation and evaluated at the *th* observation, i.e., we are willing to provide theoretical guarantees on the following cumulative loss.

$$T_0^{-1} \sum_{t=1}^{T_0} (\hat{Y}_t^0(\mathcal{F}_{t-1}) - \mu(X_t, \mathbf{0}))^2, \quad T_m^{-1} \sum_{t=T_0+m}^{T} (\hat{Y}_t^1(\mathcal{F}_{t-1}) - \mu(X_t, \mathbf{1}))^2 \tag{13}$$

where, $\hat{Y}_t^0(\mathcal{F}_{t-1})$ denotes the prediction at time t using only information at time $t-1$. Since $\hat{Y}_t^0(\mathcal{F}_{t-1})$ is estimated only on the previous data and evaluated at X_t this notion of performance is rooted in out-of-sample performance metric.

[3] Boruvka et al. (2018) defines moderated proximal effect as $\mathbb{E}[Y_t((1, D_{t-1}, D_{t-2}, \dots) - Y_t((0, D_{t-1}, D_{t-2}, \dots)|S(\mathcal{F}_{t-1})]$ where $S(\mathcal{F}_{t-1})$ contains a subset of covariates of interest observed in the past. Here we let X_t also contain past information of interest, similarly to the previous definition of $S(\mathcal{F}_{t-1})$. We compare instead potential outcomes *always* under treatment and *always* under control, also discussed in Athey and Imbens (2018) among others.

It is natural to compare the cumulative loss in (13) with the smallest cumulative loss incurred by any of the algorithms under consideration. We define such metric of comparison as the Conditional Mean Proxy Regret (CMPR).

$$\mathcal{R}^0 = T_0^{-1} \sum_{t=1}^{T_0} (\hat{Y}_t^0(\mathcal{F}_{t-1}) - \mu(X_t, \mathbf{0}))^2 - \min_{i \in \{1,\dots,p\}} T_0^{-1} \sum_{t=1}^{T_0} (\hat{g}_i^0(X_t) - \mu(X_t, \mathbf{0}))^2$$
$$\mathcal{R}^1 = T_m^{-1} \sum_{t=T_0+m}^{T} (\hat{Y}_t^1(\mathcal{F}_{t-1}) - \mu_1(X_t, \mathbf{1}))^2 - \min_{i \in \{1,\dots,p\}} T_m^{-1} \sum_{t=T_0+m}^{T} (\hat{g}_i^1(X_t) - \mu_1(X_t, \mathbf{1}))^2.$$

Our definitions above combine definitions in the literature on prediction of individual sequences (Cesa-Bianchi et al., 1999) with the literature on causal inference. The main difference with standard notions of regret is that CMPR is based on the *unobserved* deviation of the predicted counterfactual from the conditional mean evaluated at X_t, and not just on the cumulative loss of the predictor. We discuss our results without requiring stationarity conditions, but still relying on Assumption 5.

Theorem 3.4. *Let Assumption 4, 5 holds and let Y_t being bounded for any t. Consider a quadratic loss function l, and an exponential weighting scheme as in (5) with $\eta^0 \propto \sqrt{\log(p)/T_0}, \eta^1 \propto \sqrt{\log(p)/T_m}$ and carry-over effect that propagates up to $m < \infty$. Then, for T_0, T_m large enough, with probability at least $1 - 2\delta$,*

$$\mathcal{R}^0 \leq C_0 \sqrt{\frac{\log(p/\delta)}{T_0}}, \qquad \mathcal{R}^1 \leq C_0 \sqrt{\frac{\log(p/\delta)}{T_m}}$$

for C_0 being a constant independent of time or p.

Proof is presented in the Supplement.
Theorem 3.4 provides an error bound for the empirical one step ahead prediction error. Remarkably, it does not require any stationarity assumption. The bound scales logarithmically with the number of learners and it scales at square-root T with the length of the sequence.

In the next theorem we present the result for the case of subgaussian random variables.

Assumption 6. For $\psi(x) = \exp(x^2) - 1$, let $\mathbb{E}\left[\psi(\sum_{s=1}^t \varepsilon_s / c) | X_t, \mathcal{F}_{t-1}\right] \leq \tau < \infty, t \leq T$.

Theorem 3.5. *Let Assumption 4, 5, 6 hold and Y_t being sub-gaussian for any t. Consider a quadratic loss function l, and an exponential weighting scheme as in (5) with $\eta^0 \propto \sqrt{\log(p)/T_0}, \eta^1 \propto \sqrt{\log(p)/T_m}$ and carry-over effect that propagates up to $m < \infty$. Then, for T_0, T_m large enough, with probability at least $1 - 2\delta$,*

$$\mathcal{R}^0 \leq C_0 \sqrt{\frac{\log(T_0) \log(p/\delta)}{T_0}}, \qquad \mathcal{R}^1 \leq C_0 \sqrt{\frac{\log(T_m) \log(p/\delta)}{T_m}}$$

for C_0 being a constant independent of time or p.

If we are willing to assume more, in that our class of algorithms contains one learner that consistently estimates the unknown model, then, previous results imply that our synthetic learner will preserve that consistency regardless of the number of learners in the entire class. We consider below asymptotics for $T \to \infty$ and $T_0 = \lambda T$ where $\lambda \in (0, 1)$ is potentially a random variable.

Corollary. *Suppose that the number of learners is such that $\log(p)/\min\{T_0^{1/2}, T^{1/2}\} = o(1)$ and conditions in Theorem 3.4 hold, or $\max\{\log(T_0), \log(T_m)\} \log(p)/\min\{T_0^{1/2}, T^{1/2}\} = o(1)$ and conditions in Theorem 3.5 hold.*

Assume also that the following holds $\min_{i\in\{1,\dots,p\}} T_0^{-1}\sum_{t=1}^{T_0}|\mu(X_t,\mathbf{0})-\hat g_i^0(X_t)|^2 = o_p(1)$, *and* $\min_{j\in\{1,\dots,p\}} T_m^{-1}\sum_{t=T_0+m+1}^{T}|\mu(X_t,\mathbf{1})-\hat g_j^1(X_t)|^2 = o_p(1)$.

Then,

$$T_0^{-1}\sum_{t=1}^{T_0}(\hat Y_t^0(\mathcal{F}_{t-1})-\mu(X_t,\mathbf{0}))^2 = o_p(1), \qquad T_m^{-1}\sum_{t=T_0+m}^{T}(\hat Y_t^1(\mathcal{F}_{t-1})-\mu(X_t,\mathbf{1}))^2 = o_p(1)$$

for $T\to\infty$, $T_0=\lambda T$.

4. Numerical experiments

In observational studies, accurate detection of the treatment effects requires overcoming two potential sources of bias. First, we need to estimate well the underlying model for the outcome variable, and second, we need to allow for the presence of noninformative learners. We test the ability of the proposed Synthetic learner to respond to both sources of bias. We showcase that neither violates the detection properties of the proposed testing algorithm; we compare power and showcase a significant improvement over existing methods including permutation tests of Synthetic Control as well as the Difference-in-Difference method.

4.1. Experimental Setups.

We describe our experiments in terms of the outcome model as well as the model of the design of the covariates and the error terms. We very models and start from a simple linear AR model, continue with various non-linear models that include error terms that follow AR-ARCH and AR processes and a factor model. In our homogenous treatment effects cases time is fixed as T_0. We study heterogenous Difference-in-Differenceeects in Section 4.3.2.

In our first experiment, **DGP1**, we considered a simple Linear Outcome Model

$$Y_t = X_t\beta + a_tD_t + \epsilon_t$$

and tested the ability of our method to detect changes in the treatment effect a_t. This example is intended to model a setting where classical Synthetic Controlmethod is optimal. Here we set $\beta_j = 1/(1+j)^2$, $j=1,\dots,50$, with the last beta chosen such that $\sum_j \beta_j = 1$. Parameter β will be kept as is for all our experiments. We considered a simple AR model for the errors ϵ_t with $\epsilon_t = 0.6\epsilon_{t-1}+v_t$ and $v_t\sim\mathcal{N}(0,1-0.6^2)$. Control units are generated according to a factor model as

$$X_{j,t} = \mu_j + \theta_t + \lambda_j F_t + u_t$$

with unit specific term $\lambda_j=\mu_j=(1+j)/j$ a time random effect $\theta_t\sim\mathcal{N}(0,1)$ and an unobserved factors $F_t\sim\mathcal{N}(0,1)$. Errors u_t are following simple AR model $u_t = 0.6u_{t-1}+h_t$ with $h_t\sim\mathcal{N}(0,1-0.6^2)$.

In our second experiment, **DGP2**, we considered a Logistic-like Outcome Model

$$Y_t = a_tD_t + \exp(X_t\beta+\epsilon_t)/(1+\exp(X_t\beta+\epsilon_t))$$

with $\epsilon_t = 0.5\epsilon_{t-1}+0.3v_{t-1}+v_t$. This experiment has three settings: (a), (b) and (c). Setting (a) and (b) assume $v_t\sim\mathcal{N}(0,\sigma^2)$ with $(a)\sigma=0.1$ and $(b)\sigma=1$, respectively. Setting (c) assumes $\epsilon_t=0.8\epsilon_{t-1}+v_t$, with

$$v_t=\sqrt{h_t}z_t, \quad h_t = 0.001+0.99v_{t-1}^2$$

with $z_t\sim\mathcal{N}(0,1)$(AR-ARCH process). In addition we let $X_t=h_t+u_t$ with h_t being i.i.d over time with $\mathcal{N}(0,\Sigma)$ distribution with $\Sigma_{i,j}=0.5^{|i-j|}$ and $u_t=0.8u_{t-1}+k_t$ with $k_t\sim\mathcal{N}(0,1-0.8^2)$. This setting is design to test the ability of the proposed Synthetic Learner to adapt to nonlinear outcome model.

Our third experimental setting, **DGP3**, considers a interaction outcome model that is polynomial in structure

$$Y_t = a_tD_t + (X_{1,t}+X_{2,t}+\dots+X_{10,t})^2+\epsilon_t$$

with ϵ_t being the same as in **DGP2**(a) design X_t is the same as throughout **DGP2**.

Our fourth setting, **DGP4**, postulates a cosine, hence periodic, type of outcome model

$$Y_t = \cos(X_t\beta + \epsilon_t) + a_t D_t$$

with 50 features. Error and design setting have three components: (a), (b) and (c) that are following the setup of **DGP2** (a), (b) and (c), respectively.

Lastly, we consider our fifth setting, **DGP5**, that follows a factor model

$$Y_t = 0.5 + a_t D_t + \theta_t + 0.5 F_t + \epsilon_t$$

with 50 features. Error and design structures are the same as that of **DGP1**. This setting favors Synthetic Control.

4.2. **Testing.** The goal of this simulation study is to verify that Synthetic learner can be used to build asymptotically valid hypothesis tests that improve (in terms of power or generality of the setting) over Synthetic Control and difference-in-difference in finite samples.

We consider testing the following hypothesis H_0, with

$$H_0 : Y_t(1) - Y_t(0) = 0, \quad t > T_0.$$

We consider the Synthetic Learner with experts including a naive XGboost (which uses the default tuning parameter of the package XGboost in R), Support Vector Regression and ARIMA(0,1,1) with external regressors together with 50 non-informative learners. Non-informative experts are randomly drawn from a multivariate gaussian with full covariance matrix.

We compare Synthetic Learner's performance to Synthetic Control (SC) with weights being constrained to sum to one and an intercept according to Equation (7) and (8) in Chernozhukov et al. (2017), as well as the Difference in Difference (DiD) estimator, namely

$$\hat{Y}_t^{DiD} = \hat{\alpha} + (\hat{\beta} + \hat{\Delta})\mathbb{1}_{t>T_0}$$

with coefficient computed as in a standard DiD with the two periods corresponding to pre and post-treatment periods. We consider the test statistics for Synthetic Control

$$\frac{1}{\sqrt{T - T_0}} \sum_{t=T_0+1}^{T} |Y_t - a_t^o - X_t \hat{w}_{SC}^0|^2 \tag{14}$$

where $\hat{w}_{SC}^0$ are computed via constrained Least Squares, with coefficients summing to one for Synthetic Control. Finally, we consider

$$\frac{1}{\sqrt{T - T_0}} \sum_{t=T_0+1}^{T} |Y_t - a_t^o - \hat{Y}_t^{DiD}|^2 \tag{15}$$

for Difference-in-Differences. In Figures 5, 6, 7 we compare the performance of our method to permutation tests where $\hat{w}_{SC}$ and $\hat{Y}_t^{DiD}$ must be computed on the entire sample, as described in Chernozhukov et al. (2017). In Figures 10 and 9 we compute $\hat{w}_{SC}$ and $\hat{Y}_t^{DiD}$ only using information until time T_0, as discussed in Doudchenko and Imbens (2016).

4.2.1. *Power study.* We fix $T_0 = 250$ and $T_- = 125$ and we consider three different scenarios $T \in \{300, 350, 400\}$. We run the Synthetic Learner after training on the period running from 1 to $T_- = 125$ and we use the remaining observations for computing weights and bootstrap. We consider different treatment effects, denoted by α. We present power plots across different T in Figures 5, 6 and 7, corresponding to $T = 300, 350$ and 400, respectively. Across all figures, we observe a striking improvement over permutation tests with both SC and DiD methods. Even in cases designed to fit SC perfectly, **DGP1** and **DGP5**, we see that test based on our Synthetic Learner maintains power that is the order of magnitude better. This can be due to two factors: bootstrap outperforming permutation as well as Synthetic Learner's better performance in comparison to SC

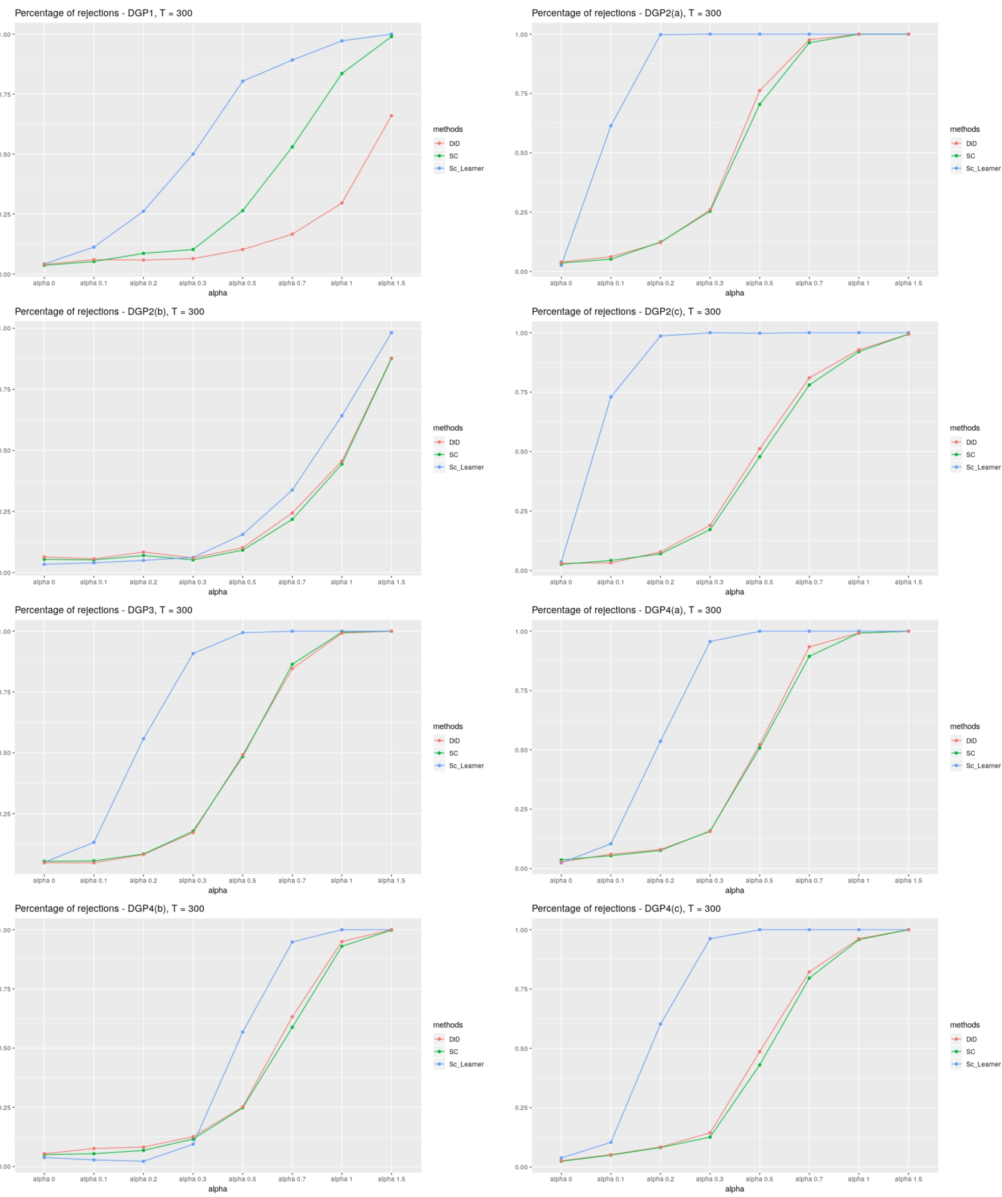

FIGURE 5. Percentage of rejections of the null hypothesis of no treatment effects over 500 repetitions. Synthetic learner has XGboost,Support Vector Regression and ARIMA(0,1,1) and 50 additional non informative predictions. Post treatment ends at $T_+ = 300$.

and DiD. With larger post-treatment period we see sharp decay in the performance of SC and DiD based permutation tests, reconfirming that permutations are not designed to work well with long post-treatment periods.

4.2.2. *Variability in the quality of the learners.* Next, we study the variability of the proposed method concerning the number and quality of learners included in the class of learners. We consider four different variations of the Synthetic Learner: Exponential and Least Squares weighting with 10 and 100 new non-informative learners. Figure 8 contains the results. There we observe that a large number of non-informative learners does little to nothing to the proposed Synthetic Learner.

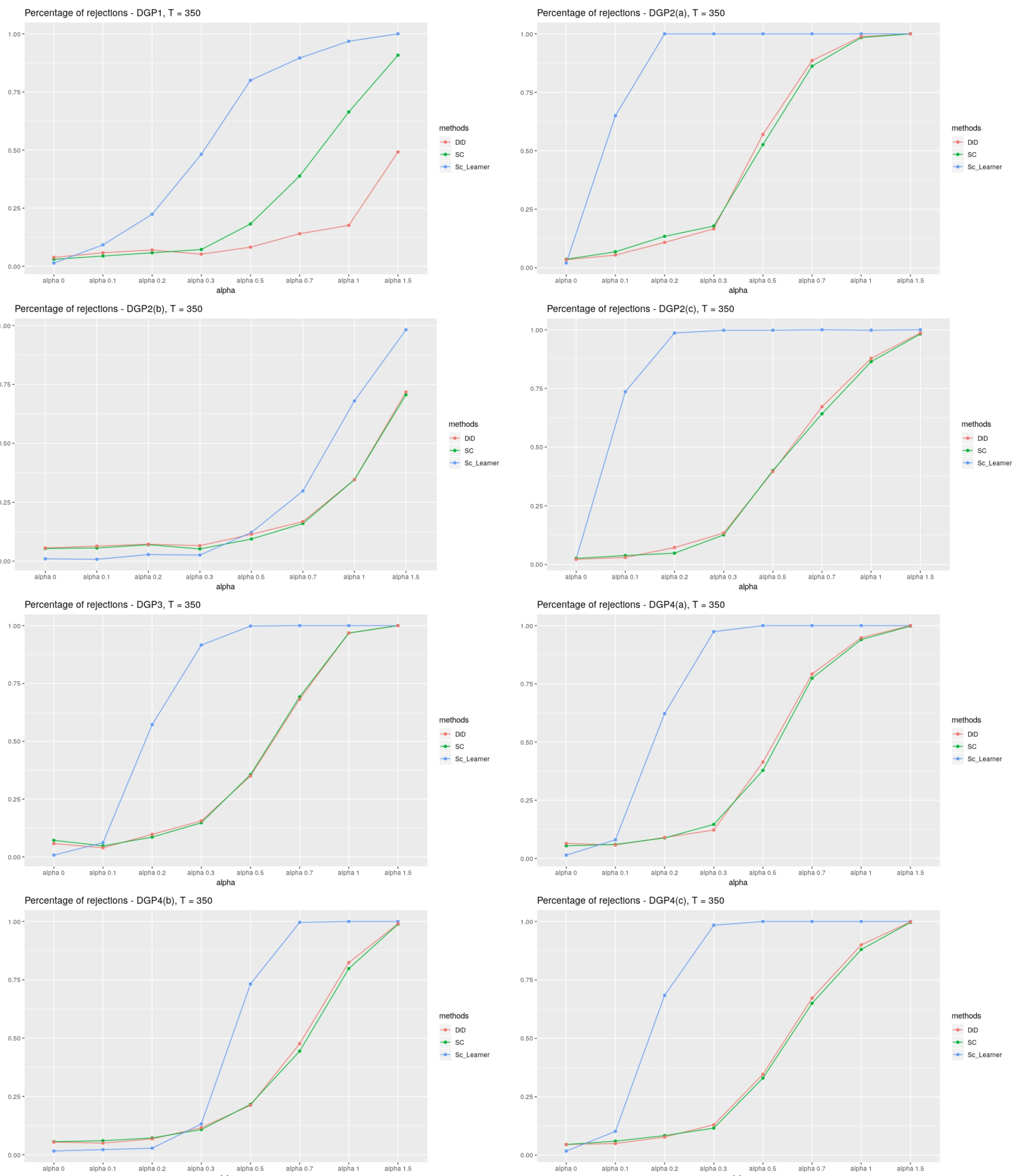

FIGURE 6. Percentage of rejections of the null hypothesis of no treatment effects over 500 repetitions. Synthetic learner has XGboost,Support Vector Regression, ARIMA(0,1,1) and 50 additional non informative predictions. Post treatment ends at $T_+ = 350$.

In sharp contrast Least Squares weighting, suffers great loss in power when the number of non-informative learners is increased.

4.2.3. *Oracle Study.* In order to understand better 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 9 collects our results. Synthetic Learner does not have uninformative learners, and the class consists of XGboost, Support Vector Regression, SC 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.

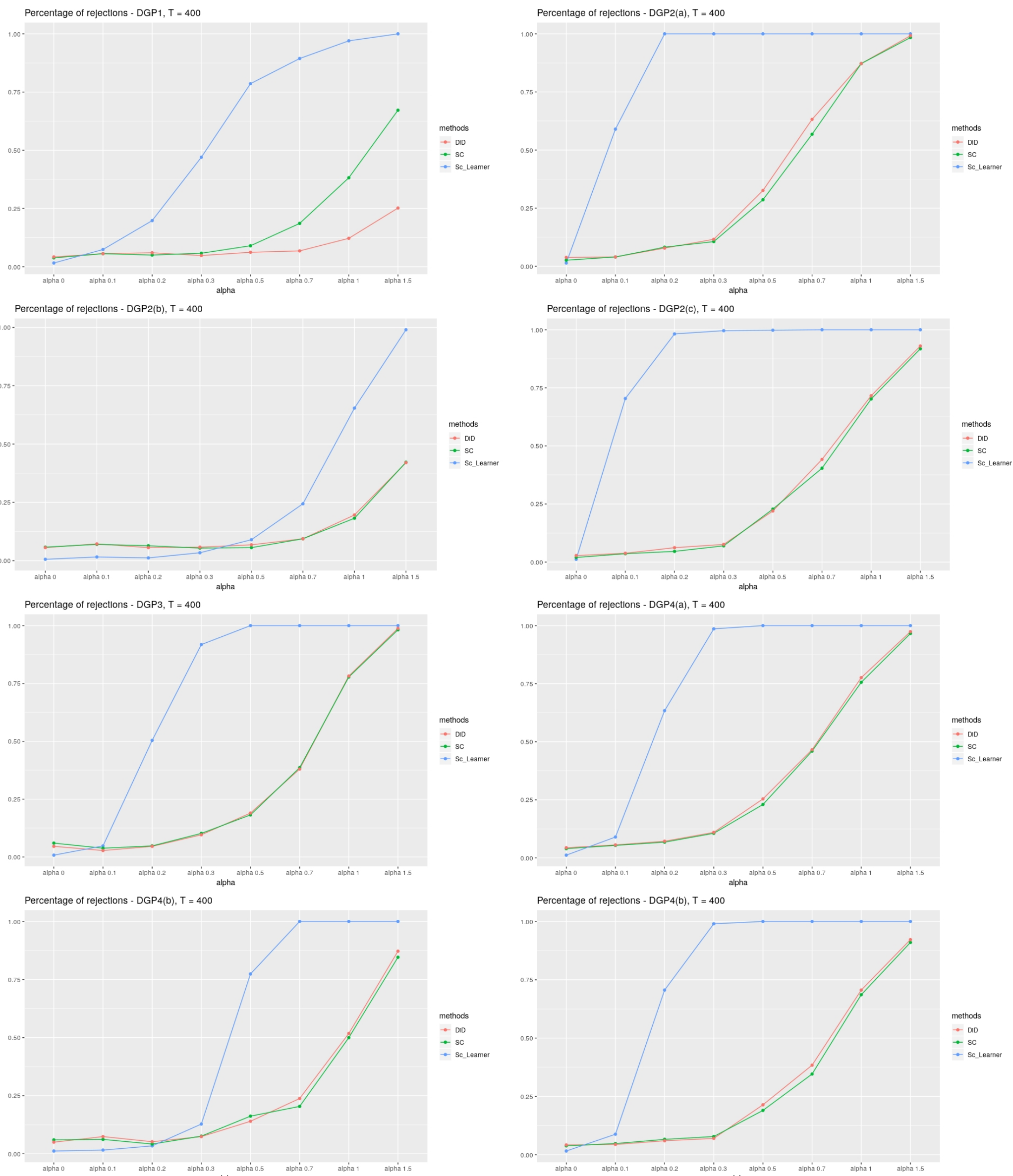

FIGURE 7. Percentage of rejections of the null hypothesis of no treatment effects over 500 repetitions. Synthetic learner has XGboost, Support Vector Regression , ARIMA(0,1,1) and 50 additional non informative predictions. Post treatment ends at $T_+ = 400$.

The proposed method outperforms uniformly DiD, and SC in almost all **DGP**s considered, while we observe an improvement of DiD and SC especially in **DGP1** and **DGP5** when compared to results using permutation methods. This result provides evidence that improvements in the power are driven not only by a better performance over permutation methods of the resampling scheme proposed (see Table 1) but also by a better performance of the predictive method.

We also present a comparison of the distribution of the two test statistics: one using SC and one using Synthetic Learner. See Figure 10 for more details. We observe that the proposed test has a much smaller variance regardless of the structure of the outcome model. This suggests a certain

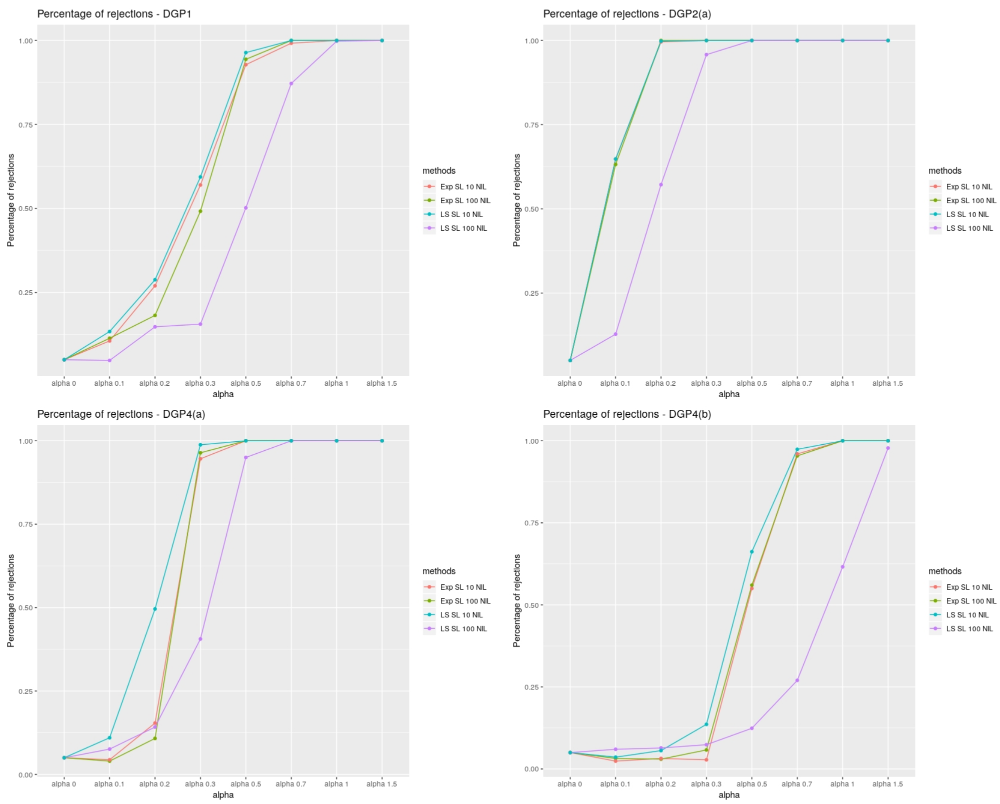

FIGURE 8. Percentage of rejections of null hypothesis H_0 over 500 repetitions with $T = 300$, $T_0 = 280$, $J = 50$ and $T_- = 140$. As base learner we consider XGboost, Support Vector Regression, ARIMA(0,1,1) and either 100 or 10 additional non informative learners(NIL). We denote exp SL as the Synthetic Learner using exponential weights over experts and LS SL the Synthetic Learner using Least Squares weights over experts. On the y-axis we report the percentage of rejection of the sharp null hypothesis of no effect.

robustness property of the proposed method, i.e., greater power when detecting deviations from the null hypothesis.

4.2.4. *Bootstrap vs permutations.* Finally, we compare the performance of the circular bootstrap against permutations, proposed in Chernozhukov et al. (2017). To make the comparison fair, we consider only one learner: Least Squares learner. Namely, for the bootstrap method, we compute the OLS coefficient using only the first $T_0/2$ observations, and we bootstrap the remaining ones. For the permutation method, we estimate the coefficient on the full sample, after imposing the null hypothesis of no effect. We consider the true effect is either $\alpha_t = 0.2$ or $\alpha_t = 0.3$. We keep $T = 300$ and we consider two scenarios with $T_0 \in \{280, 200\}$. Results are collected in Table 1. We observe that even in short post-treatment periods, bootstrap has significantly better performance whenever the model is non-linear.

4.3. **Homogeneous and heterogeneous treatment effects.**

4.3.1. *Homogeneous treatment effects.* We fix the true treatment effect to be $a_t^o = 1$ being homogenous in the population. We simulate data for different sample sizes and we let $T_0 = T/2$. The Synthetic Learner trains learners on observations $\{1, \ldots, T_0/4\}$ and evaluates the weights on

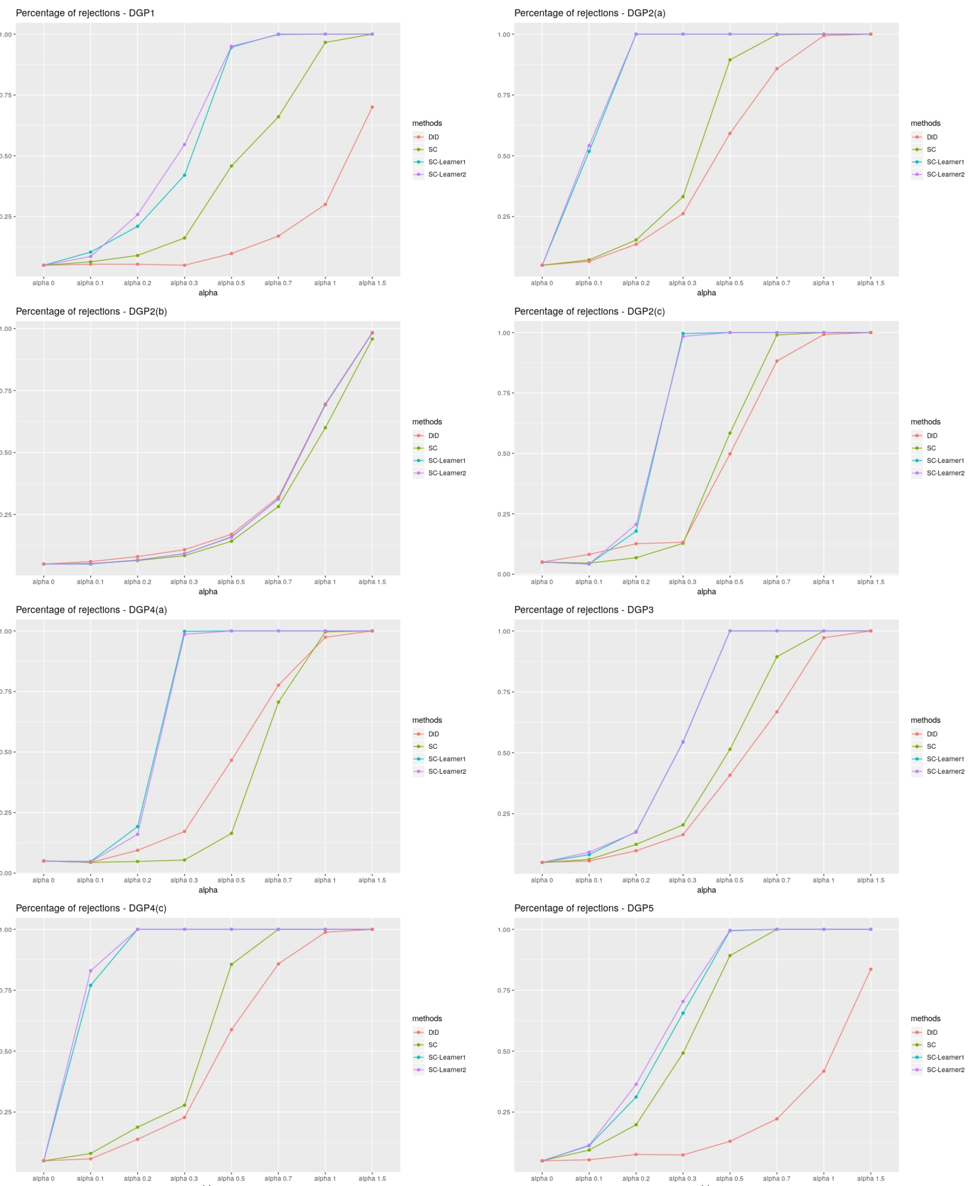

FIGURE 9. Percentage of rejections over 500 repetitions with $T_+ = 300$, $T_0 = 280$, $J = 50$ and $T_- = 140$ when the critical quantile is *known*(oracle case). SC and DiD are estimated using all information until time T_0(no permutation test required). SC-Learner1 is the Synthetic Learner trained 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.

observations $\{T_0/4 + 1, \ldots, T_0\}$. We consider Random Forest, Lasso, XGboost, Support Vector Regression, ARIMA(0,1,1).

We consider the performance to two different versions of the SC. The first version does not include any intercept and it constructs the counterfactual by taking a weighted combinations of

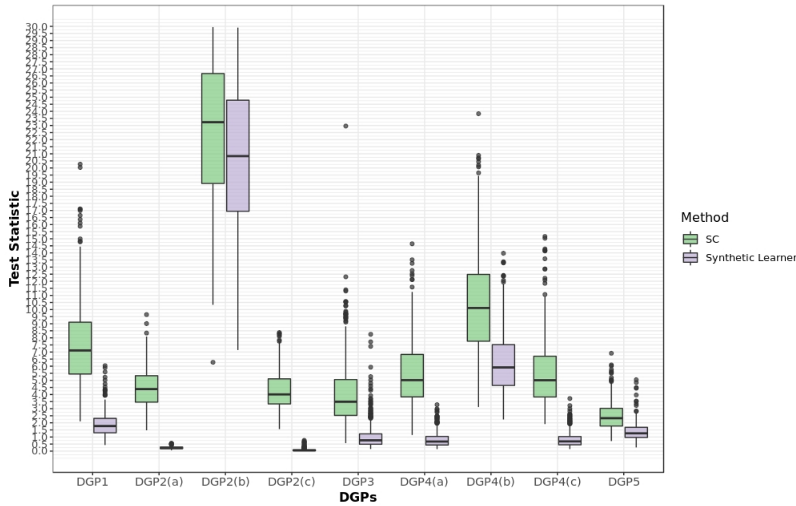

FIGURE 10. Box-Plot of test statistic computed using Synthetic Learner (purple) and Synthetic Control (green) as defined in (15). We have utilized short post-treatment period with $T_+ = 300$, $T_0 = 280$ and $T_- = 140$. Results are averaged over 500 replications. Synthetic Learner has ARIMA(0,1,1), XGboost and Support Vector Regression together with 50 non-informative learners.

TABLE 1. We compare the percentage of rejection of the sharp null hypothesis $\alpha_t^o = 0$ over 500 replications. We use the bootstrap method, while rejections via permutations are presented in the parenthesis. We predict the counterfactuals only using Least Squares. For the bootstrap method, consistently with the method proposed in the paper, we compute coefficient over the first $T_0/2$ observations and we bootstrap the remaining one. For permutations, we compute the coefficient over the full sample, after imposing the null hypothesis, as discussed in Chernozhukov et al. (2017). We consider $T = 300$ and T_0 varying. α denote the true policy effect.

	$T_0 = 280$ $\alpha = 0.2$	$\alpha = 0.3$	$T_0 = 200$ $\alpha = 0.2$	$\alpha = 0.3$
DGP1	0.258(0.138)	0.456(0.260)	0.088(0.044)	0.286(0.106)
DGP2(a)	0.976(0.942)	1(1)	0.934(0.502)	1(0.820)
DGP2(b)	0.050(0.052)	0.068(0.076)	0.012(0.040)	0.022(0.056)
DGP2(c)	0.914(0.354)	0.970(0.564)	0.820(0.204)	0.928(0.356)
DGP3	0.130(0.094)	0.316(0.244)	0.040(0.064)	0.216(0.112)
DGP4(a)	0.178(0.098)	0.542(0.406)	0.294(0.144)	0.704(0.286)
DGP4(b)	0.050(0.046)	0.164(0.096)	0.086(0.072)	0.234(0.090)
DGP4(c)	0.108(0.048)	0.238(0.160)	0.108(0.048)	0.238(0.160)
DGP5	0.284(0.136)	0.532(0.314)	0.128(0.044)	0.368(0.078)

X_t. The second version, denoted as Synthetic Control +, computes the ATE as follows

$$\widehat{\text{ATE}}_{SC+} = \frac{1}{T - T_0} \sum_{t > T_0} Y_t - X_t \hat{w}_{SC}^0 - \frac{1}{T_0} \sum_{t \leq T_0} Y_t - X_t \hat{w}_{SC}^0 \tag{16}$$

where $\hat{w}_{SC}^0$ are computed via constraint Least Squares from the beginning of the period to T_0 with weights summing up to one. Similarly, for the Synthetic Learner, we consider

$$\widehat{\text{ATE}}_{SL} = \frac{1}{T-T_0} \sum_{t>T_0} Y_t - \mathbf{g}(X_t)\widehat{\mathbf{w}_0} - \frac{1}{T_0} \sum_{t\leq T_0} Y_t - \mathbf{g}(X_t)\widehat{\mathbf{w}_0}. \tag{17}$$

An important point worth noting, is that the treatment assignment indicator $D_t \perp X_s$ for all s, t. Henceforth, any method that subtracts the mean of the outcome over the post-treatment period from the mean of the outcome on the pre-treatment period is an unbiased estimator. To see this, notice that since $Y_t = f(X_t) + a^o D_t + \epsilon_t$ and $X_t \perp D_s$ for all s, t, $\mathbb{E}[\bar{Y}_{t>T_0} - \bar{Y}_{t\leq T_0}] = a^o$. Therefore, bias arises when the method is not able to capture the mean of the outcome. On the other hand, a better estimation of $f(\cdot)$ can lead to an improvement in the variance of the estimator of the ATE. Indeed this is what we observe in finite samples. The variance of the Synthetic Learner is the lowest, after excluding **DGP1** and **DGP5**. The bias of the estimators tends to be close to zero in most of the cases considered. These two phenomena are visible in Figure 11.

4.3.2. *Heterogeneous treatment effects.* We set up examples below according to **DGP1** with *heterogenous* treatment effect sbeing either linear or quadratic, namely

(1) (Linear) $a_t = \tau(X_t) = \sum_j X_{j,t}$;
(2) (Quadratic) $a_t = \tau(X_t) = \sum_j X_{j,t}^2$.

We compare the cases with $\hat{\tau}(\cdot)$ being estimated via X-learning motivated by Künzel et al. (2019) and T-Learning. T-Learning is done using the algorithm as described in the main text. X-learning adds a modification to the original algorithm as described below in Algorithm 5.

Algorithm 5 Synthetic X-Learner

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

1: Estimate $\widehat{\mathbf{w}_1}, \widehat{\mathbf{w}_0}, \widehat{\mathbf{g}}_0, \widehat{\mathbf{g}}_1$ as discussed in Section 2 and 3.3.

2: Compute pre-treatment and post-treatment residuals

$$\tilde{W}_t^1 = Y_t^1 - \widehat{\mathbf{g}}_0(X_t)\widehat{\mathbf{w}_0} \text{ for } t > T_0;$$
$$\tilde{W}_t^0 = Y_t^0 - \widehat{\mathbf{g}}_1(X_t)\widehat{\mathbf{w}_1} \text{ for } t \leq T_0.$$

3: Estimate $\tilde{\tau}^1(X_t)$ and $\tilde{\tau}^0(X_t)$ by training respectively $(\tilde{W}_t^1, X_t)$, $t > T_0$ and $(\tilde{W}_t^0, X_t)$, $t \leq T_0$ with Synthetic Learner;

return Estimate the treatment effect

$$\hat{\tau}(X_t) = \frac{T_0}{T}\tilde{\tau}^0(X_t) + \frac{T-T_0}{T}\tilde{\tau}^1(X_t).$$

X-Learning, computes $\tilde{\tau}$ and $\tilde{\tau}^0$ using Lasso only (with cross-validation) and X-learning+ uses averages of learners using the exponential weighting scheme. Learners include Lasso, XGboost, Support Vector Regression and ARIMA(0,1,1), Random Forest and Least Squares. We study the performance of the algorithms in terms of the square root of the MSE, namely

$$\sqrt{\text{MSE}(T)} = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(\hat{\tau}(X_t) - \tau(X_t))^2}. \tag{18}$$

Figure 12 collects the results. The two pictures at the top discuss the case of a linear effect. On the top-right we show the case of a balanced data set and on the top left we consider an unbalanced

FIGURE 11. On the y-axis we report the median bias (first 5 plots) or variance (last 5 plots) for estimating the ATE for different sample size (x-axis). SC and SC+ differ in that the latter has a bias correction factor. We used 800 replications.

case. In the case of a linear effect, X-Learning and X-Learning+ performs better than T-Learning for T large enough.

The pictures at the bottom of Figure 12 show the behavior of the MSE when the effects are quadratic. In simulations T-Learning performs better than X-Learning and X-Learning+. In addition, the MSE for X-Learning remains constant with the sample size, i.e. it does not decrease.

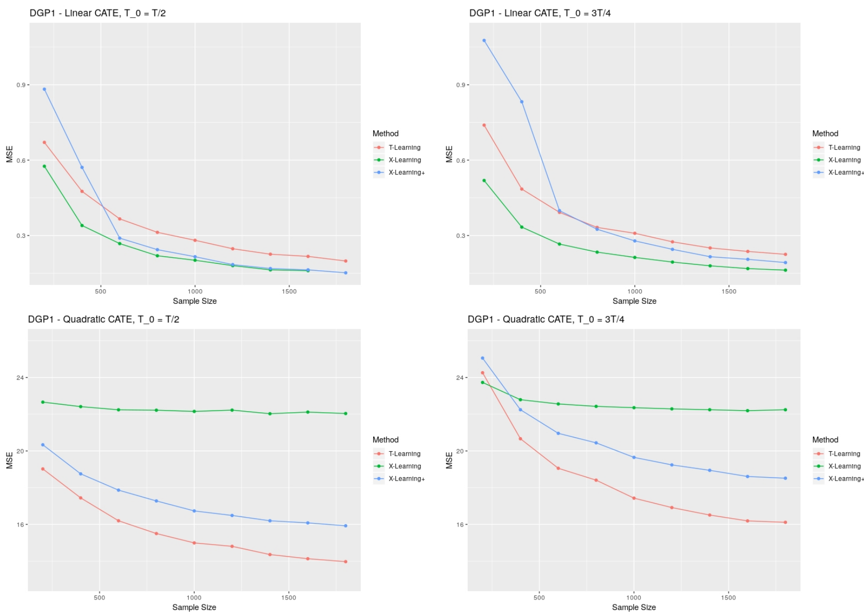

FIGURE 12. The y-axis reports the median $\sqrt{\text{MSE}}$ over 500 replications while the x-axis reports the sample size. We compare the Synthetic Learner with either T or X learning step. In simulations we consider **DGP1** with linear and quadratic CATE. X-learner performs Lasso on the residuals while X-Learning+ performs averaging over learners: Lasso, Support Vector Regression, ARIMA, XGboost.

The simulation shows a simple yet important point. If the learner in the second step of the X-learning regression is not able to capture the true relation between covariates and treatment effect, X-Learning leads to poorer performance when compared to T-Learning.

In Figure 13 we study the convergence rate of the mean squared error. In each figure we plot

$$\mathcal{S}(T) = \log\left(\sqrt{\text{MSE}(T)/\text{MSE}(\tilde{T})}\right), \quad T = \tilde{T} + 200. \tag{19}$$

For T large enough we would expect the ratio of the two MSE is close to the ratio of the rates. For instance, if

$$\text{MSE}(T) \leq \frac{C_0}{T^{2\alpha}} \Rightarrow \mathcal{S}(T) \leq \alpha(\log(T-200) - \log(T)).$$

For $\alpha = 1/2$ then MSE(T) decays at rate $1/T$.

To make comparisons with the rate of convergence we report $\mathcal{S}(T)$ computed for each algorithm; see the first two plots at the bottom of Figure 13. In addition we plot $\alpha(\log(T-200) - \log(T))$ for different values α that approximately match the observed convergence rate. For the case of linear effects, since we include Least Squares in the class of learners, the learner with optimal rate is Least Squares, and it achieves the parametric rate $1/T$. The MSE achieves the same rate which is $1/T$, providing suggestive evidence that the weighting scheme as proposed in the current paper achieves the rate of the optimal estimator in the class under consideration. This statement is stronger than our theoretical guarantees, and future research should check its validity.

Quadratic effects are presented in the last row of Figure 13. Linear regressors are no longer correctly specified and the convergence rate of the MSE crucially depends on the unknown convergence rate of the non-parametric methods such are XGboost or Random Forest. We observe that the rate of the MSE is closer to $1/\sqrt{T}$, far slower than the parametric rate. The slow convergence might be attributed to the fact that none of the learners achieves a slower rate than $1/\sqrt{T}$.

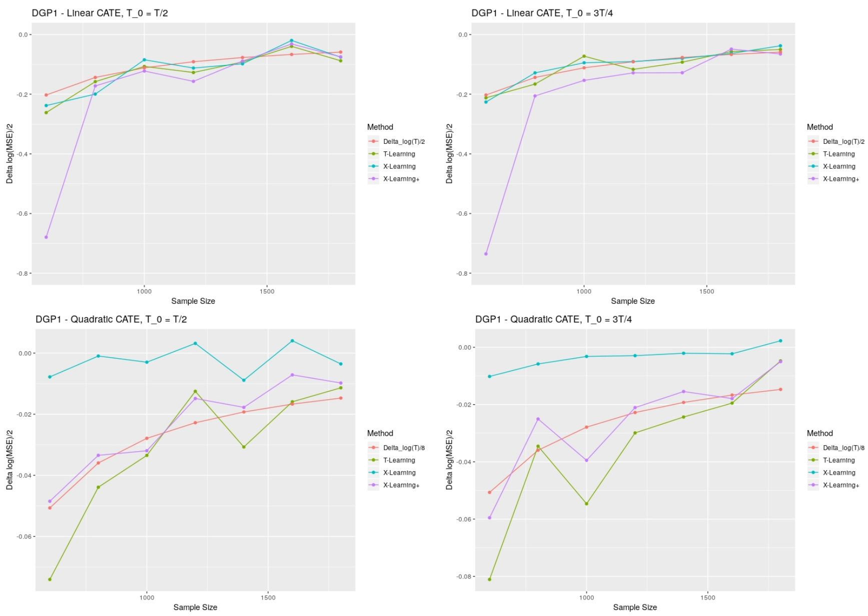

FIGURE 13. The ratio of mean square errors(corresponding to the difference of the logs, denoted as Delta log in the figure) for T and $T+ = 200$ as discussed in (19), for different T. We compare Synthetic Learner with T and X learning step. X-learning step is performed via Lasso, for X-Learning+ it is performed via the Synthetic Learner. We consider 500 replications.

5. The Effect of Public Health Insurance Ineligibility on Access to Medical Care

As mentioned in Section 1, we study the effects of the TennCare disenrolment program on the affordability of health-care expenses. Understanding the effect of public health insurance coverage on health care access is a major concern in health economics. There is a large literature in health economics that studies the relationship between public health insurance eligibility on health outcomes and access to health care. Examples are Kolstad and Kowalski (2012), Long et al. (2009), Baicker et al. (2013), Anderson et al. (2012) among others. The large literature reflects a keen interest of health economists in the following question: does public health insurance coverage improve access to medical care? The absence of large experimental studies has challenged the researcher in finding quasi-experimental designs for answering this question. Examples were Oregon and the Massachutes Medicaid enrollment program together with the TennCare disenrolment program, studied, among others, in Garthwaite et al. (2014) and Tello-Trillo (2016).

The TennCare disenrolment program represents the largest reduction in public health insurance coverage ever experienced in the US. Between 2005 and 2006 approximately 170,000 individuals lost public health insurance coverage. Most of these individuals were childless adults, who gained public health insurance coverage approximately ten years before, in 1994, during the expansion of the Medicaid program in Tennessee. In this section, we study the effect of the reform over childless adults on delayed access to medical care due to medical costs. This population is of particular interest since most of the Affordable Care Act expansions target childless adults. Tello-Trillo (2016) estimates that the TennCare disenrolment "significantly decreased the likelihood of having health insurance between 2 and 5 percent". The author estimates a non-significant increase of 1.3 percentage points on the probability of not seeing a doctor because of medical costs[4]. Our analysis provides formal evidence of a significant effect of the disenrolment program on health care access,

[4] The reader might refer to Panel B, Table 6 in Tello-Trillo (2016).

after controlling for state-level variability. Using the only Southern States, we estimate an average treatment effect of 1.7 percentage points. We reject the sharp null hypothesis of no effect at 95% level controlling for either only Southern States or for all other US states that did not experience sharp changes in their public health insurance program. Placebo tests in the other Southern States fail to reject the null hypothesis of interest, providing suggestive evidence of control of the nominal size[5].

5.1. **Data.** We use BFRSS data for investigating the effect of the reform on the percentage of people that cannot afford health-care expenses for medical costs. BFRSS is a national survey which is continuously run over the years since 1984. The survey contains individual-specific information, including residence, state of health, access to health coverage and others. We focus our analysis on childless adults[6]. The dataset can be organized as a long sequence of monthly observations since we can cluster observations by date of the interview. On average, we observe 150 childless adults between 18 and 64 years old in Tennessee per each month from 2017 to 1993. The outcome variable is the monthly percentage of childless adults who answered yes to the following survey question:"Was there a time in the past 12 months when you needed to see a doctor but could not because of the cost?".

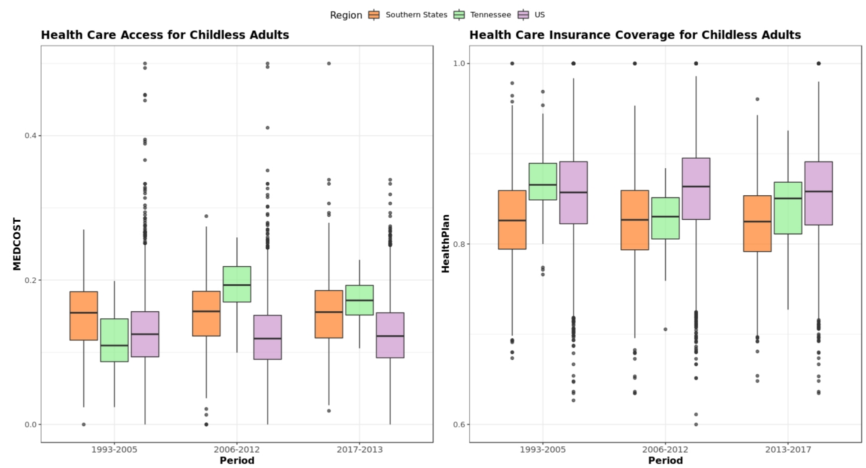

FIGURE 14. Sample distribution of chidless adults between 18 and 64 years old in Tennessee, the Southern States and the US who were not able to afford health care expenses(left panel) and who are covered by health insurance(right panel). BFRSS data.

In Figure 15 we report a box-plot on the number of monthly observations of childless adults in Tennessee between 8 and 64 years old over each year. As shown in the figure, for most of the months there are enough observations to construct a valid estimate of the proportion of the population of childless adults who would answers respectively yes to the questions above. On the other hand, there are a few months, such as one month in 2004 where we have no or very few observations. For these specific cases, we use linear interpolation. In all plots, we will smooth the time series using a local polynomial smoother.

In Figure 14 we report the distribution of respondents who were not able to afford medical cost in the past 12 months(left panel) and who are covered by health insurance[7](right panel), after

[5]For replication of the results and a more comprehensive set of plots, the reader might visit dviviano.github.io.

[6]In this sense, our estimates are valid for this sub-population.

[7]For the latter questions we count the number of individuals who answer yes to the question: "Do you have any kind of health care coverage, including health insurance, prepaid plans such as HMOs, or government plans such as Medicare or Indian Health Service?" We consider observations who answer "I do not know" as not having a plan.

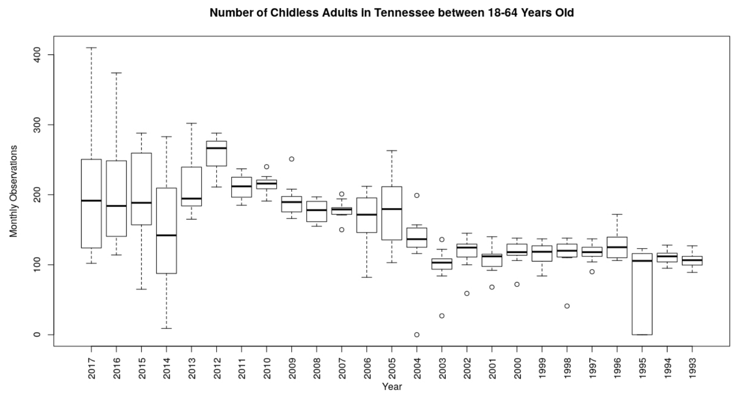

FIGURE 15. Number of monthly observations of chidless adults who are resident in Tennessee between 2017 and 1993. BFRSS data.

clustering over the period 1993-2005, 2006-2012 and 2013-2017 for Tennessee, other Southern States and the United States. We observe a shift in the mean of the outcome of Tennessee over these three periods, with a larger shift in the period just after the policy, between 2006 and 2012 while the variance remains approximately stable.

As proposed in the Synthetic Control literature (Abadie et al., 2010), we impute the potential outcome under no disenrolment using as a set of control variables the other states. In particular, we use all other states, after excluding Puerto Rico, since it does not contain enough observations, Oregon and Massachutes since both states experienced impactful Medicaid reform over the years under consideration. To control for confounders, we replicate our analysis using the only Southern States as discussed in the next lines. Throughout our analysis, we consider December 2005 the starting date of the policy, corresponding to six months after the beginning of the disenrolment program[8].

5.2. **Results.** We construct the "Synthetic Control" using the Synthetic Learner described in the current paper. We consider the share of individuals in other countries who were not able to afford necessary health care expenses as control variables, after removing Puerto Rico, Massachutes and Oregon for the reasons described in the previous subsection. We train Lasso, XGboost, Support Vector Regression and standard Synthetic Control method (Abadie et al., 2010) as base learners. Hyperparameters are chosen via cross-validation. In particular, we use the built-in function to the package glmnet to cross-validate Lasso. We validate the choice of the hyperparameter of the weights (η) using a two-sample splitting rule. The validation step is also performed in the bootstrap, and compute weights using the exponential weighting scheme. We consider the following sample splitting rule. Observations between 2000 and 2006 are used for the training of the algorithms; observations between 1993 and 1999 are used to compute the weights and for the bootstrap. Observations from January 2006 onwards are used to compute the test statistic.

[8]The overall disenrolment started in July 2005 and it lasted until June 2006. Most Childless adults who disenroled during this period were not able to requalify for Medicare (Garthwaite et al., 2014).

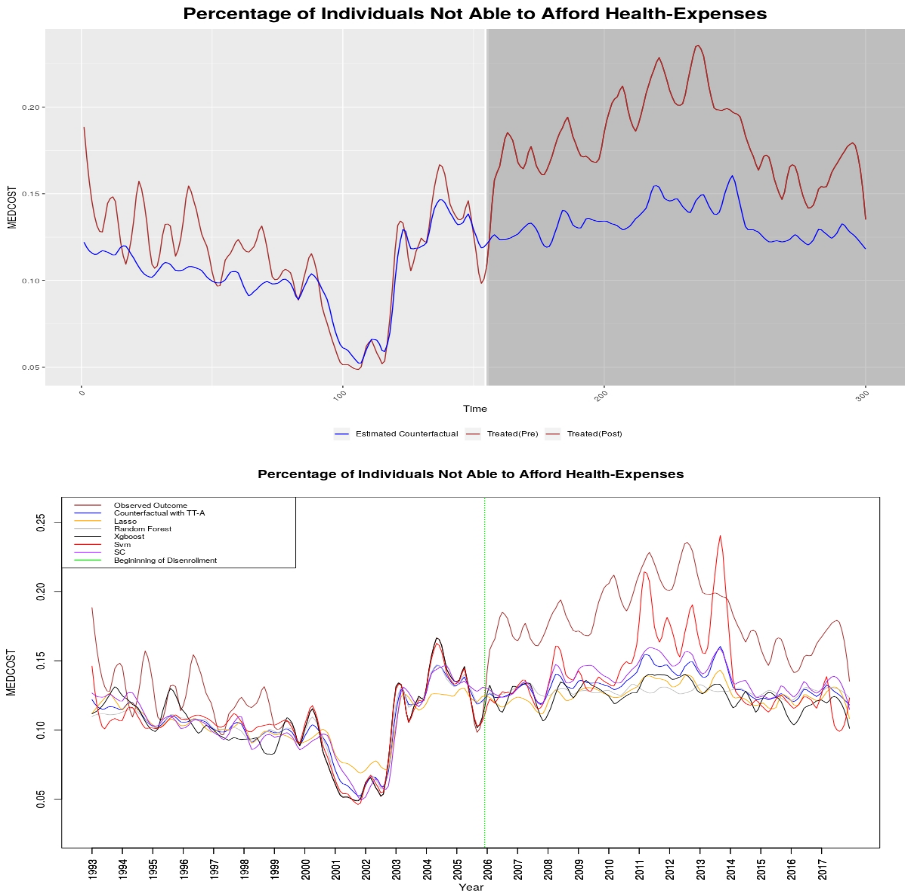

FIGURE 16. Observed and predicted counterfactual of percentage of childless adults who are not able to afford health-care expenses. The brown line is the observed time series in Tennessee and blue line is the estimated counterfactual under no-disenrollment. Gray region denote the period under treatment. The picture below reports the same quantities and the prediction of the counterfactual for each base-learner used. In the plot the time series is smoothed using a local polynomial.

In Figure 16 we plot the observed outcome and the estimated counterfactual for the percentage of individuals not having economic access to health-care. The picture at the bottom shows the prediction of each learner. The Synthetic Learner predicts an effect which is larger than the one predicted by Synthetic Control but smaller than other algorithms such as Random Forest, as shown in Figure 16.

To check for stationarity of observed time-series, we test for unit roots at 95% confidence level. We reject the null hypothesis of a unit root in the time series of interest displayed in Figure 16. We use an Augmented Dickey-Fuller test, with constant and without time trend and we include one, two and three lags. P-values are respectively < 0.01 for the first two tests and 0.05 for the latter. In Table 2 we report the estimated test statistic for testing the null hypothesis of no effect, namely $H_0 : Y_t(\mathbf{0}) - Y_t(\mathbf{1}) = 0,\ t > T_0$. We test the long-run effect for different levels of carry-over effects, observing that the effect falls always outside the 95% confidence region.

The spike in the series in 2014 is likely to be attributed to Obamacare's launch in 2014. The Affordable Health Care Act, also known as Obama Care, was officially approved in 2010 but the major changed entered into force in 2014. The reform drastically changed the individual insurance market and Medicaid expansion. Whereas the former was implemented at a country-level, Medicaid

Table 2. T-Statistic and 95% acceptance region for testing the increase in the percentage of childless adults not able to afford health care expenses in Tennessee.Different carry-over effects, m, indicating the number of months of carry-over effects is considered. The first two rows report results when observations from all the States, except for Oregon, Massachutes and Puerto Rico are used as covariates. The third and fourth row report results when only observations of Southern States that did not adopt the state-level ObamaCare Medicaid expansion are used as covariates. The last two rows report the corresponding estimates of the average treatment effects.

HealthCare Unaffordability	$m = 0$	$m = 12$	$m = 48$	$m = 84$
Test-Statistic all States	0.043	0.041	0.035	0.021
95% Accept. Region	$(0.039, 0.026)$	$(0.035, 0.023)$	$(0.031, 0.016)$	$(0.018, 0.009)$
Test-Statistic Southern	0.026	0.024	0.020	0.011
95% Accept. Region	$(0.018, 0.011)$	$(0.017, 0.010)$	$(0.015, 0.0008)$	$(0.012, 0.005)$
ATE all States	0.03	0.03	0.032	0.02
ATE Southern	0.017	0.018	0.018	0.008

expansion must be approved by individual states. Tennessee, together with the majority of southern states did not approve Medicaid expansion, whereas changes in the private individual insurance market might have had substantial effects on health insurance coverage.

To control for potential confounders, we repeat the study on the effects of TennCare disenrolment program on childless adults controlling only for southern states that did not expand Medicaid between 2010 and 2017, namely South and North Carolina, Mississippi, Alabama, Florida and Georgia. Results when using both all the US states and only southern states are reported in Table 2. Acceptance regions are estimated via bootstrap as discussed in previous sections.

Results on the weights of the learners for both cases are reported in Table 3. When using all the states, we observe that the performance of the learner is comparable to the pre-treatment period. One intuitive explanation is that in this case, each learner has access to enough information in order to accurately predict the time series, while regularization methods avoid overfitting. On the other hand, when we consider only few control units, namely the only Southern States, the performance of the methods differ, suggesting that in the presence of limited information using individual methods can lead to very different predictions and hence different results.

Table 3. Exponential weights over learners estimated over the period 1993-1999.

learners	All US states	Southern states
Lasso	0.20	0.23
Random Forest	0.19	0.27
XGboost	0.20	0.08
Support Vector Regression	0.20	0.03
Synthetic Control	0.20	0.37

In Figure 17 we report the test statistics and the acceptance regions Tennessee and for placebo tests performed on the other southern states that did not adopt Medicaid expansion. A placebo test consists in testing the effect of a policy from 2006 to 2017 in a state different from Tennessee. Since none of the other southern states had significant changes in the Medicaid system, we would expect no rejections for all southern states except for Tennessee. This is shown in Figure 17. Results are consistent with previous literature showing that the disenrollment program had a significant effect on health care unaffordability by approximately two percentage points, after controlling for other states. Results are robust to different choices of carry-over effects.

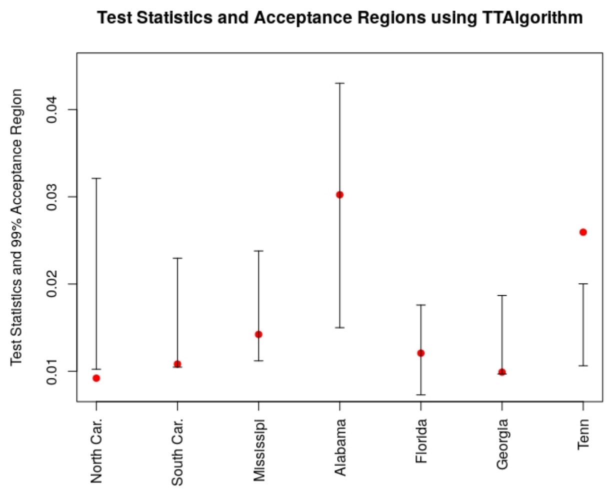

FIGURE 17. We test significant changes in the percentage of childless adults who are not able to afford medical expensense in those southern States that did not adopt the Medicaid Obama-Care. We report the test statistic(red dot) and 99% confidence region for each of the state, including Tennessee.

6. Discussion

In this paper, we have introduced a novel strategy for estimating treatment effects and testing the null hypothesis of interest in the presence of time-dependent observations. Motivated by applications in the social sciences, we have considered the scenario of one treated unit observed before and after the treatment with controls serving as covariates. We developed a novel algorithm, denoted as Synthetic Learner that predicts the counterfactual building on multiple regression methods. In practice, practitioners seek to predict counterfactuals using many models with unknown theoretical properties. Our framework provides a starting point for performing inference which is valid regardless of the class of models under consideration. Building on the Neyman-Rubin potential outcome framework, we outlined the importance of considering carry-over effects, i.e., treatment effects that propagate over time. We use a simple yet effective strategy that takes into account these effects. Carryover effects, first discussed in Robins (1986), are often ignored in the literature on Synthetic Control, while they can bring substantial bias to standard estimators. The presence of one single treated unit with a given time of the adoption of the policy brings substantial challenges from an identification perspective. We considered three scenarios of increasing complexity. First we consider the case of the adoption date being deterministic, T_0 and fixed treatment effects similarly to Chernozhukov et al. (2017), Chernozhukov et al. (2018), Arkhangelsky et al. (2018) among others. We show that, under stationarity and mixing conditions, our algorithm controls the nominal size regardless of the class of base-algorithm under consideration, even in the presence of misspecification bias. The case of deterministic time of the adoption of the policy is necessary for theoretical guarantees with one single treated unit. Extending this result to non-deterministic T_0 is conceptually feasible in the presence of multiple units treated at different points in time. Moreover, we show that the estimator for the average treatment effect is consistent under weak assumptions, letting T_0 be non-deterministic. Third, in the same spirit of Boruvka et al. (2018), we considered the case of treatment effects being potentially heterogeneous in the population. We provide bounds on the predictive performance under this complex scenario, and we pioneer the idea of estimating heterogeneous treatment effect via T-Learning and X-learning for time-dependent observations.

Introducing heterogenous treatment effects in the population requires to carefully re-define standard causal estimands of interest (e.g., the propensity score). The presence of only a few (or just one) treated units under staggered adoption induces additional challenges compared to the standard framework of causal inference in longitudinal studies (Robins et al., 1999, 2000). We leave to future research addressing this question. Our paper also opens new questions on constructing valid machine learning methods for causal inference when units exhibit dependence. Such dependence can either be time dependence or dependence induced by a network structure, such as in the case of spillover effects. Double machine learning methods proposed in Chernozhukov et al. (2018) requires exchangeable observations while the theoretical validity of the causal forest algorithm (Wager and Athey, 2018) crucially relies on the assumption of *iid* observations. In our paper, we briefly discussed alternatives to account for the time dependence while leaving to future research a more comprehensive study on this topic.

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.

Abadie, A., A. Diamond, and J. Hainmueller (2015). Comparative politics and the synthetic control method. *American Journal of Political Science 59*(2), 495–510.

Abbring, J. H. and J. J. Heckman (2007). Econometric evaluation of social programs, part iii: Distributional treatment effects, dynamic treatment effects, dynamic discrete choice, and general equilibrium policy evaluation. *Handbook of econometrics 6*, 5145–5303.

Abraham, S. and L. Sun (2018). Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. *arXiv preprint arXiv:1804.05785*.

Amjad, M., D. Shah, and D. Shen (2018). Robust synthetic control. *The Journal of Machine Learning Research 19*(1), 802–852.

Anderson, M., C. Dobkin, and T. Gross (2012). The effect of health insurance coverage on the use of medical services. *American Economic Journal: Economic Policy 4*(1), 1–27.

Arkhangelsky, D., S. Athey, D. A. Hirshberg, G. W. Imbens, and S. Wager (2018). Synthetic difference in differences. *arXiv preprint arXiv:1812.09970*.

Athey, S. (2018). The impact of machine learning on economics. In *The Economics of Artificial Intelligence: An Agenda*. University of Chicago Press.

Athey, S., M. Bayati, N. Doudchenko, G. Imbens, and K. Khosravi (2018). Matrix completion methods for causal panel data models. Technical report, National Bureau of Economic Research.

Athey, S. and G. Imbens (2016). Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences 113*(27), 7353–7360.

Athey, S. and G. W. Imbens (2015). Machine learning methods for estimating heterogeneous causal effects. *stat 1050*(5).

Athey, S. and G. W. Imbens (2018). Design-based analysis in difference-in-differences settings with staggered adoption. Technical report, National Bureau of Economic Research.

Athey, S., G. W. Imbens, and S. Wager (2018). Approximate residual balancing: debiased inference of average treatment effects in high dimensions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 80*(4), 597–623.

Athey, S., J. Tibshirani, S. Wager, et al. (2019). Generalized random forests. *The Annals of Statistics 47*(2), 1148–1178.

Bai, J. (2009). Panel data models with interactive fixed effects. *Econometrica 77*(4), 1229–1279.

Baicker, K., S. L. Taubman, H. L. Allen, M. Bernstein, J. H. Gruber, J. P. Newhouse, E. C. Schneider, B. J. Wright, A. M. Zaslavsky, and A. N. Finkelstein (2013). The oregon experiment—effects of medicaid on clinical outcomes. *New England Journal of Medicine 368*(18), 1713–1722.

Belloni, A., V. Chernozhukov, I. Fernández-Val, and C. Hansen (2017). Program evaluation and causal inference with high-dimensional data. *Econometrica 85*(1), 233–298.

Belloni, A., V. Chernozhukov, and C. Hansen (2014). High-dimensional methods and inference on structural and treatment effects. *Journal of Economic Perspectives 28*(2), 29–50.

Ben-Michael, E., A. Feller, and J. Rothstein (2018). The augmented synthetic control method. *arXiv preprint arXiv:1811.04170*.

Bishop, C. M. et al. (1995). *Neural networks for pattern recognition.* Oxford university press.

Blackwell, M. and A. N. Glynn (2018). How to make causal inferences with time-series cross-sectional data under selection on observables. *American Political Science Review 112*(4), 1067–1082.

Boruvka, A., D. Almirall, K. Witkiewitz, and S. A. Murphy (2018). Assessing time-varying causal effect moderation in mobile health. *Journal of the American Statistical Association 113*(523), 1112–1121.

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

Breiman, L. (2001). Random forests. *Machine learning 45*(1), 5–32.

Brodersen, K. H., F. Gallusser, J. Koehler, N. Remy, S. L. Scott, et al. (2015). Inferring causal impact using bayesian structural time-series models. *The Annals of Applied Statistics 9*(1), 247–274.

Burges, C. J. (1998). A tutorial on support vector machines for pattern recognition. *Data mining and knowledge discovery 2*(2), 121–167.

Carrasco, M. and X. Chen (2002). Mixing and moment properties of various garch and stochastic volatility models. *Econometric Theory 18*(1), 17–39.

Cesa-Bianchi, N., Y. Freund, D. Haussler, D. P. Helmbold, R. E. Schapire, and M. K. Warmuth (1997). How to use expert advice. *Journal of the ACM (JACM) 44*(3), 427–485.

Cesa-Bianchi, N. and G. Lugosi (2006). *Prediction, learning, and games.* Cambridge university press.

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

Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal 21*(1), C1–C68.

Chernozhukov, V., K. Wuthrich, and Y. Zhu (2017). An exact and robust conformal inference method for counterfactual and synthetic controls. *arXiv preprint arXiv:1712.09089*.

Chernozhukov, V., K. Wuthrich, and Y. Zhu (2018). Inference on average treatment effects in aggregate panel data settings. *arXiv preprint arXiv:1812.10820*.

Doudchenko, N. and G. W. Imbens (2016). Balancing, regression, difference-in-differences and synthetic control methods: A synthesis. Technical report, National Bureau of Economic Research.

Drucker, H., C. J. Burges, L. Kaufman, A. J. Smola, and V. Vapnik (1997). Support vector regression machines. In *Advances in neural information processing systems*, pp. 155–161.

Fang, Z. and A. Santos (2018). Inference on directionally differentiable functions. *The Review of Economic Studies 86*(1), 377–412.

Farrell, M. H., T. Liang, and S. Misra (2018). Deep neural networks for estimation and inference: Application to causal effects and other semiparametric estimands. *arXiv preprint arXiv:1809.09953*.

Firpo, S. and V. Possebom (2018). Synthetic control method: Inference, sensitivity analysis and confidence sets. *Journal of Causal Inference 6*(2).

Fithian, W., D. Sun, and J. Taylor (2014). Optimal inference after model selection. *arXiv preprint arXiv:1410.2597*.

Garthwaite, C., T. Gross, and M. J. Notowidigdo (2014). Public health insurance, labor supply, and employment lock. *The Quarterly Journal of Economics 129*(2), 653–696.

Green, D. P. and H. L. Kern (2010). Modeling heterogeneous treatment effects in large-scale experiments using bayesian additive regression trees. In *The annual summer meeting of the society of political methodology*.

Hahn, J. and R. Shi (2017). Synthetic control and inference. *Econometrics 5*(4), 52.

Hahn, P. R., J. Murray, and C. M. Carvalho (2017). Bayesian regression tree models for causal inference: regularization, confounding, and heterogeneous effects. *Confounding, and Heterogeneous Effects (October 5, 2017)*.

Hamilton, J. D. (1994). *Time series analysis*, Volume 2. Princeton university press Princeton, NJ.

Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica: Journal of the Econometric Society*, 1029–1054.

Hartford, J., G. Lewis, K. Leyton-Brown, and M. Taddy (2017). Deep iv: A flexible approach for counterfactual prediction. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 1414–1423. JMLR. org.

Hazlett, C. and Y. Xu (2018). Trajectory balancing: A general reweighting approach to causal inference with time-series cross-sectional data. *https://ssrn.com/abstract=3214231*.

Hernán, M. and J. Robins (2015). Causal inference book.

Hsiao, C., H. Steve Ching, and S. Ki Wan (2012). A panel data approach for program evaluation: measuring the benefits of political and economic integration of hong kong with mainland china. *Journal of Applied Econometrics 27*(5), 705–740.

Iavor Bojinov, N. S. (2019, Forthcoming). Time series experiments and causal estimands: exact randomization tests and trading. *Journal of the American Statistical Association*.

Imai, K. and I. S. Kim (2019, Forthcoming). When should we use linear fixed effects regression models for causal inference with longitudinal data? *American Political Science Review*.

Imai, K., I. S. Kim, and E. Wang (2018). Matching methods for causal inference with time-series cross-section data. *https://imai.fas.harvard.edu/research/files/tscs.pdf*.

Imai, K., M. Ratkovic, et al. (2013). Estimating treatment effect heterogeneity in randomized program evaluation. *The Annals of Applied Statistics 7*(1), 443–470.

Imbens, G. W. and D. B. Rubin (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

Kallus, N. (2018). Deepmatch: Balancing deep covariate representations for causal inference using adversarial training. *arXiv preprint arXiv:1802.05664*.

Knaus, M., M. Lechner, and A. Strittmatter (2018). Machine learning estimation of heterogeneous causal effects: Empirical monte carlo evidence. *arXive preprint arXiv:1810.13237*.

Kolstad, J. T. and A. E. Kowalski (2012). The impact of health care reform on hospital and preventive care: evidence from massachusetts. *Journal of Public Economics 96*(11-12), 909–929.

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

Künzel, S. R., J. S. Sekhon, P. J. Bickel, and B. Yu (2019). Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences 116*(10), 4156–4165.

Lahiri, S. N. et al. (1999). Theoretical comparisons of block bootstrap methods. *The Annals of Statistics 27*(1), 386–404.

Lange, T., A. Rahbek, and S. T. Jensen (2011). Estimation and asymptotic inference in the ar-arch model. *Econometric Reviews 30*(2), 129–153.

Lee, O. (2005). Probabilistic properties of a nonlinear arma process with markov switching. *Communications in Statistics-Theory and Methods 34*(1), 193–204.

Long, S. K., K. Stockley, and A. Yemane (2009). Another look at the impacts of health reform in massachusetts: evidence using new data and a stronger model. *American Economic Review 99*(2), 508–11.

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

McCaffrey, D. F., G. Ridgeway, and A. R. Morral (2004). Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological methods 9*(4), 403.

McDonald, D. J. and C. R. Shalizi (2011). Rademacher complexity of stationary sequences. *arXiv preprint arXiv:1106.0730*.

Pham, T. D. and L. T. Tran (1985). Some mixing properties of time series models. *Stochastic processes and their applications 19*(2), 297–303.

Politis, D. N. and J. P. Romano (1992). A circular block-resampling procedure for stationary data. *Exploring the limits of bootstrap 2635270*.

Politis, D. N. and J. P. Romano (1994). The stationary bootstrap. *Journal of the American Statistical association 89*(428), 1303–1313.

Polley, E. C. and M. J. Van Der Laan (2010). Super learner in prediction.

Powers, S., J. Qian, K. Jung, A. Schuler, N. H. Shah, T. Hastie, and R. Tibshirani (2017). Some methods for heterogeneous treatment effect estimation in high-dimensions. *arXiv preprint arXiv:1707.00102*.

Rambachan, A. and N. Shephard (2019). A nonparametric dynamic causal model for macroeconometrics. *arXiv preprint arXiv:1903.01637*.

Rigollet, P., A. B. Tsybakov, et al. (2012). Sparse estimation by exponential weighting. *Statistical Science 27*(4), 558–575.

Rinaldo, A., L. Wasserman, M. G'Sell, J. Lei, and R. Tibshirani (2016). Bootstrapping and sample splitting for high-dimensional, assumption-free inference. *arXiv preprint arXiv:1611.05401*.

Robins, J. (1986). A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Mathematical modelling 7*(9-12), 1393–1512.

Robins, J. M. (1989). The analysis of randomized and non-randomized aids treatment trials using a new approach to causal inference in longitudinal studies. *Health service research methodology: a focus on AIDS*, 113–159.

Robins, J. M., S. Greenland, and F.-C. Hu (1999). Estimation of the causal effect of a time-varying exposure on the marginal mean of a repeated binary outcome. *Journal of the American Statistical Association 94*(447), 687–700.

Robins, J. M., M. A. Hernan, and B. Brumback (2000). Marginal structural models and causal inference in epidemiology.

Schapire, R. E. and Y. Freund (2012). *Boosting: Foundations and algorithms.* MIT press.

Taddy, M., M. Gardner, L. Chen, and D. Draper (2016). A nonparametric bayesian analysis of heterogenous treatment effects in digital experimentation. *Journal of Business & Economic Statistics 34*(4), 661–672.

Tello-Trillo, D. S. (2016). Effects of losing public health insurance on health care access, utilization, and health outcomes. Technical report, Working paper, University of Virginia. Accessed 14 October 2017. Available . . . .

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

Varian, H. R. (2016). Causal inference in economics and marketing. *Proceedings of the National Academy of Sciences 113*(27), 7310–7315.

Vaver, J. and J. Koehler (2011). Measuring ad effectiveness using geo experiments. *Google Inc.*

Wager, S. and S. Athey (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association 113*(523), 1228–1242.

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

White, H. and I. Domowitz (1984). Nonlinear regression with dependent observations. *Econometrica: Journal of the Econometric Society*, 143–161.

Xu, Y. (2017). Generalized synthetic control method: Causal inference with interactive fixed effects models. *Political Analysis 25*(1), 57–76.

Yu, B. (1994). Rates of convergence for empirical processes of stationary mixing sequences. *The Annals of Probability*, 94–116.

Zhu, Y., D. L. Coffman, and D. Ghosh (2015). A boosting algorithm for estimating generalized propensity scores with continuous treatments. *Journal of causal inference 3*(1), 25–40.

Supplementary Materials

Appendix Appendix A Theorem 3.1

A.1 Definitions.

Definition A.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 $h \to \infty$

Theoretical analysis of the Synthetic Learner builds on literature on empirical processes and the proof builds on Lunde and Shalizi (2017). 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

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

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

Definition A.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

$$\sup_{f\in BL_1} |\mathbb{E}[f(\mathbb{X}_n)] - \mathbb{E}[f(\mathbb{X})]| = o_\mathbb{P}(1).$$

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 A.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 $n \ge 1$ 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). The notion is stronger than necessary for the Functional Delta Method, but it is necessary for the consistency of the bootstrap (Fang and Santos, 2018). 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 A.4. Let the weights be $w_0(\cdot) : \mathcal{A} \subset l^\infty(\mathbb{R}^p, \mathbb{R}) \to \mathcal{W}$, where $\mathcal{W} = [0,1]^p$.

The weights are now reparametrized to be the function of the cdf function instead of the observations directly. Observer that whenever the loss function is $l(x,y) = (x-y)^2$ then the ith entry of the weights computed using exponential weighting scheme

$$w_0^{(i)}(F) = \frac{\exp(-\eta \int (z_1 - z_i)^2 dF)}{\sum_{j=1}^p \exp(-\eta \int (z_1 - z_j)^2 dF)}$$

where $z \in \mathbb{R}^{p+1}$ and z_i is the $i-$th coordinate of the integral.

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 A.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.

We define $W_t = (Y_t^o, \hat{\mathbf{g}}(X_t))$. We derive our proof conditional $t > 1$, namely conditional on the sample of observations used to train the learners. Notice that we can write $\hat{g}_i(X_t) = h_i(X_t, X_0, Y_0, \ldots, X_{T_-}, Y_{T_-})$ for some function h_i. We will condition on the filtration $\mathcal{F}_0$. Conditional on $\mathcal{F}_0$, $\hat{g}_i(X_t)$ is only a function of X_t. By definition of strict stationarity, stationarity of $\{W_t\}$ follows by assumption. Let denote the empirical measure for control and treatment period, respectively,

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

Similarly $\mathcal{P}_T^*, \mathcal{P}_{T_0}^*$ denote the bootstrapped counterparts. 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 $F_{T_0}(s) = \mathcal{P}_{T_0} f_s$ and $F_T(s) = \mathcal{P}_T f_s$ respectively and similarly this hold for bootstrap measures. Notice that we can see $F_{T_0}(\cdot)$ and $F_T(\cdot)$ 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

$$\mathbb{H}_T(f_s) = \sqrt{T_0}\begin{bmatrix} \mathcal{P}_{T_0} f_s - \mathcal{P} f_s \\ \mathcal{P}_T f_s - \mathcal{P} f_s \end{bmatrix}, \quad \mathbb{H}_T^*(f_s) = \sqrt{T_0}\begin{bmatrix} \mathcal{P}_{T_0}^* f_s - \mathcal{P}_{T_0} f_s \\ \mathcal{P}_T^* f_s - \mathcal{P}_T f_s \end{bmatrix}. \tag{20}$$

We express the test statistics of interest as functionals of $\mathbb{H}_T$ and a fixed null trajectory $\{a_t^o\}$. That is, for (7) we can define $(A,B) \to T_S(A,B)$ with

$$T_S(A,B) = \int S_0(z, w_0(A)) dB.$$

Similarly, for (24) we can define $(A,B) \mapsto T_A(A,B)$ with $T_A(A,B) = \left(\int z_1 dB - \int w_0(A) z_2 dB - a^o\right)^2$.

A.2 Auxiliary Lemmas.

Lemma A.1. *(Functional Delta Method for the Bootstrap, Kosorok (2008), Theorem 12.1) For normed spaces $\mathbb{D}$ and $\mathbb{E}$ let $\phi : \mathbb{D}_\phi \subset \mathbb{D} \to \mathbb{E}$ be an Hadamard differentiable map at θ, tangential to $\mathbb{D}_0 \subset \mathbb{D}$ with derivative ϕ'_θ. Let $\mathbb{X}_n$ and $\mathbb{X}_n^*$ have values in $\mathbb{D}_\phi$ with $r_n(\mathbb{X}_n - \theta) \rightsquigarrow \mathbb{X}$ where $\mathbb{X}$ is tight and takes values in $\mathbb{D}_0$ for some sequence of constants $0 < r_n \to \infty$, the maps $W_n \to h(\mathbb{X}_n)$ are measurable for every $h \in \mathbb{C}_b(D)$ almost surely and where $r_n c(\mathbb{X}_n^* - \mathbb{X}_n) \to \mathbb{X}$ in a weakly sense for $0 < c < \infty$, then $r_n c(\phi(\mathbb{X}_n^*) - \phi(\mathbb{X}_n)) \rightsquigarrow \phi'_\theta(\mathbb{X})$.*

The functional Delta Method for the bootstrap suggests that if we can prove Hadamard differentiability of the function of interest, as well as that the empirical and bootstrapped process converge to the same process that is tight and if further regularity conditions hold, then the bootstrap is consistent. To use Lemma A.1, we need to define the parameters of interest as functional that maps from a space of bounded functions to a Banach space. We now give a list of lemmas that will be used for applying the Functional Delta Method for the Bootstrap.

Lemma A.2. *(Lemma 3.8 in Lunde and Shalizi (2017)) Let Y be a β-mixing process with mixing rate that decays at least at a cubic rate. Consider $\mathbb{H}_t$ defined in (20). 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

$$\Gamma(f,g) = \lim_{k\to\infty} \sum_{i=1}^{\infty} \{\mathbb{E}[f(Z_k)g(Z_i)] - \mathbb{E}[f(Z_k)]\mathbb{E}[g(Z_i)]\}, \quad \forall f,g \in \mathcal{F}. \tag{21}$$

Lemma A.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 A.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 A.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*

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

Proof of Lemma A.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

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

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. Henceforht, exploting the definition of β-mixing given in Bradley et al. (2005) we have

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

□

A.3 Proof of the theorem.

Proof of Theorem 3.1. First, consider the case of no carry-over effects, namely $m = 0$ for the sake of clarity. The proof is organized as follow. First prove that $\mathbb{H}_T^*$ and $\mathbb{H}_T$ which we define as the empirical and bootstrapped measures as in (20), converge to the same process $\mathbb{H}$. We let $2T_0 = T$ for notational convenience. We study convergence properties conditional on $\mathcal{F}_0$ and we treat $\hat{g}$ as fixed as discussed in previous paragraphs. We study the distribution of $(Y_t^o, \hat{\mathbf{g}}(X_t))$ conditional on $\mathcal{F}_0$. By Lemma A.5 beta-mixing conditions on (Y_t^o, X_t) imply the same conditions on the beta-mixing coefficients of $(Y_t^o, \hat{\mathbf{g}}(X_t))$. Similarly, stationarity of (Y_t^o, X_t) also implies stationarity of $(Y_t^o, \hat{\mathbf{g}}(X_t))$.

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 A.2 $\mathbb{H}_T \to_d \mathbb{H}$. By Lemma A.3, $\mathcal{P}_{T_0}^* - \mathcal{P}_{T_0}$ and $\mathcal{P}_T^* - \mathcal{P}_T$ converges marginally to a Gaussian Process with covariance matrix described in Lemma A.3 [9]. Under the same argument as in A.2 in Lunde and Shalizi (2017) $\mathcal{P}_{T_0}^* - \mathcal{P}_{T_0}$ and $\mathcal{P}_T^* - \mathcal{P}_T$ are asymptotically independent, which implies that $\mathbb{H}^* \to_d \mathbb{H}$. The same lemma goes through if we consider $\mathcal{P}_{T_0}$ and $\mathcal{P}_T^{-m}$, where $\mathcal{P}_T^{-m}$ excludes observations from $T_0 + 1$ to $T_0 + m$, for fixed m, namely for fixed carryover effects m.

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) and it follows similarly as in Lunde and Shalizi (2017). We can see $T_S(\cdot, \cdot)$ as the composition of maps

$$T_S : (A, B) \overset{(a)}{\to} (B, w_0(A)) \overset{(b)}{\to} (B, S_0(z, w_0(A))) \overset{(c)}{\to} \int S_0(z, w_0(A)) dB.$$

The map (c) is Hadamard differentiable by Lemma A.4. We have to show that $S_0(z, w)$ is itself Hadamard differentiable in w at $w_0(F)$ that we write as $\bar{w}$ for short, where F denotes the distribution of the process. 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

$$\left\| \frac{S(., \bar{w} + t_n h_n) - S(., \bar{w})}{t_n} - S'_{\bar{w}}(., h) \right\|_\infty \to 0 \tag{22}$$

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

$$||2(z_1 - z_2'\bar{w})z_2'(h_n - h)||_\infty + |t_n|||(z_2'h_n)^2||_\infty. \tag{23}$$

For the first term in (23) 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.

[9]Conditions in Lemma A.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.

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.
Consider now $T_A(\cdot, \cdot)$ as

$$T_A(A, B) = \left(\int z_1 - z_2 w_0(A) dB - a^o \right)^2. \tag{24}$$

for a fixed a^o. Hadamard differentiability holds by the chain rule of the Hadamard derivative once we show differentiability of $w_0(A)$. In fact, $z_1 - w_0(A)z_2$ is bounded by the compact support assumption and it is easy to show that the function $f(x) = x^2$ for x being bounded is Hadamard differentiable.

We are left to show that $A \mapsto w_0(A)$ is Hadamard differentiable, for w_0 being either computed via Least Squares or using the exponential weighting scheme. 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. 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

$$\hat{w}_0^{(i)}(F) = \frac{\exp(-\eta \int l(z_1, z_i) dF)}{\sum_{j=2}^{p+1} \exp(-\eta \int l(z_1, z_j) dF)} \tag{25}$$

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 $w_0^{(i)}(F)$ is the composition of mappings:

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

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 $||h_n - h||_\infty \to 0$, $t_n \to 0$

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

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 $\bar{h}_n \in [h_n, h]$,

$$\begin{aligned} \frac{\exp(-\eta(x + t_n h_n)) - \exp(-\eta(x))}{t_n} &= \frac{\exp(-\eta x)[\exp(-\eta t_n h_n) - 1]}{t_n} \\ &= \frac{\exp(-\eta x)[-\eta t_n \exp(-\eta t_n \bar{h}_n) h_n]}{t_n} \\ &= -\eta \exp(-\eta x - \eta t_n \bar{h}_n) h_n \to -\eta \exp(-\eta x) h. \end{aligned} \tag{26}$$

Hence we have

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

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

$$\left\|\frac{\frac{x+t_nh_{n,1}}{y+t_nh_{n,2}}-\frac{x}{y}}{t_n}-\frac{h_1}{y}+h_2\frac{x}{y^2}\right\|_\infty$$
$$=\sup_{0\le x\le 1,y\ge re^{-\eta C}}\left\|\frac{(h_{n,1}-h_1)-(h_{n,2}-h_2)x/y-h_1t_nh_{n,2}/y+h_2xh_{n,2}t_n/y^2}{y+t_nh_{n,2}}\right\|_\infty$$
$$\le\sup_{0\le x\le 1,y\ge re^{-\eta C}}\frac{|h_{n,1}-h_1|}{|y+t_nh_{n,2}|}+\frac{|h_{n,2}-h_2||\frac{x}{y}|}{|y+t_nh_{n,2}|}+|t_n|\frac{|h_1h_{n,2}|/|y|}{|y+t_nh_{n,2}|}+\frac{|h_2h_{n,2}x/y^2|}{|y+t_nh_{n,2}|}|t_n| \tag{27}$$
$$\le\frac{|h_{n,1}-h_1|}{pe^{-\eta C}-|t_nh_{n,2}|}+\frac{|h_{n,2}-h_2|}{pe^{-\eta C}(pe^{-\eta C}-|t_nh_{n,2}|)}+|t_n|\frac{|h_1h_{n,2}|}{pe^{-\eta C}(pe^{-\eta C}-|t_nh_{n,2}|)}$$
$$+\frac{|h_2h_{n,2}|}{pe^{-2\eta C}(pe^{-\eta C}-|t_nh_{n,2}|)}|t_n|$$
$$\to 0.$$

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.

Finally we discuss Hadamard differentiability when weights are computed using generic potential weights. Clearly, whenever $t=1$ potential weights are equal to a constant, hence they are Hadamard differentiable. With a slight abuse of notation let $w_0(F_{t-1})$ be the weights computed at time $t-1$. At time t potential weights can be written as follow:

$$w_0^{(i)}(F_t)=\frac{\phi'(\eta_t\sum_{s=1}^t l(Y_t,\hat{\mathbf{g}}(X_t)w_0(F_{t-1})^{(i)})-l(Y_t,\hat{g}_i(X_t)))}{\sum_{j=1}^p\phi'(\eta_t\sum_{s=1}^t l(Y_t,\hat{\mathbf{g}}(X_t)w_0(F_{t-1})^{(j)})-l(Y_t,\hat{g}_j(X_t)))}.$$

Since $w_0(F_{t-1})$ is Hadamard differentiable, by the chain rule for the Hadamard derivative, following the same argument as discussed for exponential weights $w_0(F_t)^{(i)}$ is Hadamard differentiable if $\phi'(\cdot)$ is Hadamard differentiable. Henceforth, by the functional delta method for the bootstrap,consitency of the bootstrap follows conditional on $\mathcal{F}_0$. By Jensen's inequality and the law of iterated expectations, the result follows also unconditional on $\mathcal{F}_0$. The proof is complete[10].

□

[10]Further extensions of this proof can consider the case of (Y_t,X_t) stationary unconditional on $\mathcal{F}_0$ and study uniform convergence over $\mathcal{F}_0$.

Appendix Appendix B Proofs of the results in Section 3.2 and 3.3

B.1 Auxiliary Lemmas.

Lemma B.1. *Consider a measurable function* $\phi(Y_t, X_t, Y_0, X_0, Y_{-1}, X_{-1}, \ldots, Y_{T_-}, X_{T_-}) = \tilde{\phi}(Y_t, X_t, F_-)$ *where* $F_- \in \mathcal{D}$ *being the empirical distribution of* $(Y_0, X_0, Y_{-1}, X_{-1}, \ldots, Y_{T_-}, X_{T_-}) \in [-M, M]^{|T_-|} \times [-M, M]^{|T_-|}$*, with* $\mathcal{D}$ *being the space of probability measure on* $[-M, M]^{|T_-|} \times [-M, M]^{|T_-|}$*. Let* (X_t, Y_t) *being stationary and* α*-mixing. Assume that* $\mathbb{E}[\tilde{\phi}(Y_t, X_t, F_-)] < \infty$ *uniformly in* F_-*. Then*

$$\sup_{A \in \mathcal{D}} \left| \frac{1}{T} \sum_{t=1}^{T} \tilde{\phi}(Y_t, X_t, A) - \mathbb{E}[\tilde{\phi}(X_t, A)] \right| = o_p(1)$$

as $T \to \infty$*, for fixed* $T_- > -\infty$

Proof of Lemma B.1. Consider an underlying probability space $(\Omega, \mathcal{F}, Pr)$. By assumption, $\beta(\omega) = (Y_0(\omega), \ldots, Y_{T_-}(\omega), X_0(\omega), \ldots, X_{T_-}(\omega)) \in P$ with P being compact. Let (P, π) be the corresponding metric space. We will study uniform convergence of the process

$$\frac{1}{T} \sum_{t=1}^{T} \phi(X_t, Y_t, \beta(\omega))$$

over each possible realization of $\beta(\omega) \in P$. For simplicity in notation we omit the first two arguments in ϕ whenever it is clear from the context. Throughout the analysis we omit the argument in β for the sake of brevity. We now introduce the following notation.

$$\text{Mod}_\phi(\delta, \beta) = \sup\{|\phi(Y_t, X_t, \beta) - \phi(Y_t, X_t, \beta^*)| : \beta \in P \text{ and } \pi(\beta, \beta^*) < \delta\}.$$

Notice that since the expected value of ϕ is bounded, for each $\beta \in P$ there is a $\delta_\beta > 0$ such that

$$\mathbb{E}[\text{Mod}_\phi(\delta_\beta, \beta)] < \infty \Rightarrow \lim_{\delta \downarrow 0} \mathbb{E}[\text{Mod}_\phi(\delta, \beta)] = 0$$

where the second result follows from the dominated convergence theorem. Hence, there must exist $\delta^*(\beta, j)$ such that $\mathbb{E}[\text{Mod}_\phi(\delta^*(\beta, j), \beta)] < 1/j$. Therefore,

$$|\mathbb{E}[\phi(\beta^*) - \mathbb{E}[\phi(\beta)]| \leq \mathbb{E}|\phi(\beta) - \phi(\beta^*)| \leq \mathbb{E}[\text{Mod}_\phi(\delta^*(\beta, j), \beta)] < 1/j$$

where in addition $\delta^*(\beta, j)$ can be chosen independently of β by compactness of P. We can conclude that there exist a positive function $\delta^*(\beta, j)$ such that

$$|\mathbb{E}[\phi(\beta^*)] - \mathbb{E}[\phi(\beta)]| < 1/j$$

for all $\beta^* \in P$ such that $\pi(\beta^*, \beta) < \delta^*(\beta, j)$ for all $j \geq 1$.

Next, we show that there must exist an integer value function $T^+(\omega, \beta, j)$[11] and a positive function $\delta^+(\beta, j)$ and an indexed set $\Lambda^+(\beta) \in \mathcal{F}$ with $Pr(\Lambda^+(\beta)) = 1$ such that

$$|\frac{1}{T} \sum_{t=1}^{T} \phi(\beta) - \phi(\beta^*)]| < 1/j$$

for all $\beta^* \in P$ such that $\pi(\beta, \beta^*) < \delta^+(\beta, j)$, $T \geq T^+(\omega, \beta, j)$, $\omega \in \Lambda^+(\beta)$ and $j \geq 1$.

By the ergodic theorem (Hansen, 1982), $\frac{1}{T} \sum_{t=1}^{T} \phi(\beta) \to_{a.s.} \mathbb{E}[\phi(\beta)]$ pointwise. For some positive integer n, $\text{Mod}_\phi(\delta_\beta, 1/n)$ has finite first moment and therefore the Law of Large Numbers also applies to the time series average of $\text{Mod}_\phi(\delta_\beta, \beta)$ converging to its expectation on a set $\Lambda^-(\beta, j)$ having probability one for some $j \geq n$. Take

$$\Lambda^+(\beta) = \cap_{j \geq n} \Lambda^-(\beta, j).$$

[11]Here T^+ is a function of some sample path $\omega \in \Omega$, $\beta(\tilde{\omega})$ for $\beta : \Omega \to P$ and integer j.

$\Lambda^+(\beta)$ is measurable and $Pr(\Lambda^+(\beta)) = 1$. For each j, choose $[1/\delta^+(\beta, j)]$ to equal some integer greater than or equal to n such that $\mathbb{E}[\text{Mod}_\phi(\delta_\beta, 1/j)] < 1/(2j)$. Since the time series average of $Mod_\phi(\beta, \delta^+(\beta, j))$ converges almost surely to its expectation, we can pick T large enough such that

$$|1/T \sum_{t=1}^{T} Mod_\phi(\beta, \delta^+(\beta, j)) - \mathbb{E}[Mod_\phi(\beta, \delta^+(\beta, j))]| < 1/(2j).$$

Hence, for such $T \geq T^+(\omega, \beta, j)$

$$\frac{1}{T}|\sum_{t=1}^{T} \phi(\beta) - \phi(\beta^*)| < |1/T \sum_{t=1}^{T} Mod_\phi(\beta, \delta^+(\beta, j))| + 1/(2j)$$

for all $\beta^* \in P$ such that $\pi(\beta, \beta^*) < \delta^+(\beta, j)$ and $j \geq 1$, since $\mathbb{E}[Mod_\phi(\beta, \delta^+(\beta, j))]$ can be arbitrary close to zero.

Finally we show that the result implies almost sure convergence of $\frac{1}{T}\sum_{t=1}^{T} \phi(\beta)$ to its expectation uniformly on $\beta \in P$.

Consider

$$Q(\beta, j) = \{\beta^* \in P : \pi(\beta, \beta^*) < min\{\delta^*(\beta, j), \delta^+(\beta, j)\}\}.$$

For each $j \geq 1$,

$$P = \cup_{\beta \in P} Q(\beta, j).$$

Since P is compact a finite number of β_i can be selected so that

$$P = \cup_{i \geq 1}^{N(j)} Q(\beta_i, j)$$

where $N(j)$ is integer value and β_i is a sequence in P. Construct

$$\Lambda = \cap_{i \geq 1}[\Lambda^*(\beta_i) \cap \Lambda^+(\beta_i)].$$

Then for $\Lambda \in \mathcal{F}$ and $Pr(\Lambda) = 1$ let

$$T(\omega, j) = \max\{T^*(\omega, \beta_1, j), T^*(\omega, \beta_2, j), \ldots, T^*(\omega, \beta_{N(j)}, j), T^+(\omega, \beta_1, j), \ldots, T^+(\omega, \beta_{N(j)}, j)\}.$$

For $T \geq T(\omega, j)$,

$$\sup_{\beta \in P} |\frac{1}{T}\sum_{t=1}^{T} \phi(\beta) - \mathbb{E}[\phi(\beta)]|$$
$$\leq |\frac{1}{T}\sum_{t=1}^{T} \phi(\beta) - \phi(\beta_i)| + |\frac{1}{T}\sum_{t=1}^{T} \phi(\beta_i) - \mathbb{E}[\phi(\beta_i)]| + |\mathbb{E}[\phi(\beta_i) - \mathbb{E}[\phi(\beta)]| < 3/j$$

where β_i is chosen so that $\beta \in Q(\beta_i, j)$ for some $1 \leq i \leq N(j)$.

□

With an abuse of notation we define the prediction of the potential outcome at time t as $\hat{Y}_t^0(\mathcal{F}_{t-1}) := \hat{m}_0(X_t, w^{t-1})$ where w^{t-1} are exponential weights as in (5) computed using information available only until time $t-1$.

Lemma B.2. *Consider the Synthetic Learner with the exponential weighting scheme with* $\eta = \sqrt{\frac{8\log(p)}{M^2 T_0}}$ *and* $\hat{g}_i : \mathcal{X} \to [-\frac{M}{2}, \frac{M}{2}]$. *Then*

$$\frac{1}{T_0}\sum_{t=1}^{T_0}(Y_t - \hat{m}_0(X_t, w^{t-1}))^2 - min_{i\in\{1,\ldots,p\}} \frac{1}{T_0}\sum_{t=1}^{T_0}(Y_t - \hat{g}_i(X_t))^2 \leq C\sqrt{\frac{\log(p)}{2T_0}} \quad (28)$$

where $C = 2M(\max_{t\in\{1,\ldots,T_0\}} |Y_t| + M)$.

Proof of Lemma B.2 . The proof follows similarly to Cesa-Bianchi et al. (1999). Let

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

We denote $\hat{g}(X_t) := \hat{g}_{i,t}$ for expositional convenience.

$$\begin{aligned}\log(W_{T_0}/W_0) = \log(W_{T_0}) - \log(p) &= \log(\sum_{i=1}^{p} \exp(-\eta \sum_{t=1}^{T_0}(Y_t - \hat{g}_{i,t})^2) - \log(p) \\ &\geq \log(\max_{i\in\{1,\dots,p\}} \exp(-\eta \sum_{t=1}^{T_0}(Y_t - \hat{g}_{i,t})^2) - \log(p) \\ &= -\eta \min_{i\in\{1,\dots,p\}} \sum_{t=1}^{T}(Y_t - \hat{g}_{i,t})^2 - \log(p).\end{aligned}$$

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

$$\begin{aligned}\log(W_{T_0}/W_0) = \log(\prod_{t=1}^{T_0} W_t/W_{t-1}) &= \sum_{t=1}^{T_0} \log(\sum_{i=1}^{p} \frac{w_{i,t-1}}{W_{t-1}} \exp(-\eta(Y_t - \hat{g}_{i,t})^2) \\ &= \sum_{t=1}^{T_0} \log(\mathbb{E}_{\hat{\mathbf{g}}\sim Q_t}[\exp(-\eta(Y_t - \hat{g}_{i,t})^2)])\end{aligned}$$

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

$$\begin{aligned}\log(\mathbb{E}_{\hat{\mathbf{g}}\sim Q_t}[\exp(-\eta(Y_t - \hat{g}_{i,t})^2)]) &\leq -\eta \mathbb{E}_{\hat{\mathbf{g}}\sim Q_t}[(Y_t - \hat{g}_{i,t})^2] + \frac{\eta^2 C^2}{8} \\ \leq -\eta(Y_t - \mathbb{E}_{\hat{\mathbf{g}}\sim Q_t}[\hat{g}_{i,t}])^2 + \frac{\eta^2 C^2}{8} &= -\eta(Y_t - \sum_{i=1}^{p} \frac{w_{i,t}}{W_t}\hat{g}_{i,t})^2 + \frac{\eta^2 C^2}{8} \\ = -\eta(Y_t - \hat{m}_0(X_t))^2 + \frac{\eta^2 C^2}{8}\end{aligned}$$

where $C = \max_t |\hat{g}_{i,t}^2 - 2\hat{g}_{i,t}Y_t| \leq M^2 + 2M \max_t |Y_t|$. Hence we have

$$-\eta \min_{i\in\{1,\dots,p\}} \sum_{t=1}^{T}(Y_t - \hat{m}_0(X_t, w^{t-1}))^2 - \log(p) \tag{29}$$

$$\leq \log(W_{T_0}/W_0) \leq \sum_{t=1}^{T_0} -\eta(Y_t - \sum_{i=1}^{p} \frac{w_{i,t}}{W_t}\hat{g}_{i,t})^2 + \frac{T_0 \eta^2 C^2}{8}. \tag{30}$$

Rearranging terms we get

$$\begin{aligned}&\sum_{t=1}^{T_0}(Y_t - \hat{m}_0(X_t, w^{t-1}))^2 - \min_{i\in\{1,\dots,p\}} \sum_{t=1}^{T_0}(Y_t - \hat{g}_{i,t})^2 \leq \frac{\log(p)}{\eta} + \frac{T_0 C^2 \eta}{8} \\ \Rightarrow &\frac{1}{T_0}\sum_{t=1}^{T_0}(Y_t - \hat{m}_0(X_t, w^{t-1}))^2 - \min_{i\in\{1,\dots,p\}} \frac{1}{T_0}\sum_{t=1}^{T_0}(Y_t - \hat{g}_{i,t})^2 \leq C\sqrt{\frac{\log(p)}{2T_0}}\end{aligned} \tag{31}$$

where $\eta = \sqrt{\frac{8\log(p)}{C^2 T_0}}$. □

B.2 Proof of Theorem 3.2. Here we index observations by t, with $t \geq 0$. We will show convergence in probability conditional on $\mathcal{F}_0$. We discuss uniform convergence in Theorem 3.3. Notice here that in our analysis we let $T_0 = \lambda T_+$ where λ is potentially random and strictly bounded between 0 and 1. We consider the asymptotic regime where $T_+ \to \infty$. Such regime is equivalent to the classical *iid* setting where the number of treated and control units is let to be proportional to the sample size. We keep T_- fixed in our analysis. We can write

$$
\widehat{\text{ATE}} = \underbrace{\frac{\sum_{t=T_0+m}^{T_+} Y_t}{T_+ - T_0 - m} - \frac{2\sum_{t=T_0/2+1}^{T_0} Y_t}{T_0}}_{(I)} \underbrace{- \frac{\sum_{t=T_0+m}^{T_+} \hat{\mathbf{g}}(X_t)\hat{\mathbf{w}}_0(1,T_0)}{T_+ - T_0 - m} + \frac{2\sum_{t=T_0/2+1}^{T_0} \hat{\mathbf{g}}(X_t)\hat{\mathbf{w}}_0(1,T_0/2)}{T_0}}_{(II)} \tag{32}
$$

where $\hat{\mathbf{w}}_0(u,v)$ is given in Equation (5). Hence

$$
|\widehat{\text{ATE}} - \mathbb{E}[Y_t(\mathbf{1})] + \mathbb{E}[Y_t(\mathbf{0})]| \leq |(I) - \mathbb{E}[Y_t(\mathbf{1})] - \mathbb{E}[Y_t(\mathbf{0})]| + |(II)|
$$

where (I) and (II) are as in (32).

We start our analysis by studying the first term (I). Under the assumption stated, $Y_t(\mathbf{0})$ and $Y_t(\mathbf{1})$ are ergodic sequences; see for example Proposition 3.44 in White (2014). By the ergodic theorem (e.g. Theorem 3.34 White (2014)), conditional on $\mathcal{F}_0$,

$$
\frac{\sum_{t=T_0+m}^{T_+} Y_t}{T_+ - T_0 - m} \to_p \mathbb{E}[Y_t|\mathcal{F}_0, t \geq T_0+m] = \mathbb{E}[Y_t(\mathbf{1})|\mathcal{F}_0, t \geq T_0+m] = \mathbb{E}[Y_t(\mathbf{1})|\mathcal{F}_0]
$$

where the second equality follows from the fact that under the potential outcome notation,

$$
\begin{aligned}
Y_t = {} & Y_t(\mathbf{0})1\{t \leq T_0\} + Y_t(\mathbf{1})1\{t > T_0+m\} \\
& + Y_t((1,0,0,\dots))1\{t = T_0+1\} + Y_t((1,1,0,0,\dots))1\{t = T_0+2\} + \dots.
\end{aligned}
$$

The last equality follows from the fact that $T_0 \perp Y_t(\mathbf{1})$. The same argument applies to the second term, hence conditional on $\mathcal{F}_0$,

$$
\frac{2\sum_{t=T_0/2+1}^{T_0} Y_t}{T_0} \to_p \mathbb{E}[Y_t|t \leq T_0, \mathcal{F}_0] = \mathbb{E}[Y_t(\mathbf{0})|\mathcal{F}_0, t \leq T_0] = \mathbb{E}[Y_t(\mathbf{0})|\mathcal{F}_0].
$$

By the Slutsky theorem it follows that conditional on $\mathcal{F}_0$

$$
\frac{\sum_{t=T_0+m}^{T_+} Y_t}{T_+ - T_0 - m} - \frac{2\sum_{t=T_0/2+1}^{T_0} Y_t}{T_0} \to_p \mathbb{E}[Y_t(\mathbf{1})|\mathcal{F}_0] - \mathbb{E}[Y_t(\mathbf{0})|\mathcal{F}_0].
$$

By the Portmanteau theorem combined with Jensen's inequality, the result holds unconditionally on $\mathcal{F}_0$, namely

$$
\frac{\sum_{t=T_0+m}^{T_+} Y_t}{T_+ - T_0 - m} - \frac{2\sum_{t=T_0/2+1}^{T_0} Y_t}{T_0} \to_p \mathbb{E}[Y_t(\mathbf{1})] - \mathbb{E}[Y_t(\mathbf{0})].
$$

Consider now (II). Since $\hat{g}_i^0(X_t) \in [-M, M]$ for $M < \infty$, by the triangular inequality, the following hold

$$
\begin{aligned}
\mathbb{E}[|Y_t - \hat{\mathbf{g}}_0(X_t)\hat{\mathbf{w}}_0^o|^q|\mathcal{F}_0] &\leq \mathbb{E}[|Y_t|^q|\mathcal{F}_0] + \mathbb{E}[|\hat{\mathbf{g}}_0(X_t)\hat{\mathbf{w}}_0^o|^q|\mathcal{F}_0] \\
&\leq \mathbb{E}[|Y_t|^q|\mathcal{F}_0] + M^q < \infty, \quad q \in \{1,2\}
\end{aligned}
$$

for any positive weights that sum to one $\hat{\mathbf{w}}_0^o$. Hence, $\mathbb{E}[l(Y_t, \hat{\mathbf{g}}_0(X_t)\hat{\mathbf{w}}_0^o)|\mathcal{F}_0] < \infty$, where l denotes the loss function of interest. Under stationarity of (Y_t, X_t) conditional on $\mathcal{F}_0$, $l(Y_t, \hat{g}(X_t))$ is stationary.

In addition, by Lemma 2.1 in White and Domowitz (1984), $l(Y_t, \hat{g}(X_t))$ is also alpha-mixing. By the ergodic theorem,

$$\frac{1}{T_0}\sum_{t=1}^{T_0} l(Y_t, \hat{g}_j(X_t)) \to_p \mathbb{E}[l(Y_t, \hat{g}_j(X_t))|\mathcal{F}_0]$$

and $\frac{2}{T_0}\sum_{t=1}^{T_0/2} l(Y_t, \hat{g}_j(X_t)) \to_p \mathbb{E}[l(Y_t, \hat{g}_j(X_t))|\mathcal{F}_0]$ converge to the same probability limit. By the Continuous Mapping Theorem

$$\begin{aligned}
&\exp(-\frac{2}{T_0}\sum_{t=1}^{T_0/2} l(Y_t, \hat{g}_j(X_t))), \exp(-\frac{1}{T_0}\sum_{t=1}^{T_0} l(Y_t, \hat{g}_j(X_t))) \\
&\quad \to_p \exp(-\mathbb{E}[l(Y_t, \hat{g}_j(X_t))|t \leq T_0, \mathcal{F}_0]) \\
&\quad = \exp(-\mathbb{E}[l(Y_t(\mathbf{0}), \hat{g}_j(X_t))|\mathcal{F}_0]) > 0.
\end{aligned}$$

Also,

$$\begin{aligned}
&\sum_{j=1}^{p}\exp(-\frac{1}{T_0}\sum_{t=1}^{T_0} l(Y_t, \hat{g}_j(X_t))), \sum_{j=1}^{p}\exp(-\frac{2}{T_0}\sum_{t=1}^{T_0/2} l(Y_t, \hat{g}_j(X_t))) \\
&\quad \to_p \sum_{j=1}^{p}\exp(-\mathbb{E}[l(Y_t(\mathbf{0}), \hat{g}_j(X_t))|\mathcal{F}_0]) > 0.
\end{aligned}$$

Hence, conditional on $\mathcal{F}_0$, by Slutsky theorem each entry $\hat{\mathbf{w}}_0^{(j)}(1, T_0), \hat{\mathbf{w}}_0^{(j)}(1, T_0/2) \to_p \mathbf{w}_0^{o(j)}$ converge to the same probability limit, where

$$\mathbf{w}_0^{o(j)} = \frac{\exp(-\mathbb{E}[l(Y_t(\mathbf{0}), \hat{g}_j(X_t))|\mathcal{F}_0])}{\sum_{i=1}^{p}\exp(-\mathbb{E}[l(Y_t(\mathbf{0}), \hat{g}_i(X_t))|\mathcal{F}_0])}.$$

From the union bound, $\hat{\mathbf{w}}_0(1, T_0), \hat{\mathbf{w}}_0(1, T_0/2) \to_p \mathbf{w}_0^o$.

Finally, under the stationarity and mixing conditions stated

$$\begin{aligned}
\frac{\sum_{t=T_0+m}^{T_+}\hat{\mathbf{g}}_0(X_t)}{T_+ - T_0 - m} &\to_p \mathbb{E}[\hat{\mathbf{g}}_0(X_t)|\mathcal{F}_0, t \geq T_0 + m] = \mathbb{E}[\hat{\mathbf{g}}_0(X_t)|\mathcal{F}_0] \\
\frac{2\sum_{t=T_0/2+1}^{T_0}\hat{\mathbf{g}}_0(X_t)}{T_0} &\to_p \mathbb{E}[\hat{\mathbf{g}}_0(X_t)|\mathcal{F}_0, t \leq T_0] = \mathbb{E}[\hat{\mathbf{g}}_0(X_t)|\mathcal{F}_0]
\end{aligned}$$

where the second equality in each equation follows from $T_0 \perp X_t$. By Slutsky theorem $(II) = o_p(1)$ pointwise in $\mathcal{F}_0$.

B.3 Proof of Theorem 3.3. Here we index observations by t, with $t \geq 0$. Notice here that in our analysis we let $T_0 = \lambda T_+$ where λ is potentially random and strictly bounded between 0 and 1. We consider the asymptotic regime where $T_+ \to \infty$. Such regime is equivalent to the classical *iid* setting where the number of treated and control units is let to be proportional to the sample size. We keep T_- fixed in our analysis. We can write

$$\begin{aligned}
\widehat{\text{ATE}} = &\underbrace{\frac{\sum_{t=T_0+m}^{T_+} Y_t}{T_+ - T_0 - m} - \frac{2\sum_{t=T_0/2+1}^{T_0} Y_t}{T_0}}_{(I)} \\
&- \underbrace{\frac{\sum_{t=T_0+m}^{T_+}\hat{\mathbf{g}}(X_t)\hat{\mathbf{w}}_0(1, T_0)}{T_+ - T_0 - m} + \frac{2\sum_{t=T_0/2+1}^{T_0}\hat{\mathbf{g}}(X_t)\hat{\mathbf{w}}_0(1, T_0/2)}{T_0}}_{(II)}
\end{aligned} \tag{33}$$

where $\hat{\mathbf{w}}_0(u, v)$ is given in Equation (5). Hence

$$|\widehat{\text{ATE}} - \mathbb{E}[Y_t(\mathbf{1})] - \mathbb{E}[Y_t(\mathbf{0})]| \leq |(I) - \mathbb{E}[Y_t(\mathbf{1})] - \mathbb{E}[Y_t(\mathbf{0})]| + |(II)|$$

where (I) and (II) are as in (33). We start our analysis by studying the first term (I). Under the assumption stated, $Y_t(\mathbf{0})$ and $Y_t(\mathbf{1})$ are ergodic sequences; see for example Proposition 3.44 in White (2014). By the ergodic theorem (e.g. Theorem 3.34 White (2014)),

$$\frac{\sum_{t=T_0+m}^{T_+} Y_t}{T_+ - T_0 - m} \to_p \mathbb{E}[Y_t | t \geq T_0 + m] = \mathbb{E}[Y_t(\mathbf{1}) | t \geq T_0 + m] = \mathbb{E}[Y_t(\mathbf{1})]$$

where the second equality follows from the fact that under the potential outcome notation,

$$\begin{aligned} Y_t = \; & Y_t(\mathbf{0}) 1\{t \leq T_0\} + Y_t(\mathbf{1}) 1\{t > T_0 + m\} + Y_t((1, 0, 0, \dots)) 1\{t = T_0 + 1\} \\ & + Y_t((1, 1, 0, 0, \dots)) 1\{t = T_0 + 2\} + \dots . \end{aligned}$$

The last equality follows from the fact that $T_0 \perp Y_t(\mathbf{1})$. The same argument applies to the second term, hence

$$\frac{2 \sum_{t=T_0/2+1}^{T_0} Y_t}{T_0} \to_p \mathbb{E}[Y_t | t \leq T_0] = \mathbb{E}[Y_t(\mathbf{0}) | t \leq T_0] = \mathbb{E}[Y_t(\mathbf{0})]$$

By the Slutsky theorem it follows that

$$\frac{\sum_{t=T_0+m}^{T_+} Y_t}{T_+ - T_0 - m} - \frac{2 \sum_{t=T_0/2+1}^{T_0} Y_t}{T_0} \to_p \mathbb{E}[Y_t(\mathbf{1})] - \mathbb{E}[Y_t(\mathbf{0})].$$

Consider now (II). With an abuse of notation we write $\hat{g}_i^0(X_t) = h_i(X_t, F_-)$ where F_- denote the empirical distribution of $(X_{T_-}, Y_{T_-}, \dots, X_0, Y_0)$, $T_- > -\infty$. We study convergence of (II) uniformly over F_-, for fixed $T_- > -\infty$. First, we study convergence properties of $\frac{1}{T_0} \sum_{t=1}^{T_0} l(Y_t, h_i(X_t, F_-))$ uniformly in F_-. Notice that $\hat{g}_i^0(X_t) \in [-M, M]$ for $M < \infty$, by the triangular inequality, the following hold

$$|Y_t - h_i(X_t, F_-)|^q \leq |Y_t|^q + M^q < \infty \quad q \in \{1, 2\}$$

for any F_-. Hence, by Lemma B.1,

$$\frac{1}{T_0} \sum_{t=1}^{T_0} l(Y_t, h_i(X_t, F_-)) \to_p \mathbb{E}[l(Y_t, h_i(X_t, F_-)) | t \leq T_0] = \mathbb{E}[l(Y_t, F_-)] < \infty$$

uniformly over F_-. By the Continuous Mapping Theorem

$$\begin{aligned} \exp(-\frac{2}{T_0} \sum_{t=1}^{T_0/2} l(Y_t, h_j(X_t, F_-))), \exp(-\frac{1}{T_0} \sum_{t=1}^{T_0} l(Y_t, h_j(X_t, F_-))) \\ \to_p \exp(-\mathbb{E}[l(Y_t, h_j(X_t, F_-)) | t \leq T_0]) \\ = \exp(-\mathbb{E}[l(Y_t(\mathbf{0}), h_j(X_t, F_-))]) > 0. \end{aligned}$$

Also, for $p < \infty$ fixed

$$\begin{aligned} \sum_{j=1}^{p} \exp(-\frac{1}{T_0} \sum_{t=1}^{T_0} l(Y_t, h_j(X_t, F_-))), \sum_{j=1}^{p} \exp(-\frac{2}{T_0} \sum_{t=1}^{T_0/2} l(Y_t, h_j(X_t, F_-))) \\ \to_p \sum_{j=1}^{p} \exp(-\mathbb{E}[l(Y_t(\mathbf{0}), h_j(X_t, F_-))]). \end{aligned}$$

Hence, by Slutsky theorem each entry $\hat{\mathbf{w}}_0^{(j)}(1,T_0,F_-),\hat{\mathbf{w}}_0^{(j)}(1,T_0/2,F_-)\to_p \mathbf{w}_0^{o(j)}(F_-)$ converge uniformly over F_- to the same probability limit, where

$$\mathbf{w}_0^{o(j)}(F_-)=\frac{\exp(-\mathbb{E}[l(Y_t(\mathbf{0}),h_j(X_t,F_-))])}{\sum_{i=1}^p\exp(-\mathbb{E}[l(Y_t(\mathbf{0}),j_j(X_t,F_-))])}.$$

From the union bound, $\hat{\mathbf{w}}_0(1,T_0,F_-),\hat{\mathbf{w}}_0(1,T_0/2,F_-)\to_p \mathbf{w}_0^o(F_-)$ uniformly over F_-.

Similarly, by Lemma B.1, under the stationarity and mixing conditions stated

$$\frac{\sum_{t=T_0+m}^{T_+}h_j(X_t,F_-)}{T_+-T_0-m}\to_p \mathbb{E}[\hat{\mathbf{g}}(X_t)|t\geq T_0+m]=\mathbb{E}[h_j(X_t,F_-)]$$
$$\frac{2\sum_{t=T_0/2+1}^{T_0}h_j(X_t,F_-)}{T_0}\to_p \mathbb{E}[h_j(X_t,F_-)|t\leq T_0]=\mathbb{E}[h_j(X_t,F_-)]$$

uniformly over F_-. By Slutsky theorem $(II)=o_p(1)$ uniformly over F_-.

B.4 Proof of Theorem 3.4. We let w^{t-1} denote the exponential weights computed using information only from time 1 up to time $t-1$ and $w_{T_0}^s$ being weights that use information from time T_0 to time s. With an abuse of notation, we let $\hat{m}_d(\cdot,\cdot)$ be our prediction (i.e. $\hat{Y}_t^0(\mathcal{F}_{t-1}=\hat{m}_0(X_t,w^{t-1}))$ which takes as second argument the weights used for prediction. We start by decomposing the average loss as follows.

$$\frac{1}{T_0}\sum_{t=1}^{T_0}(Y_t-\hat{m}_0(X_t,w^{t-1}))^2=\frac{1}{T_0}\sum_{t=1}^{T_0}(\hat{m}_0(X_t,w^{t-1})-\mu_0(X_t))^2 + \varepsilon_t(0)^2+2(\hat{m}_0(X_t,w^{t-1})-\mu_0(X_t))\varepsilon_t(0) .$$
$$\frac{1}{T_0}\sum_{t=1}^{T_0}(Y_t-\hat{g}_i(X_t))^2=\frac{1}{T_0}\sum_{t=1}^{T_0}(\hat{g}_i(X_t)-\mu_0(X_t))^2+\varepsilon_t(0)^2+2(\hat{g}_i(X_t)-\mu_0(X_t))\varepsilon_t(0)$$

Henceforth

$$\begin{aligned}&\frac{1}{T_0}\{\sum_{t=1}^{T_0}(Y_t-\hat{m}_0(X_t,w^{t-1}))^2-\min_{i\in\{1,\ldots,p\}}\sum_{t=1}^{T_0}(\hat{g}_{i,t}-Y_t)^2\}\\ &=\frac{1}{T_0}\sum_{t=1}^{T_0}(\mu_0(X_t)-\hat{m}_0(X_t,w^{t-1}))^2+2\min_i\frac{1}{T_0}\sum_{t=1}^{T_0}(\hat{m}_0(X_t,w^{t-1})\\ &\quad-\hat{g}_i(X_t))\varepsilon_t(0)-(\hat{g}_{i,t}-\mu_0(X_t))^2.\end{aligned}\tag{34}$$

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

Using Lemma B.2 and the decomposition in (34) we write

$$\begin{aligned}&\frac{1}{T_0}\{\sum_{t=1}^{T_0}(\mu_0(X_t)-\hat{m}_0(X_t,w^{t-1}))^2\\ &\leq\min_i\frac{1}{T_0}\sum_{t=1}^{T_0}\{(\hat{g}_{i,t}-\hat{m}_0(X_t,w^{t-1}))\varepsilon_t(0)+(\hat{g}_{i,t}-\mu_0(X_t))^2\}+C\sqrt{\frac{2log(p)}{T_0}}\\ &\leq\underbrace{\max_j\frac{1}{T_0}\sum_{t=1}^{T_0}(\hat{g}_{j,t}-\hat{m}_0(X_t,w^{t-1}))\varepsilon_t(0)}_{(I)}+\underbrace{\min_i\frac{1}{T_0}\sum_{t=1}^{T_0}(\hat{g}_{i,t}-\mu_0(X_t))^2+C\sqrt{\frac{2log(p)}{T_0}}}_{(II)}.\end{aligned}\tag{35}$$

We discuss the term (I). Notice that

$$\mathbb{E}[\varepsilon_t(\mu(X_t) - \hat{g}_i(X_t))|X_t, \mathcal{F}_{t-1}] = \mathbb{E}[\varepsilon_t(\mu(X_t) - \hat{g}_i(X_t))|X_t, \mathcal{F}_{t-1}, t \le T_0] = 0,$$

since $\varepsilon_t \perp D_t | X_t, \mathcal{F}_{t-1}$. In addition by assumption $|\varepsilon_t(\mu(X_t) - \hat{g}_i(X_t))| < M^2 < \infty$. By Hoeffding-Azuma inequality, with probability $\ge 1 - \delta$

$$\frac{1}{T_0} \sum_{t=1}^{T_0} \varepsilon_t(X_t - \hat{g}_i(X_t)) \le C_0 \sqrt{\frac{\log(1/\delta)}{T_0}}$$

for a universal constant C_0. After subsitituting into the previous expression, since $C \le M$ by boundeness assumpion on Y_t, the result follows. For the regret bound for $t > T_0$ the result follows in the similar way after appropriately indixing the period of interest.

B.5 Proof of Theorem 3.5. The proof follows the same step as in Theorem 3.4 until (35). We will bound (I) differently since we cannot use Hoeffding-Azuma. Notice first that the following inequality hold.

$$\mathbb{E}[\psi(\frac{\sum_{s=1}^t \varepsilon_s}{c})|X_t, \mathcal{F}_{t-1}] < \tau \Rightarrow \mathbb{E}[\psi(\frac{\sum_{s=1}^t \varepsilon_s(\mu(X_s) - \hat{g}_{i,s})}{c})|X_t, \mathcal{F}_{t-1}] < \tilde{\tau} < \infty$$

where the right hand side follows by boundeness of μ and $\hat{g}_{i,t}$. By Theorem 14 in McDonald and Shalizi (2011), since $\mathbb{E}[\varepsilon_t(\mu(X_t) - g_{i,t})] = \mathbb{E}[\varepsilon_t(\mu(X_t) - g_{i,t})|X_t, \mathcal{F}_{t-1}] = 0 \ \forall t$ and since $\varepsilon_t \perp D_t | X_t, \mathcal{F}_{t-1}$ by assumption, $\mathbb{E}[\varepsilon_t(\mu(X_t) - g_{i,t})|X_t, \mathcal{F}_{t-1}, t \le T_0] = 0$, hence we have for T_0 large enough

$$\mathbb{P}\left(\frac{1}{T_0} \sum_{t=1}^{T_0} \varepsilon_t(\mu(X_t) - \hat{g}_{i,t}) > t\right) \le \exp(-\frac{v^2 T_0}{32(\tilde{\tau}+1)^2 \tilde{c}^2}) \tag{36}$$

where $\tilde{c} = 2cM^2$. Henceforth with probability at least $1 - \delta$, by the union bound,

$$\max_j \frac{1}{T_0} \sum_{t=1}^{T_0} (\hat{g}_{j,t} - \hat{m}_0(X_t, w^{t-1}))\varepsilon_t(0) \le 4(\tilde{\tau}+1)\tilde{c} \sqrt{\frac{2\log(p/\delta)}{T_0}} \tag{37}$$

which implies together with Equation 35 that the following hold

$$\mathcal{R}^0 \le C_0 \max_{t=1,\dots,T_0} |Y_t| \sqrt{\frac{\log(p/\delta)}{T_0}}$$

for a universal constant C_0. By definition of subgaussianity of Y_t, we have for some $u, v > 0$,

$$P(\max_{t=1,\dots,T_0} |Y_t| > t) \le T_0 u \exp(-vt^2) \Rightarrow \max_{t=1,\dots,T_0} |Y_t| \le \sqrt{\frac{\log(uT_0/\delta)}{v}}, \quad w.p. \ge 1 - \delta.$$

Hence by previous arguments, by the union bound, $\mathcal{R}^0 \le \tilde{C}_0 \sqrt{\log(2T_0 p/\delta)/T_0}$ with probability $1 - \delta$. For the regret bound for $t > T_0$ the result follows in the similar way after appropriately indixing the period of interest.

Appendix Appendix C Additional numerical experiments

C.1 Bootstrap with Carry-over Effects. We study the robustness of the bootstrap method to different levels of carryover effects. We consider the case where the researcher knows the true level of carry-over effects. We consider the scenario with $T = 300$, $T_0 = 250$, $T_- = 125$ and the Synthetic Learner trained with learners XGboost, ARIMA(0,1,1) and Support vector regression. We collect results in Figure 18. Remarkably, the bootstrap is robust to all choices of m considered. In practice researchers should check the robustness of their results over different levels of carry-over effects. This result together with our theoretical guarantees suggests that the resampling scheme proposed in this paper is a valid method to perform this task.

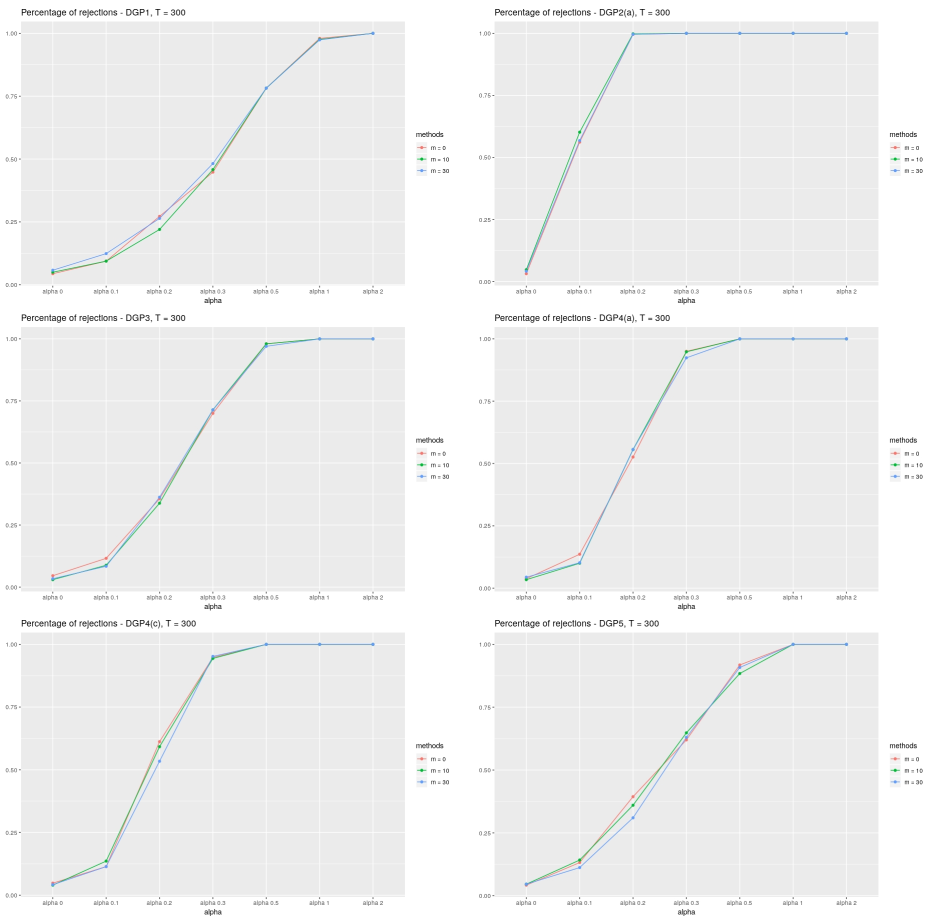

FIGURE 18. Percentage of rejections using the Synthetic Learner with learners XG-boost, Support Vector Regression and ARIMA(0,1,1). We consider $T = 300$ and $T_0 = 50$ and use 500 replications. We study different levels of carry-over effects, denoted with $m \in \{0, 10, 30\}$.