SEMIPARAMETRIC INFERENCE FOR IMPULSE RESPONSE FUNCTIONS USING DOUBLE/DEBIASED MACHINE LEARNING

Daniele Ballinari[*] [†]
Swiss National Bank
daniele.ballinari@snb.ch

Alexander Wehrli[*†]
Swiss National Bank
alexander.wehrli@snb.ch

December 17, 2025
First version: November 18, 2024

ABSTRACT

We introduce a double/debiased machine learning estimator for the impulse response function in settings where a time series of interest is subjected to multiple discrete treatments, assigned over time, which can have a causal effect on future outcomes. The proposed estimator can rely on fully nonparametric relations between treatment and outcome variables, opening up the possibility to use flexible machine learning approaches to estimate impulse response functions. To this end, we extend the theory of double machine learning from an *i.i.d.* to a time series setting and show that the proposed estimator is consistent and asymptotically normally distributed at the parametric rate, allowing for semiparametric inference for dynamic effects in a time series setting. The properties of the estimator are validated numerically in finite samples by applying it to learn the impulse response function in the presence of serial dependence in both the confounder and observation innovation processes. We also illustrate the methodology empirically by applying it to the estimation of the effects of macroeconomic shocks.

JEL Classification C14 · C22 · C51 · C53 · C55

Keywords Impulse response function · Double machine learning · Time series · Average treatment effect

1 Introduction

The estimation of the response of a time series to an external impulse is a common task in many scientific disciplines. For example, in economics, one might be interested in the reaction of the economy to a change in a central bank's monetary policy (Angrist, Jordà, & Kuersteiner, 2018). In the analysis of their trading costs, financial professionals are interested in the causal effect that their trades have on an asset's price (Bouchaud, Bonart, Donier, & Gould, 2018). In medicine, when administering a drug to a patient over time, one is interested in measuring its effect on the health of the patient (Bica, Alaa, & Van Der Schaar, 2020). Readers are referred to the surveys in Runge, Gerhardus, Varando, Eyring, and Camps-Valls (2023) and Raha et al. (2021) for more examples.

The quantity of interest in these applications is commonly referred to as the *impulse response function* (IRF). Ideally, the IRF measures the *causal* effect that an action (or "treatment") has on the time series of interest. Recently, IRFs and ideas stemming from the causal inference framework have been related (Jordà, 2023). In particular, Rambachan and Shephard (2021) provided assumptions under which IRFs coincide with classical *average treatment effects* (ATE) analyzed in the potential outcomes framework of causal inference (Robins, 1986; Rubin, 1974). Given this relation

[*]The views, opinions, findings, and conclusions or recommendations expressed in this paper are strictly those of the authors. They do not necessarily reflect the views of the Swiss National Bank (SNB). The SNB takes no responsibility for any errors or omissions in, or for the correctness of, the information contained in this paper.

[†]We thank Nora Bearth, Jonathan Chassot and Victor Chernozhukov for helpful comments and suggestions. We are also grateful to Guido Kuersteiner for providing the data used in the empirical application.

between ATE and IRFs, it seams natural to adopt estimation procedures from the causal inference literature for the problem of IRF estimation. Traditionally, IRFs have primarily been estimated by modelling the entire dynamic system under consideration, e.g. using vector autoregressive processes (Sims, 1980). The seminal work of Jordà (2005) later showed how to directly estimate univariate conditional expectations using local projections (Jordà & Taylor, 2024). This approach compares the conditional expectation of an outcome variable, once conditional on a shock (treatment) and once conditional on no shock. As such, the local projection framework is directly related to the regression adjustment approach used for ATE estimation. The approach of Jordà (2005) allows for some flexibility in the estimation of the impulse response function, as it can easily incorporate polynomial and interaction terms of the regressors, state dependence (Gonçalves, Herrera, Kilian, & Pesavento, 2024), or instrumental variables (Stock & Watson, 2018). More recently, Adamek, Smeekes, and Wilms (2024) extended the local projection approach to high-dimensional settings using penalized local projections. Jordà and Taylor (2016) and Angrist et al. (2018) are examples of applications that use propensity score weighting for the estimation of IRFs, another common estimation approach coming from the causal inference literature, additionally accommodating asymmetric and nonlinear responses.

While estimators for IRFs have become more flexible in recent years, they still require the definition of a functional form relating treatment and outcome variables. Here, we introduce an estimator for the IRF that can rely on fully nonparametric relations between treatment and outcome variables, opening up the possibility to use flexible machine learning approaches to estimate IRFs. We consider a setting where a single time series is subjected to a discrete treatment at multiple points in time and one is interested in the average (causal) effect that these treatments have at different prediction horizons. Inspired by the approach of Chernozhukov et al. (2018) for the *i.i.d.* setting, the proposed estimator leverages the efficient influence function for the IRF in combination with cross-fitting, which makes the IRF estimation insensitive – or, formally, (Neyman-)orthogonal – to the biased estimation of the conditional expectation and treatment probability functions by machine learners. Moreover, the proposed estimator avoids over-fitting by using a cross-fitting procedure. These two ingredients, orthogonality and cross-fitting, eliminate regularization and over-fitting bias to which machine learning algorithms are prone to, an approach coined *double/debiased machine learning* (Chernozhukov et al., 2017). Our theoretical results show that the proposed IRF estimator is consistent and asymptotically normally distributed at the parametric rate, building the basis for semiparametric inference for dynamic effects in time series settings. The problem studied in this paper relates to classical semiparametric estimation techniques for dependent data, e.g. using kernel (Li & Racine, 2006; Robinson, 1983), series (X. Chen & Christensen, 2015; J. Lee & Robinson, 2016) or general sieve estimators (X. Chen & Liao, 2015; X. Chen & Shen, 1998). Apart from such classical approaches, our contribution relates to a nascent but growing body of literature on the application of machine learning for semi- and nonparametric causal inference in time series problems. Lewis and Syrgkanis (2021) e.g. provide a method of estimating and conducting inference for dynamic treatment effects, based on a sequential regression peeling process, focusing on a fixed-length time series panel setup. Bica et al. (2020) propose an estimator for time-varying treatment effects in the presence of hidden confounders, building on a recurrent neural network. Their setting also considers a fixed-length time series observed for multiple individuals. Using panel data, Paranhos (2025) employs a generalized random forest to obtain locally linear impulse response functions. Hauzenberger, Huber, Klieber, and Marcellino (2025) estimate impulse response functions using Bayesian neutral networks. Grecov et al. (2021) consider a multivariate time series setting where some units become and remain treated at a specific point in time. The counterfactual outcomes are obtained from a global forecasting model based on a recurrent neural network. Others have employed flexible machine learning approaches for causal discovery in time series; see, among others, Bussmann, Nys, and Latré (2021); Nauta, Bucur, and Seifert (2019); Yin and Barucca (2022).

The rest of the paper is organized as follows. Section 2 sets up the problem. Section 3 presents the double machine learning (DML) estimator and our main theoretical results on its asymptotic properties. Proofs are relegated to the Appendix for legibility. Section 4 offers recommendations for the practical implementation of the estimator. Section 5 validates the developed theory in a simulation study and Section 6 provides a comparison to local projections. Section 7 applies the proposed estimator to policy decisions in a macrodynamic setting and Section 8 concludes.

2 Notation and identification

2.1 Notation

Let $S^{(h)} = \{Z_t^{(h)} : t \in \mathcal{T}\}$ be stochastic processes generated from a distribution $\mathcal{P}$ with $Z_t^{(h)} = (Y_{t+h}, X_t, D_t)$, where Y_{t+h} is a scalar real-valued random variable, D_t a binary treatment variable, and $X_t \in \mathcal{X} \subseteq \mathbb{R}^n$ a random vector which may contain also lagged variables, including lagged values of Y_t and D_t. If not specified otherwise, $\mathcal{T}$ is a collection of ordered time indices (also referred to as the index set) with cardinality $|\mathcal{T}| = T$. The quantity of interest is the impulse response function at horizon $h \in \mathbb{N}_0$ for a binary impulse variable D_t on the outcome variable

Y_{t+h}, defined as (Rambachan & Shephard, 2021)

$$\theta_0^{(h)} = \mathbb{E}\left[\mathbb{E}[Y_{t+h}|D_t = 1, X_t] - \mathbb{E}[Y_{t+h}|D_t = 0, X_t]\right]. \tag{1}$$

We focus on the case where $\theta_0^{(h)}$ and the conditional first moments of the outcome and treatment random variables are time-invariant.

Assumption 1. *For all $s, t \in \mathcal{T}$ and $h \in \mathbb{N}_0$, the following holds.*

1. *The impulse response function is time-invariant, i.e. $\theta_0^{(h)} = \mathbb{E}[\mathbb{E}[Y_{t+h}|D_t = 1, X_t] - \mathbb{E}[Y_{t+h}|D_t = 0, X_t]] = \mathbb{E}[\mathbb{E}[Y_{s+h}|D_s = 1, X_s] - \mathbb{E}[Y_{s+h}|D_s = 0, X_s]]$.*
2. *The conditional first moments of Y_{t+h} and D_t are time-invariant, i.e. $\mu_0(d, x, h) = \mathbb{E}[Y_{t+h}|D_t = d, X_t = x] = \mathbb{E}[Y_{s+h}|D_s = d, X_s = x]$ and $e_0(x) = \Pr(D_t = 1|X_t = x) = \Pr(D_s = 1|X_s = x)$.*

While we present results for a binary treatment D_t, this can be generalized to multivariate discrete treatments (Angrist et al., 2018), which we will also do in our empirical application. In fact, discrete treatments can be interpreted as pairwise binary treatment comparisons, and as such, the theoretical results presented in the sequel extend directly to multivariate treatments. Moreover, even in settings where the treatment variable is continuous, one can obtain estimation results by discretizing the treatment of interest (see, e.g. Knaus, 2021). We refer to $\Gamma_0 = (\mu_0(d, x, h), e_0(x))$ as *nuisance functions*.

Much of traditional estimation of IRFs relies on regression adjustment, i.e., the estimation of $\theta_0^{(h)}$ as the average difference between $\mu_0(1, X_t, h)$ and $\mu_0(0, X_t, h)$ (Cochran, 1968; Jordà, 2005; Pearl, 2009; Robins, 1986). However, regression adjustment estimators typically tend to be rather sensitive to small amounts of misspecification in the conditional expectation models. Alternatively, approaches using inverse probability weighting (Angrist et al., 2018; Rosenbaum & Rubin, 1983; Tsiatis, 2006) have been devised, but are also sensitive to misspecification of the propensity score models. For the estimator presented in Section 3, as in e.g. Chernozhukov et al. (2018), we instead rely on the efficient influence function (Hahn, 1998; Robins & Rotnitzky, 1995) to estimate the IRF, namely $g\left(Z_t^{(h)}, h; \Gamma_0\right) - \theta_0^{(h)}$ where

$$\begin{aligned} g\left(Z_t^{(h)}, h; \Gamma_0\right) = \mu_0(1, X_t, h) - \mu_0(0, X_t, h) + \frac{D_t}{e_0(X_t)}(Y_{t+h} - \mu_0(1, X_t, h)) \\ - \frac{1 - D_t}{1 - e_0(X_t)}(Y_{t+h} - \mu_0(0, X_t, h)) \end{aligned} \tag{2}$$

and it can be shown that

$$\theta_0^{(h)} = \mathbb{E}\left[g\left(Z_t^{(h)}, h; \Gamma_0\right)\right].$$

An influence function measures how a small perturbation of the data affects an estimator. The *efficient* influence function is the particular influence function that (among all regular estimators) achieves the lowest possible asymptotic variance allowed by the semiparametric model. Readers are referred to Hines, Dukes, Diaz-Ordaz, and Vansteelandt (2022), Kennedy (2024), Fisher and Kennedy (2021) and Tsiatis (2006) for a review of influence functions and semiparametric theory. In contrast to regression adjustment or inverse probability weighting, an estimator relying on the above influence function is *Neyman orthogonal* (Chernozhukov et al., 2017; Neyman, 1959, 1979). Technically, the efficient influence function of the IRF is Neyman orthogonal since the (Gateaux) derivative of its expected value with respect to either nuisance function equals zero (for a detailed discussion, see Chernozhukov et al., 2018). This property ensures that small deviations from the true nuisance functions have no first-order effect on the estimation of $\theta_0^{(h)}$. Loosely speaking, if the estimated nuisance functions are "close enough" to their true values, estimation errors only have a vanishing impact on the IRF estimator. While the theory presented in Section 3 leverages this Neyman orthogonality property, it is worth mentioning that an estimator based on Equation (2) is also *doubly robust*, in the sense that it remains consistent if only one of the nuisance functions is correctly specified. In the sequel, we use standard notations $O(\cdot)$ and $o(\cdot)$ to indicate rates of convergence for sequences. In particular, if $\{x_t\}_1^\infty$ is any real sequence, $\{a_t\}_1^\infty$ a sequence of positive real numbers, and there exists a finite constant B such that $|x_t|/a_t \leq B$ for all t, we write $x_t = O(a_t)$. If x_t/a_t converges to zero, we write $o(a_t)$. We use $\|\cdot\|_q$ to denote the L_q-norm; e.g. we write $\|f\|_q = \|f(Z)\|_q = \left(\int |f(z)|^q \mathrm{d}\mathcal{P}(z)\right)^{1/q}$.

2.2 Identification

While the focus of this paper in on the estimation of the quantity introduced in Equation (1), we here briefly present assumptions under which $\theta_0^{(h)}$ conveys the interpretation of the average *causal* effect that the binary impulse variable

D_t has on the outcome variable Y_{t+h}. Identification is formulated within the potential outcomes framework of causal inference (Robins, 1986; Rubin, 1974). Let $Y_{t+h}(d)$ be the potential outcome, i.e., the random variable one would observe at time $t+h$ if the treatment at time t would have been $D_t = d$. The following assumption ensures identification of the causal effect.

Assumption 2 (Angrist et al. (2018); Rambachan and Shephard (2021))**.** *For all $t \in \mathcal{T}$ and $h \in \mathbb{N}_0$ the following holds.*

1. *The potential outcomes are conditionally independent of the treatment, i.e. $Y_{t+h}(1), Y_{t+h}(0) \perp D_t | X_t$.*
2. *The observed outcome is $Y_{t+h} = D_t Y_{t+h}(1) + (1 - D_t) Y_{t+h}(0)$.*
3. *For all $x \in \mathcal{X}$ it holds that $\eta < e_0(x) < 1 - \eta$, for some $0 < \eta < 1$.*

Assumption 2.1 requires conditional independence between the treatment at time t and the potential outcomes. Notably, it is not necessary for D_t to be conditionally independent from future treatment assignments. However, in the case where $D_t \not\perp D_{t+1}, \ldots, D_{t+h} | X_t$, the identified effect corresponds to the effect of a treatment including potential future treatments caused by D_t (Jordà, 2023). Assumption 2.3 imposes that at each point in time, the treatment assignment is not deterministic. In other words, there are no situations in which either $D_t = 1$ or $D_t = 0$ with (conditional) probability of one. In essence, these assumptions require the treatment variable of interest (or a sufficiently informative proxy) to be observed, and a set of control variables to be available such that the treatment assignment is (conditionally) as good as random. In macroeconomics, this requirement corresponds to identification strategies that rely on constructed shocks, such as narrative monetary policy shocks (e.g., Ramey, 2016; Romer & Romer, 2004), or "direct causal inference" approaches based on externally constructed measures of structural shocks (Nakamura & Steinsson, 2018). For approaches that use continuous shock measures, the shock can also be incorporated into our framework by discretizing them. This is however only required to define the treatment variable within our setup, not by the identification strategy itself. For a more rigorous discussion of the identification of treatment effects with time-dependent data – and in particular for the connection between Assumption 2 and classical macroeconomic shocks – we refer to Rambachan and Shephard (2021). The following theorem finally establishes identification of the average treatment effect.

Theorem 1. *Under Assumptions 1 and 2 the ATE is identified as*

$$\theta_0^{(h)} = \mathbb{E}\left[Y_{t+h}(1) - Y_{t+h}(0)\right].$$

As pointed out by Chernozhukov et al. (2018), the DML estimator will yield unbiased results also in a setting where the causal identification assumptions presented in this section fail to hold. In this case however, the estimated IRF has to be interpreted as a prediction difference rather than a causal effect.

3 Estimation

This section outlines the estimator for $\theta_0^{(h)}$ and its asymptotic properties when the nuisance functions are estimated with flexible, nonparametric machine learning algorithms. The estimator is developed in three steps. First, we provide results for the (hypothetical) case where the nuisance functions are known. In a second step, nuisance functions are estimated, but multiple independent stochastic processes generated from the same distribution are available. Lastly, we provide results for the case where nuisance functions have to be estimated and only a single stochastic process is available.

3.1 An oracle estimator

In case the nuisance functions are known, we can estimate the effect of interest by simply averaging the stochastic processes $\mathcal{G}^{(h)} = \left\{ g\left(Z_t^{(h)}, h; \Gamma_0\right) : t \in \mathcal{T} \right\}$ over the index set. We refer to this estimator as the *oracle estimator* for the IRF. The following assumption and theorem provide conditions under which the oracle estimator is asymptotically normally distributed.

Assumption 3. *For some $\beta > 2$ and all $h \in \mathbb{N}_0$, the following conditions hold.*

1. *The stochastic processes $\mathcal{G}^{(h)}$ are weakly stationary.*
2. *The variance satisfies that $0 < V_0^{(h)} = \lim_{T\to\infty} \mathrm{Var}\left[\frac{1}{\sqrt{T}} \sum_{t\in\mathcal{T}} g\left(Z_t^{(h)}, h; \Gamma_0\right)\right]$.*

3. $\mathcal{G}^{(h)}$ *is uniformly* L_β*-bounded, i.e.* $\sup_{t\in\mathcal{T}} \mathbb{E}\left[\left|g\left(Z_t^{(h)}, h; \Gamma_0\right)\right|^\beta\right] < \infty$.

4. $\mathcal{G}^{(h)}$ *is* α*-mixing, with coefficients* $\alpha(s)$, $s \in \mathbb{N}$, *satisfying* $\sum_{s=1}^{\infty} \alpha(s)^{(\beta-2)/\beta} < \infty$.

Theorem 2. *Let the oracle IRF estimator be given by*

$$\tilde{\theta}^{(h)} = \frac{1}{T}\sum_{t\in\mathcal{T}} g\left(Z_t^{(h)}, h; \Gamma_0\right).$$

Under Assumptions 1 and 3, we have that

$$\sqrt{T}(\tilde{\theta}^{(h)} - \theta_0^{(h)}) \xrightarrow{\mathrm{d}} \mathcal{N}\left(0, V_0^{(h)}\right),$$

with $V_0^{(h)} = \sum_{s=-\infty}^{+\infty} \mathrm{Cov}\left[g\left(Z_t^{(h)}, h; \Gamma_0\right), g\left(Z_{t-s}^{(h)}, h; \Gamma_0\right)\right]$.

Given that $g(z, h; \Gamma)$ is a measurable function as long as the nuisance functions are measurable, Assumption 3.4 is satisfied if the stochastic processes $S^{(h)}$ are α-mixing for some $\beta > 2$ (Davidson, 2021). This is however not necessary, and $S^{(h)}$ can exhibit less favorable dependence structures as long as $\mathcal{G}^{(h)}$ adheres to Assumptions 1 and 3. Moreover, while the above assumptions are standard in the application of functional central limit theory, variations are possible that still lead to the desired asymptotics. For example, weak stationarity in Assumption 3.1 can be relaxed to a constant mean, permitted that additionally $V_0^{(h)} < \infty$ (see e.g. the discussion in Phillips (1987)). Weak stationarity is however required in our setting in order to obtain a tractable estimator for $V_0^{(h)}$.

It is important to remark that the two assumptions 3.3 and 3.4 represent an inherent trade-off. The more absolute moments of $\mathcal{G}^{(h)}$ are required to exist, the more dependence is acceptable in the stochastic processes to still reach asymptotic normality. For the sum in Assumption 3.4 to converge, we need $\alpha(s) = O(s^{-\phi})$ for some $\phi > \phi_0 = \beta/(\beta-2)$, i.e. the process needs to be α-mixing of size $-\phi_0$. As $\beta \to \infty$ so that all moments are finite, the required mixing size $\phi_0 \to 1$. Because the mixing coefficients also determine bounds for the (absolute) autocovariance function of the process (see Davidson (2021), Corollary 15.3), this directly implies that with all moments existing, $\mathrm{Cov}\left[g\left(Z_t^{(h)}, h; \Gamma_0\right), g\left(Z_{t-s}^{(h)}, h; \Gamma_0\right)\right] = O(s^{-1})$ for the assumptions to be satisfied.

3.2 The double machine learning estimator with multiple independent stochastic processes

We now provide asymptotic results for the case where the nuisance functions are estimated and $K \geq 2$ independent stochastic processes generated from the same distribution $\mathcal{P}$ are available. Denote the individual stochastic processes as $S_i^{(h)} = \{Z_t^{(h)} : t \in \mathcal{T}_i\}$, where, without loss of generality, we assume $|\mathcal{T}_i| = T/K$ for all $i = 1, ..., K$. The estimation procedure is outlined in Procedure 1. Asymptotics for the DML estimator $\hat{\theta}^{(h)}$ from Procedure 1 are

For each forecast horizon h, follow the subsequent procedure.

1. For each $i = 1, ..., K$
   (a) Fit appropriate machine learners $\hat{\Gamma}_{S_{-i}^{(h)}} = (\hat{\mu}_{S_{-i}^{(h)}}(d, x, h),\ \hat{e}_{S_{-i}^{(h)}}(x))$ on the sample $S_{-i}^{(h)} = \bigcup_{j=1, j\neq i}^{K} S_j^{(h)}$.
   (b) Compute the average of $g\left(z, h; \hat{\Gamma}_{S_{-i}^{(h)}}\right)$ on $S_i^{(h)}$ as
   $$\hat{\theta}_{S_i^{(h)}}^{(h)} = \frac{1}{|\mathcal{T}_i|}\sum_{t\in\mathcal{T}_i} g\left(Z_t^{(h)}, h; \hat{\Gamma}_{S_{-i}^{(h)}}\right).$$
2. Compute the IRF estimator at horizon h as
   $$\hat{\theta}^{(h)} = \sum_{i=1}^{K} \frac{|\mathcal{T}_i|}{T} \hat{\theta}_{S_i^{(h)}}^{(h)}.$$

Procedure 1: DML estimator for the IRF with cross-fitting on multiple independent stochastic processes

obtained by introducing assumptions under which $\hat{\theta}^{(h)}$ has the same asymptotic distribution as the oracle estimator $\tilde{\theta}^{(h)}$ in Theorem 2. To this end, we impose standard assumptions on the convergence rates of the learners used in step 1. of Procedure 1. In particular, we assume that the machine learners are consistent and that the product of the estimation errors decays fast enough.

Assumption 4. *Let the realization set be $\Xi_T^{(h)}$, which is a shrinking neighborhood of the true nuisance functions $\Gamma_0 = (\mu_0(d, x, h), e_0(x))$. Let $\{\Delta_T\}_{T\geq 1}$ and $\{\delta_T\}_{T\geq 1}$ be sequences of positive constants converging to zero.*

Define the statistical rates $r_{\mu,T} = \sup_{t\in\mathcal{T}} \sup_{\mu\in\Xi_T^{(h)}} \|\mu(D_t, X_t, h) - \mu_0(D_t, X_t, h)\|_2$ and $r_{e,T} = \sup_{t\in\mathcal{T}} \sup_{e\in\Xi_T^{(h)}} \|e(X_t) - e_0(X_t)\|_2$. Let C be a fixed strictly positive constant. For all $i = 1, ..., K$ and $h \in \mathbb{N}_0$, the following conditions hold.

1. *The nuisance function estimators $\hat{\Gamma}_{S_{-i}^{(h)}}$ belong to $\Xi_T^{(h)}$ with probability at least $1 - \Delta_T$.*

2. *For $q > 2$, we have $\sup_{t\in\mathcal{T}} \sup_{\mu\in\Xi_T^{(h)}} \|\mu(D_t, X_t, h) - \mu_0(D_t, X_t, h)\|_q \leq C < \infty$ and $\sup_{t\in\mathcal{T}} \sup_{e\in\Xi_T^{(h)}} \|e(X_t) - e_0(X_t)\|_q \leq C < \infty$.*

3. *$r_{\mu,T} \leq \delta_T$, $r_{e,T} \leq \delta_T$ and $r_{\mu,T} \cdot r_{e,T} \leq T^{-1/2}\delta_T$.*

4. *$\sup_{t\in\mathcal{T}} \sup_{e\in\Xi_T^{(h)}} \|e(X_t) - 1/2\|_\infty \leq 1/2 - \eta$ for $0 < \eta < 1$.*

5. *$\sup_{t\in\mathcal{T}} \mathbb{E}\left[(Y_{t+h} - \mu_0(d, X_t, h))^2 | X_t, D_t = d\right] \leq \epsilon_d^2 < \infty$*

The statistical rates in Assumption 4 are defined in terms of uniform L_2-norms. Under the additional assumptions of strict stationarity of $S^{(h)}$ and measurability of the nuisance functions, these uniform norms would reduce to simple L_2-norms. Assumption 4.2 requires that, with probability approaching one, the estimation errors are uniformly L_q-bounded for $q > 2$. Assumption 4.3 requires the estimation errors to converge uniformly in L_2-norm to zero and their products to converge at least at the rate $\sqrt{T}$ with probability approaching one. Our assumptions on convergence rates are set up to accommodate the application of a broad range of machine learning estimators for the nuisance functions. There is a rich literature deriving convergence rates of machine learners under more strict conditions than used here. Wong, Li, and Tewari (2020), for example, provide Lasso convergence rates for stochastic processes under the assumption of exact sparsity. For α-mixing Gaussian processes, they find the L_2 convergence rate to be of order $O\left(A(T)\sqrt{\log \dim(\mathcal{X})/T}\right)$, where $A(T) = \sum_{s=0}^{T} \alpha(s)$. Assumption 4.3 is satisfied for $A(T) = o(T^{1/4})$, imposing a restriction on how fast the dependence in the data has to decay. We provide more references in Section 4.2. Next, Assumption 4.4 implies that the estimated propensity scores remain uniformly bounded away from zero and one with probability approaching one. Finally, Assumption 4.5 requires that the conditional variance of the outcome variable is a bounded random variable.

We furthermore impose the following sequential conditional exogeneity condition.

Assumption 5. *For all $t \in \mathcal{T}$ and $h \in \mathbb{N}_0$ we have that $\mathbb{E}[Y_{t+h}|X_t, D_t = d, \{Z_u^{(0)} : u \in \mathcal{T}, u < t\}] = \mathbb{E}[Y_{t+h}|X_t, D_t = d]$ and $\mathbb{E}[D_t|X_t, \{Z_u^{(0)} : u \in \mathcal{T}, u < t\}] = \mathbb{E}[D_t|X_t]$.*

Assumption 5 implies that the residuals $D_t - e_0(X_t)$ and $Y_{t+h} - D_t\mu_0(1, X_t, h) - (1 - D_t)\mu_0(0, X_t, h)$ are mean independent of past information on (Y_t, X_t, D_t). In line with standard assumptions in the literature (e.g. Olea, Plagborg-Møller, Qian, & Wolf, 2024; Semenova, Goldman, Chernozhukov, & Taddy, 2023), this in practice requires a rich enough set of control variables, which can also contain past values of Y_t and D_t.

The following theorem finally establishes that the DML estimator for the IRF is asymptotically unbiased and normally distributed. In particular, the estimator retains the parametric $\sqrt{T}$ convergence rate. The proof is relegated to the Appendix. The main idea of the proof is to show that the IRF estimator using estimated nuisance functions converges to the oracle IRF estimator $\tilde{\theta}^{(h)}$, which itself is asymptotically normally distributed as shown in Theorem 2, at rate $\sqrt{T}$.

Theorem 3. *Let* $S_i^{(h)} = \left\{ Z_t^{(h)} : t \in \mathcal{T}_i \right\}$ *for* $i = 1, ..., K \geq 2$ *and* $h \in \mathbb{N}_0$ *be independent stochastic processes generated from the same distribution* $\mathcal{P}$ *such that* $|\mathcal{T}_i| = T/K$ *for all* i*. Define the estimator as*

$$\hat{\theta}^{(h)} = \sum_{i=1}^{K} \frac{|\mathcal{T}_i|}{T} \hat{\theta}^{(h)}_{S_i^{(h)}} \quad \textit{with}$$

$$\hat{\theta}^{(h)}_{S_i^{(h)}} = \frac{1}{|\mathcal{T}_i|} \sum_{t \in \mathcal{T}_i} g\left(Z_t^{(h)}, h; \hat{\Gamma}_{S_{-i}^{(h)}} \right),$$

where the nuisance functions $\hat{\Gamma}_{S_{-i}^{(h)}} = (\hat{\mu}_{S_{-i}^{(h)}}, \hat{e}_{S_{-i}^{(h)}})$ *are estimated on* $S_{-i}^{(h)} = \bigcup_{j=1, j \neq i}^{K} S_j^{(h)}$*. Then under Assumptions 1 and 3 - 5 it holds that*

$$\sqrt{T}(\hat{\theta}^{(h)} - \theta_0^{(h)}) \xrightarrow{\mathrm{d}} \mathcal{N}\left(0, V_0^{(h)}\right),$$

with $V_0^{(h)}$ *as in Theorem 2.*

3.3 The double machine learning estimator with one stochastic process

In practice, multiple independent stochastic processes generated from the same distribution are often not available. Instead, a sample from a single stochastic process $S^{(h)}$ is observed. In the same spirit as approaches used for cross-validating models with dependent data (Bergmeir & Benítez, 2012; Racine, 2000) and the cross-fitting approach proposed by Semenova et al. (2023) for panel data, we split the single stochastic process into sub-sequences, removing a block of k_T coordinates where the process is split. The resulting sub-sequences replace the independent stochastic processes $S_i^{(h)}$ from Procedure 1. To this end, let $\{\mathcal{T}_i : i = 1, \ldots, K\}$ be a partition of the index set $\mathcal{T}$ such that the order of the time indices within each $\mathcal{T}_i$ and across all subsets follows the original order in $\mathcal{T}$. Without loss of generality, we continue to assume that $|\mathcal{T}_i| = T/K$ for all i. The estimation procedure is described in Procedure 2.

For each forecast horizon h, follow the subsequent procedure.

1. For each $i = 1, ..., K$
   (a) Define $\mathcal{T}_{-i} = \{t : t \in \mathcal{T} \wedge (t < \inf(\mathcal{T}_i) - k_T \vee t > \sup(\mathcal{T}_i) + k_T)\}$
   (b) Fit appropriate machine learners $\hat{\Gamma}_{S_{-i}^{(h)}} = (\hat{\mu}_{S_{-i}^{(h)}}(d, x, h), \hat{e}_{S_{-i}^{(h)}}(x))$ on the sample $S_{-i}^{(h)} = \{Z_t^{(h)} : t \in \mathcal{T}_{-i}\}$.
   (c) Compute the average of $g\left(z, h; \hat{\Gamma}_{S_{-i}^{(h)}}\right)$ on $S_i^{(h)}$ as

$$\hat{\theta}^{(h)}_{S_i^{(h)}} = \frac{1}{|\mathcal{T}_i|} \sum_{t \in \mathcal{T}_i} g\left(Z_t^{(h)}, h; \hat{\Gamma}_{S_{-i}^{(h)}} \right).$$

2. Compute the IRF estimator at horizon h as

$$\hat{\theta}^{(h)} = \sum_{i=1}^{K} \frac{|\mathcal{T}_i|}{T} \hat{\theta}^{(h)}_{S_i^{(h)}}.$$

Procedure 2: DML estimator for the IRF with cross-fitting on a single stochastic processes

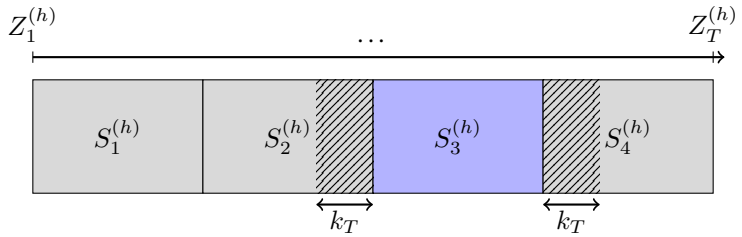
An illustration of the cross-fitting approach for $K = 4$ is given in Figure 1. The time series is divided into four sub-sequences. The nuisance functions are estimated on the union of sub-sequences S_1, S_2 and S_3, where k_T coordinates are removed at the boundaries of S_3. Sub-sequence S_3 is used to compute $\hat{\theta}^{(h)}_{S_3}$. This procedure is repeated such that each sub-sequence is used once to conduct inference.

The asymptotic results for the estimator described in Procedure 2 require some additional assumptions.

Assumption 6. *For* $d \in \{0, 1\}$*,* $h \in \mathbb{N}_0$ *and some scalar constant* $p \geq 1$ *the following conditions hold.*

1. $k_T = O(T)$.
2. *The nuisance functions* $\mu_0(d, x, h)$*,* $e_0(x)$ *and the functions* $\mu(d, x, h), e(x) \in \Xi_T^{(h)}$ *are measurable.*

Figure 1: Illustration of the cross-fitting procedure



NOTE: The figure illustrates the cross-fitting procedure for $K = 4$. The nuisance functions are estimated using appropriate machine learners on the union of sub-sequences $S_1^{(h)}$, $S_2^{(h)}$ and $S_4^{(h)}$ after dropping k_T observation at the boundaries to $S_3^{(h)}$. Sub-sequence $S_3^{(h)}$ is used to compute $\hat{\theta}^{(h)}_{S_3^{(h)}}$. This procedure is repeated such that each sub-sequence is used once to conduct inference.

3. *For for $r > p$ and $1/r = 1/r' + 1/r''$, we have* $\sup_{t\in\mathcal{T}} \|\sup_{\mu\in\Xi_T^{(h)}}(\mu(d, X_t, h) - \mu_0(d, X_t, h))\|_{2r'} < \infty$ *and* $\sup_{t\in\mathcal{T}} \|e_0(X_t) - D_t\|_{2r''} < \infty$.
4. *For $q > p$ and $1/q = 1/q' + 1/q''$, we have* $\sup_{t\in\mathcal{T}} \|\sup_{e\in\Xi_T^{(h)}}(e(X_t) - e_0(X_t))\|_{2q'} < \infty$ *and* $\sup_{t\in\mathcal{T}} \|Y_t - \mu_0(D_t, X_t, h)\|_{2q''} < \infty$.
5. *The stochastic processes $S^{(h)} = \{Z_t^{(h)} : t \in \mathcal{T}\}$ are α-mixing with coefficients $\alpha(s)$, satisfying for $T \to \infty$ that $\alpha\,(k_T)^{\psi} = o(T^{-1})$, where $\psi = 1/p - 1/\min(r, q)$.*

Assumption 6 imposes restrictions on the dependence structure in the stochastic processes $S^{(h)}$. As we can no longer rely on the independence between $S_i^{(h)}$ and $S_{-i}^{(h)}$, we require the dependence in the stochastic process to decay fast enough. Intuitively, after removing k_T coordinates at the boundaries of $S_i^{(h)}$ it should (asymptotically) become independent of $S_{-i}^{(h)}$. The choice of k_T hereby represents a trade-off. The larger k_T, the smaller is the effective estimation sample size. Indeed, Assumption 6.1 requires k_T to not increase more rapidly than the sample size. However, k_T needs to be large enough to satisfy Assumption 6.5. The stronger the dependence in the stochastic processes $S^{(h)}$, the larger k_T needs to be, in turn reducing the effective estimation sample size. For the sake of exposition, assume that $k_T = O(T^{\vartheta})$ for $0 < \vartheta \leq 1$ and let us look at different exemplary assumptions on the dependence structure of $S^{(h)}$.

(i) α-mixing process: If the stochastic processes $S^{(h)}$ are α-mixing of size $-\phi_0$, i.e. $\alpha(s) = O(s^{-\phi})$ for some $\phi > \phi_0$, then Assumption 6.5 is satisfied for $\vartheta > (\phi\psi)^{-1}$ and $0 < \psi < 1$. Put differently, the slower the decay in the dependence (smaller ϕ), the larger ϑ and thus k_T has to be.

(ii) Persistent process: For stochastic processes $S^{(h)}$ with very slowly decaying dependence of the form $\alpha(s) = O(s^{2d-1})$ with $0 < d < 1/2$, we have additionally that $0 < \phi < 1$ and as a consequence Assumption 6.5 is indeed never satisfied.

(iii) Weakly dependent processes: If the stochastic processes $S^{(h)}$ have mixing coefficients $\alpha(s) = O(\rho^s)$ for $0 < \rho < 1$, then Assumption 6.5 is already satisfied with $\vartheta > 0$.

(iv) Independent process: In this case, i.e. when it is assumed that the stochastic processes $S^{(h)}$ are a collection of independent random variables, then $\alpha(s) = 0$ for all $s > 0$ and Assumption 6.5 is already satisfied for $k_T = 0$.

These examples intuit two things that are required for our theory to hold in practice. First, if only one stochastic process $S^{(h)}$ is available for estimation, enough coordinates must be removed when splitting, so that no influence of past coordinates exists in coordinates of a subsequent split. Second, the stochastic processes $S^{(h)}$ itself have to exhibit fast enough decaying temporal dependence. If this is not the case, suitable transformations of the original process need to be found to achieve this. Lastly, the boundedness conditions in the above assumptions also represent a trade-off. The more moments of the residuals and estimation errors are finite (larger r and q), the less stringent are the conditions on the dependence decay, and vice versa. In general, with only one stochastic process, the conditions on the estimation errors are more restrictive than those imposed in Assumption 4.2. The proposed cross-fitting approach is closely related to the methodology introduced by Semenova et al. (2023) for panel data, wherein the data is partitioned into folds along the temporal dimension. The primary distinction lies in the selection of observations omitted between the estimation and inference samples. While Semenova et al. (2023) advocate for a K-fold partitioning of the sample,

removing an entire fold between estimation and inference samples, i.e. $k_T = \lfloor T/K \rfloor$ (see Section 4.1 for further discussion), the present study adopts a more flexible strategy. Specifically, as outlined above, the choice of k_T can be informed by the underlying dependence structure of the data. The following theorem finally establishes asymptotic properties for the case when nuisance functions are estimated based on a single stochastic process.

Theorem 4. *Given the stochastic processes $S^{(h)} = \{Z_t^{(h)} : t \in \mathcal{T}\}$ for $h \in \mathbb{N}_0$, define $K \geq 2$ sub-sequences $S_i^{(h)} = \left\{Z_t^{(h)} : t \in \mathcal{T}_i\right\}$ such that $\{\mathcal{T}_i : i = 1, \ldots, K\}$ is a partition of the index set $\mathcal{T}$ where the order of the time indices within each $\mathcal{T}_i$ and across all sub-sequences with length $|\mathcal{T}_i| = T/K$ follows the original order in $\mathcal{T}$. Define the estimator as*

$$\hat{\theta}^{(h)} = \sum_{i=1}^{K} \frac{|\mathcal{T}_i|}{T} \hat{\theta}^{(h)}_{S_i^{(h)}} \quad \textit{with}$$

$$\hat{\theta}^{(h)}_{S_i^{(h)}} = \frac{1}{|\mathcal{T}_i|} \sum_{t \in \mathcal{T}_i} g\left(Z_t^{(h)}, h; \hat{\Gamma}_{S_{-i}^{(h)}}\right),$$

where the nuisance functions $\hat{\Gamma}_{S_{-i}^{(h)}} = (\hat{\mu}_{S_{-i}^{(h)}}, \hat{e}_{S_{-i}^{(h)}})$ are estimated using the sub-sequences $S_{-i}^{(h)} = \{Z_t^{(h)} : t \in \mathcal{T}_{-i}\}$ for $\mathcal{T}_{-i} = \{t : t \in \mathcal{T} \wedge (t < \inf(\mathcal{T}_i) - k_T \vee t > \sup(\mathcal{T}_i) + k_T)\}$ respectively. Then under Assumptions 1 and 3 - 6 it holds that

$$\sqrt{T}(\hat{\theta}^{(h)} - \theta_0^{(h)}) \xrightarrow{\mathrm{d}} \mathcal{N}\left(0, V_0^{(h)}\right),$$

with $V_0^{(h)}$ as in Theorem 2.

3.4 Variance estimation and inference

The variances $V_0^{(h)}$ can be estimated using standard long-run variance estimators for time series, such as the one proposed by Newey and West (1987).

Assumption 7. *The following conditions hold.*

1. *There are some fixed finite constants C and $r > 4$ such that $\sup_{t \in \mathcal{T}} \mathbb{E}\left[\left|g\left(Z_t^{(h)}, h; \Gamma_0\right)\right|^r\right] < C$.*
2. *There exists a measurable function $m(z)$ such that $\sup_{\Gamma \in \Xi_T} |g(z, h; \Gamma)| < m(z)$, where for some finite constant D, $\sup_{t \in \mathcal{T}} \mathbb{E}[m(Z_t)^2] < D$.*
3. *For $q > 2$ and some fixed strictly positive and finite constant C we have $\sup_{t \in \mathcal{T}} \|Y_t\|_q \leq C$.*
4. *For some scalar $0 < b_m < b_r \leq 1/2$ it holds that:*
   (a) *The bandwidth m_T is a function of the sample size such that $\lim_{T \to \infty} m_T = \infty$ and for $T \to \infty$ it holds that $T^{-b_m} m_T = o(1)$.*
   (b) *$r_{\mu,T} \leq \delta_T T^{-b_r}$ and $r_{e,T} \leq \delta_T T^{-b_r}$.*

Assumptions 7.4a and 7.4b represent an inherent trade-off. The slower the convergence rate of the machine learner, the slower the bandwidth is allowed to grow. If the learner converges at the parametric rate, Assumption 7.4a reduces to the usual assumption $m_T = o(T^{1/2})$ (see, e.g. Kool, 1988). Note that by Assumption 4.3 we have that $b_r \geq 1/4$.

The following theorem establishes consistency of a variance estimator resulting from averaging Newey and West (1987) type estimators on each $S_i^{(h)}$.

Theorem 5. *Given the stochastic processes $S^{(h)} = \{Z_t^{(h)} : t \in \mathcal{T}\}$, define the sub-sequences $S_i^{(h)}$ for $i = 1, ..., K \geq 2$ and $h \in \mathbb{N}_0$ as in Theorem 4. Furthermore, define $v_t^{(h)} = g\left(Z_t^{(h)}, h; \Gamma_0\right) - \theta_0^{(h)}$ and the corresponding estimated quantities as $\hat{v}^{(h)}_{S_i^{(h)},t} = g\left(Z_t^{(h)}, h; \hat{\Gamma}_{S_{-i}}\right) - \hat{\theta}^{(h)}$. $\hat{\theta}^{(h)}$ and $S_{-i}^{(h)}$ are defined as in Theorem 4 and the nuisance functions $\hat{\Gamma}_{S_{-i}^{(h)}} = (\hat{\mu}_{S_{-i}^{(h)}}, \hat{e}_{S_{-i}^{(h)}})$ are estimated on $S_{-i}^{(h)}$. Let $w(s, m_T) = 1 - s/(m_T + 1)$, where m_T is a bandwidth parameter and define the additional index sets $\mathcal{T}_{i,s} = \{t \in \mathcal{T}_i : t - s \geq \inf(\mathcal{T}_i)\}$. Moreover, define the following Newey and*

West (1987) type variance estimators as

$$\widehat{V}^{(h)}_{S_i^{(h)}} = \frac{1}{|\mathcal{T}_i|}\left(\sum_{t\in\mathcal{T}_i}(\hat{v}^{(h)}_{S^{(h)}_{-i},t})^2 + 2\sum_{s=1}^{m_T} w(s,m_T)\sum_{t\in\mathcal{T}_{i,s}}\hat{v}^{(h)}_{S^{(h)}_{-i},t}\hat{v}^{(h)}_{S^{(h)}_{-i},t-s}\right).$$

The variance estimator is finally defined as

$$\widehat{V}^{(h)} = \sum_{i=1}^{K}\frac{|\mathcal{T}_i|}{T}\widehat{V}^{(h)}_{S_i^{(h)}}$$

Then for $V_0^{(h)}$ as in Theorem 2 and under Assumptions 1 and 3 - 7 as $T\to\infty$ it holds that

$$\left|\widehat{V}^{(h)} - V_0^{(h)}\right| \xrightarrow{p} 0$$

with measure $\mathcal{P}$.

Note that while Theorem 5 is formulated in terms of a specific weight function $w(s,m)$, as in Newey and West (1987), the variance estimator is consistent for any weight function additionally satisfying for each s that $\lim_{m(T)\to\infty} w(s,m_T) = 1$ and $|w(s,m_T)| < \infty$. Using the estimators in Theorems 4 and 5, inference can be conducted by constructing level-α confidence bounds for $\theta_0^{(h)}$ as

$$\theta_0^{(h)} \in \left(\hat{\theta}^{(h)} \pm \frac{1}{\sqrt{T}}\Phi^{-1}\left(1-\frac{\alpha}{2}\right)\sqrt{\widehat{V}^{(h)}}\right), \tag{3}$$

where $\Phi^{-1}(\cdot)$ represents the inverse cumulative distribution function of the standard normal distribution.

4 Considerations on the practical implementation of the estimator

Here, we gather practical recommendations for the time series DML estimator.

4.1 Cross-fitting and small samples

For small samples, we recommend setting K to a rather large value (e.g. $K = 10$ or $K = 20$). This increases the number of observations available to estimate the nuisance functions. Regarding the choice of k_T, it is important to note that this is directly affected by the definition of the outcome variable. In cases where the effect on the outcome variable h periods after the treatment is of interest, k_T has to be chosen such that samples are not overlapping, i.e. $k_T \geq h$. Similarly, k_T has to take into account possible lagged values in X_t. As a guideline, we recommend that practitioners set K to either 10 or 20, and, following Semenova et al. (2023), use $k_T = \lfloor T/K \rfloor$ as an initial choice. This satisfies Assumptions 6.1 and 6.5 across a wide range of dependence structures. Sensitivity analysis for variations in k_T can then be done to ensure robustness of results.

4.2 Estimators for the nuisance functions

As in the *i.i.d.* setting (Chernozhukov et al., 2018), our theory requires the estimators for the nuisance functions to be consistent with fast enough convergence rates. Following Theorems 3 and 4, this needs to extend to estimation on time-dependent observations. Consistency and convergence rates on α-mixing sequences are derived for Lasso in Wong et al. (2020), for random forests in Goehry, Benjamin (2020) and Davis and Nielsen (2020), for boosting algorithms in Lozano, Kulkarni, and Schapire (2014), for support vector machines in Steinwart, Hush, and Scovel (2009), for kernel and nearest-neighbour regressions in Irle (1997) and for spline and wavelet series regression estimators in X. Chen and Christensen (2015). Consistency of deep feed-forward neural networks with ReLU activation functions on exponentially α-mixing processes was recently shown in Ma and Safikhani (2022).

The selection of an appropriate machine learning algorithm ultimately has to consider the specific problem and the characteristics of the data. For example, the random forest algorithm has been shown to perform effectively in macroeconomic contexts, even with relatively small sample sizes (Beck & Wolf, 2025; Goulet Coulombe, 2024; Goulet Coulombe, Leroux, Stevanovic, & Surprenant, 2022; Medeiros, Vasconcelos, Álvaro Veiga, & Zilberman, 2021). Our numerical experiments, as well as the empirical application, also find random forests to be effective on representative sample sizes and data generating processes. In contexts with typically larger sample sizes, recurrent

neural networks have e.g. demonstrated success in modelling high-frequency market data (Lucchese, Pakkanen, & Veraart, 2024; Zhang, Zohren, & Roberts, 2019). For an overview of applying machine learning algorithms to time series, we also refer to the recent survey by Masini, Medeiros, and Mendes (2023).

Our numerical experiments suggest that, as in the *i.i.d.* case (Bach, Schacht, Chernozhukov, Klaassen, & Spindler, 2024), properly tuning hyperparameters of the chosen estimators plays an important role in the application of double machine learning also for time series data. In summary, we recommend to estimate and tune multiple different estimators and select the best in terms of the relevant loss function for the problem at hand (e.g. predictive mean squared error).

4.3 Modelling multiple forecast horizons

Our theory is agnostic to how the nuisance functions for different forecast horizons h are modelled. In analogy to classical impulse response function estimation using local projections, each forecast horizon (and nuisance function) can be estimated separately, and potentially with different learning algorithms. Depending on the application, it is however also possible to estimate $\mu_0(d, x, h)$ with one model for all h, e.g. using sequence-to-sequence approaches (Mariet & Kuznetsov, 2019), provided they exhibit appropriate convergence rates. Applying approaches from multitask learning (Caruana, 1997), it is finally also possible to estimate $\mu_0(d, x, h)$ and $e_0(x)$ in one model if both problems can be learned using a shared representation. This approach has e.g. been explored by Shi, Blei, and Veitch (2019) in the context of *i.i.d.* data.

4.4 Inference in finite samples

Theorem 5 requires the choice of a kernel bandwidth m_T that fulfils Assumption 7. A valid choice would be, for example, $m_T = \gamma T^{1/3}$ where γ is determined by the procedure proposed by Newey and West (1994). The variance estimator, when combined with standard normal critical values, is asymptotically valid under general forms of heteroskedasticity and autocorrelation. However, extensive research has shown that this approach can perform poorly in finite samples (Kiefer & Vogelsang, 2005; Sun, 2014). In response, the literature has proposed fixed-bandwidth asymptotics, where the ratio of bandwidth and sample size m_T/T is held fixed as the sample size grows, rather than shrinking to zero (as in traditional small-bandwidth asymptotics). This framework yields non-standard limiting distributions and requires the use of fixed-bandwidth critical values (see Kiefer & Vogelsang, 2005), which better approximate the finite-sample behavior of test statistics. Our numerical experiments confirm this finding also for the estimator proposed in this manuscript. For practical applications, we thus advise to use fixed-bandwidth critical values when performing inference.

4.5 Extreme propensity scores

Assumption 2 and 4 require the true propensity scores $e_0(x)$ to be bounded away from zero and one. In applications, however, certain treatments may have essentially zero probability for particular regions of the covariate space. For example policy interventions that are infeasible under extreme macroeconomic conditions. When such limited support is a concern, we recommend employing a propensity-score-based trimming procedure, following Crump, Hotz, Imbens, and Mitnik (2009). This approach systematically excludes observations with extreme propensity scores to improve overlap between treated and control units and to ensure that the resulting estimand pertains to a subpopulation with adequate support. A related but different issue that may arise in finite samples are numerical instabilities of the reciprocal of the propensity score. In other words, while it may hold that at the population level the propensity score is bounded away from zero and one, in finite samples, the estimated propensity scores can still be close to zero or one. A simple approach to address this instability is to winsorize the estimated propensity scores to a small number, e.g. 0.01 and 0.99 (B. K. Lee, Lessler, & Stuart, 2011). Alternatively, one can calibrate the estimated propensity score (Ballinari & Bearth, 2025; Klaassen, Rabenseifner, Kueck, & Bach, 2025).

5 Simulation experiments

To validate our theoretical results from the previous sections in finite samples, we conduct simulation experiments. We compare the DML estimator from Procedures 1 and 2 to a regression adjustment estimator (RA) that estimates the nuisance functions $\mu_0(1, x, h)$ and $\mu_0(0, x, h)$ separately on the full sample and takes their difference (a T-learner in the terminology of Künzel, Sekhon, Bickel, and Yu (2019)). To disentangle the effect of cross-fitting and usage of a Neyman orthogonal estimator, we also compute the IRF using the doubly robust influence function (2) without relying on cross-fitting (DR). In addition, we estimate the impulse response functions using standard local projections (LP) (Jordà, 2005). Results are presented throughout using random forests (Breiman, 2001) as machine learning estimators

for the nuisance functions. In the Online Appendix, Table C6, we also include results using a gradient boosting algorithm estimator (T. Chen & Guestrin, 2016), supporting the validity of our asymptotic theory for alternative machine learning estimators.[3] More details on the hyperparameter tuning schemes for both considered estimators are also given in the Online Appendix A.1. In all simulations, we generate data according to the following data generating process (DGP), which is a modification of the setup in Nie and Wager (2020). For some noise level σ_ϵ, a propensity score $e_0(X_t)$, a baseline effect $b(X_t)$ and a (conditional) treatment effect function $\tau(X_t)$, the outcome process is defined as

$$Y_t = b(X_t) + (D_t - 0.5)\,\tau(X_t) + \gamma Y_{t-1} + \epsilon_t,$$

where the innovations ϵ_t are generated from a GARCH(1,1) process $\epsilon_t = \sigma_t \zeta_t$, with $\zeta_t \sim \mathcal{N}(0,1)$ and $\sigma_t^2 = \omega + \beta_1 \zeta_{t-1}^2 + \beta_2 \sigma_{t-1}^2$. Following Jordà (2005) we set $\beta_1 = 0.3$ and $\beta_2 = 0.5$. ω is set to ensure that the $\mathbb{E}[\epsilon_t^2] = \sigma_\epsilon^2$, that is $\omega = (1 - \beta_1 - \beta_2)\sigma_\epsilon^2$. We set $D_t | X_t \sim \text{Ber}\left(e_0(X_t)\right)$ with

$$\begin{aligned} e_0(X_t) &= \left(1 + e^{-X_{1,t}} + e^{-X_{2,t}}\right)^{-1} \\ b(X_t) &= 0.5\left((X_{1,t} + X_{2,t} + X_{3,t})^+ + (X_{4,t} + X_{5,t})^+\right) \\ \tau(X_t) &= (X_{1,t} + X_{2,t} + X_{3,t})^+ - (X_{4,t} + X_{5,t})^+ \end{aligned}$$

where $(x)^+ = \max(0, x)$. The confounder process is modelled as a n-dimensional, zero mean VARMA(p,q) process

$$X_t = \sum_{i=1}^{p} A_i X_{t-i} + \sum_{j=1}^{q} M_j u_{t-j} + u_t,$$

where u_t is a zero mean white noise random variable with nonsingular covariance matrix, parameterized as $\Sigma_u = \sigma_u^2 I_n$ using some scalar σ_u and the n-dimensional identity matrix I_n. In the spirit of Adamek et al. (2024), the coefficient matrices are defined as $A_i = \alpha_A^{i-1}\Gamma^A$ and $M_j = \alpha_M^{j-1}\Gamma^M$, where α_A, α_M are some scalars, Γ^A (and Γ^M correspondingly) is a tapered Toeplitz matrix with $\Gamma^A_{i,j} = \rho_A^{|i-j|+1}$ and $\Gamma^A_{i,j} = 0$ for $|i-j| \geq n/2$. We finally scale the process X_t so that the confounders have unit variance. The baseline parametrization for our simulations is $\gamma = 0.6$, $\sigma_\epsilon = \sigma_u = 1$, $n = 12$, $\alpha_A = \alpha_B = 0.3$, $p = 2$, $q = 1$, $\rho_A = 0.35$ and $\rho_M = 0.7$. The simulation procedure is described in Procedure 3. We perform the numerical experiments for two settings. A first one, where

1. Draw a realization from the DGP with T observations.
2. For each evaluated forecast horizon $h = 0, 1, ..., H$:
   a) Construct $\{S_i^{(h)} : i = 1, \ldots, K\}$ from the realization.
   b) Find optimal hyperparameters for the estimators for $\mu_0(0, X, h)$, $\mu_0(1, X, h)$, and $e_0(X)$ by cross-validation using $\{S_i^{(h)} : i = 1, \ldots, K\}$ as folds and removing k_T observations at the boundary between estimation and inference sample.
   c) Train each of the four learners (DML, RA, DR, LP); and for each learner
      i. compute the IRF estimator $\hat{\theta}^{(h)}$ according to Procedure 1 or Procedure 2
      ii. compute the variance estimator $\widehat{V}^{(h)}$ from Theorem 5. Following the arguments outlined in Section 4.4, we use the approach in Newey and West (1994) to determine the bandwidth m_T.
3. Repeat steps 1. and 2. N times.

Procedure 3: Setup of the simulation study

in step 2.a) the realizations for $S_i^{(h)}$ are in fact drawn separately by simulating K independent realizations from the DGP in step 1., each with T/K observations. In a second setting, the sub-samples are constructed from the one single realization drawn in step 1. In this setting, we remove $k_T = T/K$ coordinates at the boundary of the estimation and inference samples. Following our practical recommendation, we set $K = 10$. Results for the baseline parametrization of the DGP for the DML, RA, DR and LP estimators for the setting with one stochastic process and for sample sizes $T \in \{125, 250, 500, 1'000, 8'000\}$ are shown in Table 1. Results for the setting with independent stochastic processes

[3]The gradient boosting algorithm is found to require slightly larger sample sizes than random forests to reach a consistent estimate, highlighting that estimators can exhibit varying sample size requirements with respect to our asymptotic theory. For practical applications we thus suggest, cf. Section 4.2, to select the estimator yielding the best predictive performance on the available sample size.

and for some parameter variations (different number of confounders, higher noise in the outcome process, empirically calibrated parameters) are deferred to Tables C2-C5 in the Online Appendix.

Overall, our simulations support the validity of our theory. Compared to the RA and DR estimators, the DML estimator exhibits the smallest average bias in all considered settings, converging at the expected $\sqrt{T}$-rate. Importantly, this holds true for both Procedure 1 relying on independent realizations and Procedure 2 using a single realization. Being linear estimators, local projections do not estimate the true nonlinear average treatment effect, but a weighted average of marginal effects (Kolesár & Plagborg-Møller, 2025). This is highlighted by the observation that the bias of the LP estimator does not decrease and its coverage deteriorates with increasing sample sizes. On these points also refer to Section 6. Finally, the DML estimator produces valid confidence intervals, while the regression adjustment, doubly robust and local projection estimators fail to allow valid inference. As expected, when K truly independent realizations are available (cf. Table C2 in the Online Appendix), the bias of the DML estimator is lower than in the one-realization setting.

In Table 1, for the DML estimator, we report both the coverage using asymptotic and fixed-bandwidth critical values (cf. Section 4.4). In small samples, coverage is better when using fixed-bandwidth critical values. As sample sizes increase, the improvement over asymptotic critical values becomes negligible.[4] In most settings, the coverage of the DML estimator is marginally too low, which likely reflects a finite-sample bias in estimating the variance of $\hat{\theta}^{(h)}$. Unreported results indeed show that, on average, our variance estimator is slightly smaller than the empirical variance of the IRF estimates across realizations. This downward bias is expected because, while the variance estimator is derived under the assumption of known nuisance functions, it is in practice constructed using estimates thereof. This introduces sampling variability that is not fully accounted for in finite samples.

Figure 2 provides a visual summary of our results by depicting the true and estimated IRF. The regression adjustment estimator is biased, overestimating on average the true impact of the treatment D_t, in particular for longer horizons. When estimating the IRF with the doubly robust influence function (DR), the bias is reduced. Only when using an estimator that is Neyman orthogonal and uses cross-fitting (DML), the distribution of the estimated IRFs across simulation replications is centered around the true IRF. In the inset of the top panel of Figure 2, for the DML estimator, normalized biases as a function of T are plotted for all forecast horizons h and contrasted to the $\sqrt{T}$ scaling implied by Theorem 3, showing that the bias of the DML estimator follows the $\sqrt{T}$ scaling implied by Theorem 3 quite well. The LP estimator finally exhibits a bias of similar magnitude as the RA estimator, as the linear estimator fails to capture the highly nonlinear and heterogeneous relation between D_t, X_t and Y_{t+h} in the DGP.

6 Comparison to local projections

The simulation study in the previous section provided evidence that, in the presence of nonlinearities, local projections are asymptotically biased, while the proposed DML estimator consistently estimates the true impulse response function. Here, we contrast these two estimators and their underlying assumptions. For a comparison of local projections and VARs, see Plagborg-Møller and Wolf (2021).

6.1 An illustrative example

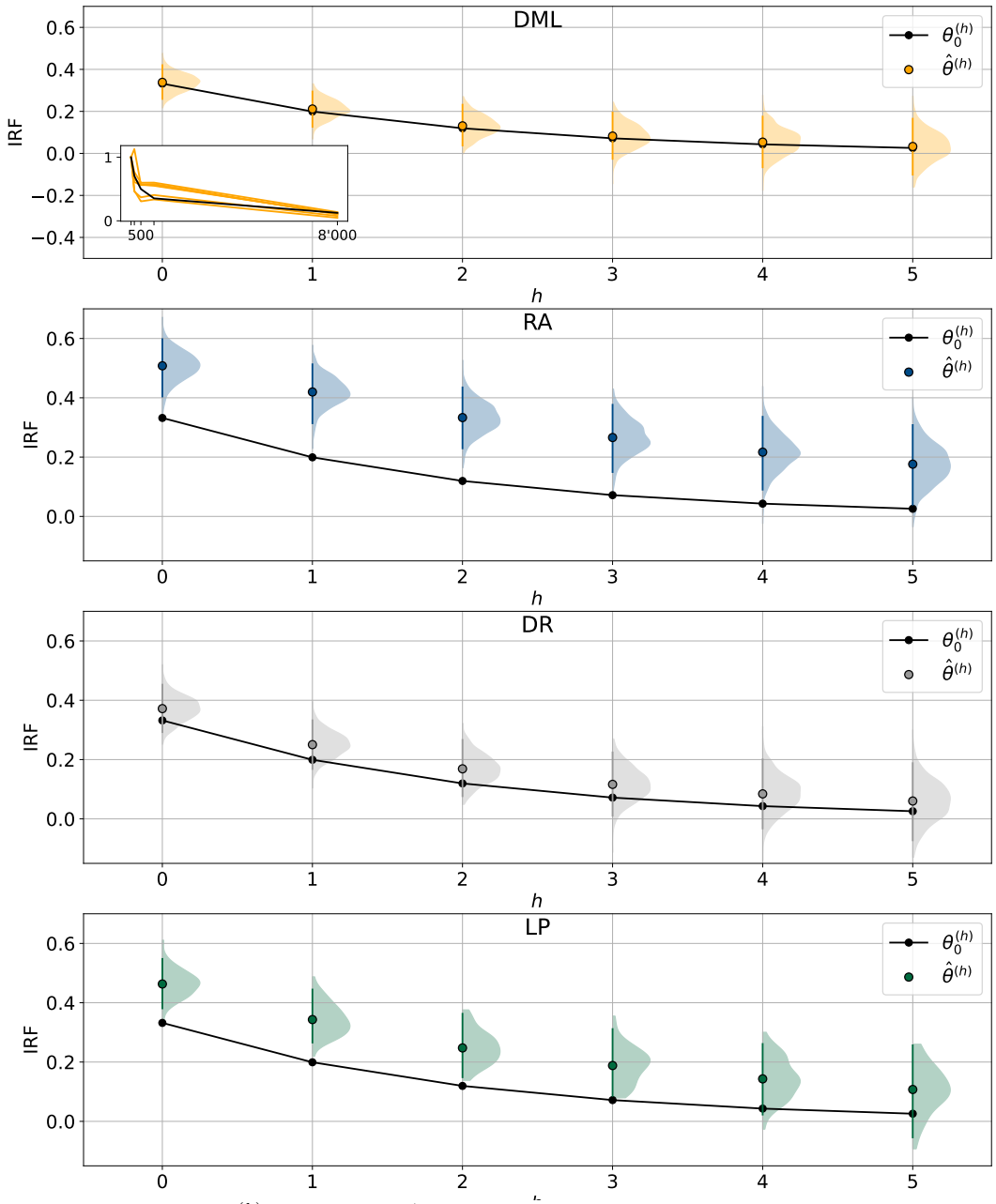
Consider the following stochastic processes

$$\begin{aligned} Y_{t+h} &= f^{(h)}(D_t, X_t) + \epsilon_{t+h} \\ D_t &= q(X_t, \eta_t), \end{aligned} \tag{4}$$

where ϵ_{t+h} and η_t are *i.i.d.* noise terms with zero mean and finite variance. The functions $f^{(h)}(\cdot)$ and $q(\cdot)$ are measurable and possibly nonlinear. Let $X_t = (V_t, V_{t-1})'$ be a two-dimensional vector with $V_t = \phi V_{t-1} + u_t$, where $|\phi| < 1$ and u_t is an *i.i.d.*, mean zero random variable with finite variance. The quantity of interest is the impulse response function at horizon h, i.e. $\theta_0^{(h)} = \mathbb{E}[\tau^{(h)}(X_t)]$ with $\tau^{(h)}(X_t) = \mathbb{E}[f^{(h)}(1, X_t)|X_t] - \mathbb{E}[f^{(h)}(0, X_t)|X_t]$. In the following, we illustrate how assumptions of local projection and DML estimators are satisfied in the example process (4).

[4] Simulation results are qualitatively unchanged when the bandwidth is determined alternatively using the rule of thumb by Wooldridge (2016) or Lazarus, Lewis, Stock, and Watson (2018). Results are available from the authors upon request.

Figure 2: Distribution of impulse response function estimates for a baseline nonlinear DGP with $n = 12$, $\sigma_\epsilon = 1.0$ and random forest nuisance function estimates



NOTE: Comparison of the true $\theta_0^{(h)}$ with estimates $\hat{\theta}^{(h)}$ of the IRF obtained for the setting with one stochastic process of length $T = 8'000$ from Table 1. Except for the LP estimator, nuisance functions are estimated with random forests. For the DML estimator, we use 10-fold cross-fitting and set $k_T = T/10$. The parameters of the data generating process are $n = 12$, $\sigma_\epsilon = 1.0$, $\gamma = 0.6$, $p = 2$, $q = 1$, $\sigma_u = 1.0$, $\alpha_A = 0.3$, $\alpha_M = 0.3$, $\rho_A = 0.35$, $\rho_M = 0.7$, $\beta_1 = 0.3$, $\beta_2 = 0.5$. For individual h, we show kernel density estimates of the distribution of $\hat{\theta}^{(h)}$ across $N = 1'000$ realizations. The dots indicate the average, and the vertical lines the (2.5%, 97.5%)-quantile range of the distribution. In the inset of the top panel, for the DML estimator, normalized biases as a function of T are plotted for all forecast horizons h and contrasted to the $\sqrt{T}$ scaling implied by Theorem 3 in black.

Table 1: Simulation results for a baseline nonlinear DGP with $n = 12$, $\sigma_\epsilon = 1.0$ and random forest nuisance function estimates

$h = 0$, $\theta_0^{(h)} = 0.3321$

	DML					RA				DR				LP			
T	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)	C_a(95%)	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)
125	0.053	0.797	0.799	0.945	0.944	0.360	0.455	0.580	0.511	0.263	0.413	0.490	0.707	0.090	0.357	0.368	0.871
250	0.060	0.398	0.402	0.949	0.941	0.285	0.300	0.413	0.514	0.208	0.278	0.347	0.732	0.128	0.251	0.282	0.847
500	0.030	0.220	0.222	0.954	0.947	0.219	0.196	0.294	0.464	0.142	0.183	0.232	0.776	0.135	0.173	0.220	0.829
1'000	0.029	0.144	0.147	0.931	0.928	0.182	0.139	0.229	0.422	0.103	0.131	0.167	0.784	0.138	0.126	0.187	0.730
8'000	0.006	0.042	0.042	0.966	0.965	0.176	0.048	0.182	0.016	0.040	0.042	0.058	0.853	0.131	0.046	0.139	0.141

$h = 1$, $\theta_0^{(h)} = 0.1992$

	DML					RA				DR				LP			
T	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)	C_a(95%)	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)
125	0.088	0.857	0.861	0.930	0.920	0.379	0.446	0.586	0.480	0.275	0.407	0.491	0.708	0.060	0.362	0.367	0.923
250	0.067	0.496	0.500	0.947	0.940	0.312	0.301	0.434	0.394	0.222	0.277	0.355	0.692	0.111	0.265	0.287	0.914
500	0.053	0.236	0.242	0.955	0.952	0.250	0.196	0.317	0.321	0.161	0.183	0.243	0.731	0.123	0.193	0.229	0.887
1'000	0.053	0.144	0.153	0.942	0.941	0.218	0.135	0.257	0.226	0.126	0.127	0.179	0.708	0.139	0.133	0.193	0.792
8'000	0.012	0.044	0.045	0.955	0.954	0.220	0.049	0.226	0.001	0.051	0.042	0.066	0.759	0.144	0.053	0.153	0.183

$h = 2$, $\theta_0^{(h)} = 0.1195$

	DML					RA				DR				LP			
T	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)	C_a(95%)	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)
125	0.107	0.815	0.822	0.925	0.916	0.353	0.455	0.576	0.478	0.256	0.424	0.495	0.748	0.031	0.410	0.411	0.931
250	0.065	0.517	0.521	0.932	0.932	0.278	0.313	0.419	0.425	0.197	0.296	0.356	0.731	0.078	0.306	0.316	0.927
500	0.060	0.259	0.266	0.946	0.942	0.232	0.212	0.314	0.346	0.151	0.201	0.252	0.737	0.104	0.221	0.244	0.911
1'000	0.060	0.167	0.178	0.918	0.917	0.210	0.153	0.260	0.220	0.125	0.146	0.193	0.717	0.128	0.160	0.204	0.854
8'000	0.011	0.050	0.051	0.951	0.951	0.214	0.052	0.220	0.001	0.049	0.048	0.069	0.793	0.128	0.055	0.140	0.352

$h = 3$, $\theta_0^{(h)} = 0.0717$

	DML					RA				DR				LP			
T	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)	C_a(95%)	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)
125	0.105	0.899	0.905	0.934	0.926	0.305	0.482	0.571	0.483	0.225	0.452	0.505	0.767	0.012	0.448	0.448	0.937
250	0.069	0.527	0.532	0.920	0.916	0.248	0.339	0.420	0.467	0.175	0.322	0.366	0.766	0.058	0.340	0.345	0.934
500	0.060	0.287	0.294	0.935	0.931	0.208	0.226	0.307	0.362	0.140	0.220	0.260	0.774	0.087	0.249	0.264	0.915
1'000	0.061	0.185	0.195	0.921	0.919	0.193	0.164	0.253	0.242	0.118	0.160	0.199	0.765	0.112	0.178	0.210	0.902
8'000	0.011	0.057	0.058	0.940	0.940	0.194	0.058	0.203	0.003	0.045	0.055	0.071	0.831	0.117	0.061	0.131	0.479

$h = 4$, $\theta_0^{(h)} = 0.0430$

	DML					RA				DR				LP			
T	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)	C_a(95%)	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)
125	0.138	0.949	0.959	0.929	0.908	0.270	0.500	0.569	0.538	0.205	0.475	0.517	0.773	0.021	0.496	0.497	0.935
250	0.064	0.671	0.674	0.910	0.907	0.217	0.369	0.428	0.462	0.159	0.352	0.387	0.764	0.049	0.377	0.380	0.922
500	0.051	0.322	0.327	0.940	0.926	0.177	0.246	0.304	0.390	0.119	0.239	0.267	0.806	0.062	0.270	0.277	0.929
1'000	0.056	0.201	0.209	0.938	0.936	0.170	0.174	0.244	0.280	0.106	0.170	0.201	0.803	0.092	0.192	0.213	0.921
8'000	0.011	0.063	0.064	0.946	0.946	0.174	0.061	0.184	0.007	0.041	0.061	0.074	0.874	0.100	0.067	0.121	0.662

$h = 5$, $\theta_0^{(h)} = 0.0258$

	DML					RA				DR				LP			
T	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)	C_a(95%)	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)
125	0.147	1.015	1.025	0.919	0.900	0.238	0.539	0.589	0.534	0.190	0.515	0.549	0.782	0.032	0.551	0.552	0.921
250	0.069	0.647	0.651	0.913	0.910	0.185	0.381	0.423	0.485	0.135	0.367	0.392	0.755	0.036	0.394	0.395	0.918
500	0.045	0.340	0.343	0.936	0.934	0.150	0.256	0.297	0.454	0.103	0.251	0.271	0.805	0.047	0.283	0.287	0.943
1'000	0.049	0.213	0.219	0.935	0.933	0.147	0.183	0.234	0.337	0.092	0.179	0.202	0.845	0.071	0.202	0.214	0.933
8'000	0.007	0.070	0.070	0.948	0.947	0.150	0.067	0.165	0.017	0.034	0.067	0.075	0.899	0.082	0.078	0.113	0.789

NOTE: The table depicts simulation results across $N = 1'000$ draws obtained for the setting with one stochastic process. Except for the LP estimator, nuisance functions are estimated with random forests. For the DML estimator, we use 10-fold cross-fitting and set $k_T = T/10$. For sample size $T = 125$, probabilities are winsorized at 1%. The parameters of the data generating process are $n = 12$, $\sigma_\epsilon = 1.0$, $\gamma = 0.6$, $p = 2$, $q = 1$, $\sigma_u = 1.0$, $\alpha_A = 0.3$, $\alpha_M = 0.3$, $\rho_A = 0.35$, $\rho_M = 0.7$, $\beta_1 = 0.3$, $\beta_2 = 0.5$. $C_a(\cdot)$ and $C_b(\cdot)$ in the tables denote the coverage at the given confidence level using asymptotic and fixed-bandwidth critical values respectively.

Local projection estimator

Local projections estimate the impulse response function via the coefficient $\hat{\beta}^{(h)}$ in the linear regression model

$$Y_{t+h} = \hat{\beta}^{(h)} D_t + \hat{\alpha}' X_t + \hat{r}_{t+h}.$$

By the projection theorem, the population coefficient $\beta_0^{(h)}$ is given by

$$\beta_0^{(h)} = \frac{\mathbb{E}[f^{(h)}(D_t, X_t)(D_t - \lambda_0' X_t])}{\mathbb{E}[(D_t - \lambda_0' X_t)^2]},$$

where $\lambda_0 = \arg\min \mathbb{E}[(D_t - \lambda' X_t)^2]$. Asymptotically, under certain regularity conditions on the time dependence and moments of the stochastic processes $S^{(h)}$, we have $\sqrt{T}(\hat{\beta}^{(h)} - \beta_0^{(h)}) \xrightarrow{d} \mathcal{N}(0, \mathrm{Var}(\beta_0^{(h)}))$. To ensure this result, D_t and X_t must be assumed to be stationary and ergodic for second moments, imposing restrictions on their time dependence, similar to – although generally weaker then – those in Assumption 3 for the DML estimator. For the

process in (4), this assumption is indeed fulfilled, as D_t and X_t are stationary and geometrically strong mixing.[5] Furthermore, D_t and X_t are assumed to be uncorrelated with the innovation ϵ_{t+h}, which corresponds to Assumption 5 for the DML estimator. Lastly, similar to Assumptions 3 and 6, the first four moments of D_t, X_t, and ϵ_{t+h} need to be finite. Note that since the linear regression estimator converges at the $\sqrt{T}$-rate, Assumption 4 is satisfied, even though this is not required for $\hat{\beta}^{(h)}$ to converge to $\beta_0^{(h)}$.

DML estimator

Under Theorem 3, the DML estimator consistently recovers $\theta_0^{(h)}$ provided that Assumptions 1 and 3-6 hold. Assumption 1 is met by construction of the example process (4). Since V_t is geometrically strong mixing and all involved functions are measurable, it follows that $S^{(h)}$ and $g(Z_t^{(h)}, h; \Gamma_0)$ are also geometrically strong mixing (see Theorem 15.1 in Davidson, 2021) and thus Assumption 3 is satisfied. Assumption 4 requires that the L_2-norm of the estimation error of the nuisance function estimators converges to zero at least at rate $T^{1/4}$. For the example process (4) this requirement is met for example by random forests, which converge at least at rate $T^{1/3}$ (Davis & Nielsen, 2020). Neural networks would also satisfy this condition, provided that the functions $f^{(h)}(\cdot)$ and $q(\cdot)$ are sufficiently smooth (for more details, see Ma & Safikhani, 2022). Lasso can achieve an even faster rate of $T^{1/2}$ (Wong et al., 2020), provided that the nuisance functions can be well approximated by polynomials of the conditioning variables. Assumption 5 requires the set of covariates to be sufficiently rich, so that past information on $Z_t^{(h)}$ cannot predict Y_{t+h} and D_t. For the example process (4), this assumption holds directly if the covariates X_t include both V_t and V_{t-1}. This would still hold if additional covariates or lagged values of $Z_t^{(h)}$ were included in the estimation. Assumption 6 finally holds for any value of ψ in the setting of (4), since S is geometrically strong mixing. Therefore, it is also sufficient that the residuals and estimation errors possess finite $4 + \nu$ moments, for some $\nu > 0$.

To conduct inference, Assumption 7 additionally requires choosing a bandwidth m_T depending on the convergence rate of the nuisance function estimators. For the example process (4), random forests estimators would permit m_T to be $o(T^{1/3})$ (e.g. Newey & West, 1994). Using Lasso estimators, m_T can be $o(T^{1/2})$ (e.g. Lazarus et al., 2018).

6.2 Linear and nonlinear processes

In case the function $f^{(h)}(D_t, X_t)$ is linear, then it can be shown that $\beta_0^{(h)} = \theta_0^{(h)}$ and thus linear projections recover the true impulse response function. However, if $f^{(h)}(D_t, X_t)$ is a nonlinear function, then the coefficients $\beta_0^{(h)}$ will no longer recover the true impulse response function, but an expected, weighted average of marginal effects $\tau^{(h)}(x)$. For an extensive discussion and derivation of these results, refer to Kolesár and Plagborg-Møller (2025).

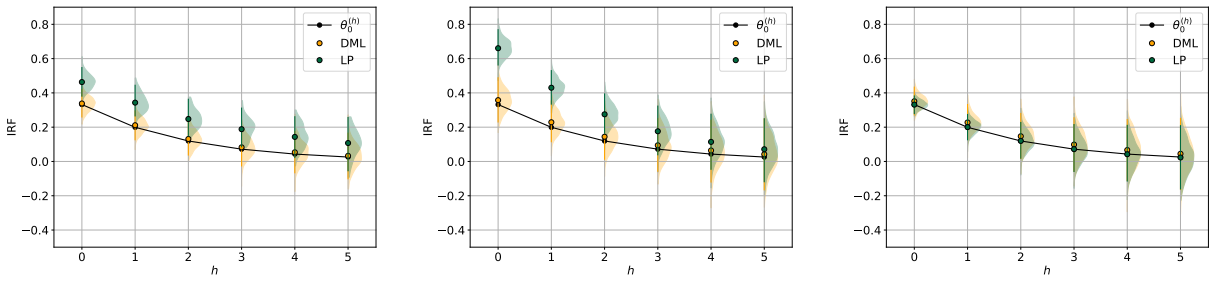
These effects are illustrated in Figure 3, which compares the distribution of impulse response function estimates obtained with DML and local projection estimators for three different DGPs: a nonlinear process, a linear process with interactions, and a purely linear process. Detailed results are deferred to Tables C7 and C8 in the Online Appendix. As already seen in Section 5 and again shown in Figure 3a, estimation using local projections is biased on a nonlinear process, whereas the DML estimator recovers $\theta_0^{(h)}$. This also holds in a setting where the outcome variable is generated from a linear function, but there is interaction between X_t and D_t (see Figure 3b). Local projections only recover $\theta_0^{(h)}$ for a purely linear process where its functional form is correctly specified (see Figure 3c). In this case, local projections have lower finite sample bias and variance than the DML estimator.

7 Empirical Application

As an illustration, we apply the proposed methodology to the empirical study conducted in Angrist et al. (2018). Based on the same data set, we revisit the estimation of the effect of U.S. monetary policy decisions on macroeconomic aggregates using modern machine learning estimators. Our monthly observations cover the period from July 1989 to December 2008 and we estimate the effect of federal funds target rate changes on a set of macroeconomic outcome variables. As predictors, we consider the same futures-based expectation measure for the federal funds rate as in Angrist et al. (2018), as well as the level of the target rate at the end of the prior month and its change, a scale factor that accounts for when within the month the Federal Reserve's Open Market Committee (FOMC) meeting was scheduled, dummies for months with a scheduled FOMC meeting, as well as measures for inflation and unemployment

[5] Since V_t is a linear process, it can be shown to be geometrically strong mixing with coefficients $\alpha(s) = O(|\phi|^s)$ (see Theorem 15.9 in Davidson, 2021).

Figure 3: Comparison of the distribution of DML and LP impulse response function estimates for nonlinear and linear DGPs



(a) Nonlinear DGP (b) Linear DGP with interactions (c) Linear DGP

NOTE: The figure compares the true $\theta_0^{(h)}$ with estimates $\hat{\theta}^{(h)}$ of the IRF obtained for the setting with one stochastic process of length $T = 8'000$ generated from three different DGPs: the nonlinear DGP studied in Section 5 (Figure 3a), a linear DGP with interaction terms where $b(X_t) = 0.5 \sum_{i=1}^{5} X_{i,t}$ and $\tau(X_t) = \theta_0^{(0)} + \sum_{i=1}^{3} X_{i,t} - \sum_{i=4}^{5} X_{i,t}$ (Figure 3b), and a linear DGP with $b(X_t) = 0.5 \sum_{i=1}^{5} X_{i,t}$ and $\tau(X_t) = \theta_0^{(0)}$ (Figure 3c). For the DML estimator, nuisance functions are estimated with random forests, we use 10-fold cross-fitting and set $k_T = T/10$. The parameters of the data generating processes are $n = 12$, $\sigma_\epsilon = 1.0$, $\gamma = 0.6$, $p = 2$, $q = 1$, $\sigma_u = 1.0$, $\alpha_A = 0.3$, $\alpha_M = 0.3$, $\rho_A = 0.35$, $\rho_M = 0.7$, $\beta_1 = 0.3$, $\beta_2 = 0.5$. For individual h, we show kernel density estimates of the distribution of $\hat{\theta}^{(h)}$ across $N = 1'000$ realizations. The dots indicate the average, and the vertical lines the (2.5%, 97.5%)-quantile range of the distribution.

(including lagged values).[6] Compared to the setup in Angrist et al. (2018), we make the following modifications to accommodate for the change in estimation technique from linear models to flexible nonparametric machine learners. First, we exclude dummies for monthly fixed effects and special events like Y2K and the September 11, 2001 attacks, because IRF estimates are based on out-of-sample predictions in our approach. Second, we drop manually constructed interaction variables, as machine learning estimators are able to infer these effects from the data, if they are present. Third, we include up to four lags of inflation, unemployment rate and of the target variable. Our treatment variable D_t can assume one of five discrete values $d \in \{-0.5\%, -0.25\%, 0.0\%, 0.25\%, 0.5\%\}$. The propensity score model for $e_0(d, x) = \Pr(D_t = d | X_t = x)$ is implemented as an ordinal classification (Frank & Hall, 2001). To estimate the conditional mean nuisance functions $\mu_0(d, x, h)$, we include the treatments as dummy variables in X_t and estimate/tune one joint model for all types of treatments (an S-learner in the terminology of Künzel et al. (2019)). As in our simulation experiments in Section 5, we explore random forests and a gradient boosted trees algorithm as estimators for the nuisance functions. Details on tuning of the machine learners are provided in the Online Appendix A.2. We apply our cross-fitting approach with $K = 10$. In line with Angrist et al. (2018), we estimate impulse responses up to $H = 24$ months and given the limited sample size set $k_T = 24$ to retain as much data as possible for estimation. As advocated in Section 4.1, we remove the k_T observations from the data used to estimate the machine learners.

We present estimated IRFs for the federal funds rate and the unemployment rate in Figure 4. Additionally, estimates for the effects on the bond yield curve are provided in Figure C5 in the Online Appendix. Predictive performance of both types of machine learners explored are comparable, but random forests appear to produce slightly smoother impulse responses, which, similar to standard local projection estimation (Barnichon & Brownlees, 2019), is advantageous. We thus focus on results using random forests and for 25 basis point changes of the target rate.[7] Overall, we identify similar dynamics in the outcome variables as in Angrist et al. (2018). However, for the federal funds rate, we find a larger absolute effect for target rate decreases of around 50 basis points that remains significant at the 5% level until around one year after the target rate decrease. In comparison, the peak effect of a target rate increase is around one percentage point, occurring one year after the increase, though it is accompanied by greater uncertainty. Furthermore, we do not find that either expansionary or tightening monetary policy has a significant effect on the unemployment rate. Looking at the effects on the yield curve, in line with Angrist et al. (2018), changes in the federal funds target rate have a higher initial impact on short-term yields than on long-term yields, as expected. Moreover, significant effects are observed only for shorter tenures. In contrast to Angrist et al. (2018), however, our estimates do not suggest that term rates are less sensitive to policy accommodation than to tightening. Finally, the empirical application also provides evidence that even in settings with limited sample sizes commonly encountered in macroeconomic studies, sufficiently accurate estimates of the nonparametric nuisance functions can be obtained in order to produce IRF

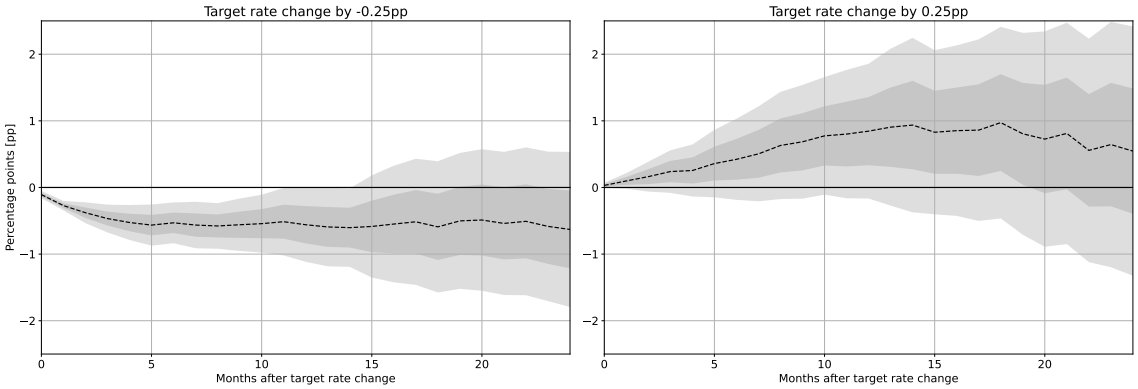
[6] This corresponds to the model specification labelled OP_{F2} in Angrist et al. (2018).

[7] Additional results are available from the authors on request.

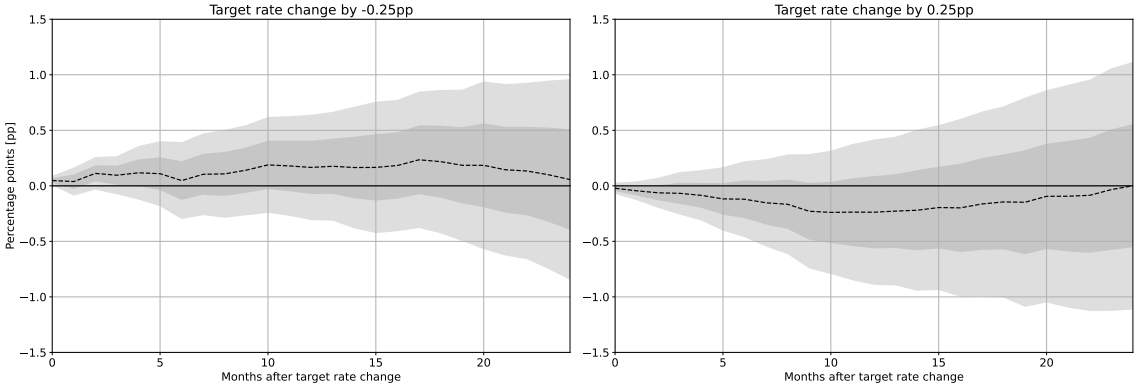
estimates comparable to ones obtained with conventional techniques, without having to construct all of the (interaction) variables and nonlinearities manually.

Figure 4: Estimated cumulative effects of target rate changes on the federal funds rate and the unemployment rate

(a) Federal Funds Rate



(b) Unemployment Rate



NOTE: The figure shows the estimated cumulative effects of target rate changes on (a) the federal funds rate and (b) the unemployment rate for the time period July 1989 to December 2008. The left (right) column shows the effect of decreasing (increasing) the target rate by 25 basis points. The nuisance functions are estimated by random forests using 10-fold cross-fitting removing $k_T = 24$ observations from the estimation sample at the boundary to the inference sample. The shaded areas represent 68% and 95% confidence intervals with fixed-bandwidth critical values (Kiefer & Vogelsang, 2005). The variances are estimated using bandwidths determined by the procedure of Newey and West (1994).

8 Conclusion

We have shown how to adopt recent ideas from the causal inference framework to flexibly estimate IRFs. This presents a novel estimator that can rely on fully nonparametric relations between treatment and outcome variables, opening up the possibility to use flexible machine learning approaches to estimate IRFs. Our theoretical results outline conditions for this estimator to be consistent and asymptotically normally distributed at the parametric rate. Simulations where a highly nonlinear time series is treated over time corroborate these results. Alternative estimators often used in practice estimate the IRF with a larger bias and fail to allow valid inference. Finally, we have illustrated the proposed methodology empirically by applying it to the estimation of the effects of macroeconomic shocks, allowing us to estimate IRFs of U.S. monetary policy decisions on macroeconomic aggregates using modern machine learning estimators. For future work, several semiparametric techniques available in the *i.i.d.* setting could be extended to time series settings in order to develop our approach further. This includes approaches for continuous treatments (Colangelo & Lee, 2025), instrumental variables (Chernozhukov et al., 2018), the estimation of other moments (Chernozhukov, Newey, & Singh, 2022a, 2022b), or conditional treatment effects (Kennedy, 2023; Qingliang Fan & Zhang, 2022; Semenova & Chernozhukov, 2020; Zimmert & Lechner, 2019) to estimate generalized IRFs. In general, future research could extend

the theoretical results developed in this paper to a broader class of estimands relying on linear scores (Chernozhukov et al., 2018).

Proofs

Proof of Theorem 1. We have that

$$\mathbb{E}\left[Y_{t+h}(d)\right] = \mathbb{E}\left[\mathbb{E}\left[Y_{t+h}(d)|D_t = d, X_t\right]\right] = \mathbb{E}\left[\mathbb{E}\left[Y_{t+h}|D_t = d, X_t\right]\right] = \mathbb{E}\left[\mu_0(d, X_t, h)\right],$$

where the second equality follows from Assumption 2.1 and the last equality from Assumption 2.2. It thus follows that $\mathbb{E}\left[Y_{t+h}(1) - Y_{t+h}(0)\right] = \mathbb{E}\left[\mu_0(1, X_t, h) - \mu_0(0, X_t, h)\right]$. □

Proof of Theorem 2. Under Assumption 3 the stochastic process $\mathcal{G}^{(h)}$ satisfies the assumptions for the Central Limit Theorem for α-mixing processes (Herrndorf, 1984). □

Proof of Theorem 3. The proof follows a similar strategy to the one in Wager (2022) and Chernozhukov et al. (2018) for the *i.i.d.* case. For the sake of legibility, we ease the notation and drop the reference to the forecast horizon h. Let $\tilde{\theta}$ be the oracle IRF estimator as defined in Theorem 2. Denote by $\mathcal{E}_T$ the event that $(\hat{\mu}_{S_{-i}}(d, x, h), \hat{e}_{S_{-i}}(x)) \in \Xi_T$ for all $i = 1, ..., K$. We have that

$$\begin{aligned}
\sqrt{T}(\hat{\theta} - \theta_0) &= \sqrt{T}(\hat{\theta} - \tilde{\theta}) + \sqrt{T}(\tilde{\theta} - \theta_0) \\
&= \sqrt{T}\left(\sum_{i=1}^{K} \frac{|\mathcal{T}_i|}{T}\hat{\theta}_{S_i} - \tilde{\theta}\right) + \sqrt{T}(\tilde{\theta} - \theta_0) \\
&= \sum_{i=1}^{K} \sqrt{T}\left(\frac{|\mathcal{T}_i|}{T}\hat{\theta}_{S_i} - \frac{|\mathcal{T}_i|}{T}\tilde{\theta}_{S_i}\right) + \sqrt{T}(\tilde{\theta} - \theta_0),
\end{aligned}$$

where $\tilde{\theta}_{S_i} = |\mathcal{T}_i|^{-1} \sum_{t \in \mathcal{T}_i} g(Z_t; \Gamma_0)$. We have to show that the summation converges to zero in probability. Note that since K is a finite integer, it suffices to show convergence for one summand. We begin by expanding a summand for some arbitrary i as

$$\begin{aligned}
\frac{|\mathcal{T}_i|}{T}\hat{\theta}_{S_i} - \frac{|\mathcal{T}_i|}{T}\tilde{\theta}_{S_i} &= \frac{1}{|\mathcal{T}_i|}\sum_{t \in \mathcal{T}_i} g(Z_t; \hat{\Gamma}_{S_{-i}}) - g(Z_t; \Gamma_0) \\
&= \frac{1}{|\mathcal{T}_i|}\sum_{t \in \mathcal{T}_i}\Bigg(\hat{\mu}_{S_{-i}}(1, X_t) - \mu_0(1, X_t) + \frac{D_t}{\hat{e}_{S_{-i}}(X_t)}(Y_t - \hat{\mu}_{S_{-i}}(1, X_t)) \\
&\qquad - \frac{D_t}{e_0(X_t)}(Y_t - \mu_0(1, X_t))\Bigg) \\
&\quad - \frac{1}{|\mathcal{T}_i|}\sum_{t \in \mathcal{T}_i}\Bigg(\hat{\mu}_{S_{-i}}(0, X_t) - \mu_0(0, X_t) + \frac{1 - D_t}{1 - \hat{e}_{S_{-i}}(X_t)}(Y_t - \hat{\mu}_{S_{-i}}(0, X_t)) \\
&\qquad - \frac{1 - D_t}{1 - e_0(X_t)}(Y_t - \mu_0(0, X_t))\Bigg).
\end{aligned}$$

We will prove convergence for the first summation, the second summation can be treated analogously. The first summation can be decomposed as

$$\begin{aligned}
&\frac{1}{|\mathcal{T}_i|}\sum_{t \in \mathcal{T}_i}\left(\hat{\mu}_{S_{-i}}(1, X_t) - \mu_0(1, X_t) + \frac{D_t}{\hat{e}_{S_{-i}}(X_t)}(Y_t - \hat{\mu}_{S_{-i}}(1, X_t)) - \frac{D_t}{e_0(X_t)}(Y_t - \mu_0(1, X_t))\right) \\
&\quad = \underbrace{\frac{1}{|\mathcal{T}_i|}\sum_{t \in \mathcal{T}_i}\left(\hat{\mu}_{S_{-i}}(1, X_t) - \mu_0(1, X_t)\right)\left(1 - \frac{D_t}{e_0(X_t)}\right)}_{=P_1} \\
&\qquad + \underbrace{\frac{1}{|\mathcal{T}_i|}\sum_{t \in \mathcal{T}_i} D_t\left(Y_t - \mu_0(1, X_t)\right)\left(\frac{1}{\hat{e}_{S_{-i}}(X_t)} - \frac{1}{e_0(X_t)}\right)}_{=P_2} \\
&\qquad + \underbrace{\frac{1}{|\mathcal{T}_i|}\sum_{t \in \mathcal{T}_i} D_t\left(\hat{\mu}_{S_{-i}}(1, X_t) - \mu_0(1, X_t)\right)\left(\frac{1}{\hat{e}_{S_{-i}}(X_t)} - \frac{1}{e_0(X_t)}\right)}_{=P_3}.
\end{aligned}$$

We will show that $P_k = o_p(T^{-1/2})$ for $k \in \{1, 2, 3\}$.

Term P_1: From the squared L_2-norm of P_1 we have that

$$\begin{aligned}
&\mathbb{E}\left[\left(\frac{1}{|\mathcal{T}_i|}\sum_{t\in\mathcal{T}_i}\left(\hat{\mu}_{S_{-i}}(1,X_t)-\mu_0(1,X_t)\right)\left(1-\frac{D_t}{e_0(X_t)}\right)\right)^2\right] \\
&\quad= \frac{1}{|\mathcal{T}_i|^2}\mathbb{E}\left[\mathbb{E}\left[\left(\sum_{t\in\mathcal{T}_i}\left(\hat{\mu}_{S_{-i}}(1,X_t)-\mu_0(1,X_t)\right)\left(1-\frac{D_t}{e_0(X_t)}\right)\right)^2\middle| S_{-i}\right]\right] \\
&\quad= \frac{1}{|\mathcal{T}_i|^2}\mathbb{E}\Bigg[\sum_{t\in\mathcal{T}_i}\sum_{s\in\mathcal{T}_i}\mathbb{E}\Bigg[\left(\hat{\mu}_{S_{-i}}(1,X_t)-\mu_0(1,X_t)\right)\left(1-\frac{D_t}{e_0(X_t)}\right)\times \\
&\qquad\qquad\left(\hat{\mu}_{S_{-i}}(1,X_s)-\mu_0(1,X_s)\right)\left(1-\frac{D_s}{e_0(X_s)}\right)\Bigg| S_{-i}\Bigg]\Bigg] \\
&\quad= \frac{1}{|\mathcal{T}_i|^2}\sum_{t\in\mathcal{T}_i}\mathbb{E}\left[\left(\hat{\mu}_{S_{-i}}(1,X_t)-\mu_0(1,X_t)\right)^2\left(\frac{1}{e_0(X_t)}-1\right)\right] \\
&\quad\leq \frac{1}{\eta|\mathcal{T}_i|^2}\sum_{t\in\mathcal{T}_i}\mathbb{E}\left[\left(\hat{\mu}_{S_{-i}}(1,X_t)-\mu_0(1,X_t)\right)^2\right] = \frac{o_p(1)}{|\mathcal{T}_i|} = o_p(T^{-1}).
\end{aligned}$$

The third equality follows from the fact that the sum has mean zero and the inequality at the end follows from Assumption 4.4. The step from the second to the third equality follows from Assumption 5, which gives for $t < s$ that

$$\begin{aligned}
&\mathbb{E}\left[\left(\hat{\mu}_{S_{-i}}(1,X_t)-\mu_0(1,X_t)\right)\left(1-\frac{D_t}{e_0(X_t)}\right)\left(\hat{\mu}_{S_{-i}}(1,X_s)-\mu_0(1,X_s)\right)\left(1-\frac{D_s}{e_0(X_s)}\right)\middle| S_{-i}\right] \\
&\quad= \mathbb{E}\Bigg[\left(\hat{\mu}_{S_{-i}}(1,X_t)-\mu_0(1,X_t)\right)\left(1-\frac{D_t}{e_0(X_t)}\right)\left(\hat{\mu}_{S_{-i}}(1,X_s)-\mu_0(1,X_s)\right)\times \\
&\qquad\underbrace{\mathbb{E}\left[1-\frac{D_s}{e_0(X_s)}\middle| X_s, \{(X_u, Y_u, D_u) : u \in \mathcal{T}_i, u < s)\}, S_{-i}\right]}_{=0}\Bigg| S_{-i}\Bigg] = 0.
\end{aligned}$$

The same argument can be made for $t > s$. Convergence in the last step finally follows from the fact that K is a fixed and finite integer and thus $\lim_{T\to\infty}|\mathcal{T}_i| = \lim_{T\to\infty} T/K = \infty$ for all i and by noting that conditional on S_{-i}, the nuisance function estimator is non-stochastic, and thus conditional on the event $\mathcal{E}_T$, we have that

$$\begin{aligned}
\sup_{t\in\mathcal{T}_i}\mathbb{E}\left[\left(\hat{\mu}_{S_{-i}}(1,X_t)-\mu_0(1,X_t)\right)^2\middle| S_{-i}\right] &\leq \sup_{t\in\mathcal{T}_i}\sup_{\mu\in\Xi_T}\mathbb{E}\left[\left(\mu(1,X_t)-\mu_0(1,X_t)\right)^2\middle| S_{-i}\right] \\
&\leq \sup_{t\in\mathcal{T}_i}\sup_{\mu\in\Xi_T}\|\mu(1,X_t)-\mu_0(1,X_t)\|_2^2 = o_p(1)
\end{aligned}$$

since by Assumption 4 $\sup_{t\in\mathcal{T}_i}\sup_{\mu\in\Xi_T}\|\mu(D_t,X_t)-\mu_0(D_t,X_t)\|_2^2 = (r_{\mu,T})^2 = o_p(1)$. By Lemma 6.1 in Chernozhukov et al. (2018) it follows that $\sup_{t\in\mathcal{T}_i}\mathbb{E}[(\hat{\mu}_{S_{-i}}(1,X_t)-\mu_0(1,X_t))^2] = o_p(1)$, and we thus conclude that P_1 is $o_p(T^{-1/2})$.

Term P_2: Similarly, from the squared L_2-norm of P_2 we have that

$$\begin{aligned}
&\mathbb{E}\left[\left(\frac{1}{|\mathcal{T}_i|}\sum_{t\in\mathcal{T}_i} D_t\left(Y_t-\mu_0(1,X_t)\right)\left(\frac{1}{\hat{e}_{S_{-i}}(X_t)}-\frac{1}{e_0(X_t)}\right)\right)^2\right] \\
&\quad=\frac{1}{|\mathcal{T}_i|^2}\mathbb{E}\left[\mathbb{E}\left[\left(\sum_{t\in\mathcal{T}_i} D_t\left(Y_t-\mu_0(1,X_t)\right)\left(\frac{1}{\hat{e}_{S_{-i}}(X_t)}-\frac{1}{e_0(X_t)}\right)\right)^2\middle| S_{-i}\right]\right] \\
&\quad=\frac{1}{|\mathcal{T}_i|^2}\mathbb{E}\left[\sum_{t\in\mathcal{T}_i}\sum_{s\in\mathcal{T}_i}\mathbb{E}\left[D_t\left(Y_t-\mu_0(1,X_t)\right)\left(\frac{1}{\hat{e}_{S_{-i}}(X_t)}-\frac{1}{e_0(X_t)}\right)\times\right.\right. \\
&\qquad\qquad\left.\left. D_s\left(Y_s-\mu_0(1,X_s)\right)\left(\frac{1}{\hat{e}_{S_{-i}}(X_s)}-\frac{1}{e_0(X_s)}\right)\middle| S_{-i}\right]\right] \\
&\quad=\frac{1}{|\mathcal{T}_i|^2}\sum_{t\in\mathcal{T}_i}\mathbb{E}\left[D_t\left(Y_t-\mu_0(1,X_t)\right)^2\left(\frac{1}{\hat{e}_{S_{-i}}(X_t)}-\frac{1}{e_0(X_t)}\right)^2\right] \\
&\quad\leq\frac{1}{|\mathcal{T}_i|^2}\frac{\epsilon_1^2}{\eta^2}\sum_{t\in\mathcal{T}_i}\mathbb{E}\left[\left(\hat{e}_{S_{-i}}(X_t)-e_0(X_t)\right)^2\right]=\frac{o_p(1)}{|\mathcal{T}_i|}=o_p(T^{-1}).
\end{aligned}$$

The third equality follows from the fact that the sum has mean zero, and the inequality follows from Assumptions 4.4 and 4.5. The crucial step is again to establish that the variance of the sum equals the sum of the variances (from the second to the third equality). This follows from Assumption 5, which gives that whenever $t<s$, it holds that

$$\begin{aligned}
&\mathbb{E}\left[D_t\left(Y_t-\mu_0(1,X_t)\right)\left(\frac{1}{\hat{e}_{S_{-i}}(X_t)}-\frac{1}{e_0(X_t)}\right)\times\right. \\
&\qquad\left. D_s\left(Y_s-\mu_0(1,X_s)\right)\left(\frac{1}{\hat{e}_{S_{-i}}(X_s)}-\frac{1}{e_0(X_s)}\right)\middle| S_{-i}\right] \\
&=\mathbb{E}\left[D_t\left(Y_t-\mu_0(1,X_t)\right)\left(\frac{1}{\hat{e}_{S_{-i}}(X_t)}-\frac{1}{e_0(X_t)}\right)\times\right. \\
&\qquad\underbrace{\mathbb{E}\left[D_s(Y_s-\mu_0(1,X_s))\middle| X_s,\{(X_u,Y_u,D_u):u\in\mathcal{T}_i,u<s\},S_{-i}\right]}_{=0}\left.\left(\frac{1}{\hat{e}_{S_{-i}}(X_s)}-\frac{1}{e_0(X_s)}\right)\middle| S_{-i}\right]=0,
\end{aligned}$$

and the same argument can be made for $t>s$. Convergence in the last step finally follows from the fact that K is a fixed and finite integer and thus $\lim_{T\to\infty}|\mathcal{T}_i|=\lim_{T\to\infty}T/K=\infty$ for all i and by noting that conditional on S_{-i}, the nuisance function estimator is non-stochastic, and thus conditional on the event $\mathcal{E}_T$, we have that

$$\begin{aligned}
\sup_{t\in\mathcal{T}_i}\mathbb{E}\left[\left(\hat{e}_{S_{-i}}(X_t)-e_0(X_t)\right)^2\middle| S_{-i}\right]&\leq\sup_{t\in\mathcal{T}_i}\sup_{e\in\Xi_T}\mathbb{E}\left[\left(e(X_t)-e_0(X_t)\right)^2\middle| S_{-i}\right] \\
&\leq\sup_{t\in\mathcal{T}_i}\sup_{e\in\Xi_T}\|e(X_t)-e_0(X_t)\|_2^2=(r_{e,T})^2
\end{aligned}$$

by definition of the rate $r_{e,T}$ in Assumption 4. By Lemma 6.1 in Chernozhukov et al. (2018) and Assumption 4 it follows that $\sup_{t\in\mathcal{T}_i}\mathbb{E}\left[\left(\hat{e}_{S_{-i}}(X_t)-e_0(X_t)\right)^2\right]=o_p(1)$, and we thus conclude that P_2 is $o_p(T^{-1/2})$.

Term P_3: Finally, from the L_1-norm of P_3 we get that

$$\begin{aligned}
&\mathbb{E}\left[\left|\frac{1}{|\mathcal{T}_i|}\sum_{t\in\mathcal{T}_i} D_t\left(\hat{\mu}_{S_{-i}}(1,X_t)-\mu_0(1,X_t)\right)\left(\frac{1}{\hat{e}_{S_{-i}}(X_t)}-\frac{1}{e_0(X_t)}\right)\right|\right] \\
&\quad\leq \mathbb{E}\left[\frac{1}{|\mathcal{T}_i|}\sum_{t\in\mathcal{T}_i} D_t\left|\hat{\mu}_{S_{-i}}(1,X_t)-\mu_0(1,X_t)\right|\left|\frac{1}{\hat{e}_{S_{-i}}(X_t)}-\frac{1}{e_0(X_t)}\right|\right] \\
&\quad= \frac{1}{|\mathcal{T}_i|}\sum_{t\in\mathcal{T}_i}\mathbb{E}\left[D_t\left|\hat{\mu}_{S_{-i}}(D_t,X_t)-\mu_0(D_t,X_t)\right|\left|\frac{1}{\hat{e}_{S_{-i}}(X_t)}-\frac{1}{e_0(X_t)}\right|\right] \\
&\quad\leq \frac{1}{\eta}\frac{1}{|\mathcal{T}_i|}\sum_{t\in\mathcal{T}_i}\mathbb{E}\left[\left|\hat{\mu}_{S_{-i}}(D_t,X_t)-\mu_0(D_t,X_t)\right|\left|\hat{e}_{S_{-i}}(X_t)-e_0(X_t)\right|\right] = \frac{o_p(1)}{T^{1/2}},
\end{aligned}$$

where the last inequality follows from Assumption 4.4. Convergence in the last equality finally follows from the fact that K is a fixed and finite integer and thus $\lim_{T\to\infty}|\mathcal{T}_i| = \lim_{T\to\infty} T/K = \infty$ for all i and by noting that conditional on S_{-i}, the nuisance function estimators are non-stochastic, and thus conditional on the event $\mathcal{E}_T$, we have that

$$\begin{aligned}
&\sup_{t\in\mathcal{T}_i}\mathbb{E}\left[\left|\hat{\mu}_{S_{-i}}(D_t,X_t)-\mu_0(D_t,X_t)\right|\left|\hat{e}_{S_{-i}}(X_t)-e_0(X_t)\right|\middle| S_{-i}\right] \\
&\leq \sup_{t\in\mathcal{T}_i}\mathbb{E}\left[\left|\hat{\mu}_{S_{-i}}(D_t,X_t)-\mu_0(D_t,X_t)\right|^2\middle| S_{-i}\right]^{1/2}\sup_{t\in\mathcal{T}_i}\mathbb{E}\left[\left|\hat{e}_{S_{-i}}(X_t)-e_0(X_t)\right|^2\middle| S_{-i}\right]^{1/2} \\
&\leq \sup_{t\in\mathcal{T}_i}\sup_{\mu\in\Xi_T}\mathbb{E}\left[\left|\mu(D_t,X_t)-\mu_0(D_t,X_t)\right|^2\middle| S_{-i}\right]^{1/2}\sup_{t\in\mathcal{T}_i}\sup_{e\in\Xi_T}\mathbb{E}\left[\left|e(X_t)-e_0(X_t)\right|^2\middle| S_{-i}\right]^{1/2} \\
&\leq \sup_{t\in\mathcal{T}_i}\sup_{\mu\in\Xi_T}\left\|\mu(D_t,X_t)-\mu_0(D_t,X_t)\right\|_2\sup_{t\in\mathcal{T}_i}\sup_{e\in\Xi_T}\left\|e(X_t)-e_0(X_t)\right\|_2 = r_{\mu,T}\cdot r_{e,T}
\end{aligned}$$

by Cauchy-Schwarz and the definition of the rates $r_{\mu,T}$ and $r_{e,T}$ in Assumption 4. By Lemma 6.1 in Chernozhukov et al. (2018) and Assumption 4.3 it follows that $\sup_{t\in\mathcal{T}_i}\mathbb{E}\left[\left|\hat{\mu}_{S_{-i}}(1,X_t)-\mu_0(1,X_t)\right|\left|\hat{e}_{S_{-i}}(X_t)-e_0(X_t)\right|\right] = o_p(T^{-1/2})$.

We have shown that P_1, P_2 and P_3 are $o_p(T^{-1/2})$. It follows that $\frac{|\mathcal{T}_i|}{T}\hat{\theta}_{S_i} - \frac{|\mathcal{T}_i|}{T}\tilde{\theta}_{S_i} = o_p(T^{-1/2})$ for all $i = 1, ..., K$ and we can thus conclude that

$$\sqrt{T}(\hat{\theta}-\theta) = \sum_{i=1}^{K}\underbrace{\sqrt{T}\left(\frac{|\mathcal{T}_i|}{T}\hat{\theta}_{S_i}-\frac{|\mathcal{T}_i|}{T}\tilde{\theta}_{S_i}\right)}_{=o_p(1)} + \underbrace{\sqrt{T}(\tilde{\theta}-\theta_0)}_{\xrightarrow{\mathrm{d}}\mathcal{N}(0,V_0)}.$$

□

Lemma 1. *Let Ξ be a convex subset of some normed vector space, $g:\mathbb{R}^n\times\Xi\to\mathbb{R}$ be a measurable function, $\{Z_t : t\in\mathcal{T}\}$ an α-mixing stochastic process with mixing coefficient $\alpha(m)$ and Z_t a real-valued random vector. For $s\geq t$ denote by $\mathcal{F}_t^s=\sigma(Z_t,\ldots,Z_s)$ the smallest σ-field such that $Z_t,\ldots,Z_s$ are measurable. If $\|\sup_{\Gamma\in\Xi} g(Z_t,\Gamma)\|_r<\infty$ for some $r\geq p\geq 1$, then for all $t\in\mathcal{T}$*

$$\sup_{\Gamma\in\Xi}\mathbb{E}\left[g(Z_t,\Gamma)|\mathcal{F}_{-\infty}^{t-k}\right] = \sup_{\Gamma\in\Xi}\mathbb{E}\left[g(Z_t,\Gamma)\right] + O_p(\alpha(k)^{1/p-1/r}).$$

Proof of Lemma 1. We will prove the statement by bounding the L_p-norm. First notice that

$$\left\|\sup_{\Gamma\in\Xi}\mathbb{E}\left[g(Z_t,\Gamma)|\mathcal{F}_{-\infty}^{t-k}\right]-\sup_{\Gamma\in\Xi}\mathbb{E}\left[g(Z_t,\Gamma)\right]\right\|_p \leq \left\|\mathbb{E}\left[\sup_{\Gamma\in\Xi}\left(g(Z_t,\Gamma)-\mathbb{E}\left[g(Z_t,\Gamma)\right]\right)|\mathcal{F}_{-\infty}^{t-k}\right]\right\|_p$$

By Theorem 15.2 in Davidson (2021) we have

$$\begin{aligned}
&\left\| \mathbb{E}\left[\sup_{\Gamma\in\Xi} \left(g(Z_t,\Gamma) - \mathbb{E}\left[g(Z_t,\Gamma)\right]\right) | \mathcal{F}_{-\infty}^{t-k} \right] \right\|_p \\
&\quad \leq 2(2^{1/2}+1)\alpha(k)^{1/p-1/r} \left\| \sup_{\Gamma\in\Xi} \left(g(Z_t,\Gamma) - \mathbb{E}\left[g(Z_t,\Gamma)\right]\right) \right\|_r \\
&\quad \leq 2(2^{1/2}+1)\alpha(k)^{1/p-1/r} \left(\left\| \sup_{\Gamma\in\Xi} g(Z_t,\Gamma) \right\|_r + \left| \sup_{\Gamma\in\Xi} \mathbb{E}\left[g(Z_t,\Gamma)\right] \right| \right) = O(\alpha(k)^{1/p-1/r})
\end{aligned}$$

where the second inequality follows from the Minkowski inequality. □

Proof of Theorem 4. The proof builds on the proof of Theorem 3 and we continue to omit the horizon h for the sake of legibility. Following Davidson (2021), let the smallest σ-field on which the stochastic process $S_{-i} = \{Z_t : t \in \mathcal{T}_{-i}\}$ is measurable, be denoted as $\mathcal{F}_{-i} = \sigma\left(Z_t : t \in \mathcal{T} \wedge (t < \inf(\mathcal{T}_i) - k_T \vee t > \sup(\mathcal{T}_i) + k_T)\right)$. Consider again the three summations P_1, P_2 and P_3 from the proof of Theorem 3.

Term P_1: From the squared L_2-norm of P_1 we to obtain

$$\begin{aligned}
&\mathbb{E}\left[\left(\frac{1}{|\mathcal{T}_i|} \sum_{t\in\mathcal{T}_i} \left(\hat{\mu}_{S_{-i}}(1,X_t) - \mu_0(1,X_t) \right) \left(1 - \frac{D_t}{e_0(X_t)} \right) \right)^2 \right] \\
&\quad = \frac{1}{|\mathcal{T}_i|^2} \sum_{t\in\mathcal{T}_i} \mathbb{E}\left[\left(\hat{\mu}_{S_{-i}}(1,X_t) - \mu_0(1,X_t) \right)^2 \left(1 - \frac{D_t}{e_0(X_t)} \right)^2 \right] \\
&\qquad + \frac{1}{|\mathcal{T}_i|^2} \sum_{t\in\mathcal{T}_i} \sum_{s\in\mathcal{T}_i, s\neq t} \mathbb{E}\Bigg[\left(\hat{\mu}_{S_{-i}}(1,X_t) - \mu_0(1,X_t) \right) \left(1 - \frac{D_t}{e_0(X_t)} \right) \times \\
&\qquad\qquad \left(\hat{\mu}_{S_{-i}}(1,X_s) - \mu_0(1,X_s) \right) \left(1 - \frac{D_s}{e_0(X_s)} \right) \Bigg].
\end{aligned} \tag{A1}$$

First, note that by applying Hölder's inequality twice we have that for $r > p \geq 1$ and $1/r = 1/r' + 1/r''$

$$\begin{aligned}
&\sup_{t,s\in\mathcal{T}_i} \sup_{\mu\in\Xi_T} \left\| (\mu(1,X_t) - \mu_0(1,X_t)) \left(1 - \frac{D_t}{e_0(X_t)} \right) (\mu(1,X_s) - \mu_0(1,X_s)) \left(1 - \frac{D_s}{e_0(X_s)} \right) \right\|_r \\
&\quad \leq \sup_{t\in\mathcal{T}_i} \sup_{\mu\in\Xi_T} \left\| (\mu(1,X_t) - \mu_0(1,X_t))^2 \left(1 - \frac{D_t}{e_0(X_t)} \right)^2 \right\|_r \\
&\quad \leq \sup_{t\in\mathcal{T}_i} \sup_{\mu\in\Xi_T} \left\| (\mu(1,X_t) - \mu_0(1,X_t))^2 \right\|_{r'} \left\| \left(1 - \frac{D_t}{e_0(X_t)} \right)^2 \right\|_{r''}
\end{aligned}$$

which is bounded by Assumption 6.3. Next, for the first summand in (A1), note that conditional on $\mathcal{F}_{-i}$ the estimator is non-stochastic, and thus conditional on the event $\mathcal{E}_T$ we have that

$$\begin{aligned}
&\sup_{t\in\mathcal{T}_i} \mathbb{E}\left[\left(\hat{\mu}_{S_{-i}}(1,X_t) - \mu_0(1,X_t) \right)^2 \left(1 - \frac{D_t}{e_0(X_t)} \right)^2 \middle| \mathcal{F}_{-i} \right] \\
&\quad \leq \sup_{t\in\mathcal{T}_i} \sup_{\mu\in\Xi_T} \mathbb{E}\left[(\mu(1,X_t) - \mu_0(1,X_t))^2 \left(1 - \frac{D_t}{e_0(X_t)} \right)^2 \middle| \mathcal{F}_{-i} \right] \\
&\quad \leq \sup_{t\in\mathcal{T}_i} \sup_{\mu\in\Xi_T} \mathbb{E}\left[(\mu(1,X_t) - \mu_0(1,X_t))^2 \left(1 - \frac{D_t}{e_0(X_t)} \right)^2 \right] + O_p\left(\alpha(k_T)^\psi \right) \\
&\quad = o_p(1) + O_p\left(\alpha(k_T)^\psi \right)
\end{aligned}$$

by Lemma 1 in combination with Assumption 6, and the definition of the rate $r_{\mu,T}$ in Assumption 4. By Lemma 6.1 in Chernozhukov et al. (2018) and Assumptions 4.3 and 6.5 it follows that

$$\frac{1}{|\mathcal{T}_i|^2} \sum_{t\in\mathcal{T}_i} \mathbb{E}\left[\left(\hat{\mu}_{S_{-i}}(1,X_t) - \mu_0(1,X_t) \right)^2 \left(1 - \frac{D_t}{e_0(X_t)} \right)^2 \right] = \frac{o_p(1)}{|\mathcal{T}_i|} = o_p(T^{-1}).$$

Similarly, for the second summand of the L_2-norm of P_1, conditional on the event $\mathcal{E}_T$ we have that

$$\begin{aligned}
&\sup_{t,s\in\mathcal{T}_i} \mathbb{E}\Bigg[\left(\hat{\mu}_{S_{-i}}(1,X_t) - \mu_0(1,X_t)\right)\left(1-\frac{D_t}{e_0(X_t)}\right)\times \\
&\qquad\qquad \left(\hat{\mu}_{S_{-i}}(1,X_s) - \mu_0(1,X_s)\right)\left(1-\frac{D_s}{e_0(X_s)}\right)\Bigg|\mathcal{F}_{-i}\Bigg] \\
&\leq \sup_{t,s\in\mathcal{T}_i}\sup_{e\in\Xi_T} \mathbb{E}\Bigg[\left(\mu(1,X_t) - \mu_0(1,X_t)\right)\left(1-\frac{D_t}{e_0(X_t)}\right)\times \\
&\qquad\qquad \left(\mu(1,X_s) - \mu_0(1,X_s)\right)\left(1-\frac{D_s}{e_0(X_s)}\right)\Bigg|\mathcal{F}_{-i}\Bigg] \\
&\leq \sup_{t,s\in\mathcal{T}_i}\sup_{e\in\Xi_T} \mathbb{E}\Bigg[\left(\mu(1,X_t) - \mu_0(1,X_t)\right)\left(1-\frac{D_t}{e_0(X_t)}\right)\times \\
&\qquad\qquad \left(\mu(1,X_s) - \mu_0(1,X_s)\right)\left(1-\frac{D_s}{e_0(X_s)}\right)\Bigg] + O_p\left(\alpha(k_T)^{\psi}\right)
\end{aligned}$$

by Lemma 1 in combination with Assumption 6, and the expectation in the last inequality is zero. By Lemma 6.1 in Chernozhukov et al. (2018) and Assumptions 4.3 and 6.5 it follows that

$$\begin{aligned}
&\frac{1}{|\mathcal{T}_i|^2}\sum_{t\in\mathcal{T}_i}\sum_{s\in\mathcal{T}_i, s\neq t} \mathbb{E}\Bigg[\left(\hat{\mu}_{S_{-i}}(1,X_t) - \mu_0(1,X_t)\right)\left(1-\frac{D_t}{e_0(X_t)}\right)\times \\
&\qquad\qquad \left(\hat{\mu}_{S_{-i}}(1,X_s) - \mu_0(1,X_s)\right)\left(1-\frac{D_s}{e_0(X_s)}\right)\Bigg] = O_p\left(\alpha(k_T)^{\psi}\right).
\end{aligned}$$

By Assumption 6.5 it follows that P_1 is $o_p(T^{-1/2})$. Following a similar argument, it can be shown that P_2 is $o_p(T^{-1/2})$.

Convergence for P_3 can again be shown by bounding its L_1-norm as

$$\begin{aligned}
&\mathbb{E}\left[\left|\frac{1}{|\mathcal{T}_i|}\sum_{t\in\mathcal{T}_i} D_t\left(\hat{\mu}_{S_{-i}}(1,X_t) - \mu_0(1,X_t)\right)\left(\frac{1}{\hat{e}_{S_{-i}}(X_t)} - \frac{1}{e_0(X_t)}\right)\right|\right] \\
&\quad = \frac{1}{|\mathcal{T}_i|}\mathbb{E}\left[\left|\sum_{t\in\mathcal{T}_i} D_t\left(\hat{\mu}_{S_{-i}}(D_t,X_t) - \mu_0(D_t,X_t)\right)\left(\frac{1}{\hat{e}_{S_{-i}}(X_t)} - \frac{1}{e_0(X_t)}\right)\right|\right] \\
&\quad \leq \frac{1}{\eta}\frac{1}{|\mathcal{T}_i|}\sum_{t\in\mathcal{T}_i}\mathbb{E}\left[\left|\hat{\mu}_{S_{-i}}(D_t,X_t) - \mu_0(D_t,X_t)\right|\left|\hat{e}_{S_{-i}}(X_t) - e_0(X_t)\right|\right].
\end{aligned}$$

Noting that conditional on $\mathcal{F}_{-i}$, the nuisance function estimators are non-stochastic, and thus conditional on the event $\mathcal{E}_T$ we have that

$$\begin{aligned}
&\sup_{t\in\mathcal{T}_i}\mathbb{E}\left[\left|\hat{\mu}_{S_{-i}}(D_t,X_t) - \mu_0(D_t,X_t)\right|\left|\hat{e}_{S_{-i}}(X_t) - e_0(X_t)\right|\Bigg|\mathcal{F}_{-i}\right] \\
&\quad \leq \sup_{t\in\mathcal{T}_i}\sup_{\mu,e\in\Xi_T}\mathbb{E}\left[\left|\mu(D_t,X_t) - \mu_0(D_t,X_t)\right|\left|e(X_t) - e_0(X_t)\right|\Bigg|\mathcal{F}_{-i}\right] \\
&\quad \leq \sup_{t\in\mathcal{T}_i}\sup_{\mu,e\in\Xi_T}\mathbb{E}\left[\left|\mu(D_t,X_t) - \mu_0(D_t,X_t)\right|\left|e(X_t) - e_0(X_t)\right|\right] + O_p(\alpha(k_T)^{\psi}) \\
&\quad \leq \sup_{t\in\mathcal{T}_i}\sup_{\mu\in\Xi_T}\left\|\mu(D_t,X_t) - \mu_0(D_t,X_t)\right\|_2 \sup_{t\in\mathcal{T}_i}\sup_{e\in\Xi_T}\left\|e(X_t) - e_0(X_t)\right\|_2 + O_p(\alpha(k_T)^{\psi}) \\
&\quad = r_{\mu,T}\cdot r_{e,T} + O_p(\alpha(k_T)^{\psi})
\end{aligned}$$

by Lemma 1 in combination with Assumption 6, Cauchy-Schwarz and the definition of the rates $r_{\mu,T}$ and $r_{e,T}$ in Assumption 4. By Lemma 6.1 in Chernozhukov et al. (2018) and Assumptions 4.3 and 6.5 it follows that

$\sup_{t\in\mathcal{T}_i} \mathbb{E}\left[\left|\hat{\mu}_{S_{-i}}(1,X_t)-\mu_0(1,X_t)\right|\left|\hat{e}_{S_{-i}}(X_t)-e(X_t)\right|\right] = o_p(T^{-1/2})$. This concludes the proof as we can now apply the same arguments as in the proof of Theorem 3. □

Proof of Theorem 5. Define $v_t(\theta,\Gamma) = g(Z_t;\Gamma)-\theta$, for some nuisance functions $\Gamma=(\mu,e)$, $g(Z_t;\Gamma)$ is the influence function (2) evaluated using the nuisance functions in Γ and we will drop the forecast horizon h everywhere for legibility. Let

$$V_T = \text{Var}\left(\frac{1}{\sqrt{T}}\sum_{t\in\mathcal{T}} g(Z_t;\Gamma_0)\right) = \frac{1}{T}\left(\sum_{t=1}^{T}\mathbb{E}[v_t^2] + 2\sum_{s=1}^{T-1}\sum_{t=s+1}^{T}\mathbb{E}[v_t v_{t-s}]\right)$$

with $v_t = g(Z_t;\Gamma_0)-\theta_0$ and $\Gamma_0=(\mu_0,e_0)$, and thus $\lim_{T\to\infty} V_T = V_0$. Next, define

$$V_{S_i} = \frac{1}{|\mathcal{T}_i|}\left(\sum_{t\in\mathcal{T}_i}\mathbb{E}[v_t^2] + 2\sum_{s\in\mathcal{T}_i}\sum_{t\in\mathcal{T}_{i,s}}\mathbb{E}[v_t v_{t-s}]\right)$$

with $\mathcal{T}_{i,s} = \{t\in\mathcal{T}_i \,|\, t-s\geq \inf(\mathcal{T}_i)\}$ for $i=1,...,K$. Then we have that $\lim_{T\to\infty} V_{S_i} = \lim_{T\to\infty} V_T = V_0$, so given K being a finite integer, for $|\widehat{V}-V_T| \xrightarrow{p} 0$ it will be sufficient to show that $|\widehat{V}_{S_i}-V_{S_i}| = o_p(1)$, where

$$\widehat{V}_{S_i} = \frac{1}{|\mathcal{T}_i|}\left(\sum_{t\in\mathcal{T}_i}\hat{v}_t^2 + 2\sum_{s=1}^{m_T} w(s,m_T)\sum_{t\in\mathcal{T}_{i,s}}\hat{v}_t\hat{v}_{t-s}\right)$$

with

$$\hat{v}_t = g(Z_t;\hat{\Gamma}_t)-\hat{\theta} \qquad \text{and} \qquad \hat{\Gamma}_t = \{\hat{\Gamma}_{S_{-i}} : i\in\{1,...,K\}, t\in\mathcal{T}_i\}.$$

Moreover, let

$$V_{S_i}^m = \frac{1}{|\mathcal{T}_i|}\left(\sum_{t\in\mathcal{T}_i} v_t^2 + 2\sum_{s=1}^{m_T} w(s,m_T)\sum_{t\in\mathcal{T}_{i,s}} v_t v_{t-s}\right).$$

Applying the triangle inequality, we will follow Newey and West (1987) and prove that the following three terms converge to zero in probability

$$|\widehat{V}_{S_i}-V_{S_i}| \leq \underbrace{|\widehat{V}_{S_i}-V_{S_i}^m|}_{=P_1} + \underbrace{|V_{S_i}^m-\mathbb{E}[V_{S_i}^m]|}_{=P_2} + \underbrace{|\mathbb{E}[V_{S_i}^m]-V_{S_i}|}_{=P_3}. \tag{A2}$$

Since terms P_2 and P_3 do not contain any estimated quantities, they are $o_p(1)$ following the same arguments as in the proof of Theorem 2 in Newey and West (1987) and Kool (1988), provided that Assumptions 7.1-7.4a hold.

For the term P_1, define a function

$$\begin{aligned} f(\theta,r) =& \frac{1}{|\mathcal{T}_i|}\sum_{t\in\mathcal{T}_i} v_t(\theta,\Gamma_0+r(\hat{\Gamma}_t-\Gamma_0))^2 \\ &+ \frac{2}{|\mathcal{T}_i|}\sum_{s=1}^{m_T} w(s,m_T)\sum_{t\in\mathcal{T}_{i,s}} v_t(\theta,\Gamma_0+r(\hat{\Gamma}_t-\Gamma_0))v_{t-s}(\theta,\Gamma_0+r(\hat{\Gamma}_{t-s}-\Gamma_0)), \end{aligned}$$

so that $\widehat{V}_{S_i} = f(\hat{\theta},1)$ and $V_{S_i}^m = f(\theta_0,0)$. By the multivariate mean-value theorem, for some $(\tilde{\theta},\tilde{r})$ on the line segment from $(\theta_0,0)$ to $(\hat{\theta},1)$, on the event $\mathcal{E}_T$ we have

$$\widehat{V}_{S_i} - V_{S_i}^m = f(\hat{\theta},1)-f(\theta_0,0) = \frac{\partial f}{\partial\theta}(\tilde{\theta},\tilde{r})(\hat{\theta}-\theta_0) + \frac{\partial f}{\partial r}(\tilde{\theta},\tilde{r})$$

and by the triangle inequality

$$|\widehat{V}_{S_i}-V_{S_i}^m| \leq \underbrace{\left|\frac{\partial f}{\partial\theta}(\tilde{\theta},\tilde{r})(\hat{\theta}-\theta_0)\right|}_{=P_{11}} + \underbrace{\left|\frac{\partial f}{\partial r}(\tilde{\theta},\tilde{r})\right|}_{=P_{12}}. \tag{A3}$$

We will show that both terms on the right hand side of (A3) are $o_p(1)$.

Term P_{11}: The partial derivative with respect to θ in P_{11} is

$$\begin{aligned}\frac{\partial f}{\partial \theta}(\tilde{\theta}, \tilde{r}) = &-\frac{2}{|\mathcal{T}_i|}\sum_{t\in\mathcal{T}_i} v_t(\tilde{\theta}, \Gamma_0 + \tilde{r}(\hat{\Gamma}_t - \Gamma_0)) \\ &- \frac{2}{|\mathcal{T}_i|}\sum_{s=1}^{m_T} w(s, m_T) \sum_{t\in\mathcal{T}_{i,s}} \Big[v_t(\tilde{\theta}, \Gamma_0 + \tilde{r}(\hat{\Gamma}_t - \Gamma_0)) + v_{t-s}(\tilde{\theta}, \Gamma_0 + \tilde{r}(\hat{\Gamma}_{t-s} - \Gamma_0))\Big]\end{aligned}$$

and thus

$$\left|\frac{\partial f}{\partial \theta}(\tilde{\theta}, \tilde{r})\right| \leq 2\sup_{t\in\mathcal{T}_i}|v_t(\tilde{\theta}, \Gamma_0 + \tilde{r}(\hat{\Gamma}_t - \Gamma_0))| + 4m_T \sup_{t\in\mathcal{T}_i}|v_t(\tilde{\theta}, \Gamma_0 + \tilde{r}(\hat{\Gamma}_t - \Gamma_0))|.$$

Then we get for P_{11} that

$$\begin{aligned}\left|\frac{\partial f}{\partial \theta}(\tilde{\theta}, \tilde{r})\right| |\hat{\theta} - \theta_0| \leq &\, 2\sup_{t\in\mathcal{T}_i}|v_t(\tilde{\theta}, \Gamma_0 + \tilde{r}(\hat{\Gamma}_t - \Gamma_0))| \cdot |\hat{\theta} - \theta_0| \\ &+ 4\frac{m_T}{\sqrt{T}} \sup_{t\in\mathcal{T}_i}|v_t(\tilde{\theta}, \Gamma_0 + \tilde{r}(\hat{\Gamma}_t - \Gamma_0))| \cdot \sqrt{T}|\hat{\theta} - \theta_0|.\end{aligned} \tag{A4}$$

By Theorem 4, on the event $\mathcal{E}_T$, $|\hat{\theta} - \theta_0| = o_p(1)$ and $\sqrt{T}(\hat{\theta} - \theta_0) = O_p(1)$. Moreover, by Assumption 7.4a, $\lim_{T\to\infty} m_T/\sqrt{T} = 0$. Thus it remains to show that on the event $\mathcal{E}_T$, $\sup_{t\in\mathcal{T}_i}|v_t(\tilde{\theta}, \Gamma_0 + \tilde{r}(\hat{\Gamma}_t - \Gamma_0))|$ remains bounded in probability. By the triangular inequality we have that

$$\left|v_t(\tilde{\theta}, \Gamma_0 + \tilde{r}(\hat{\Gamma}_t - \Gamma_0))\right| = \left|g(Z_t; \Gamma_0 + \tilde{r}(\hat{\Gamma}_t - \Gamma_0)) - \tilde{\theta}\right| \leq \left|g(Z_t; \Gamma_0 + \tilde{r}(\hat{\Gamma}_t - \Gamma_0))\right| + |\tilde{\theta}|.$$

For the first term after the inequality we have, for some $q > 2$

$$\begin{aligned}&\left\|g(Z_t; \Gamma_0 + \tilde{r}(\hat{\Gamma}_t - \Gamma_0))\right\|_q \\ &= \Big\|\mu_0(1, X_t) + \tilde{r}(\hat{\mu}_t(1, X_t) - \mu_0(1, X_t)) \\ &\quad - \mu_0(0, X_t) - \tilde{r}(\hat{\mu}_t(0, X_t) - \mu_0(0, X_t)) \\ &\quad + \frac{D_t}{e_0(X_t) + \tilde{r}(\hat{e}_t(X_t) - e_0(X_t))}(Y_t - \mu_0(1, X_t) - \tilde{r}(\hat{\mu}_t(1, X_t) - \mu_0(1, X_t))) \\ &\quad - \frac{1 - D_t}{1 - e_0(X_t) - \tilde{r}(\hat{e}_t(X_t) - e_0(X_t))}(Y_t - \mu_0(0, X_t) - \tilde{r}(\hat{\mu}_t(0, X_t) - \mu_0(0, X_t)))\Big\|_q \\ &\leq \|\mu_0(1, X_t)\|_q + \tilde{r}\,\|\hat{\mu}_t(1, X_t) - \mu_0(1, X_t)\|_q + \|\mu_0(0, X_t)\|_q + \tilde{r}\,\|\hat{\mu}_t(0, X_t) - \mu_0(0, X_t)\|_q \\ &\quad + \left\|\frac{D_t}{e_0(X_t) + \tilde{r}(\hat{e}_t(X_t) - e_0(X_t))}(Y_t - \mu_0(1, X_t) - \tilde{r}(\hat{\mu}_t(1, X_t) - \mu_0(1, X_t)))\right\|_q \\ &\quad + \left\|\frac{1 - D_t}{1 - e_0(X_t) - \tilde{r}(\hat{e}_t(X_t) - e_0(X_t))}(Y_t - \mu_0(0, X_t) - \tilde{r}(\hat{\mu}_t(0, X_t) - \mu_0(0, X_t)))\right\|_q\end{aligned} \tag{A5}$$

where

$$\hat{\mu}_t = \{\hat{\mu}_{S_{-i}} : i \in \{1, ..., K\}, t \in \mathcal{T}_i\} \qquad \text{and} \qquad \hat{e}_t = \{\hat{e}_{S_{-i}} : i \in \{1, ..., K\}, t \in \mathcal{T}_i\}.$$

From Assumptions 4 and 7.3 we have

$$\|\hat{\mu}_t(d, X_t) - \mu_0(d, X_t)\|_q \leq \frac{C}{\eta^{1/q}} \qquad \text{and} \qquad \|\mu_0(d, X_t)\|_q \leq \frac{C}{\eta^{1/q}}, \tag{A6}$$

for $q > 2$, any $\hat{\Gamma}_t \in \Xi_T$ and $d \in \{0, 1\}$, as shown in the proof of Theorem 5.1 in Chernozhukov et al. (2018). Using these results, we get the following bound for $|\tilde{\theta}|$

$$|\tilde{\theta}| \leq |\tilde{\theta} - \theta_0| + |\theta_0| \leq |\hat{\theta} - \theta_0| + |\theta_0| \leq |\hat{\theta} - \theta_0| + 2\frac{C}{\eta^{1/q}},$$

where $|\theta_0| \leq 2C/\eta^{1/q}$ follows from Assumption 4, as shown in the proof of Theorem 5.1 in Chernozhukov et al. (2018). From Theorem 4, it follows that $|\hat{\theta} - \theta_0| = o_p(1)$, so $|\tilde{\theta}| = O_p(1)$.

We continue with the term in the second line of the last inequality in (A5). For $q > 2$, conditional on $\mathcal{E}_T$, we get

$$\begin{aligned}
&\left\| \frac{D_t}{e_0(X_t) + \tilde{r}(\hat{e}_t(X_t) - e_0(X_t))} (Y_t - \mu_0(1, X_t) - \tilde{r}(\hat{\mu}_t(1, X_t) - \mu_0(1, X_t))) \right\|_q \\
&\leq \frac{1}{\eta} \left\| D_t (Y_t - \mu_0(1, X_t) - \tilde{r}(\hat{\mu}_t(1, X_t) - \mu_0(1, X_t))) \right\|_q \\
&\leq \frac{1}{\eta} \left\| Y_t - \mu_0(1, X_t) - \tilde{r}(\hat{\mu}_t(1, X_t) - \mu_0(1, X_t)) \right\|_q \\
&\leq \frac{1}{\eta} \|Y_t\|_q + \frac{1}{\eta} \|\mu_0(1, X_t)\|_q + \frac{\tilde{r}}{\eta} \|\hat{\mu}_t(1, X_t) - \mu_0(1, X_t)\|_q \leq \frac{1}{\eta} C + \frac{1}{\eta} \frac{C}{\eta^{1/q}} + \frac{\tilde{r}}{\eta} \frac{C}{\eta^{1/q}},
\end{aligned} \tag{A7}$$

where the first inequality follows from Assumption 4. The second inequality follows from $D_t \in \{0, 1\}$, and the third from the Minkowski inequality. The fourth follows from (A6).

For the term in the third line of the last inequality in (A5), we get for $q > 2$ and conditional on $\mathcal{E}_T$ that

$$\begin{aligned}
&\left\| \frac{1 - D_t}{1 - e_0(X_t) - \tilde{r}(\hat{e}_t(X_t) - e_0(X_t))} (Y_t - \mu_0(0, X_t) - \tilde{r}(\hat{\mu}_t(0, X_t) - \mu_0(0, X_t))) \right\|_q \\
&\leq \frac{1}{\eta} \left\| Y_t - \mu_0(0, X_t) - \tilde{r}(\hat{\mu}_t(0, X_t) - \mu_0(0, X_t)) \right\|_q \leq \frac{1}{\eta} C + \frac{1}{\eta} \frac{C}{\eta^{1/q}} + \frac{\tilde{r}}{\eta} \frac{C}{\eta^{1/q}},
\end{aligned} \tag{A8}$$

using the same arguments as for the term in (A7). Since all of the terms in the first line of the last inequality in (A5) are L_q-bounded, they are also $O_p(1)$ by Markov's inequality. The same holds for (A7) and (A8), so we can conclude that all of the terms on the right-hand-side of the last inequality in (A5) are $O_p(1)$. As a consequence, $\sup_{t \in \mathcal{T}_i} |v_t(\tilde{\theta}, \Gamma_0 + \tilde{r}(\hat{\Gamma}_t - \Gamma_0))| = O_p(1)$ in (A4) and thus $P_{11} = o_p(1)$.

Term P_{12}: From the partial derivative with respect to r in P_{12} we find

$$\begin{aligned}
\left| \frac{\partial f}{\partial r}(\tilde{\theta}, \tilde{r}) \right| = \Bigg| &\frac{2}{|\mathcal{T}_i|} \sum_{t \in \mathcal{T}_i} v_t(\tilde{\theta}, \Gamma_0 + \tilde{r}(\hat{\Gamma}_t - \Gamma_0)) \frac{\partial v_t}{\partial r}(\tilde{\theta}, \Gamma_0 + \tilde{r}(\hat{\Gamma}_t - \Gamma_0)) \\
&+ \frac{2}{|\mathcal{T}_i|} \sum_{s=1}^{m_T} w(s, m_T) \sum_{t \in \mathcal{T}_{i,s}} v_t(\tilde{\theta}, \Gamma_0 + \tilde{r}(\hat{\Gamma}_t - \Gamma_0)) \frac{\partial v_{t-s}}{\partial r}(\tilde{\theta}, \Gamma_0 + \tilde{r}(\hat{\Gamma}_{t-s} - \Gamma_0)) \\
&+ \frac{2}{|\mathcal{T}_i|} \sum_{s=1}^{m_T} w(s, m_T) \sum_{t \in \mathcal{T}_{i,s}} v_{t-s}(\tilde{\theta}, \Gamma_0 + \tilde{r}(\hat{\Gamma}_{t-s} - \Gamma_0)) \frac{\partial v_t}{\partial r}(\tilde{\theta}, \Gamma_0 + \tilde{r}(\hat{\Gamma}_t - \Gamma_0)) \Bigg| \\
\leq 2 &\sup_{t \in \mathcal{T}_i} |v_t(\tilde{\theta}, \Gamma_0 + \tilde{r}(\hat{\Gamma}_t - \Gamma_0))| \sup_{t \in \mathcal{T}_i} \left| \frac{\partial v_t}{\partial r}(\tilde{\theta}, \Gamma_0 + \tilde{r}(\hat{\Gamma}_t - \Gamma_0)) \right| \\
&+ 4 \frac{m_T}{T^{b_m}} \sup_{t \in \mathcal{T}_i} |v_t(\tilde{\theta}, \Gamma_0 + \tilde{r}(\hat{\Gamma}_t - \Gamma_0))| \sup_{t \in \mathcal{T}_i} \left| \frac{\partial v_t}{\partial r}(\tilde{\theta}, \Gamma_0 + \tilde{r}(\hat{\Gamma}_t - \Gamma_0)) \right| T^{b_m}.
\end{aligned}$$

From proving $P_{11} = o_p(1)$ we know that $\sup_{t \in \mathcal{T}_i} |v_t(\tilde{\theta}, \Gamma_0 + \tilde{r}(\hat{\Gamma}_t - \Gamma_0))| = O_p(1)$. Moreover, $m_T / T^{b_m} = o(1)$ by Assumption 7.4a. Thus it remains to show that $\sup_{t \in \mathcal{T}_i} |\frac{\partial v_t}{\partial r}(\tilde{\theta}, \Gamma_0 + \tilde{r}(\hat{\Gamma}_t - \Gamma_0))| = o_p(T^{-b_m})$. Since by Assumption 7 $b_m < b_r$, we will show that the quantity is actually $o_p(T^{-b_r})$. By the triangle inequality, we have

$$\sup_{t \in \mathcal{T}_i} \left| \frac{\partial v_t}{\partial r}(\tilde{\theta}, \Gamma_0 + \tilde{r}(\hat{\Gamma}_t - \Gamma_0)) \right| \leq P_{12,A} + P_{12,B} + P_{12,C} + P_{12,D}, \tag{A9}$$

where

$$
\begin{aligned}
P_{12,A} &= \sup_{t\in\mathcal{T}_i} |\hat{\mu}_t(1,X_t) - \mu_0(1,X_t) - (\hat{\mu}_t(0,X_t) - \mu_0(0,X_t))| \\
P_{12,B} &= \sup_{t\in\mathcal{T}_i} \Bigg| \frac{D_t}{\left(e_0(X_t) + \tilde{r}(\hat{e}_t(X_t) - e_o(X_t))\right)^2} \times \\
&\qquad (Y_t - \mu_0(1,X_t) - \tilde{r}(\hat{\mu}_t(1,X_t) - \mu_0(1,X_t)))\,(\hat{e}_t(X_t) - e_0(X_t)) \Bigg| \\
P_{12,C} &= \sup_{t\in\mathcal{T}_i} \left| \frac{1-D_t}{1 - e_0(X_t) - \tilde{r}(\hat{e}_t(X_t) - e_0(X_t))}\,(\hat{\mu}_t(0,X_t) - \mu_0(0,X_t)) \right| \\
P_{12,D} &= \sup_{t\in\mathcal{T}_i} \Bigg| \frac{1-D_t}{\left((1-\tilde{r})(1-e_0(X_t)) + \tilde{r}(1-\hat{e}_t(X_t))\right)^2} \times \\
&\qquad (Y_t - \mu_0(0,X_t) + \tilde{r}(\hat{\mu}_t(0,X_t) - \mu_0(0,X_t)))\,(\hat{e}_t(X_t) - e_0(X_t)) \Bigg|.
\end{aligned}
$$

For $P_{12,A}$, we first establish that on the event $\mathcal{E}_T$ for $d \in \{0,1\}$ we have

$$
\begin{aligned}
&\sup_{t\in\mathcal{T}_i} \mathbb{E}\left[|\hat{\mu}_t(d,X_t) - \mu_0(d,X_t)|^2 \,\middle|\, S_{-i} \right] \\
&\quad \leq \sup_{t\in\mathcal{T}_i} \sup_{\mu\in\Xi_T} \mathbb{E}\left[\left(\mu(d,X_t) - \mu_0(d,X_t)\right)^2 \,\middle|\, S_{-i} \right] \\
&\quad \leq \sup_{t\in\mathcal{T}_i} \sup_{\mu\in\Xi_T} \mathbb{E}\left[\left(\mu(d,X_t) - \mu_0(d,X_t)\right)^2 \right] + O_p(\alpha(k_T)^\psi) \\
&\quad \leq r_{\mu,T}^2 + O_p(\alpha(k_T)^\psi) = o_p(T^{-2b_r}).
\end{aligned}
\tag{A10}
$$

The second inequality follows from Lemma 1, the third by the definition of the rate $r_{\mu,T}$ in Assumption 4. The final equality follows by Lemma 6.1 in Chernozhukov et al. (2018) and Assumptions 4.3, 6.5 and 7.4b. We thus conclude that on the event $\mathcal{E}_T$,

$$
\sup_{t\in\mathcal{T}_i} |\hat{\mu}_t(d,X_t) - \mu_0(d,X_t)| = o_p(T^{-b_r}) \tag{A11}
$$

and as a consequence

$$
P_{12,A} \leq \sup_{t\in\mathcal{T}_i} |\hat{\mu}_t(1,X_t) - \mu_0(1,X_t)| + \sup_{t\in\mathcal{T}_i} |\hat{\mu}_t(0,X_t) - \mu_0(0,X_t)| = o_p(T^{-b_r}).
$$

For $P_{12,B}$, we first establish that on the event $\mathcal{E}_T$ we have

$$
\begin{aligned}
\sup_{t\in\mathcal{T}_i} \mathbb{E}\left[|\hat{e}_t(X_t) - e_0(X_t)|^2 \,|S_{-i}\right] &\leq \sup_{t\in\mathcal{T}_i} \sup_{e\in\Xi_T} \mathbb{E}\left[(\hat{e}_t(X_t) - e_0(X_t))^2 | S_{-i} \right] \\
&\leq \sup_{t\in\mathcal{T}_i} \sup_{e\in\Xi_T} \mathbb{E}\left[(\hat{e}_t(X_t) - e_0(X_t))^2 \right] + O_p(\alpha(k_T)^\psi) \\
&\leq r_{e,T}^2 + O_p(\alpha(k_T)^\psi) = o_p(T^{-2b_r})
\end{aligned}
$$

by the same arguments as for (A10) and thus on the event $\mathcal{E}_T$,

$$
\sup_{t\in\mathcal{T}_i} |\hat{e}_t(X_t) - e_0(X_t)| = o_p(T^{-b_r}). \tag{A12}
$$

This allows us to write

$$
\begin{aligned}
P_{12,B} &\leq \sup_{t\in\mathcal{T}_i} \frac{1}{\eta^2} |Y_t - (1-\tilde{r})\mu_0(1,X_t) - \tilde{r}\hat{\mu}_t(1,X_t)|\,|\hat{e}_t(X_t) - e_0(X_t)| \\
&\leq \sup_{t\in\mathcal{T}_i} \frac{1-\tilde{r}}{\eta^2} |Y_t - \mu_0(1,X_t)|\,|\hat{e}_t(X_t) - e_0(X_t)| \\
&\quad + \sup_{t\in\mathcal{T}_i} \frac{\tilde{r}}{\eta^2} |Y_t - \hat{\mu}_t(1,X_t)|\,|\hat{e}_t(X_t) - e_0(X_t)| \\
&\leq \sup_{t\in\mathcal{T}_i} \frac{2}{\eta^2} |Y_t - \mu_0(1,X_t)|\,|\hat{e}_t(X_t) - e_0(X_t)| \\
&\quad + \sup_{t\in\mathcal{T}_i} \frac{1}{\eta^2} |\hat{\mu}_t(1,X_t) - \mu_0(1,X_t)|\,|\hat{e}_t(X_t) - e_0(X_t)|,
\end{aligned}
\tag{A13}
$$

where the first inequality follows from Assumption 4 and $D_t \in \{0,1\}$. From Assumption 7.3 and (A6), we have for $q > 2$, $d \in \{0,1\}$ and any $\mu \in \Xi_T$ that

$$\sup_{t\in\mathcal{T}_i} \|Y_t - \mu(d, X_t)\|_q \leq \sup_{t\in\mathcal{T}_i} \|Y_t\|_q + \sup_{t\in\mathcal{T}_i} \|\mu(d, X_t)\|_q \leq C\left(1 + \frac{1}{\eta^{1/q}}\right). \tag{A14}$$

Combined with (A12) it follows that for $d \in \{0,1\}$ and on the event $\mathcal{E}_T$ we have

$$\begin{aligned}
&\sup_{t\in\mathcal{T}_i} \||Y_t - \mu(d, X_t)||\hat{e}_t(X_t) - e_0(X_t)|\|_1 \\
&\quad \leq \sup_{t\in\mathcal{T}_i} \|Y_t - \mu(d, X_t)\|_2 \|\hat{e}_t(X_t) - e_0(X_t)\|_2 \\
&\quad \leq \sup_{t\in\mathcal{T}_i} C\left(1 + \frac{1}{\eta^{1/q}}\right) \|\hat{e}_t(X_t) - e_0(X_t)\|_2 = o_p(T^{-b_r})
\end{aligned}$$

for any $\mu \in \Xi_T$, and as a consequence, combined with (A11), we have $P_{12,B} = o_p(T^{-b_r})$.

For $P_{12,C}$, we have by Assumption 4 and $D_t \in \{0,1\}$ that $P_{12,C} \leq \sup_{t\in\mathcal{T}_i} \frac{1}{\eta}|\hat{\mu}_t(0, X_t) - \mu_0(0, X_t)|$ and thus $P_{12,C} = o_p(T^{-b_r})$ by (A11).

For $P_{12,D}$ finally, we have

$$\begin{aligned}
P_{12,D} &\leq \sup_{t\in\mathcal{T}_i} \frac{1}{\eta^2} \left|(1-\tilde{r})(Y_t - \mu_0(0, X_t)) + \tilde{r}(Y_t - \hat{\mu}_t(0, X_t))\right| \left|\hat{e}_t(X_t) - e_0(X_t)\right| \\
&\leq \sup_{t\in\mathcal{T}_i} \frac{2}{\eta^2} \left|Y_t - \mu_0(0, X_t)\right| \left|\hat{e}_t(X_t) - e_0(X_t)\right| \\
&\quad + \sup_{t\in\mathcal{T}_i} \frac{1}{\eta^2} \left|\hat{\mu}_t(0, X_t) - \mu_0(0, X_t)\right| \left|\hat{e}_t(X_t) - e_0(X_t)\right|,
\end{aligned}$$

where the first inequality again follows from Assumption 4 and $D_t \in \{0,1\}$, and the second from the same arguments as in (A13). Using (A11), (A12) and (A14) lets us conclude that $P_{12,D} = o_p(T^{-b_r})$. This shows that all terms on the right hand side of (A9) are $o_p(T^{-b_r})$ and thus $\sup_{t\in\mathcal{T}_i} |\frac{\partial v_t}{\partial r}(\tilde{\theta}, \Gamma_0 + \tilde{r}(\hat{\Gamma}_t - \Gamma_0))| = o_p(T^{-b_r})$. As a consequence, $P_{11} = o_p(1)$ and $P_{12} = o_p(1)$ in (A3). We conclude that $|\widehat{V}_{S_i} - V^m_{S_i}| = o_p(1)$ and thus $|\widehat{V}_{S_i} - V_{S_i}| = o_p(1)$ so that $|\widehat{V} - V_T| \xrightarrow{p} 0$. □

References

Adamek, R., Smeekes, S., & Wilms, I. (2024). Local Projection Inference in High Dimensions. *The Econometrics Journal*, 1–20. doi:10.1093/ectj/utae012

Angrist, J. D., Jordà, Ò., & Kuersteiner, G. M. (2018). Semiparametric Estimates of Monetary Policy Effects: String Theory Revisited. *Journal of Business & Economic Statistics*, *36*(3), 371–387. doi:10.1080/07350015.2016.1204919

Bach, P., Schacht, O., Chernozhukov, V., Klaassen, S., & Spindler, M. (2024). Hyperparameter Tuning for Causal Inference with Double Machine Learning: A Simulation Study. In F. Locatello & V. Didelez (Eds.), *Proceedings of the third conference on causal learning and reasoning* (Vol. 236, pp. 1065–1117). PMLR.

Ballinari, D., & Bearth, N. (2025). *Improving the finite sample estimation of average treatment effects using double/debiased machine learning with propensity score calibration.* (Preprint (arXiv:2409.04874))

Barnichon, R., & Brownlees, C. (2019). Impulse Response Estimation by Smooth Local Projections. *The Review of Economics and Statistics*, *101*(3), 522–530. doi:10.1162/rest_a_00778

Beck, E., & Wolf, M. (2025). *Forecasting inflation with the hedged random forest* (Swiss National Bank Working Paper No. 07/2025).

Bergmeir, C., & Benítez, J. M. (2012). On the use of cross-validation for time series predictor evaluation. *Information Sciences*, *191*, 192–213. doi:10.1016/j.ins.2011.12.028

Bica, I., Alaa, A., & Van Der Schaar, M. (2020). Time Series Deconfounder: Estimating Treatment Effects over Time in the Presence of Hidden Confounders. In H. D. III & A. Singh (Eds.), *Proceedings of the 37th international conference on machine learning* (Vol. 119, pp. 884–895). PMLR. doi:10.5555/3524938.3525021

Bouchaud, J.-P., Bonart, J., Donier, J., & Gould, M. (2018). *Trades, quotes and prices: Financial markets under the microscope*. Cambridge University Press. doi:10.1017/9781316659335

Breiman, L. (2001). Random forests. *Machine Learning*, *45*, 5–32. doi:10.1023/A:1010933404324

Bussmann, B., Nys, J., & Latré, S. (2021). Neural Additive Vector Autoregression Models for Causal Discovery in Time Series. In C. Soares & L. Torgo (Eds.), *Discovery Science* (pp. 446–460). Cham: Springer International Publishing. doi:10.1007/978-3-030-88942-5_35

Caruana, R. (1997). Multitask Learning. *Machine Learning*, *28*, 41–75. doi:10.1023/A:1007379606734

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (p. 785—794). New York, NY, USA: Association for Computing Machinery. doi:10.1145/2939672.2939785

Chen, X., & Christensen, T. M. (2015). Optimal uniform convergence rates and asymptotic normality for series estimators under weak dependence and weak conditions. *Journal of Econometrics*, *188*(2), 447–465. doi:10.1016/j.jeconom.2015.03.010

Chen, X., & Liao, Z. (2015). Sieve semiparametric two-step GMM under weak dependence. *Journal of Econometrics*, *189*(1), 163–186. doi:10.1016/j.jeconom.2015.07.001

Chen, X., & Shen, X. (1998). Sieve Extremum Estimates for Weakly Dependent Data. *Econometrica*, *66*(2), 289–314. doi:10.2307/2998559

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., & Newey, W. (2017). Double/debiased/neyman machine learning of treatment effects. *American Economic Review*, *107*(5), 261–265. doi:10.1257/aer.p20171038

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., & Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, *21*(1), C1–C68. doi:10.1111/ectj.12097

Chernozhukov, V., Newey, W. K., & Singh, R. (2022a). Automatic Debiased Machine Learning of Causal and Structural Effects. *Econometrica*, *90*(3), 967–1027. doi:10.3982/ECTA18515

Chernozhukov, V., Newey, W. K., & Singh, R. (2022b). Debiased machine learning of global and local parameters using regularized Riesz representers. *The Econometrics Journal*, *25*(3), 576–601. doi:10.1093/ectj/utac002

Cochran, W. G. (1968). The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics*, *24*(2), 295–313. doi:10.2307/2528036

Colangelo, K., & Lee, Y.-Y. (2025). Double Debiased Machine Learning Nonparametric Inference with Continuous Treatments. *Journal of Business & Economic Statistics*, *0*(0), 1–13. doi:10.1080/07350015.2025.2505487

Crump, R. K., Hotz, V. J., Imbens, G. W., & Mitnik, O. A. (2009, 01). Dealing with limited overlap in estimation of average treatment effects. *Biometrika*, *96*(1), 187-199. doi:10.1093/biomet/asn055

Davidson, J. (2021). *Stochastic Limit Theory: An Introduction for Econometricians* (2nd ed.). Oxford University Press. doi:10.1093/oso/9780192844507.001.0001

Davis, R. A., & Nielsen, M. S. (2020). Modeling of time series using random forests: Theoretical developments. *Electronic Journal of Statistics*, *14*(2), 3644-3671.

Fisher, A., & Kennedy, E. H. (2021). Visually Communicating and Teaching Intuition for Influence Functions. *The*

American Statistician, *75*(2), 162–172. doi:10.1080/00031305.2020.1717620
Frank, E., & Hall, M. (2001). A Simple Approach to Ordinal Classification. In L. De Raedt & P. Flach (Eds.), *Machine learning: Ecml 2001* (pp. 145–156). Springer Berlin Heidelberg.
Goehry, Benjamin. (2020). Random forests for time-dependent processes. *ESAIM: PS*, *24*, 801–826. doi:10.1051/ps/2020015
Gonçalves, S., Herrera, A., Kilian, L., & Pesavento, E. (2024). State-dependent local projections. *Journal of Econometrics*, *244*(2), 105702. doi:10.1016/j.jeconom.2024.105702
Goulet Coulombe, P. (2024). The macroeconomy as a random forest. *Journal of Applied Econometrics*, *39*(3), 401-421. doi:https://doi.org/10.1002/jae.3030
Goulet Coulombe, P., Leroux, M., Stevanovic, D., & Surprenant, S. (2022). How is machine learning useful for macroeconomic forecasting? *Journal of Applied Econometrics*, *37*(5), 920-964. doi:https://doi.org/10.1002/jae.2910
Grecov, P., Bandara, K., Bergmeir, C., Ackermann, K., Campbell, S., Scott, D., & Lubman, D. (2021). Causal Inference Using Global Forecasting Models for Counterfactual Prediction. In K. Karlapalem et al. (Eds.), *Advances in Knowledge Discovery and Data Mining* (pp. 282–294). Cham: Springer International Publishing. doi:10.1007/978-3-030-75765-6_23
Hahn, J. (1998). On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects. *Econometrica*, *66*(2), 315–331. doi:10.2307/2998560
Hauzenberger, N., Huber, F., Klieber, K., & Marcellino, M. (2025). Machine learning the macroeconomic effects of financial shocks. *Economics Letters*, *250*, 112260. doi:https://doi.org/10.1016/j.econlet.2025.112260
Herrndorf, N. (1984). A functional central limit theorem for weakly dependent sequences of random variables. *The Annals of Probability*, 141–153. doi:10.1214/aop/1176993379
Hines, O., Dukes, O., Diaz-Ordaz, K., & Vansteelandt, S. (2022). Demystifying statistical learning based on efficient influence functions. *The American Statistician*, *76*(3), 292–304. doi:10.1080/00031305.2021.2021984
Irle, A. (1997). On Consistency in Nonparametric Estimation under Mixing Conditions. *Journal of Multivariate Analysis*, *60*(1), 123–147. doi:10.1006/jmva.1996.1647
Jordà, Ò. (2005). Estimation and Inference of Impulse Responses by Local Projections. *The American Economic Review*, *95*(1), 161–182. doi:10.1257/0002828053828518
Jordà, Ò. (2023). Local Projections for Applied Economics. *Annual Review of Economics*, *15*, 607–631. doi:10.1146/annurev-economics-082222-065846
Jordà, Ò., & Taylor, A. M. (2016). The Time for Austerity: Estimating the Average Treatment Effect of Fiscal Policy. *The Economic Journal*, *126*(590), 219–255. doi:10.1111/ecoj.12332
Jordà, Ò., & Taylor, A. M. (2024, August). *Local Projections* (Working Paper No. 32822). National Bureau of Economic Research. doi:10.3386/w32822
Kennedy, E. H. (2023). Towards optimal doubly robust estimation of heterogeneous causal effects. *Electronic Journal of Statistics*, *17*(2), 3008–3049. doi:10.1214/23-EJS2157
Kennedy, E. H. (2024). Semiparametric Doubly Robust Targeted Double Machine Learning: A Review. In E. Laber, B. Chakraborty, E. Moodie, T. Cai, & M. Laan (Eds.), *Handbook of Statistical Methods for Precision Medicine* (1st ed.). New York: Chapman and Hall/CRC. doi:10.1201/9781003216223-10
Kiefer, N. M., & Vogelsang, T. J. (2005). A new asymptotic theory for heteroskedasticity-autocorrelation robust tests. *Econometric Theory*, *21*(6), 1130–1164.
Klaassen, S., Rabenseifner, J., Kueck, J., & Bach, P. (2025). *Calibration Strategies for Robust Causal Estimation: Theoretical and Empirical Insights on Propensity Score-Based Estimators.* (Preprint (arXiv:2503.17290))
Knaus, M. C. (2021). A double machine learning approach to estimate the effects of musical practice on student's skills. *Journal of the Royal Statistical Society Series A: Statistics in Society*, *184*(1), 282–300. doi:10.1111/rssa.12623
Kolesár, M., & Plagborg-Møller, M. (2025). Dynamic Causal Effects in a Nonlinear World: the Good, the Bad, and the Ugly. *Journal of Business & Economic Statistics*, *43*(4), 737–754. doi:10.1080/07350015.2025.2539478
Kool, J. (1988). *A note on consistent estimation of heteroskedastic and autocorrelated covariance matrices* (Serie Research Memoranda No. 0021). VU University Amsterdam, Faculty of Economics, Business Administration and Econometrics.
Künzel, S., Sekhon, J., Bickel, P., & Yu, B. (2019). Metalearners for estimating heterogeneous treatment effects using machine learning. *Proc. Nat. Acad. Sci.*, *116*(10), 4156–4165. doi:10.1073/pnas.1804597116
Lazarus, E., Lewis, D. J., Stock, J. H., & Watson, M. W. (2018). HAR Inference: Recommendations for Practice. *Journal of Business & Economic Statistics*, *36*(4), 541–559. doi:10.1080/07350015.2018.1506926
Lee, B. K., Lessler, J., & Stuart, E. A. (2011, 03). Weight trimming and propensity score weighting. *PLOS ONE*, *6*(3), 1-6. doi:10.1371/journal.pone.0018174
Lee, J., & Robinson, P. M. (2016). Series estimation under cross-sectional dependence. *Journal of Econometrics*, *190*(1), 1–17. doi:10.1016/j.jeconom.2015.08.001

Lewis, G., & Syrgkanis, V. (2021). Double/Debiased Machine Learning for Dynamic Treatment Effects. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, & J. W. Vaughan (Eds.), *Advances in Neural Information Processing Systems* (Vol. 34, pp. 22695–22707). Curran Associates, Inc.

Li, Q., & Racine, J. S. (2006). *Nonparametric Econometrics: Theory and Practice* (1st ed., Vol. 1) (No. 8355). Princeton University Press.

Lozano, A. C., Kulkarni, S. R., & Schapire, R. E. (2014). Convergence and Consistency of Regularized Boosting With Weakly Dependent Observations. *IEEE Transactions on Information Theory*, *60*(1), 651–660. doi:10.1109/TIT.2013.2287726

Lucchese, L., Pakkanen, M. S., & Veraart, A. E. (2024). The short-term predictability of returns in order book markets: A deep learning perspective. *International Journal of Forecasting*, *40*(4), 1587-1621. doi:https://doi.org/10.1016/j.ijforecast.2024.02.001

Ma, M., & Safikhani, A. (2022). Theoretical analysis of deep neural networks for temporally dependent observations. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, & A. Oh (Eds.), *Advances in Neural Information Processing Systems* (Vol. 35, pp. 37324–37334). Curran Associates, Inc.

Mariet, Z., & Kuznetsov, V. (2019). Foundations of Sequence-to-Sequence Modeling for Time Series. In K. Chaudhuri & M. Sugiyama (Eds.), *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics* (Vol. 89, pp. 408–417). Retrieved from `https://proceedings.mlr.press/v89/mariet19a.html`

Masini, R. P., Medeiros, M. C., & Mendes, E. F. (2023). Machine learning advances for time series forecasting. *Journal of Economic Surveys*, *37*(1), 76-111.

Medeiros, M. C., Vasconcelos, G. F. R., Álvaro Veiga, & Zilberman, E. (2021). Forecasting inflation in a data-rich environment: The benefits of machine learning methods. *Journal of Business & Economic Statistics*, *39*(1), 98–119. doi:10.1080/07350015.2019.1637745

Nakamura, E., & Steinsson, J. (2018, August). Identification in macroeconomics. *Journal of Economic Perspectives*, *32*(3), 59-86. doi:10.1257/jep.32.3.59

Nauta, M., Bucur, D., & Seifert, C. (2019). Causal Discovery with Attention-Based Convolutional Neural Networks. *Machine Learning and Knowledge Extraction*, *1*(1), 312–340. doi:10.3390/make1010019

Newey, W. K., & West, K. D. (1987). A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix. *Econometrica*, *55*(3), 703–708. doi:10.2307/1913610

Newey, W. K., & West, K. D. (1994). Automatic Lag Selection in Covariance Matrix Estimation. *The Review of Economic Studies*, *61*(4), 631–653. doi:10.2307/2297912

Neyman, J. (1959). Optimal Asymptotic Tests of Composite Statistical Hypotheses. In U. Grenander (Ed.), *Probability and Statistics* (pp. 13–34). John Wiley & Sons, New York.

Neyman, J. (1979). C(α) tests and their use. *Sankhyā: The Indian Journal of Statistics, Series A (1961-2002)*, *41*(1/2), 1–21.

Nie, X., & Wager, S. (2020). Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika*, *108*(2), 299–319. doi:10.1093/biomet/asaa076

Olea, J. L. M., Plagborg-Møller, M., Qian, E., & Wolf, C. K. (2024). *Double robustness of local projections and some unpleasant varithmetic.* (Preprint (arXiv:2405.09509))

Paranhos, L. (2025). How do firms' financial conditions influence the transmission of monetary policy? a non-parametric local projection approach. *Journal of Econometrics*, *249*, 105886. doi:https://doi.org/10.1016/j.jeconom.2024.105886

Pearl, J. (2009). *Causality* (2nd ed.). Cambridge, UK: Cambridge University Press. doi:10.1017/CBO9780511803161

Phillips, P. C. B. (1987). Time Series Regression with a Unit Root. *Econometrica*, *55*(2), 277–301. doi:10.2307/1913237

Plagborg-Møller, M., & Wolf, C. K. (2021). Local Projections and VARs Estimate the Same Impulse Responses. *Econometrica*, *89*(2), 955-980. doi:https://doi.org/10.3982/ECTA17813

Qingliang Fan, R. P. L., Yu-Chin Hsu, & Zhang, Y. (2022). Estimation of Conditional Average Treatment Effects With High-Dimensional Data. *Journal of Business & Economic Statistics*, *40*(1), 313–327. doi:10.1080/07350015.2020.1811102

Racine, J. (2000). Consistent cross-validatory model-selection for dependent data: hv-block cross-validation. *Journal of Econometrics*, *99*(1), 39–61. doi:10.1016/S0304-4076(00)00030-0

Raha, M., Sheth, P., Karami, M., Bhattacharya, A., Wang, Q., Tahir, A., … Liu, H. (2021). Causal inference for time series analysis: problems, methods and evaluation. *Knowledge and Information Systems*, *63*, 3041–3085. doi:10.1007/s10115-021-01621-0

Rambachan, A., & Shephard, N. (2021). *When do common time series estimands have nonparametric causal meaning.* (Preprint)

Ramey, V. (2016). Chapter 2 - Macroeconomic Shocks and Their Propagation. In J. B. Taylor & H. Uhlig (Eds.),

(Vol. 2, p. 71-162). Elsevier. doi:https://doi.org/10.1016/bs.hesmac.2016.03.003
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), 1393–1512. doi:10.1016/0270-0255(86)90088-6
Robins, J., & Rotnitzky, A. (1995). Semiparametric Efficiency in Multivariate Regression Models with Missing Data. *Journal of the American Statistical Association*, *90*(429), 122–129. doi:10.1080/01621459.1995.10476494
Robinson, P. M. (1983). Nonparametric estimators for time series. *Journal of Time Series Analysis*, *4*(3), 185–207. doi:10.1111/j.1467-9892.1983.tb00368.x
Romer, C. D., & Romer, D. H. (2004, September). A new measure of monetary shocks: Derivation and implications. *American Economic Review*, *94*(4), 1055-1084. doi:10.1257/0002828042002651
Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*(1), 41–55. doi:10.1093/biomet/70.1.41
Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, *66*(5), 688–701. doi:10.1037/h0037350
Runge, J., Gerhardus, A., Varando, G., Eyring, V., & Camps-Valls, G. (2023). Causal inference for time series. *Nature Reviews Earth & Environment*, *4*, 487–505. doi:10.1038/s43017-023-00431-y
Semenova, V., & Chernozhukov, V. (2020). Debiased machine learning of conditional average treatment effects and other causal functions. *The Econometrics Journal*, *24*(2), 264–289. doi:10.1093/ectj/utaa027
Semenova, V., Goldman, M., Chernozhukov, V., & Taddy, M. (2023). Inference on heterogeneous treatment effects in high-dimensional dynamic panels under weak dependence. *Quantitative Economics*, *14*(2), 471-510. doi:https://doi.org/10.3982/QE1670
Shi, C., Blei, D., & Veitch, V. (2019). Adapting Neural Networks for the Estimation of Treatment Effects. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, & R. Garnett (Eds.), *Advances in neural information processing systems* (Vol. 32). Curran Associates, Inc. doi:10.5555/3454287.3454512
Sims, C. A. (1980). Macroeconomics and Reality. *Econometrica*, *48*(1), 1–48. doi:10.2307/1912017
Steinwart, I., Hush, D., & Scovel, C. (2009). Learning from dependent observations. *Journal of Multivariate Analysis*, *100*(1), 175–194. doi:10.1016/j.jmva.2008.04.001
Stock, J. H., & Watson, M. W. (2018). Identification and Estimation of Dynamic Causal Effects in Macroeconomics Using External Instruments. *The Economic Journal*, *128*(610), 917–948. doi:10.1111/ecoj.12593
Sun, Y. (2014). Let's fix it: Fixed-b asymptotics versus small-b asymptotics in heteroskedasticity and autocorrelation robust inference. *Journal of Econometrics*, *178*, 659–677.
Tsiatis, A. A. (2006). *Semiparametric Theory and Missing Data*. Springer New York. doi:10.1007/0-387-37345-4
Wager, S. (2022). *STATS 361: Causal Inference.* Retrieved from `http://web.stanford.edu/~swager/stats361.pdf`
Wong, K. C., Li, Z., & Tewari, A. (2020). Lasso guarantees for β-mixing heavy-tailed time series. *The Annals of Statistics*, *48*(2), 1124 – 1142.
Wooldridge, J. M. (2016). *Introductory econometrics: A modern approach*. South-Western Cengage Learning.
Yin, Z., & Barucca, P. (2022). Deep recurrent modelling of Granger causality with latent confounding. *Expert Systems with Applications*, *207*, 118036. doi:10.1016/j.eswa.2022.118036
Zhang, Z., Zohren, S., & Roberts, S. (2019). Deeplob: Deep convolutional neural networks for limit order books. *IEEE Transactions on Signal Processing*, *67*(11), 3001–3012.
Zimmert, M., & Lechner, M. (2019). *Nonparametric estimation of causal heterogeneity under high-dimensional confounding.* (Preprint (arXiv:1908.08779))

Supplementary material for "Semiparametric inference for impulse response functions using double/debiased machine learning"

A Hyperparameter tuning

A.1 Tuning in the simulation study

Random forests: We fix the number of trees to 500. For each simulation replication, the maximal depth of each tree (d), the minimal number of observations in the leafs of the trees (ℓ) and the maximal fraction of features considered for each node split (m_{try}) are determined by cross-validation using the K sub-processes as folds. We perform a simple grid search over $\{\{d, \ell, m_{try}\} \mid d \in \{5, 10, 20, 50\} \wedge \ell \in \{1, 5, 10\} \wedge m_{try} \in \{0.3, 1.0\}\}$ and select the hyperparameter-combination yielding the best predictive cross-validation performance.

Gradient boosted trees: We perform a two-stage tuning procedure. In the first stage, we fix the learning rate of the tree booster to 0.1, set the number of boosting rounds to some very high number (10'000) and abort the estimation process if the predictive validation error has not decreased since 50 rounds of boosting. The maximum tree depth for the base learners (d), the minimum sum of instance weight needed in a child (w), the subsampling ratio for observations used to construct each tree (s^o), and the subsampling ratio for features when constructing each tree (s^f) are determined by cross-validation using the K sub-processes as folds. We perform a simple grid search over $\{\{d, w, s^o, s^f\} \mid d \in \{1, ..., 10\} \wedge w \in \{1, ..., 10\} \wedge s^o \in \{0.25, 0.5, 0.75, 1\} \wedge s^f \in \{0.25, 0.5, 0.75, 1\}\}$ and select the hyperparameter-combination yielding the best predictive cross-validation performance. In the second stage, using the optimal tree-hyperparameters from stage 1, we select the learning rate most frequently yielding the best predictive cross-validation performance. We perform a search over the candidate set $\{0.001, 0.002, 0.005, 0.01, 0.02, 0.05, 0.1, 0.25, 0.5\}$. Finally, the optimal number of boosting rounds is determined as the average early-stopped boosting round for the selected learning rate. For computational reasons, we repeat this procedure on $R = 50$ simulated examples and fix the hyperparameters for all simulation replications to the hyperparameter combination most frequently yielding the best cross-validation performance.

A.2 Tuning in the empirical application

Random forests: We fix the number of trees to 500. The maximal depth of each tree (d), the minimal number of observations in the leafs of the trees (ℓ) and the size of the random subsets of features to consider when splitting a node ($\bar{m}$) are determined by cross-validation using the K sub-processes as folds. We perform a simple grid search over the same candidate sets as for the simulation study (cf. Appendix A.1) and select the hyperparameter-combination yielding the best predictive cross-validation performance.

Gradient boosted trees: We perform a two-stage tuning procedure. In the first stage, we fix the learning rate of the tree booster to 0.1, the number of boosting rounds to 500 and find optimal tree parameters by cross-validation. For this, we perform a simple grid search over $\{\{d, w, s^o, s^f\} \mid d \in \{1, ..., 10\} \wedge w \in \{1, ..., 10\} \wedge s^o \in \{0.25, 0.5, 0.75, 1\} \wedge s^f \in \{0.25, 0.5, 0.75, 1\}\}$, for the same parameters as in Appendix A.1, and select the hyperparameter-combination yielding the best predictive cross-validation performance. In the second stage, using the optimal tree-hyperparameters from stage 1, we finally select the learning rate and number of boosting rounds yielding the best predictive cross-validation performance. We perform a search over all combinations of $\{0.001, 0.002, 0.005, 0.01, 0.02, 0.05, 0.1, 0.25, 0.5\}$ for the learning rate and $\{10, 110, 210, 310, 410, 510\}$ for the number of boosting rounds.

All estimators are tuned using 10-fold blocked cross-validation on the full sample, where at the boundary of the folds 24 observations are dropped to eliminate dependence between the folds.

B Simulation with an empirically calibrated data generating process

We calibrate the data generating process defined in Section 5 of the main text using monthly U.S. data from 1982 to 2012 obtained from the empirical study in Angrist et al. (2018). In more detail, the outcome process follows

$$Y_t = c + b(X_t) + (D_t - 0.5)\,\tau(X_t) + \gamma Y_{t-1} + \epsilon_t.$$

We calibrate the parameters c and γ, the innovation process ϵ_t and the process governing a set of confounder variables X_t. The vector X_t contains the changes in the federal funds rate and in the ten-year Treasury yield, percentage changes in the S&P 1000 index, the M1 money stock, civilian employment, and industrial production, as well as

percentage-point changes in the unemployment rate. The outcome variable Y_t is the monthly percentage changes in core personal consumption expenditures. The specification of $b(X_t)$, $\tau(X_t)$ and $e(X_t)$ are kept identical to those used in the simulation study in Section 5 of the main text.

The innovation process is assumed to follow a GARCH(p, q) specification. Following Adamek et al. (2024); Lazarus et al. (2018), the confounder process X_t is modelled using a dynamic factor model as

$$\begin{aligned} X_t &= \Lambda F_t + U_t \\ F_t &= \sum_{j=1}^{p_F} \Phi_j F_{t-j} + V_t, \qquad V_t \sim \mathcal{N}(0, I) \\ U_{i,t} &= \sum_{j=1}^{p_U} \phi_j U_{i,t-j} + \eta_t, \qquad \eta_{i,t} \sim \mathcal{N}(0, \sigma^2_{i,\eta}) \quad \text{for } i = 1, \ldots, \dim(X_t). \end{aligned}$$

Model orders are selected using the Bayesian Information Criterion (BIC). Specifically, we determine the GARCH(p, q) orders for ϵ_t, the number of latent factors F_t, the lag length p_F of the factor VAR, and the AR order p_U of the idiosyncratic components by minimizing the BIC. For the sample at hand, this procedure selects an ARCH(1) process for ϵ_t, one latent factor following a VAR(1) process, and univariate AR(1) processes for the components of U_t. Simulation samples are generated by drawing independently from the distributions of ϵ_t, V_t and η_t.

C Additional tables and figures

Table C2: Simulation results for K independent nonlinear baseline DGPs ($n = 12$ and $\sigma_\epsilon = 1.0$) and random forest nuisance function estimates

$h = 0$, $\theta_0^{(h)} = 0.3321$

	DML					RA				DR				LP			
T	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)	C_a(95%)	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)
125	-0.020	0.749	0.749	0.969	0.959	0.379	0.438	0.579	0.530	0.275	0.400	0.486	0.732	0.099	0.361	0.375	0.868
250	0.025	0.406	0.407	0.954	0.953	0.276	0.275	0.390	0.512	0.198	0.256	0.323	0.752	0.110	0.245	0.269	0.872
500	0.027	0.212	0.214	0.944	0.944	0.215	0.199	0.293	0.469	0.139	0.185	0.231	0.776	0.139	0.181	0.228	0.821
1'000	0.015	0.131	0.132	0.962	0.959	0.174	0.130	0.217	0.431	0.095	0.121	0.154	0.809	0.135	0.121	0.181	0.747
8'000	0.009	0.044	0.045	0.957	0.955	0.177	0.052	0.184	0.015	0.042	0.044	0.061	0.818	0.138	0.043	0.145	0.094

$h = 1$, $\theta_0^{(h)} = 0.1992$

	DML					RA				DR				LP			
T	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)	C_a(95%)	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)
125	0.017	0.762	0.762	0.958	0.953	0.412	0.439	0.602	0.467	0.296	0.402	0.499	0.729	0.072	0.369	0.376	0.904
250	0.057	0.398	0.402	0.971	0.951	0.306	0.284	0.418	0.441	0.215	0.263	0.340	0.741	0.099	0.269	0.287	0.912
500	0.053	0.220	0.226	0.949	0.941	0.251	0.197	0.319	0.330	0.163	0.182	0.244	0.708	0.135	0.190	0.233	0.872
1'000	0.037	0.142	0.147	0.951	0.951	0.211	0.130	0.248	0.236	0.119	0.123	0.171	0.731	0.128	0.134	0.186	0.829
8'000	0.014	0.045	0.047	0.934	0.934	0.219	0.052	0.225	0.001	0.052	0.045	0.069	0.748	0.142	0.046	0.149	0.134

$h = 2$, $\theta_0^{(h)} = 0.1195$

	DML					RA				DR				LP			
T	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)	C_a(95%)	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)
125	0.058	0.821	0.823	0.951	0.942	0.394	0.466	0.610	0.458	0.288	0.433	0.520	0.713	0.043	0.412	0.414	0.930
250	0.067	0.460	0.464	0.967	0.938	0.289	0.299	0.416	0.434	0.205	0.280	0.347	0.797	0.081	0.313	0.324	0.926
500	0.067	0.248	0.257	0.942	0.939	0.243	0.206	0.319	0.307	0.162	0.197	0.255	0.725	0.109	0.219	0.245	0.911
1'000	0.040	0.161	0.166	0.945	0.945	0.196	0.145	0.244	0.238	0.111	0.141	0.179	0.764	0.112	0.154	0.191	0.886
8'000	0.014	0.052	0.054	0.933	0.932	0.213	0.057	0.221	0.002	0.051	0.050	0.071	0.787	0.126	0.054	0.137	0.367

$h = 3$, $\theta_0^{(h)} = 0.0717$

	DML					RA				DR				LP			
T	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)	C_a(95%)	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)
125	0.057	0.849	0.851	0.943	0.939	0.347	0.487	0.598	0.475	0.257	0.459	0.526	0.756	0.026	0.461	0.461	0.930
250	0.072	0.483	0.489	0.954	0.936	0.257	0.326	0.416	0.416	0.188	0.310	0.363	0.816	0.060	0.350	0.355	0.929
500	0.066	0.273	0.281	0.943	0.941	0.216	0.224	0.311	0.337	0.147	0.216	0.261	0.759	0.090	0.243	0.259	0.937
1'000	0.045	0.181	0.187	0.935	0.932	0.181	0.159	0.241	0.252	0.107	0.156	0.189	0.786	0.094	0.177	0.201	0.908
8'000	0.013	0.059	0.061	0.931	0.931	0.195	0.063	0.205	0.005	0.047	0.057	0.074	0.818	0.107	0.061	0.123	0.579

$h = 4$, $\theta_0^{(h)} = 0.0430$

	DML					RA				DR				LP			
T	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)	C_a(95%)	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)
125	0.071	0.896	0.899	0.941	0.934	0.299	0.513	0.594	0.487	0.222	0.486	0.534	0.771	0.027	0.481	0.482	0.935
250	0.084	0.532	0.539	0.936	0.931	0.233	0.352	0.422	0.455	0.173	0.336	0.378	0.758	0.061	0.391	0.396	0.918
500	0.057	0.308	0.313	0.937	0.930	0.188	0.241	0.305	0.404	0.129	0.235	0.269	0.790	0.073	0.263	0.273	0.932
1'000	0.040	0.200	0.203	0.936	0.932	0.157	0.173	0.234	0.278	0.093	0.170	0.193	0.807	0.071	0.193	0.206	0.937
8'000	0.014	0.066	0.068	0.931	0.930	0.176	0.067	0.189	0.009	0.044	0.064	0.078	0.846	0.088	0.067	0.111	0.740

$h = 5$, $\theta_0^{(h)} = 0.0258$

	DML					RA				DR				LP			
T	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)	C_a(95%)	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)
125	0.069	0.951	0.954	0.933	0.929	0.254	0.542	0.598	0.519	0.192	0.516	0.551	0.775	0.026	0.523	0.524	0.922
250	0.081	0.634	0.639	0.931	0.929	0.199	0.364	0.415	0.493	0.151	0.352	0.383	0.791	0.064	0.420	0.425	0.921
500	0.056	0.337	0.342	0.936	0.931	0.165	0.264	0.312	0.414	0.117	0.257	0.283	0.805	0.060	0.285	0.292	0.941
1'000	0.039	0.217	0.221	0.941	0.939	0.141	0.186	0.233	0.340	0.086	0.183	0.202	0.834	0.051	0.208	0.214	0.938
8'000	0.013	0.070	0.072	0.931	0.931	0.155	0.072	0.171	0.027	0.039	0.069	0.079	0.875	0.069	0.073	0.100	0.844

NOTE: The table depicts simulation results across $N = 1'000$ draws obtained for the scenario with K independent stochastic processes, with $K = 10$. Except for the LP estimator, nuisance functions are estimated with random forest. For the DML estimator we set $k_T = T/10$ to obtain estimation samples of the same size as in the setting with one stochastic process. For sample size $T = 125$, probabilities are winsorized at 1%. The parameters of the data generating process are $n = 12$, $\sigma_\epsilon = 1.0$, $\gamma = 0.6$, $p = 2$, $q = 1$, $\sigma_u = 1.0$, $\alpha_A = 0.3$, $\alpha_M = 0.3$, $\rho_A = 0.35$, $\rho_M = 0.7$, $\beta_1 = 0.3$, $\beta_2 = 0.5$. $C_a(\cdot)$ and $C_b(\cdot)$ in the tables denote the coverage at the given confidence level using asymptotic and fixed-bandwidth critical values, respectively.

Table C3: Simulation results for a nonlinear DGP ($n = 20$, $\sigma_\epsilon = 1.0$) and random forest nuisance function estimates

$h = 0$, $\theta_0^{(h)} = 0.3333$

	DML					RA				DR				LP			
T	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)	C_a(95%)	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)
125	0.180	0.791	0.811	0.937	0.920	0.382	0.454	0.593	0.524	0.293	0.418	0.511	0.712	0.101	0.370	0.383	0.823
250	0.082	0.385	0.393	0.959	0.945	0.279	0.289	0.402	0.519	0.210	0.269	0.342	0.719	0.117	0.248	0.274	0.857
500	0.037	0.213	0.216	0.953	0.940	0.199	0.194	0.278	0.529	0.136	0.183	0.228	0.780	0.122	0.175	0.213	0.850
1'000	0.031	0.133	0.136	0.952	0.949	0.165	0.127	0.208	0.465	0.098	0.120	0.155	0.807	0.126	0.118	0.173	0.769
8'000	0.010	0.044	0.045	0.934	0.934	0.156	0.048	0.163	0.018	0.038	0.044	0.058	0.835	0.133	0.043	0.140	0.123

$h = 1$, $\theta_0^{(h)} = 0.2000$

	DML					RA				DR				LP			
T	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)	C_a(95%)	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)
125	0.207	0.827	0.852	0.935	0.920	0.418	0.449	0.613	0.488	0.317	0.411	0.519	0.698	0.058	0.381	0.386	0.884
250	0.103	0.393	0.406	0.965	0.952	0.301	0.283	0.413	0.447	0.220	0.261	0.341	0.724	0.088	0.268	0.282	0.900
500	0.061	0.224	0.232	0.948	0.945	0.235	0.194	0.304	0.381	0.158	0.183	0.242	0.712	0.114	0.186	0.218	0.898
1'000	0.053	0.144	0.154	0.931	0.929	0.203	0.134	0.243	0.283	0.121	0.127	0.175	0.734	0.126	0.132	0.182	0.827
8'000	0.020	0.045	0.049	0.925	0.924	0.202	0.047	0.207	0.001	0.052	0.043	0.068	0.738	0.139	0.047	0.147	0.153

$h = 2$, $\theta_0^{(h)} = 0.1200$

	DML					RA				DR				LP			
T	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)	C_a(95%)	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)
125	0.222	0.878	0.905	0.906	0.898	0.378	0.477	0.609	0.488	0.283	0.439	0.522	0.728	0.019	0.424	0.425	0.903
250	0.127	0.433	0.451	0.937	0.926	0.294	0.293	0.415	0.434	0.218	0.272	0.349	0.747	0.075	0.296	0.305	0.936
500	0.075	0.249	0.260	0.941	0.937	0.224	0.209	0.307	0.357	0.153	0.199	0.251	0.716	0.095	0.214	0.234	0.927
1'000	0.049	0.164	0.171	0.928	0.927	0.185	0.145	0.235	0.260	0.106	0.140	0.175	0.764	0.098	0.149	0.179	0.901
8'000	0.022	0.051	0.056	0.919	0.919	0.198	0.052	0.205	0.001	0.052	0.049	0.072	0.758	0.127	0.054	0.138	0.362

$h = 3$, $\theta_0^{(h)} = 0.0720$

	DML					RA				DR				LP			
T	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)	C_a(95%)	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)
125	0.228	0.987	1.013	0.924	0.901	0.353	0.486	0.601	0.534	0.271	0.450	0.526	0.747	0.022	0.454	0.454	0.914
250	0.137	0.487	0.506	0.932	0.924	0.266	0.328	0.422	0.467	0.201	0.311	0.370	0.746	0.053	0.340	0.344	0.925
500	0.083	0.289	0.301	0.941	0.936	0.216	0.228	0.314	0.376	0.152	0.218	0.266	0.735	0.085	0.244	0.259	0.925
1'000	0.058	0.176	0.186	0.940	0.939	0.180	0.152	0.236	0.287	0.109	0.149	0.185	0.784	0.089	0.165	0.187	0.926
8'000	0.024	0.058	0.062	0.917	0.917	0.185	0.056	0.194	0.001	0.051	0.055	0.075	0.807	0.111	0.061	0.127	0.563

$h = 4$, $\theta_0^{(h)} = 0.0432$

	DML					RA				DR				LP			
T	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)	C_a(95%)	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)
125	0.225	1.014	1.038	0.912	0.902	0.315	0.520	0.608	0.555	0.248	0.491	0.550	0.752	0.022	0.500	0.500	0.894
250	0.130	0.543	0.558	0.927	0.906	0.246	0.350	0.427	0.499	0.186	0.334	0.383	0.753	0.038	0.365	0.367	0.922
500	0.079	0.321	0.331	0.933	0.933	0.194	0.241	0.310	0.431	0.139	0.233	0.271	0.771	0.062	0.261	0.268	0.935
1'000	0.058	0.194	0.202	0.948	0.944	0.166	0.166	0.234	0.309	0.101	0.162	0.191	0.811	0.072	0.181	0.195	0.937
8'000	0.022	0.065	0.068	0.930	0.929	0.168	0.062	0.179	0.014	0.047	0.061	0.077	0.842	0.092	0.067	0.113	0.736

$h = 5$, $\theta_0^{(h)} = 0.0259$

	DML					RA				DR				LP			
T	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)	C_a(95%)	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)
125	0.208	1.055	1.075	0.914	0.903	0.250	0.526	0.582	0.583	0.196	0.496	0.533	0.801	-0.002	0.515	0.515	0.911
250	0.136	0.593	0.608	0.919	0.902	0.217	0.361	0.421	0.548	0.170	0.343	0.383	0.778	0.027	0.381	0.382	0.942
500	0.070	0.339	0.346	0.925	0.925	0.167	0.253	0.303	0.482	0.119	0.245	0.272	0.802	0.042	0.277	0.280	0.937
1'000	0.055	0.213	0.220	0.934	0.932	0.151	0.178	0.234	0.356	0.094	0.175	0.198	0.838	0.055	0.199	0.206	0.935
8'000	0.021	0.070	0.073	0.934	0.933	0.152	0.067	0.166	0.025	0.043	0.067	0.079	0.859	0.075	0.071	0.103	0.822

NOTE: The table depicts simulation results across $N = 1'000$ draws obtained for the scenario with one stochastic process. Except for the LP estimator, nuisance functions are estimated with random forest. For the DML estimator we use 10-fold cross-fitting and set $k_T = T/10$. For sample size $T = 125$, probabilities are winsorized at 1%. The parameters of the data generating process are $n = 20$, $\sigma_\epsilon = 1.0$, $\gamma = 0.6$, $p = 2$, $q = 1$, $\sigma_u = 1.0$, $\alpha_A = 0.3$, $\alpha_M = 0.3$, $\rho_A = 0.35$, $\rho_M = 0.7$, $\beta_1 = 0.3$, $\beta_2 = 0.5$. $C_a(\cdot)$ and $C_b(\cdot)$ in the tables denote the coverage at the given confidence level using asymptotic and fixed-bandwidth critical values, respectively.

Table C4: Simulation results for a nonlinear DGP ($n = 12$, $\sigma_\epsilon = 3.0$) and random forest nuisance function estimates

$h = 0$, $\theta_0^{(h)} = 0.3321$

	DML					RA				DR				LP			
T	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)	C_a(95%)	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)
125	0.055	1.287	1.288	0.941	0.930	0.364	0.789	0.869	0.497	0.263	0.750	0.794	0.758	0.082	0.713	0.717	0.907
250	0.060	0.779	0.781	0.953	0.943	0.273	0.534	0.600	0.528	0.203	0.516	0.555	0.800	0.142	0.488	0.508	0.917
500	0.015	0.429	0.429	0.951	0.945	0.188	0.368	0.413	0.471	0.120	0.357	0.376	0.824	0.142	0.342	0.370	0.917
1'000	0.019	0.285	0.285	0.946	0.945	0.161	0.264	0.309	0.429	0.088	0.260	0.274	0.838	0.146	0.245	0.285	0.871
8'000	0.003	0.091	0.091	0.951	0.950	0.161	0.089	0.184	0.106	0.033	0.088	0.094	0.918	0.143	0.083	0.165	0.599

$h = 1$, $\theta_0^{(h)} = 0.1992$

	DML					RA				DR				LP			
T	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)	C_a(95%)	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)
125	0.101	1.430	1.433	0.918	0.918	0.370	0.846	0.924	0.523	0.271	0.815	0.859	0.759	0.060	0.782	0.784	0.922
250	0.047	1.051	1.052	0.942	0.936	0.293	0.580	0.650	0.459	0.210	0.565	0.603	0.769	0.122	0.555	0.568	0.931
500	0.043	0.474	0.476	0.946	0.945	0.214	0.388	0.443	0.429	0.134	0.381	0.404	0.832	0.128	0.392	0.412	0.920
1'000	0.049	0.310	0.314	0.949	0.947	0.198	0.279	0.342	0.345	0.111	0.275	0.297	0.842	0.153	0.271	0.311	0.914
8'000	0.010	0.102	0.102	0.947	0.946	0.207	0.097	0.228	0.047	0.045	0.098	0.108	0.913	0.153	0.096	0.181	0.642

$h = 2$, $\theta_0^{(h)} = 0.1195$

	DML					RA				DR				LP			
T	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)	C_a(95%)	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)
125	0.159	1.382	1.391	0.932	0.921	0.352	0.847	0.917	0.512	0.258	0.816	0.856	0.784	0.033	0.814	0.815	0.935
250	0.044	1.106	1.107	0.932	0.925	0.258	0.600	0.653	0.497	0.182	0.592	0.619	0.777	0.080	0.599	0.604	0.925
500	0.059	0.503	0.507	0.941	0.938	0.202	0.414	0.460	0.417	0.130	0.408	0.428	0.814	0.111	0.420	0.434	0.933
1'000	0.058	0.340	0.345	0.931	0.931	0.194	0.302	0.360	0.336	0.114	0.300	0.321	0.814	0.143	0.302	0.334	0.913
8'000	0.010	0.108	0.109	0.962	0.962	0.199	0.102	0.224	0.040	0.043	0.104	0.112	0.908	0.138	0.104	0.173	0.721

$h = 3$, $\theta_0^{(h)} = 0.0717$

	DML					RA				DR				LP			
T	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)	C_a(95%)	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)
125	0.116	1.482	1.486	0.936	0.923	0.284	0.852	0.898	0.556	0.201	0.821	0.845	0.795	-0.010	0.825	0.825	0.936
250	0.074	0.860	0.864	0.930	0.928	0.233	0.602	0.646	0.481	0.160	0.586	0.608	0.795	0.058	0.600	0.603	0.934
500	0.062	0.536	0.540	0.931	0.923	0.191	0.432	0.472	0.409	0.127	0.429	0.447	0.800	0.102	0.448	0.459	0.929
1'000	0.060	0.353	0.358	0.924	0.922	0.183	0.310	0.360	0.331	0.112	0.311	0.331	0.815	0.128	0.319	0.344	0.911
8'000	0.010	0.114	0.114	0.950	0.950	0.180	0.107	0.210	0.046	0.040	0.110	0.117	0.905	0.119	0.110	0.162	0.808

$h = 4$, $\theta_0^{(h)} = 0.0430$

	DML					RA				DR				LP			
T	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)	C_a(95%)	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)
125	0.145	1.489	1.496	0.935	0.915	0.252	0.832	0.869	0.549	0.187	0.811	0.832	0.831	0.012	0.833	0.833	0.937
250	0.041	1.358	1.358	0.916	0.916	0.208	0.639	0.672	0.485	0.150	0.623	0.641	0.787	0.058	0.631	0.634	0.935
500	0.039	0.576	0.578	0.938	0.926	0.151	0.448	0.473	0.421	0.097	0.443	0.453	0.821	0.063	0.456	0.461	0.929
1'000	0.055	0.363	0.367	0.938	0.936	0.160	0.316	0.354	0.345	0.098	0.315	0.330	0.835	0.104	0.324	0.341	0.920
8'000	0.010	0.117	0.118	0.949	0.949	0.159	0.110	0.194	0.056	0.037	0.113	0.119	0.911	0.100	0.110	0.148	0.859

$h = 5$, $\theta_0^{(h)} = 0.0258$

	DML					RA				DR				LP			
T	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)	C_a(95%)	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)
125	0.167	1.619	1.627	0.915	0.898	0.253	0.897	0.932	0.558	0.199	0.874	0.897	0.813	0.053	0.913	0.914	0.922
250	0.050	1.062	1.063	0.920	0.912	0.176	0.650	0.674	0.469	0.125	0.638	0.650	0.775	0.046	0.643	0.645	0.931
500	0.025	0.587	0.588	0.950	0.941	0.122	0.446	0.463	0.450	0.078	0.444	0.451	0.829	0.045	0.458	0.461	0.938
1'000	0.042	0.363	0.366	0.947	0.941	0.129	0.313	0.339	0.357	0.076	0.313	0.322	0.855	0.074	0.323	0.331	0.950
8'000	0.003	0.121	0.121	0.942	0.941	0.134	0.112	0.175	0.082	0.026	0.115	0.118	0.919	0.077	0.113	0.136	0.912

NOTE: The table depicts simulation results across $N = 1'000$ draws obtained for the scenario with one stochastic process. Except for the LP estimator, nuisance functions are estimated with random forest. For the DML estimator we use 10-fold cross-fitting and set $k_T = T/10$. For sample size $T = 125$, probabilities are winsorized at 1%. The parameters of the data generating process are $n = 12$, $\sigma_\epsilon = 3.0$, $\gamma = 0.6$, $p = 2$, $q = 1$, $\sigma_u = 1.0$, $\alpha_A = 0.3$, $\alpha_M = 0.3$, $\rho_A = 0.35$, $\rho_M = 0.7$, $\beta_1 = 0.3$, $\beta_2 = 0.5$. $C_a(\cdot)$ and $C_b(\cdot)$ in the tables denote the coverage at the given confidence level using asymptotic and fixed-bandwidth critical values, respectively.

Table C5: Simulation results for an empirically calibrated DGP (see Section B) and random forest nuisance function estimates

	$h = 0, \theta_0^{(h)} = 0.2021$															
	DML					RA				DR				LP			
T	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)	C_a(95%)	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)
125	0.058	0.343	0.348	0.966	0.948	0.244	0.273	0.366	0.279	0.203	0.263	0.332	0.681	0.146	0.260	0.298	0.883
250	0.042	0.207	0.211	0.935	0.931	0.210	0.190	0.283	0.238	0.168	0.184	0.249	0.629	0.153	0.189	0.243	0.845
500	0.038	0.130	0.136	0.949	0.945	0.177	0.129	0.219	0.242	0.130	0.125	0.180	0.641	0.154	0.127	0.200	0.752
1'000	0.034	0.096	0.101	0.919	0.919	0.159	0.090	0.182	0.148	0.105	0.088	0.137	0.625	0.152	0.091	0.177	0.580
8'000	-0.002	0.031	0.032	0.949	0.949	0.153	0.033	0.156	0.000	0.035	0.031	0.047	0.760	0.153	0.033	0.157	0.001
	$h = 1, \theta_0^{(h)} = 0.0521$															
	DML					RA				DR				LP			
T	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)	C_a(95%)	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)
125	0.040	0.347	0.349	0.965	0.939	0.108	0.271	0.291	0.330	0.088	0.267	0.282	0.791	0.029	0.274	0.276	0.941
250	0.025	0.245	0.247	0.962	0.944	0.100	0.192	0.216	0.258	0.079	0.191	0.206	0.789	0.041	0.199	0.203	0.936
500	0.027	0.151	0.153	0.933	0.929	0.084	0.137	0.161	0.265	0.062	0.137	0.151	0.792	0.043	0.141	0.147	0.918
1'000	0.019	0.100	0.102	0.947	0.946	0.073	0.095	0.119	0.194	0.047	0.095	0.106	0.808	0.042	0.097	0.106	0.927
8'000	-0.000	0.035	0.035	0.954	0.953	0.072	0.034	0.079	0.019	0.015	0.034	0.037	0.894	0.040	0.034	0.053	0.780
	$h = 2, \theta_0^{(h)} = 0.0134$															
	DML					RA				DR				LP			
T	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)	C_a(95%)	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)
125	0.037	0.362	0.364	0.969	0.947	0.067	0.263	0.271	0.309	0.054	0.261	0.266	0.860	0.007	0.272	0.272	0.946
250	0.008	0.228	0.228	0.955	0.944	0.045	0.191	0.196	0.288	0.034	0.190	0.193	0.814	0.003	0.196	0.196	0.941
500	0.010	0.150	0.150	0.945	0.943	0.038	0.136	0.141	0.245	0.026	0.137	0.139	0.830	0.005	0.138	0.138	0.951
1'000	0.006	0.108	0.108	0.934	0.934	0.034	0.099	0.105	0.184	0.020	0.101	0.103	0.828	0.007	0.101	0.101	0.938
8'000	0.000	0.036	0.036	0.954	0.953	0.036	0.034	0.050	0.050	0.008	0.035	0.036	0.916	0.010	0.035	0.037	0.936
	$h = 3, \theta_0^{(h)} = 0.0035$															
	DML					RA				DR				LP			
T	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)	C_a(95%)	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)
125	0.032	0.362	0.363	0.950	0.936	0.042	0.282	0.285	0.331	0.034	0.277	0.279	0.811	0.001	0.285	0.285	0.932
250	0.009	0.238	0.239	0.935	0.929	0.027	0.196	0.198	0.284	0.020	0.197	0.198	0.790	-0.006	0.199	0.199	0.940
500	0.005	0.151	0.151	0.947	0.945	0.023	0.134	0.136	0.231	0.015	0.135	0.136	0.840	-0.003	0.135	0.135	0.957
1'000	0.003	0.106	0.106	0.939	0.937	0.022	0.096	0.098	0.233	0.013	0.098	0.099	0.852	0.000	0.098	0.098	0.947
8'000	0.000	0.037	0.037	0.945	0.945	0.022	0.034	0.040	0.054	0.005	0.036	0.036	0.925	0.002	0.035	0.035	0.944
	$h = 4, \theta_0^{(h)} = 0.0009$															
	DML					RA				DR				LP			
T	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)	C_a(95%)	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)
125	0.023	0.376	0.376	0.945	0.933	0.034	0.290	0.292	0.315	0.026	0.286	0.287	0.777	-0.001	0.296	0.296	0.925
250	0.006	0.228	0.228	0.944	0.938	0.025	0.190	0.192	0.316	0.020	0.191	0.192	0.810	0.003	0.195	0.195	0.940
500	0.006	0.155	0.155	0.940	0.936	0.017	0.141	0.142	0.252	0.011	0.142	0.142	0.815	-0.005	0.145	0.145	0.938
1'000	0.006	0.103	0.103	0.956	0.955	0.020	0.093	0.095	0.185	0.012	0.095	0.096	0.871	0.001	0.095	0.095	0.951
8'000	-0.000	0.036	0.036	0.953	0.953	0.016	0.033	0.037	0.065	0.003	0.035	0.035	0.936	0.000	0.035	0.035	0.956
	$h = 5, \theta_0^{(h)} = 0.0002$															
	DML					RA				DR				LP			
T	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)	C_a(95%)	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)
125	0.024	0.370	0.371	0.944	0.930	0.022	0.283	0.284	0.305	0.017	0.281	0.282	0.909	-0.001	0.285	0.285	0.935
250	0.012	0.237	0.238	0.936	0.934	0.016	0.194	0.194	0.304	0.013	0.193	0.193	0.796	-0.004	0.197	0.197	0.946
500	0.006	0.152	0.152	0.960	0.954	0.017	0.135	0.136	0.226	0.012	0.136	0.136	0.851	-0.003	0.137	0.137	0.957
1'000	0.006	0.104	0.105	0.953	0.950	0.017	0.096	0.097	0.171	0.011	0.097	0.097	0.869	0.000	0.096	0.096	0.953
8'000	0.000	0.036	0.036	0.950	0.950	0.013	0.033	0.036	0.068	0.003	0.035	0.035	0.931	0.000	0.034	0.034	0.960

NOTE: The table depicts simulation results across $N = 1'000$ draws obtained for the scenario with one stochastic process. Except for the LP estimator, nuisance functions are estimated with random forest. For the DML estimator we use 10-fold cross-fitting and set $k_T = T/10$. For sample size $T = 125$, probabilities are winsorized at 1%. The parameters of the data generating process are empirically calibrated using monthly U.S. data from 1982 to 2012 obtained from the empirical study in Angrist et al. (2018). $C_a(\cdot)$ and $C_b(\cdot)$ in the tables denote the coverage at the given confidence level using asymptotic and fixed-bandwidth critical values, respectively.

Table C6: Simulation results for a baseline nonlinear DGP ($n = 12$, $\sigma_\epsilon = 1.0$) and gradient boosting nuisance function estimates

$h = 0$, $\theta_0^{(h)} = 0.3321$

	DML					RA				DR				LP			
T	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)	C_a(95%)	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)
125	0.166	1.285	1.295	0.937	0.937	0.071	0.525	0.530	0.800	-0.019	0.514	0.515	0.900	0.090	0.357	0.368	0.871
250	0.090	0.631	0.638	0.922	0.922	0.111	0.391	0.406	0.732	0.081	0.313	0.324	0.883	0.128	0.251	0.282	0.847
500	0.043	0.269	0.273	0.961	0.955	0.280	0.242	0.370	0.398	0.158	0.199	0.254	0.805	0.135	0.173	0.220	0.829
1'000	0.028	0.158	0.160	0.942	0.939	0.073	0.189	0.203	0.662	0.038	0.141	0.146	0.919	0.138	0.126	0.187	0.730
8'000	0.008	0.040	0.041	0.960	0.960	0.115	0.042	0.122	0.089	0.014	0.040	0.043	0.952	0.131	0.046	0.139	0.141

$h = 1$, $\theta_0^{(h)} = 0.1992$

	DML					RA				DR				LP			
T	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)	C_a(95%)	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)
125	0.166	1.322	1.332	0.952	0.938	0.090	0.547	0.554	0.751	0.009	0.514	0.514	0.898	0.060	0.362	0.367	0.923
250	0.105	0.641	0.650	0.955	0.954	0.135	0.398	0.421	0.701	0.090	0.315	0.327	0.877	0.111	0.265	0.287	0.914
500	0.067	0.266	0.275	0.959	0.957	0.320	0.236	0.397	0.286	0.175	0.196	0.263	0.777	0.123	0.193	0.229	0.887
1'000	0.044	0.154	0.160	0.953	0.952	0.097	0.187	0.211	0.553	0.052	0.137	0.146	0.916	0.139	0.133	0.193	0.792
8'000	0.016	0.042	0.045	0.937	0.936	0.134	0.044	0.141	0.013	0.021	0.042	0.047	0.928	0.144	0.053	0.153	0.183

$h = 2$, $\theta_0^{(h)} = 0.1195$

	DML					RA				DR				LP			
T	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)	C_a(95%)	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)
125	0.153	1.439	1.447	0.959	0.935	0.104	0.561	0.570	0.732	-0.006	0.541	0.541	0.913	0.031	0.410	0.411	0.931
250	0.107	0.723	0.731	0.944	0.937	0.149	0.426	0.451	0.662	0.082	0.329	0.339	0.887	0.078	0.306	0.316	0.927
500	0.072	0.300	0.308	0.952	0.952	0.310	0.266	0.408	0.346	0.171	0.219	0.278	0.800	0.104	0.221	0.244	0.911
1'000	0.051	0.176	0.183	0.930	0.926	0.104	0.210	0.234	0.478	0.058	0.156	0.167	0.898	0.128	0.160	0.204	0.854
8'000	0.015	0.049	0.051	0.943	0.943	0.131	0.051	0.141	0.014	0.019	0.048	0.052	0.935	0.128	0.055	0.140	0.352

$h = 3$, $\theta_0^{(h)} = 0.0717$

	DML					RA				DR				LP			
T	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)	C_a(95%)	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)
125	0.132	1.513	1.519	0.958	0.944	0.124	0.622	0.634	0.694	0.007	0.591	0.591	0.894	0.012	0.448	0.448	0.937
250	0.097	0.797	0.803	0.933	0.930	0.143	0.476	0.497	0.611	0.074	0.364	0.371	0.901	0.058	0.340	0.345	0.934
500	0.078	0.327	0.337	0.953	0.946	0.289	0.287	0.407	0.391	0.163	0.238	0.289	0.840	0.087	0.249	0.264	0.915
1'000	0.054	0.196	0.204	0.949	0.946	0.113	0.228	0.255	0.442	0.058	0.168	0.178	0.915	0.112	0.178	0.210	0.902
8'000	0.014	0.055	0.057	0.946	0.946	0.128	0.058	0.141	0.021	0.017	0.055	0.057	0.937	0.117	0.061	0.131	0.479

$h = 4$, $\theta_0^{(h)} = 0.0430$

	DML					RA				DR				LP			
T	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)	C_a(95%)	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)
125	0.150	1.681	1.688	0.960	0.937	0.125	0.678	0.689	0.695	0.041	0.696	0.697	0.902	0.021	0.496	0.497	0.935
250	0.069	0.867	0.870	0.927	0.924	0.142	0.510	0.529	0.591	0.077	0.405	0.412	0.873	0.049	0.377	0.380	0.922
500	0.068	0.363	0.369	0.952	0.940	0.245	0.299	0.387	0.447	0.148	0.261	0.300	0.852	0.062	0.270	0.277	0.929
1'000	0.051	0.219	0.225	0.942	0.942	0.110	0.247	0.271	0.408	0.054	0.181	0.189	0.934	0.092	0.192	0.213	0.921
8'000	0.014	0.061	0.063	0.948	0.948	0.126	0.063	0.141	0.026	0.017	0.060	0.063	0.939	0.100	0.067	0.121	0.662

$h = 5$, $\theta_0^{(h)} = 0.0258$

	DML					RA				DR				LP			
T	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)	C_a(95%)	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)
125	0.099	1.794	1.796	0.949	0.927	0.148	0.705	0.720	0.669	0.082	0.729	0.734	0.898	0.032	0.551	0.552	0.921
250	0.077	0.937	0.940	0.931	0.926	0.119	0.532	0.545	0.584	0.068	0.421	0.427	0.884	0.036	0.394	0.395	0.918
500	0.063	0.384	0.389	0.951	0.943	0.216	0.308	0.376	0.491	0.131	0.271	0.301	0.862	0.047	0.283	0.287	0.943
1'000	0.039	0.233	0.236	0.946	0.938	0.097	0.256	0.274	0.445	0.047	0.193	0.198	0.927	0.071	0.202	0.214	0.933
8'000	0.010	0.068	0.068	0.951	0.951	0.116	0.071	0.136	0.040	0.013	0.067	0.068	0.941	0.082	0.078	0.113	0.789

NOTE: The table depicts simulation results across $N = 1'000$ draws obtained for the scenario with one stochastic process. Except for the LP estimator, nuisance functions are estimated with gradient boosting. For the DML estimator we use 10-fold cross-fitting and set $k_T = T/10$. For sample size $T = 125$, probabilities are winsorized at 1%. The parameters of the data generating process are $n = 12$, $\sigma_\epsilon = 1.0$, $\gamma = 0.6$, $p = 2$, $q = 1$, $\sigma_u = 1.0$, $\alpha_A = 0.3$, $\alpha_M = 0.3$, $\rho_A = 0.35$, $\rho_M = 0.7$, $\beta_1 = 0.3$, $\beta_2 = 0.5$. $C_a(\cdot)$ and $C_b(\cdot)$ in the tables denote the coverage at the given confidence level using asymptotic and fixed-bandwidth critical values, respectively.

Table C7: Simulation results for a linear DGP ($n = 12$, $\sigma_\epsilon = 1.0$) and random forest nuisance function estimates

$h = 0$, $\theta_0^{(h)} = 0.3321$

	DML					RA				DR				LP			
T	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)	C_a(95%)	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)
125	0.121	1.375	1.381	0.927	0.918	0.851	0.550	1.013	0.270	0.669	0.498	0.834	0.503	-0.006	0.219	0.219	0.911
250	0.060	1.043	1.044	0.944	0.935	0.675	0.325	0.749	0.095	0.525	0.292	0.601	0.334	0.004	0.146	0.146	0.956
500	0.069	0.348	0.355	0.959	0.941	0.537	0.203	0.574	0.031	0.388	0.185	0.430	0.240	-0.001	0.106	0.106	0.947
1'000	0.062	0.164	0.176	0.944	0.942	0.445	0.126	0.463	0.000	0.280	0.116	0.303	0.186	-0.000	0.074	0.074	0.956
8'000	0.019	0.040	0.044	0.942	0.942	0.381	0.066	0.386	0.000	0.108	0.038	0.114	0.152	-0.001	0.025	0.025	0.940

$h = 1$, $\theta_0^{(h)} = 0.1992$

	DML					RA				DR				LP			
T	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)	C_a(95%)	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)
125	0.173	1.539	1.549	0.916	0.900	0.893	0.590	1.070	0.287	0.708	0.542	0.892	0.485	-0.030	0.317	0.319	0.918
250	0.067	1.273	1.275	0.943	0.921	0.732	0.364	0.817	0.119	0.567	0.332	0.657	0.354	-0.005	0.224	0.224	0.934
500	0.105	0.372	0.386	0.945	0.936	0.597	0.223	0.638	0.023	0.433	0.207	0.480	0.262	-0.008	0.151	0.151	0.947
1'000	0.094	0.195	0.216	0.913	0.910	0.509	0.148	0.530	0.002	0.324	0.139	0.352	0.200	0.002	0.109	0.110	0.945
8'000	0.028	0.050	0.057	0.922	0.922	0.447	0.073	0.453	0.000	0.128	0.046	0.136	0.163	0.001	0.038	0.038	0.950

$h = 2$, $\theta_0^{(h)} = 0.1195$

	DML					RA				DR				LP			
T	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)	C_a(95%)	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)
125	0.217	1.495	1.511	0.917	0.892	0.818	0.659	1.050	0.372	0.649	0.615	0.894	0.610	-0.042	0.454	0.456	0.918
250	0.082	1.058	1.061	0.917	0.905	0.656	0.418	0.778	0.238	0.510	0.393	0.644	0.546	-0.014	0.326	0.326	0.931
500	0.114	0.412	0.427	0.944	0.932	0.547	0.270	0.610	0.102	0.399	0.257	0.475	0.483	-0.009	0.218	0.218	0.954
1'000	0.102	0.237	0.258	0.898	0.895	0.470	0.192	0.508	0.033	0.303	0.185	0.355	0.429	0.004	0.165	0.165	0.936
8'000	0.027	0.066	0.072	0.926	0.924	0.420	0.084	0.428	0.000	0.119	0.061	0.134	0.457	0.000	0.055	0.055	0.957

$h = 3$, $\theta_0^{(h)} = 0.0717$

	DML					RA				DR				LP			
T	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)	C_a(95%)	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)
125	0.221	1.641	1.656	0.920	0.895	0.696	0.709	0.993	0.431	0.554	0.669	0.869	0.724	-0.051	0.562	0.564	0.917
250	0.055	1.522	1.523	0.912	0.899	0.569	0.488	0.750	0.346	0.438	0.463	0.637	0.660	-0.019	0.418	0.418	0.913
500	0.113	0.462	0.475	0.951	0.937	0.478	0.310	0.570	0.220	0.352	0.301	0.463	0.642	-0.006	0.281	0.281	0.945
1'000	0.106	0.275	0.295	0.913	0.902	0.413	0.227	0.471	0.100	0.270	0.223	0.350	0.590	0.007	0.209	0.209	0.935
8'000	0.027	0.082	0.086	0.938	0.937	0.376	0.092	0.387	0.000	0.107	0.077	0.132	0.638	0.000	0.071	0.071	0.952

$h = 4$, $\theta_0^{(h)} = 0.0430$

	DML					RA				DR				LP			
T	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)	C_a(95%)	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)
125	0.283	1.587	1.612	0.892	0.892	0.594	0.744	0.952	0.530	0.478	0.714	0.860	0.807	-0.023	0.660	0.661	0.915
250	0.092	1.054	1.058	0.907	0.897	0.476	0.545	0.724	0.409	0.369	0.524	0.640	0.715	-0.016	0.493	0.493	0.917
500	0.103	0.489	0.499	0.961	0.935	0.396	0.345	0.526	0.335	0.291	0.337	0.446	0.733	-0.016	0.326	0.326	0.957
1'000	0.102	0.318	0.334	0.914	0.909	0.357	0.261	0.442	0.194	0.237	0.257	0.350	0.693	0.010	0.247	0.247	0.937
8'000	0.023	0.097	0.100	0.942	0.941	0.329	0.100	0.343	0.002	0.092	0.091	0.130	0.756	-0.001	0.084	0.084	0.948

$h = 5$, $\theta_0^{(h)} = 0.0258$

	DML					RA				DR				LP			
T	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)	C_a(95%)	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)
125	0.280	1.696	1.719	0.902	0.886	0.500	0.809	0.951	0.560	0.411	0.777	0.879	0.794	0.001	0.738	0.738	0.919
250	0.091	1.178	1.182	0.923	0.916	0.407	0.579	0.708	0.489	0.313	0.558	0.640	0.765	-0.009	0.532	0.532	0.927
500	0.091	0.531	0.539	0.958	0.938	0.327	0.377	0.499	0.402	0.241	0.370	0.442	0.779	-0.020	0.364	0.364	0.953
1'000	0.094	0.348	0.360	0.913	0.912	0.303	0.283	0.414	0.270	0.202	0.280	0.346	0.757	0.008	0.272	0.272	0.946
8'000	0.019	0.109	0.111	0.934	0.934	0.282	0.107	0.302	0.009	0.077	0.102	0.128	0.821	-0.003	0.096	0.096	0.946

NOTE: The table depicts simulation results across $N = 1'000$ draws obtained for the scenario with one stochastic process. The outcome variable is generated from a linear DGP, i.e. $b(X_t) = 0.5 \sum_{i=1}^{5} X_{i,t}$ and $\tau(X_t) = \theta_0^{(0)}$. Except for the LP estimator, nuisance functions are estimated with random forest. For the DML estimator we use 10-fold cross-fitting and set $k_T = T/10$. For sample size $T = 125$, probabilities are winsorized at 1%. The parameters of the data generating process are $n = 12$, $\sigma_\epsilon = 1.0$, $\gamma = 0.6$, $p = 2$, $q = 1$, $\sigma_u = 1.0$, $\alpha_A = 0.3$, $\alpha_M = 0.3$, $\rho_A = 0.35$, $\rho_M = 0.7$, $\beta_1 = 0.3$, $\beta_2 = 0.5$. $C_a(\cdot)$ and $C_b(\cdot)$ in the tables denote the coverage at the given confidence level using asymptotic and fixed-bandwidth critical values, respectively.

Table C8: Simulation results for a linear DGP with interactions ($n = 12$, $\sigma_\epsilon = 1.0$) and random forest nuisance function estimates

$h = 0, \theta_0^{(h)} = 0.3321$

	DML					RA				DR				LP			
T	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)	C_a(95%)	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)
125	0.115	1.926	1.930	0.810	0.786	0.900	1.076	1.403	0.404	0.714	1.020	1.245	0.513	0.556	0.858	1.023	0.579
250	0.079	1.248	1.251	0.814	0.810	0.700	0.726	1.009	0.468	0.551	0.702	0.892	0.580	0.595	0.589	0.837	0.527
500	0.085	0.618	0.624	0.839	0.830	0.571	0.511	0.767	0.473	0.419	0.495	0.649	0.633	0.624	0.404	0.743	0.364
1'000	0.084	0.395	0.403	0.856	0.849	0.479	0.367	0.603	0.487	0.312	0.359	0.476	0.684	0.632	0.297	0.698	0.206
8'000	0.025	0.066	0.070	0.947	0.945	0.416	0.086	0.425	0.000	0.124	0.064	0.140	0.460	0.329	0.053	0.333	0.000

$h = 1, \theta_0^{(h)} = 0.1992$

	DML					RA				DR				LP			
T	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)	C_a(95%)	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)
125	0.173	1.778	1.787	0.863	0.843	0.821	0.825	1.164	0.350	0.627	0.759	0.985	0.543	0.308	0.592	0.667	0.771
250	0.113	1.003	1.009	0.881	0.878	0.662	0.562	0.868	0.330	0.498	0.525	0.723	0.519	0.375	0.425	0.567	0.705
500	0.112	0.502	0.514	0.859	0.859	0.542	0.390	0.668	0.300	0.381	0.368	0.530	0.539	0.413	0.303	0.512	0.567
1'000	0.103	0.300	0.318	0.868	0.863	0.464	0.276	0.540	0.267	0.289	0.264	0.392	0.571	0.433	0.222	0.487	0.341
8'000	0.029	0.062	0.069	0.919	0.919	0.449	0.082	0.456	0.000	0.127	0.058	0.139	0.355	0.231	0.051	0.237	0.006

$h = 2, \theta_0^{(h)} = 0.1195$

	DML					RA				DR				LP			
T	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)	C_a(95%)	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)
125	0.202	1.573	1.586	0.901	0.894	0.707	0.730	1.016	0.383	0.532	0.668	0.854	0.627	0.151	0.564	0.584	0.905
250	0.042	2.297	2.297	0.920	0.912	0.555	0.497	0.745	0.312	0.412	0.465	0.621	0.607	0.225	0.415	0.472	0.895
500	0.111	0.467	0.480	0.917	0.909	0.461	0.341	0.574	0.280	0.320	0.323	0.455	0.595	0.266	0.301	0.402	0.818
1'000	0.106	0.281	0.301	0.887	0.880	0.406	0.248	0.476	0.198	0.254	0.237	0.347	0.566	0.302	0.220	0.374	0.645
8'000	0.025	0.070	0.074	0.935	0.934	0.411	0.086	0.419	0.000	0.111	0.065	0.129	0.540	0.156	0.061	0.167	0.301

$h = 3, \theta_0^{(h)} = 0.0717$

	DML					RA				DR				LP			
T	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)	C_a(95%)	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)
125	0.208	1.625	1.638	0.915	0.901	0.594	0.720	0.933	0.454	0.453	0.669	0.808	0.709	0.072	0.612	0.617	0.929
250	0.105	0.913	0.919	0.913	0.912	0.462	0.498	0.679	0.394	0.339	0.470	0.579	0.670	0.128	0.459	0.476	0.911
500	0.102	0.470	0.481	0.940	0.927	0.388	0.334	0.512	0.296	0.269	0.322	0.420	0.686	0.171	0.331	0.372	0.903
1'000	0.102	0.285	0.303	0.911	0.908	0.350	0.243	0.426	0.191	0.222	0.236	0.324	0.658	0.214	0.237	0.319	0.827
8'000	0.023	0.082	0.085	0.934	0.933	0.363	0.092	0.375	0.001	0.097	0.078	0.124	0.706	0.104	0.074	0.128	0.709

$h = 4, \theta_0^{(h)} = 0.0430$

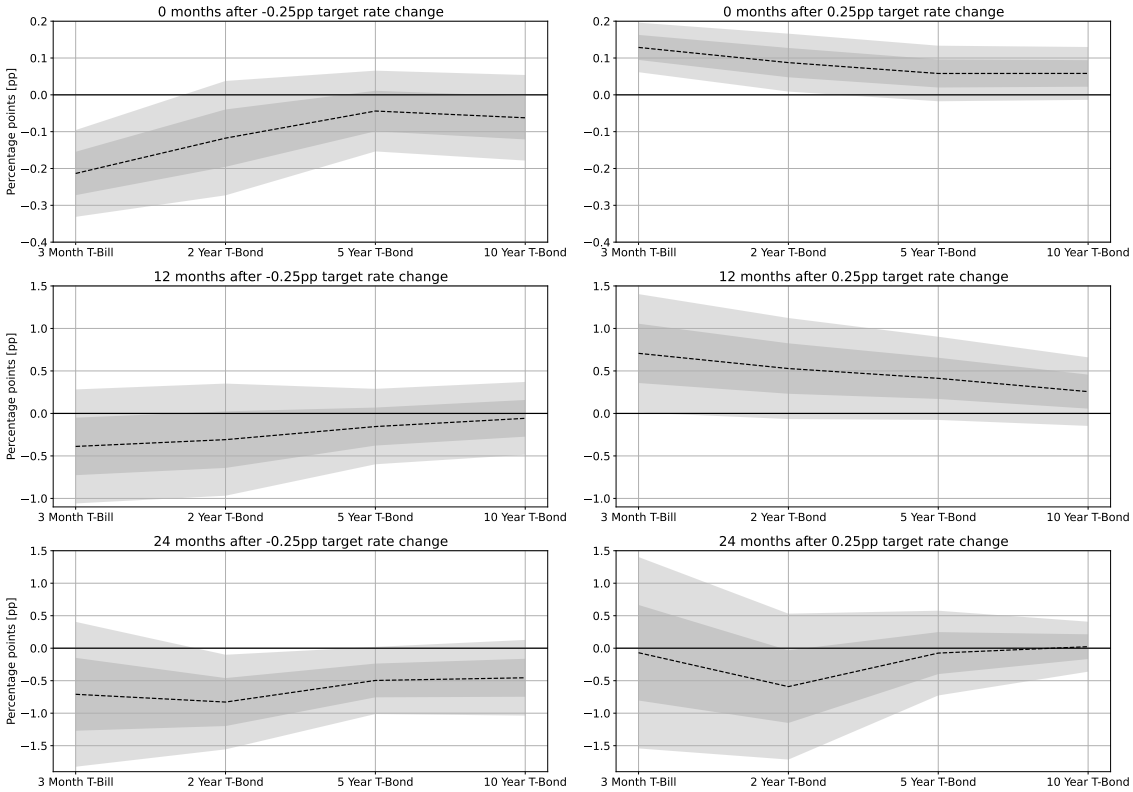
	DML					RA				DR				LP			
T	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)	C_a(95%)	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)
125	0.242	1.594	1.612	0.919	0.899	0.504	0.742	0.897	0.486	0.389	0.701	0.802	0.722	0.045	0.696	0.697	0.924
250	0.108	0.967	0.973	0.905	0.900	0.399	0.529	0.662	0.418	0.300	0.504	0.587	0.712	0.090	0.512	0.520	0.920
500	0.086	0.552	0.559	0.938	0.921	0.319	0.353	0.476	0.365	0.222	0.343	0.409	0.755	0.104	0.363	0.378	0.934
1'000	0.090	0.299	0.313	0.938	0.935	0.294	0.251	0.386	0.251	0.187	0.244	0.308	0.751	0.146	0.255	0.294	0.911
8'000	0.021	0.094	0.096	0.938	0.938	0.318	0.097	0.333	0.004	0.085	0.089	0.123	0.786	0.071	0.084	0.110	0.866

$h = 5, \theta_0^{(h)} = 0.0258$

	DML					RA				DR				LP			
T	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)	C_a(95%)	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)	Bias	std($\hat{\theta}_h$)	RMSE	C_b(95%)
125	0.242	1.669	1.686	0.899	0.899	0.428	0.786	0.894	0.516	0.342	0.745	0.820	0.757	0.043	0.770	0.771	0.912
250	0.106	0.977	0.983	0.930	0.920	0.328	0.552	0.642	0.482	0.244	0.528	0.582	0.749	0.047	0.539	0.541	0.934
500	0.084	0.533	0.540	0.943	0.923	0.269	0.370	0.457	0.403	0.190	0.362	0.409	0.775	0.067	0.386	0.392	0.941
1'000	0.084	0.317	0.328	0.942	0.937	0.251	0.265	0.365	0.303	0.162	0.260	0.306	0.797	0.102	0.274	0.292	0.939
8'000	0.016	0.104	0.105	0.942	0.942	0.273	0.103	0.292	0.009	0.071	0.099	0.121	0.829	0.045	0.095	0.106	0.924

NOTE: The table depicts simulation results across $N = 1'000$ draws obtained for the scenario with one stochastic process. The outcome variable is generated from a linear DGP, i.e. $b(X_t) = 0.5 \sum_{i=1}^{5} X_{i,t}$ and $\tau(X_t) = \theta_0^{(0)} + \sum_{i=1}^{3} X_{i,t} - \sum_{i=4}^{5} X_{i,t}$. Except for the LP estimator, nuisance functions are estimated with random forest. For the DML estimator we use 10-fold cross-fitting and set $k_T = T/10$. For sample size $T = 125$, probabilities are winsorized at 1%. The parameters of the data generating process are $n = 12$, $\sigma_\epsilon = 1.0$, $\gamma = 0.6$, $p = 2$, $q = 1$, $\sigma_u = 1.0$, $\alpha_A = 0.3$, $\alpha_M = 0.3$, $\rho_A = 0.35$, $\rho_M = 0.7$, $\beta_1 = 0.3$, $\beta_2 = 0.5$. $C_a(\cdot)$ and $C_b(\cdot)$ in the tables denote the coverage at the given confidence level using asymptotic and fixed-bandwidth critical values, respectively.

Figure C5: Estimated cumulative effects of target rate changes on the bond yield curve



NOTE: The figure shows the estimated cumulative effects of target rate changes on the bond yield curve for the time period July 1989 to December 2008. The left (right) column shows the effect of decreasing (increasing) the target rate by 25 basis points. The estimated effects on the yield curve are depicted for 0 (top row), 12 (middle row), and 24 (bottom row) months after the target rate change. The nuisance functions are estimated by random forest using 10-fold cross-fitting removing $k_T = 24$ observations from the estimation sample at the boundary to the inference sample. The shaded areas represent 68% and 95% confidence intervals with fixed-bandwidth critical values (Kiefer & Vogelsang, 2005). The variances are estimated using bandwidth determined by the procedure of Newey and West (1994).