Kitagawa, Toru; Wang, Weining; Xu, Mengshan

Working Paper

Policy choice in time series by empirical welfare maximization

cemmap working paper, No. CWP12/22

Provided in Cooperation with:
The Institute for Fiscal Studies (IFS), London

Suggested Citation: Kitagawa, Toru; Wang, Weining; Xu, Mengshan (2022) : Policy choice in time series by empirical welfare maximization, cemmap working paper, No. CWP12/22, Centre for Microdata Methods and Practice (cemmap), London, https://doi.org/10.47004/wp.cem.2022.1222

This Version is available at:
https://hdl.handle.net/10419/260393

cemmap
centre for microdata methods and practice

Policy choice in time series by empirical welfare maximization

Toru Kitagawa
Weining Wang
Mengshan Xu

The Institute for Fiscal Studies
Department of Economics, UCL

cemmap working paper CWP12/22

Economic and Social Research Council

Policy Choice in Time Series by Empirical Welfare Maximization*

Toru Kitagawa† Weining Wang‡ Mengshan Xu§

May 12, 2022

Abstract

This paper develops a novel method for policy choice in a dynamic setting where the available data is a multi-variate time-series. Building on the statistical treatment choice framework, we propose Time-series Empirical Welfare Maximization (T-EWM) methods to estimate an optimal policy rule for the current period or over multiple periods by maximizing an empirical welfare criterion constructed using nonparametric potential outcome time-series. We characterize conditions under which T-EWM consistently learns a policy choice that is optimal in terms of conditional welfare given the time-series history. We then derive a nonasymptotic upper bound for conditional welfare regret and its minimax lower bound. To illustrate the implementation and uses of T-EWM, we perform simulation studies and apply the method to estimate optimal monetary policy rules from macroeconomic time-series data.

Keywords: Causal inference, potential outcome time-series, treatment choice, regret bounds, concentration inequalities.

*We are grateful to Giuseppe Cavaliere, Yoosoon Chang, Wei Cui, Whitney Newey, Joon Park, Barbara Rossi, Frank Schorfheide, and the participants at the EEA-ESEM 2021 and seminars at Centre for Microdata Methods and Practice (CeMMAP) and Indiana University for beneficial comments. Financial support from the ESRC through the ESRC CeMMAP (grant number RES-589-28-0001) and the ERC (grant number 715940) is gratefully acknowledged.

†Department of Economics, Brown University and Department of Economics, University College London. Email: toru_kitagawa@brown.edu

‡Department of Economics and Related Studies, University of York. Email: weining.wang@york.ac.uk

§Department of Economics, University of Mannheim. Email: mengshan.xu@uni-mannheim.de

1 Introduction

A central topic in economics is the nature of the causal relationships between economic outcomes and government policies, both within and across time periods. To investigate this, empirical research makes use of time-series data, with the aim of finding desirable policy rules. For instance, a monetary policy authority may wish to use past and current macroeconomic data to learn an interest rate policy that is optimal in terms of a social welfare criterion. Building on the recent development of potential outcome time-series (Angrist et al. (2018), Bojinov and Shephard (2019), and Rambachan and Shepherd (2021)), this paper proposes a novel method to inform policy choice when the available data is a multi-variate time-series.

In contrast to the structural and semi-structural approaches that are common in macroeconomic policy analysis, such as dynamic stochastic general equilibrium (DSGE) models and structual vector autoregressions (SVAR), we set up the policy choice problem from the perspective of the statistical treatment choice proposed by Manski (2004). The existing statistical treatment choice literature typically focuses on microeconomic applications in a static setting, and the applicability of these methods to a time-series setting has yet to be explored. In this paper, we propose a novel statistical treatment choice framework for time-series data and study how to learn an optimal policy rule. Specifically, we consider extending the conditional empirical success (CES) rule of Manski (2004) and the empirical welfare maximization rule of Kitagawa and Tetenov (2018) to time-series policy choice, and characterize the conditions under which these approaches can inform welfare optimal policy. These conditions do not require functional form specifications for structural equations or the exact temporal dependence of the time-series observations, but can be connected to the structural approach under certain conditions.

In the standard microeconometric setting considered in the treatment choice literature, the planner has access to a random sample of cross-sectional units, and it is often assumed that the populations from which the sample was drawn and to which the policy will be applied are the same. These assumptions are not feasible or credible in the time-series context, which leads to several non-trivial challenges. First, the economic environment and the economy's causal response to it may be time varying. Assumptions are required to make it possible to learn an optimal policy rule for future periods based on available past data. In addition, although we can assume a stationary environment, in macroeconomics and finance applications, it is extremely unlikely that the data are obtained from a randomized control trial. We hence have to rely on observational data. Accordingly, the identifiability of social welfare under counterfactual policies becomes non-trivial and requires some conditions on how past policies were assigned over time. Second, to define an optimal policy in the time-series setting, it is reasonable to consider social welfare *conditional on* the history of observables at the time the policy decision is made. This conditional welfare contrasts

with unconditional welfare, which averages conditional welfare with respect to hypothetical realizations of the history. Third, when past data is used to inform policy, we have only a single realization of a time-series in which the observations are dependent across the periods and possibly nonstationary. Such statistical dependence complicates the characterization of the statistical convergence of the welfare performance of an estimated policy. Fourth, if the planner wishes to learn a dynamic assignment policy, which prescribes a policy in each period over multiple periods on the basis of observable information available at the beginning of every period, the policy learning problem becomes substantially more involved. This is because a policy choice in the current period may affect subsequent policy choices through the current policy assignment and a realized outcome under the assigned treatment.

Taking into account these challenges, we propose time-series empirical welfare maximization (T-EWM) methods that construct an empirical welfare criterion based on a historical average of the outcomes and obtain a policy rule by maximizing the empirical welfare criterion over a class of policy rules. We then clarify the conditions on the causal structure and data generating process under which T-EWM methods consistently estimate a policy rule that is optimal in terms of conditional welfare. Extending the regret bound analysis of Kitagawa and Tetenov (2018) to time-series dependent observations, we obtain a finite-sample uniform bound for welfare regret. We then characterize the convergence of welfare regret and establish the minimax rate optimality of the T-EWM rule.

Our development of T-EWM builds on the recent potential outcome time-series literature (Angrist et al. (2018), Bojinov and Shephard (2019), and Rambachan and Shepherd (2021)). In particular, to identify the counterfactual welfare criterion, we employ the sequential exogeneity restriction considered in Bojinov and Shephard (2019). These studies focus on retrospective evaluation of the causal impact of policies observed in historical data, and do not analyze how to perform future policy choice based on the historical evidence.

Since the seminal work of Manski (2004), statistical treatment choice and empirical welfare maximization have been active topics of research, e.g., Dehejia (2005), Stoye (2009, 2012), Qian and Murphy (2011), Tetenov (2012), Bhattacharya and Dupas (2012), Zhao et al. (2012), Kitagawa and Tetenov (2018, 2021), Kallus (2021), Athey and Wager (2021), Mbakop and Tabord-Meehan (2021), Kitagawa et al. (2021), among others. These works focus on a setting where the available data is a cross-sectional random sample obtained from an experimental or observational study with randomized treatment, possibly conditional on observable characteristics. Viviano (2021) and Ananth (2020) consider EWM approaches for treatment allocations where the training data features cross-sectional dependence due to network spillovers, but this paper is the first to consider policy choice with time-series data. As a related but distinct problem, there is a large literature on the estimation of dynamic treatment regimes, Murphy (2003), Zhao et al. (2015), Han (2021), and Sakaguchi (2021). The problem of dynamic treatment regimes assume that training data is a short panel in

which treatments have been randomized both among cross-sectional units and across time periods. The T-EWM framework, in contrast, assumes observations are drawn from a single time-series, as is common in empirical macroeconomics.

A large literature on multi-arm bandit algorithms analyzes learning and dynamic allocations of treatments when there is a trade-off between exploration and exploitation. See Lattimore and Szepesvári (2020) and references therein, Kock et al. (2020) and Kasy and Sautmann (2019) for recent works in econometrics. The setting in this paper differs from the standard multi-arm bandit setting in the following two respects. First, our framework treats the available past data as a training sample and focuses on optimizing short-run welfare. We are hence concerned with performance of the method in terms of short-term regret rather than cumulative regret over a long horizon. Second, in the standard multi-arm bandit problem, subjects to be treated are assumed to differ across rounds, which implies that the outcome generating process is independent over time. This is not the case in our setting, and we include the realization of outcomes and policies in the past periods as contextual information for the current decision. Third, suppose that bandit algorithms can be adjusted to take into account to the dependence of observations, our method is then analogous to the "pure exploration" class, involving a long exploration phase followed by a one-period exploitation at the very end. However, a major difference is that the bandit algorithm concerns data in a random experiment while our method is aimed at data in quasi-random experiments.

The analysis of welfare regret bounds is similar to the derivation of risk bounds in empirical risk minimization, as reviewed in Vapnik (1998) and Lugosi (2002). Risk bounds studied in the empirical risk minimization literature typically assume independent and identically distributed (i.i.d.) training data. A few exceptions, Jiang and Tanner (2010), Brownlees and Gudmundsson (2021), and Brownlees and Llorens-Terrazas (2021) obtain risk bounds for empirical risk minimizing predictions with time-series data, but they do not consider welfare regret bounds for causal policy learning.

The rest of the paper is organized as follows. Section 2 describes the setting using a simple illustrative model with a single discrete covariate. Section 3 discusses the general model with continuous covariates and presents the main theorems. In Section 4 we discuss extensions to our proposed framework, including nonparametric methods, a multi-period welfare function, how T-EWM is related to the Lucas critique, as well as T-EWM's links with structural VARs, Markov Decision Processes, and reinforcement learning. In Section 5.2 we show the lower bound and minimax optimality. In Section 5.3 we discuss the case with estimated propensity scores. In Section 6 and 7 we present simulation studies and an empirical application involving finding a treatment to optimize financial returns. Technical proofs and other details are presented in the supplementary material.

2 Model and illustrative example

In this section, we introduce the basic setting, notation, the conditional welfare criterion we aim to maximize, and conditions on the data generating process that are important for the learnability of an optimal policy. Then, we illustrate the main analytical tools used to bound welfare regret through a heuristic model with a simple dynamic structure.

2.1 Notation, timing, and welfare

We suppose that the social planner is at the beginning of time T. Let $W_t \in \{0,1\}$ denote a treatment or policy (e.g., nominal interest rate) implemented at time $t = 0, 1, 2, \ldots$. To simplify the analysis, we assume that W_t is binary (e.g., a high or low interest rate regime). The planner sets $W_T \in \{0,1\}$, $T \geq 1$, making use the history of observable information up to period T to inform her decision. This observable information consists of an economic outcome (e.g., GDP, unemployment rate, etc.), $Y_{0:T-1} = (Y_0, Y_1, Y_2, \cdots, Y_{T-1})$, the history of implemented policies, $W_{0:T-1} = (W_0, W_1, W_2, \cdots, W_{T-1})$, and covariates other than the policies and the outcome (e.g., inflation), $Z_{0:T-1} = (Z_0, Z_1, Z_2, \cdots, Z_{T-1})$. Z_t can be a multidimensional vector, but both Y_t and W_t are assumed to be univariate. [1]

Following Bojinov and Shephard (2019), we refer to a sequence of policies $w_{0:t} = (w_0, w_1, \ldots, w_t) \in \{0,1\}^{t+1}$, $t \geq 0$, as a treatment path. A realized treatment path observed in the data $0 \leq t \leq T-1$ is a stochastic process $W_{0:T-1} = (W_0, W_1, \ldots, W_{T-1})$ drawn from the data generating process.

Without loss of generality, we assume that Z_t is generated after the outcome Y_t is observed. The timing of realizations is therefore

$$\underbrace{W_{t-1} \to Y_{t-1} \to Z_{t-1}}_{\text{time period } t-1} \to \underbrace{W_t \to Y_t \to Z_t}_{\text{time period } t},$$

i.e., the transition between periods happens after Z_{t-1} is realised but before W_t is realised.[2]

Let

$$X_t = \{W_t, Y_t, Z_t'\}' \tag{1}$$

[1] It can be extended to the settings with multidimensional outcomes and treatments, where $W_t \in \{0,1\}^d$ and $Y_t \in \mathbb{R}^k$. For example, the treatment can be the implementation of government monetary policy, rating agency's change of grade for some asset etc. In the case of network data, treatment comes from an observed network $d = k^2$ (k is the number of nodes within the network), for example in social network such as the Facebook, w_t is the observed adjacency matrices.

[2] We do not allow Z_t to be realised between Y_t and W_t. If some Z is realised after W_t, before Y_t and is not causally effected by W_t, the causal link unaffected by placing Z before W_t and labeling it Z_{t-1}; if this Z is realised after W_t, before Y_t, and is causally effected by W_t, then Z is a bad control and should not be included in the model.

collect the observable variables for period t, and let X_t be drawn from $\mathcal{X}$ for $t \in \mathbb{N}$. We define the filtration

$$\mathcal{F}_{t-1} = \sigma(X_{0:t-1}),$$

where $\sigma(\cdot)$ denotes the Borel σ-algebra generated by the variables specified in the argument. The filtration $\mathcal{F}_{T-1}$ corresponds to the planner's information set at the time of making her decision in period T.

Following the framework of Bojinov and Shephard (2019), we introduce potential outcome time-series. At each $t = 0, 1, 2, \ldots$, and for every treatment path $w_{0:t} \in \{0,1\}^{t+1}$, let $Y_t(w_{0:t}) \in \mathbb{R}$ be the realized period t outcome if the treatment path from 0 to period t were $w_{0:t}$. Hence, we have a collection of potential outcome paths indexed by treatment path,

$$\{Y_t(w_{0:t}) : w_{0:t} \in \{0,1\}^{t+1}, t = 0, 1, 2, \ldots\},$$

which defines 2^{t+1} potential outcomes in each period t. This is an extension of the Neyman-Rubin causal model originally developed for cross-sectional causal inference. As maintained in Bojinov and Shephard (2019), the potential outcomes for each t are indexed by the current and past treatments $w_{0:t}$ only. This imposes the restriction that any future treatment w_{t+p}, $p \geq 1$, does not causally affect the current outcome, i.e., an exclusion restriction for future treatments.

For a realized treatment path $W_{0:t}$, the observed outcome Y_t and the potential outcomes satisfy

$$Y_t = \sum_{w_{0:t} \in \{0,1\}^t} 1\{W_{0:t} = w_{0:t}\} Y_t(w_{0:t})$$

for all $t \geq 0$.

The baseline setting of the current paper considers the choice of policy W_T for a single period T.[3] We denote the policy choice based on observations up to period $T-1$ by

$$g : \mathcal{X}^T \to \{0,1\}. \tag{2}$$

The period-T treatment is $W_T = g(X_{0:T-1})$, and we refer to $g(\cdot)$ as a *decision rule*. We also define the region in the space of the covariate vector for which the decision rule chooses $W_T = 1$ to be

$$G = \{X_{0:T-1} : g(X_{0:T-1}) = 1\} \subset \mathcal{X}^T, \tag{3}$$

We refer to G as a *decision set*.

We assume that the planner's preferences for policies in period-T are embodied in a social welfare criterion. In particular, we define one-period *welfare conditional on* $\mathcal{F}_{T-1}$ (conditional

[3]Section 4.2 discusses how to extend the single-period policy choice problem to multi-period settings.

welfare, for short) to be[4]

$$\mathcal{W}_T(g|\mathcal{F}_{T-1}) \;:=\; \mathsf{E}\left[Y_T(W_{0:T-1},1)g(X_{0:T-1}) + Y_T(W_{0:T-1},0)(1-g(X_{0:T-1}))|\mathcal{F}_{T-1}\right]. \quad (4)$$

With some abuse of notation, conditional welfare can be expressed with the decision set G as its argument:

$$\mathcal{W}_T(G|\mathcal{F}_{T-1}) \;:=\; \mathsf{E}\left[Y_T(W_{0:T-1},1)1\{X_{0:T-1}\in G\} + Y_T(W_{0:T-1},0)1\{X_{0:T-1}\notin G\}|\mathcal{F}_{T-1}\right] (5)$$

This welfare criterion is conditional on the planner's information set. This contrasts with the unconditonal welfare criterion common in the cross-sectional treatment choice setting, where any conditioning variables (observable characteristics of a units) are averaged out.[5] In the time series setting, it is natural for the planner's preferences to be conditional on the realized history, rather than averaging over realized and unrealized histories, as would be the case if the unconditional criterion were used.

As we clarify in Section 4.1, regret for conditional welfare and regret for unconditional welfare require different conditions for convergence, and their rates of convergence may differ. Hence, the existing results for regret convergence for unconditional welfare shown in Kitagawa and Tetenov (2018) do not immediately carry over to the time-series setting.

The planner's optimal policy g^* maximizes her one-period welfare,

$$g^* \in \arg\max_g \mathcal{W}_T(g|\mathcal{F}_{T-1}).$$

The planner does not know g^*, so she instead seeks a statistical treatment choice rule (Manski (2004)) $\hat{g}$, which is a decision rule selected on the basis of the available data $X_{0:T-1}$.

Our goal is to develop a way of obtaining $\hat{g}$ that performs well in terms of the conditional welfare criterion (5). Specifically, we assess the statistical performance of an estimated policy rule $\hat{g}$ in terms of the convergence of conditional welfare regret,

$$\mathcal{W}_T(g^*|X_{0:T-1}=x_{0:T-1}) - \mathcal{W}_T(\hat{g}|X_{0:T-1}=x_{0:T-1}) \quad (6)$$

and its convergence rate with respect to the sample size T. When evaluating realised regret, $X_{0:T-1}$ is set to its realized value in the data. On the other hand, when examining convergence, we accommodate statistical uncertainty over $\hat{g}$ by focusing on convergence with probability approaching one uniformly over a class of sampling distributions for $X_{0:T-1}$. A more precise characterization of the regret convergence results will be given below and in

[4]Throughout the paper, we acknowledge that the expectation, E, and the probability, Pr, is corresponding to the outer measure whenever a measurability issue is encountered.

[5]Manski (2004) also considers a conditional welfare criterion in the cross-sectional setting.

Section 3.

2.2 Illustrative model with a discrete covariate

We begin our analysis with a simple illustrative model, which provides a heuristic exposition of the main idea of T-EWM and its statistical properties. We cover more general settings and extensions in Sections 3 and 4.

Suppose that the data consists of a bivariate time-series $X_{0:T-1} = ((Y_t, W_t) \in \mathbb{R} \times \{0,1\} : t = 0, 1, \ldots, T-1)$ with no other covariates. To simplify exposition for the illustrative model, we impose the following restrictions on the dynamic causal structure and dependence of the observations.

Assumption 2.1. [Markov properties] The time-series of potential outcomes and observable variables satisfy the following conditions:

(i) *Markovian exclusion*: for $t = 2, \ldots, T$ and for arbitrary treatment paths $(w_{0:t-2}, w_{t-1}, w_t)$ and $(w'_{0:t-2}, w_{t-1}, w_t)$, where $w_{0:t-2} \neq w'_{0:t-2}$,

$$Y_t(w_{0:t-2}, w_{t-1}, w_t) = Y_t(w'_{0:t-2}, w_{t-1}, w_t) := Y_t(w_{t-1}, w_t) \tag{7}$$

holds with probability one.

(ii) *Markovian exogeneity*: for $t = 1, \ldots, T$ and any treatment path $w_{0:t}$,

$$Y_t(w_{0:t}) \perp X_{0:t-1} | W_{t-1}, \tag{8}$$

and for $t = 1, \ldots, T-1$,

$$W_t \perp X_{0:t-1} | W_{t-1}. \tag{9}$$

These assumptions significantly simplify the dynamic structure of the problem. Markovian exclusion, Assumption 2.1 (i), says that only the current treatment w_t and treatment in the previous period w_{t-1} can have a causal impact on the current outcome. This allows the indices of the potential outcomes to be compressed to the latest two treatments (w_{t-1}, w_t), as in (7). Markovian exogeneity, Assumption 2.1 (ii), states that once you condition on the policy implemented in the previous period W_{t-1}, the potential outcomes and treatment for the current period are statistically independent of any other past variables.

It is important to note that these assumptions do not impose stationarity: we allow the distribution of potential outcomes to vary across time periods. In addition, under Assumption 2.1, we can reduce the class of policy rules to those that map from $W_{T-1} \in \{0,1\}$ to

$W_T \in \{0,1\}$,

$$g : \{0,1\} \to \{0,1\}$$

with no loss of conditional welfare, i.e., welfare conditional on $\mathcal{F}_{T-1}$ can be simplified to

$$\mathcal{W}_T(g|W_{T-1}) = \mathsf{E}\{Y_T(W_{T-1},1)g(W_{T-1}) + Y_T(W_{T-1},0)(1-g(W_{T-1}))|W_{T-1}\}. \tag{10}$$

To make sense of Assumption 2.1 and illustrate the relationship between the potential outcome time-series and the standard structural equation modeling, we provide a toy example involving vaccination:

Example 1. Suppose the SP (monetary policy authority) is interested in setting a low or high interest rate at period T. Let W_t denote the indicator for whether the interest rate in period t is high ($W_t = 1$) or low ($W_t = 0$). Y_t denotes a measure of social welfare, which can be a function of aggregate output, inflation, and other macroeconomic variables. Let ε_t be i.i.d. shock that is statistically independent of $X_{0:t-1}$, and we assume the following structural equation for the causal relationship of Y_t on W_t (and its lag) and the regression dependence of W_t on its lag [6]

$$Y_t = \beta_0 + \beta_1 W_t + \beta_2 W_{t-1} + \varepsilon_t, \tag{11}$$

$$W_t = (1-q) + \lambda W_{t-1} + V_t, \tag{12}$$

$$\lambda = p + q - 1,$$

$$\varepsilon_t \perp (W_t, X_{0:t-1}) \qquad \forall 1 \le t \le T-1 \qquad \text{and} \qquad \varepsilon_T \perp X_{0:T-1} \tag{13}$$

$$\text{If } W_{t-1} = 1, \quad \begin{cases} V_t = 1-p & \text{with probability } p \\ V_t = -p & \text{with probability } 1-p, \end{cases} \tag{14}$$

$$\text{if } W_{t-1} = 0, \quad \begin{cases} V_t = -(1-q) & \text{with probability } q \\ V_t = q & \text{with probability } 1-q. \end{cases} \tag{15}$$

Compatibility with Assumption 2.1 can be seen as follows. Assumption 2.1(i) is implied by (11), where the structural equation of Y_t involves only (W_t, W_{t-1}) as the factors of direct cause. Assumption 2.1(ii) is implied by (11), (13), and the fact that the distribution of V_t depends solely on W_{t-1}, i.e., under (12),(14), and (15), we have $\Pr(W_t|\mathcal{F}_{t-1}) = \Pr(W_t|W_{t-1})$.

To examine the learnability of the optimal policy rule, we further restrict the data generating process. First, we impose a strict overlap condition on the propensity score:

[6] The distribution of W_t follows Hamilton (1989). However, the Markov switching model of Hamilton (1989) has unobserved W_t, which differs from this example.

Assumption 2.2. [Strict overlap] Let $e_t(w) := \Pr(W_t = 1|W_{t-1} = w)$ be the period-t propensity score. There exists a constant $\kappa \in (0, 1/2)$, such that for any $t = 1, 2, \ldots, T-1$ and $w \in \{0, 1\}$,

$$\kappa \leq e_t(w) \leq 1 - \kappa.$$

The next assumption imposes an unconfoundedness condition on observed policy assignment:

Assumption 2.3. [Unconfoundedness] For any $t = 1, 2, \ldots, T-1$ and $w \in \{0, 1\}$,

$$Y_t(W_{t-1}, w) \perp W_t | X_{0:t-1}.$$

This assumption states that the treatments observed in the data are sequentially randomized conditional on lagged observable variables. This is a key assumption to make unbiased estimation for the welfare feasible at each period in the sample, as employed in Bojinov and Shephard (2019) and others. Combining Assumption 2.1 with unconfoundedness (Assumption 2.3), we have for any measurable function f of the potential outcome $Y_t(W_{0:t-1})$ and treatment W_t, it holds

$$\mathsf{E}(f(Y_t(W_{0:t}), W_t)|\mathcal{F}_{t-1}) = \mathsf{E}(f(Y_t(W_{t-1}, W_t), W_t)|W_{t-1}) = \mathsf{E}(f(Y_t, W_t)|W_{t-1}). \tag{16}$$

Example 1 continued. Assumption 2.2 is satisfied if $0 < p < 1$ and $0 < q < 1$; Assumption 2.3 is implied by (11) and (13).

Imposing Assumption 2.2 and 2.3 and assuming propensity scores are known, consider constructing a sample analogue of (10) conditional on $W_{T-1} = w$ based on the historical average of the inverse propensity score weighted outcomes:

$$\widehat{\mathcal{W}}(g|W_{T-1} = w) = T(w)^{-1} \sum_{1 \leq t \leq T-1 : W_{t-1} = w} \left[\frac{Y_t W_t g(W_{t-1})}{e_t(W_{t-1})} + \frac{Y_t(1 - W_t)\{1 - g(W_{t-1})\}}{1 - e_t(W_{t-1})} \right], \tag{17}$$

where $T(w) = \#\{1 \leq t \leq T-1 : W_{t-1} = w\}$ is the number of observations where the policy in the previous period took value w, i.e. the subsample corresponding to $W_{t-1} = w$. Unlike the microeconometric setting considered in, e.g., Kitagawa and Tetenov (2018), we do not necessarily have $\widehat{\mathcal{W}}(g|W_{T-1} = w)$ as a direct sample analogue for the SP's social welfare objective, since we allow a non-stationary environment in which the historical average of the conditional welfare criterion can diverge from the conditional welfare in the current period. Nevertheless, we refer to $\widehat{\mathcal{W}}(g|W_{T-1} = w)$ as the empirical welfare of the policy rule g.

Denoting $(\cdot|W_{T-1} = w)$ by $(\cdot|w)$, define the true optimal policy and its empirical analogue

to be,

$$g^*(w) \in \text{argmax}_{g:\{w\}\to\{0,1\}} \mathcal{W}_T(g|w), \tag{18}$$

$$\hat{g}(w) \in \text{argmax}_{g:\{w\}\to\{0,1\}} \widehat{\mathcal{W}}(g|w). \tag{19}$$

$\hat{g}$ is constructed by maximizing empirical welfare over a class of policy rules (four policy rules in total). We call a policy rule constructed in this way the *Time-series Empirical Welfare Maximization* (T-EWM) rule. The construction of the T-EWM rule $\hat{g}$ is analogous to the conditional empirical success rule with known propensity scores considered by Manski (2004) in the i.i.d. cross-sectional setting. In the time-series setting, however, the assumptions imposed so far do not guarantee $\widehat{\mathcal{W}}(g|w)$ is an unbiased estimator of the true conditional welfare $\mathcal{W}_T(g|w)$.

2.3 Bounding the conditional welfare regret of the T-EWM rule

A major contribution of this paper is to characterizing conditions that justify the T-EWM rule $\hat{g}$ in terms of the convergence of conditional welfare. This section clarifies these points in the context of our illustrative example.

To bound conditional welfare regret, our strategy is to decompose empirical welfare $\widehat{\mathcal{W}}(g|w)$ into a conditional mean component and a deviation from it. The deviation is the sum of a martingale difference sequence (MDS), and this allows us to apply concentration inequalities for the sum of MDS. Define an intermediate welfare function,

$$\begin{aligned} & \bar{\mathcal{W}}(g|w) \\ = \; & T(w)^{-1} \sum_{1\le t\le T-1: W_{t-1}=w} \mathsf{E}\left\{Y_t(W_{t-1},1)g(W_{t-1}) + Y_t(W_{t-1},0)\left[1-g(W_{t-1})\right]|W_{t-1}\right\}. \end{aligned} \tag{20}$$

Under the strict overlap and unconfoundedness assumptions (Assumptions 2.2 and 2.3), the difference between empirical welfare and (20) is a MDS. Furthermore, we impose the assumption:

Assumption 2.4. [Invariance of the welfare ordering]
Given $w \in 0, 1$, let $g^* = g^*(w)$ defined in (18). There exist a positive constant $c > 0$, such that for any $g \in \{0, 1\}$,

$$\mathcal{W}_T(g^*|w) - \mathcal{W}_T(g|w) \le c\big[\bar{\mathcal{W}}(g^*|w) - \bar{\mathcal{W}}(g|w)\big], \tag{21}$$

with probability approaching one, i.e., noting that $\bar{\mathcal{W}}(g|w)$ is random as it depends on $W_{0:T-1}$, P_T (inequality (21) holds) $\to 1$ as $T \to \infty$, where P_T is the probability distribution for $X_{0:T-1}$.

Noting that the left-hand side of (21) is nonnegative for any g by construction and $c > 0$, this assumption implies that $\bar{\mathcal{W}}(g^*|w) - \bar{\mathcal{W}}(g|w) > 0$ must hold whenever $\mathcal{W}_T(g^*|w) - \mathcal{W}_T(g|w) > 0$. That is, optimality of g^* in terms of the conditional welfare at T is maintained in the historical average of the conditional welfares. Under this assumption, having an estimated policy $\hat{g}$ that attains a convergence of $\bar{\mathcal{W}}(g^*|w) - \bar{\mathcal{W}}(\hat{g}|w)$ to zero guarantees that $\hat{g}$ is also consistent for the optimal policy g^* in terms of the conditional welfare at T.

Remark 1. Assumption 2.4 can be restrictive in a situation where the dynamic causal structure of the current period is believed to be different from the past, but is weaker than stationarity. In the current example, if we were willing to assume,

A2.4' *The stochastic process*

$$S_t(w) = Y_t(W_{t-1}, 1)g(W_{t-1}) + Y_t(W_{t-1}, 0)\left[1 - g(W_{t-1})\right]|_{W_{t-1}=w}$$

is weakly stationary.

then, under A2.4', we would have

$$\begin{aligned} & \mathsf{E}\{Y_T(W_{T-1}, 1)g(W_{T-1}) + Y_T(W_{T-1}, 0)(1 - g(W_{T-1}))|W_{T-1} = w\} \\ = \; & \mathsf{E}\left\{Y_t(W_{t-1}, 1)g(W_{t-1}) + Y_t(W_{t-1}, 0)\left[1 - g(W_{t-1})\right]|W_{t-1} = w\right\} \\ = \; & \mathsf{E}[S_t(w)] \text{ for } 2 \leq t \leq T. \end{aligned}$$

and Assumption 2.4 would hold naturally. [7]

Furthermore, Assumption 2.4 can be satisfied by many classic non-stationary processes in linear time series models, including series with deterministic or stochastic time trends.

Example 1 continued.
(i) By Remark 1, Assumption 2.4 holds for Example 1 since

$$S_t(w) = \beta_0 + \beta_1 \cdot g + \beta_2 \cdot w + \varepsilon_t$$

is weakly stationary.

[7] Assumption 2.4 is also weaker than and can be implied by the following assumptions:
The direct one

$$\bar{\mathcal{W}}(g|w) = \mathcal{W}_T(g|w).$$

The linear one

$$\bar{\mathcal{W}}(g|w) = c_1\mathcal{W}_T(g|w) + c_2.$$

The asymptotic one

$$\bar{\mathcal{W}}(g|w) = \mathcal{W}_T(g|w) + o_p(1/\sqrt{T}).$$

ε_t remains an i.i.d. noise in the following settings.

(ii) If we replace (11) by

$$Y_t = \delta_t + \beta_1 W_t + \beta_2 W_{t-1} + \varepsilon_t,$$

where δ_t is an arbitrary deterministic time trend, the process Y_t is trend stationary (non-stationary), but Assumption 2.4 still holds with $c = 1$ since those deterministic trends are canceled out by differences, i.e.,

$$\mathcal{W}_T(g^*|w) - \mathcal{W}_T(g|w) = \beta_1 (g^* - g) = \bar{\mathcal{W}}(g^*|w) - \bar{\mathcal{W}}(g|w).$$

(iii) If we replace (11) by

$$Y_t = \beta_0 + \beta_1 W_t + \beta_2 W_{t-1} + \sum_{i=0}^{t} \varepsilon_i,$$

the process Y_t is non-stationary with stochastic trends, but Assumption 2.4 still holds with $c = 1$ since the stochastic trends are canceled out by differences.

(iv) If we replace (11) by

$$Y_t = \delta_t + \beta_{1,t} W_t + \beta_{2,t} W_{t-1} + \varepsilon_t$$

to allow heterogeneous treatment effect. Then

$$\begin{aligned} \mathcal{W}_T(g^*|w) - \mathcal{W}_T(g|w) &= \beta_{1,T} (g^* - g) \\ \bar{\mathcal{W}}(g^*|w) - \bar{\mathcal{W}}(g|w) &= \bar{\beta}_w (g^* - g), \end{aligned}$$

where $\bar{\beta}_w := \frac{1}{T(w)} \sum_{1 \le t \le T-1 : W_{t-1} = w} \beta_{1,t}$. Since $\mathcal{W}_T(g^*|w) - \mathcal{W}_T(g|w)$ is non-negative by the definition of g^* in (18), Assumption 2.4 holds if

$$\beta_{1,T} \text{ and } \bar{\beta}_w \text{ have the same sign and } c \ge \frac{|\beta_{1,T}|}{|\bar{\beta}_w|}.$$

Without loss of generality, we can assume that both $\beta_{1,T}$ and $\bar{\beta}_w$ are positive, then a sufficient condition for Assumption 2.4 is: There are positive number l and u, such that $0 < l \le \beta_{1,t} \le u$ holds for all t. In this case, $c = \frac{u}{l}$.

Assumption 2.4 implies that g^* also maximizes $\bar{\mathcal{W}}$. Hence, if empirical welfare $\widehat{\mathcal{W}}(\cdot|w)$ can approximate $\bar{\mathcal{W}}(\cdot|w)$ well, intuitively the T-EWM rule $\hat{g}$ should converge to g^*.

The motivation for Assumption 2.4 is to create a bridge between $\mathcal{W}_T(g^*|w) - \mathcal{W}_T(\hat{g}|w)$, the population regret, and $\widehat{\mathcal{W}}(\cdot|w) - \bar{\mathcal{W}}(\cdot|w)$, which is a sum of MDS with respect to filtration

$\{\mathcal{F}_{t-1} : t = 1, 2, \ldots, T\}$.[8] Specifically,

$$\begin{aligned}
&\mathcal{W}_T(g^*|w) - \mathcal{W}_T(\hat{g}|w) \\
\leq& c\left[\bar{\mathcal{W}}(g^*|w) - \bar{\mathcal{W}}(\hat{g}|w)\right] \\
=& c\left[\bar{\mathcal{W}}(g^*|w) - \widehat{\mathcal{W}}(\hat{g}|w) + \widehat{\mathcal{W}}(\hat{g}|w) - \bar{\mathcal{W}}(\hat{g}|w)\right] \\
\leq& c\left[\bar{\mathcal{W}}(g^*|w) - \widehat{\mathcal{W}}(\tilde{g}|w) + \widehat{\mathcal{W}}(\hat{g}|w) - \bar{\mathcal{W}}(\hat{g}|w)\right] \\
\leq& 2c \sup_{g:\{w\}\to\{0,1\}} |\bar{\mathcal{W}}(g|w) - \widehat{\mathcal{W}}(g|w)|, \qquad (22)
\end{aligned}$$

where the first inequality follows by Assumption 2.4. The second inequality follows from the definition of T-EWM rule $\hat{g}$ in (19).

To bound the right-hand side of (22), define

$$\widehat{\mathcal{W}}_t(g|w) = 1(W_{t-1} = w)\left[\frac{Y_t W_t g(W_{t-1})}{e_t(W_{t-1})} + \frac{Y_t(1 - W_t)\{1 - g(W_{t-1})\}}{1 - e_t(W_{t-1})}\right],$$

and

$$\bar{\mathcal{W}}_t(g|w) = 1(W_{t-1} = w)\,\mathsf{E}\{Y_t(W_{t-1}, 1)g(W_{t-1}) + Y_t(W_{t-1}, 0)(1 - g(W_{t-1}))|W_{t-1} = w\}.$$

We then express (17) and (20) as

$$\begin{aligned}
\widehat{\mathcal{W}}(g|w) =& \ \tfrac{T-1}{T(w)} \cdot \tfrac{1}{T-1} \textstyle\sum_{t=1}^{T-1} \widehat{\mathcal{W}}_t(g|w), \\
\bar{\mathcal{W}}(g|w) =& \ \tfrac{T-1}{T(w)} \cdot \tfrac{1}{T-1} \textstyle\sum_{t=1}^{T-1} \bar{\mathcal{W}}_t(g|w),
\end{aligned}$$

and

$$\widehat{\mathcal{W}}(g|w) - \bar{\mathcal{W}}(g|w) = \frac{T-1}{T(w)} \cdot \frac{1}{T-1} \sum_{t=1}^{T-1} \left[\widehat{\mathcal{W}}_t(g|w) - \bar{\mathcal{W}}_t(g|w)\right]$$

follows. By Assumptions 2.1 and 2.3, $\widehat{\mathcal{W}}_t(g|w) - \bar{\mathcal{W}}_t(g|w)$ is a MDS, so we can apply a concentration inequality for sums of MDS to obtain a high-probability bound for $\widehat{\mathcal{W}}(g|w) - \bar{\mathcal{W}}(g|w)$ that is uniform in g.

Specifically, in Appendix D.1 we show that

$$\sup_{g:\{0,1\}\to\{0,1\}} \left\{\widehat{\mathcal{W}}(g|w) - \bar{\mathcal{W}}(g|w)\right\} \lesssim_p \frac{C}{\sqrt{T-1}}, \qquad (23)$$

where C is a constant defined in the proof and $\lesssim_p$ denotes an upper bound valid with (upper) probability approaching one. That is, for a random sequence A_T that follows an

[8] Note in the current simple example, $\mathsf{E}(\cdot|\mathcal{F}_{t-1}) = \mathsf{E}(\cdot|W_{t-1})$.

(outer) probability measure P_T and a deterministic sequence ε_T, $T = 1, 2, 3, \ldots$, $A_T \lesssim_p \varepsilon_T$ means that there exists a positive constant c such that,

$$P_T(A_T > c\varepsilon_T) \overset{T\to\infty}{\to} 0.$$

Combining (22) and (23), we can conclude that the convergence rate of regret $\mathcal{W}_T(g^*|w) - \mathcal{W}_T(\hat{g}|w)$ is of order $O_p(\frac{1}{\sqrt{T-1}})$, and is uniform in the conditioning value of w.

3 Continuous covariates

This section extends the illustrative example of Section 2 by allowing X_t to contain continuous variables. For simplicity of exposition, we maintain the first-order Markovian structure similarly to the illustrative example, but it is straightforward to incorporate a higher-order Markovian structure. In this section, for ease of exposition with continuous covariates, we switch our notation from a policy rule g to its decision set G . The relationship between g and G is shown in (3).

3.1 Setting

In addition to (Y_t, W_t), we now include general covariates Z_t in $X_t \in \mathcal{X}$, which can be continuous. We maintain the Markovian dynamics, while modifying Assumptions 2.1, 2.2, and 2.3 as follows:

Assumption 3.1. [Markov properties] The time-series of potential outcomes and observable variables satisfy the following conditions:

(i) *Markovian exclusion*: same as Assumption 2.1 (i).

(ii) *Markovian exogeneity*: for $t = 1, \ldots, T$ and any treatment path $w_{0:t}$,

$$Y_t(w_{0:t}) \perp X_{0:t-1}|X_{t-1}, \tag{24}$$

and for $t = 1, \ldots, T-1$,

$$W_t \perp X_{0:t-1}|X_{t-1}. \tag{25}$$

Similarly to (10) in the illustrative example, Assumption 3.1 implies that we can reduce the conditioning information of $\mathcal{F}_{t-1}$ to X_{t-1} only and reduce the policy as a binary map of X_{t-1} with no loss of conditional welfare, i.e., partition the space of X_{t-1} into G and its complement. Following these reductions and the planner's interest in the policy choice at

period T, we can set the planner's objective function to be

$$\mathcal{W}_T(G|X_{T-1}) = \mathsf{E}\left\{Y_T(W_{T-1},1)1(X_{T-1}\in G) + Y_t(W_{T-1},0)1(X_{T-1}\notin G)|X_{T-1}\right\}. \tag{26}$$

We assume the strict overlap and unconfoundedness restrictions under the general covariates as follows:

Assumption 3.2 (Strict overlap)**.** Let $e_t(x) = \Pr(W_t = 1|X_{t-1} = x)$ be the propensity score at time t. There exists $\kappa \in (0, 1/2)$, such that

$$\kappa \leq e_t(x) \leq 1 - \kappa$$

holds for every $t = 1, \ldots, T-1$ and each $x \in \mathcal{X}$.

Assumption 3.3 (Unconfoundedness)**.** for any t and $w \in \{0, 1\}$

$$Y_t(W_{t-1}, w) \perp W_t|X_{1:t-1}.$$

It is worth noting that the above assumption together with Markov exogeneity implies that

$$Y_t(W_{t-1}, w) \perp W_t|X_{t-1}.$$

Under Assumptions 3.1 and 3.3, we can generalize (16) by including the set of covariates in the conditioning variables; for any measurable function f,

$$\mathsf{E}(f(Y_t(W_{0:t}), W_t)|\mathcal{F}_{t-1}) = \mathsf{E}(f(Y_t(W_{t-1}, W_t), W_t)|X_{t-1}) = \mathsf{E}(f(Y_t, W_t)|X_{t-1}). \tag{27}$$

For continuous conditioning covariates X_{T-1}, a simple sample analogue of the objective function is not available due to the lack of multiple observations at any single conditioning value of X_{T-1}. One approach is to use nonparametric smoothing to construct an estimate for conditional welfare. For instance, with a kernel function $K(.)$ and a bandwidth h

$$\widehat{\mathcal{W}}(G|x) = \frac{\sum_{t=1}^{T-1} K(\frac{X_{t-1}-x}{h})\left[\frac{Y_t W_t}{e_t(X_{t-1})}1(X_{t-1}\in G) + \frac{Y_t(1-W_t)}{1-e_t(X_{t-1})}1(X_{t-1}\notin G)\right]}{\sum_{t=1}^{T-1} K(\frac{X_{t-1}-x}{h})}, \tag{28}$$

where we denote $(\cdot|X_{T-1} = x)$ by $(\cdot|x)$. Theorem 4.1 of in Section 4.1 provides a regret bound for (28). The kernel method is a direct way to estimate an optimal policy with the conditional welfare criterion. However, the localization by bandwidth slows down the speed of learning; the regret of conditional welfare can only achieve a $\frac{1}{\sqrt{(T-1)h}}$-rate of convergence rather than a $\frac{1}{\sqrt{T-1}}$-rate. We defer discussion of the statistical properties of the regret bound of the kernel approach to Section 4.1. In this section, we instead pursue an alternative

approach that estimates an optimal policy rule by maximizing an empirical analogue of unconditional welfare over a specified class of decision sets $\mathcal{G}$.

3.2 Bounding the conditional regret: continuous covariate case

We first clarify how a maximizer of conditional welfare (26) can be linked to a maximizer of unconditional welfare. With this result in hand, we can focus on estimating unconditional welfare and choosing a policy by maximizing it. We will show that this approach can attain a $\frac{1}{\sqrt{T-1}}$ rate of convergence. Faster convergence relative to the kernel approach comes at the cost of imposing an additional restriction on the data generating process, as we spell out in Assumption 3.4 below.

A sufficient condition for the equivalence of maximizing conditional and unconditional welfare is that the specified class of policy rules $\mathcal{G}$ includes a first best policy for the *unconditional* problem. Unconditional welfare under policy G is defined as

$$\mathcal{W}_T(G) = \mathsf{E}\left\{Y_T(W_{T-1},1)1(X_{T-1} \in G) + Y_T(W_{T-1},0)1(X_{T-1} \notin G)\right\},$$

and an optimal policy within the class of feasible unconditional decision sets, $\mathcal{G}$, is

$$G_* \quad \in \quad \text{argmax}_{G\in\mathcal{G}}\mathcal{W}_T(G). \tag{29}$$

We define the first-best decision set by

$$G^*_{FB} := \{x \in \mathcal{X} : \mathsf{E}[Y_T(W_{T-1},1) - Y_T(W_{T-1},0)|X_{T-1} = x] \geq 0\}. \tag{30}$$

To ensure that maximizing unconditional welfare corresponds to maximizing conditional welfare over $\mathcal{G}$, we impose the following key assumption:

Assumption 3.4. [Correct specification]

Let $\mathcal{W}_T(G|x)$ be the conditional welfare as defined in (26), we have, at every $x \in \mathcal{X}$,

$$\arg\sup_{G\in\mathcal{G}} \mathcal{W}_T(G) \subset \arg\sup_{G\in\mathcal{G}} \mathcal{W}_T(G|x).$$

With this assumption, we can shift the focus to maximizing unconditional welfare, even when the planner's ultimate objective function is conditional welfare. A condition that implies Assumption 3.4 holds is

$$G^*_{FB} \in \mathcal{G}.$$

This sufficient condition states that the class of policy rules over which unconditional empirical welfare is maximized contains the set of points in $\mathcal{X}$ where the conditional average

treatment effect $\mathsf{E}[Y_T(W_{T-1},1) - Y_T(W_{T-1},0)|X_{T-1} = x]$ is positive. This assumption thus restricts the distribution of potential outcomes at T and its dependence on X_{T-1}. We refer to Assumption 3.4 as 'correct specification'.[9] If the specified class of policies $\mathcal{G}$ is sufficiently rich, the correct specification assumption is credible. If the restrictions placed on $\mathcal{G}$ are motivated by some constraints that the planner has to meet for policy practice (e.g., fairness and interpretability), the correct speicification assumption is restrictive.

The following proposition directly results from Assumption 3.4 and the definition of G^*_{FB}. Under some conditions, we show that the value of the conditional welfare function evaluated at G_* defined in (29) attains the maximum.

Proposition 3.1. Under Assumptions 3.1 (i) and 3.4, an optimal policy rule G_* in terms of the unconditional welfare maximizes the conditional welfare function,

$$G_* \in \text{argmax}_{G\in\mathcal{G}}\mathcal{W}_T(G|X_{T-1}).$$

Furthermore, if the first best solution belongs to the class of feasible policies rules, $G^*_{FB} \in \mathcal{G}$, then we have

$$G^*_{FB} \in \text{argmax}_{G\in\mathcal{G}}\mathcal{W}_T(G|X_{T-1}).$$

Having assumed the relationship of optimal policies between the two welfare criteria, we now show how the unconditional welfare function can formally bound the conditional function. For $G \in \mathcal{G}$ and $X_{T-1} = x$, define conditional regret as

$$R_T(G|x) := \mathcal{W}_T(G_*|x) - \mathcal{W}_T(G|x).$$

Note that the unconditional regret can be expressed as an integral of conditional regret,

$$\begin{aligned}\mathcal{W}_T(G) &= \int \mathcal{W}_T(G|x)dF_{X_{T-1}}(x),\\ R_T(G) := \mathcal{W}_T(G_*) - \mathcal{W}_T(G) &= \int R_T(G|x)dF_{X_{T-1}}(x).\end{aligned}$$

For $x' \in \mathcal{X}$, define

$$\begin{aligned}A(x',G) &:= \{x \in \mathcal{X} : R_T(G|x) \geq R_T(G|x')\},\\ p_{T-1}(x',G) &:= \Pr\left(X_{T-1} \in A(x',G)\right) = \int_{x\in A(x',G)} dF_{X_{T-1}}(x),\end{aligned}$$

and denote by x^{obs} the observed value of X_{T-1}. We assume the following:

[9] Kitagawa and Tetenov (2018), Kitagawa et al. (2021), and Sakaguchi (2021) consider correct specification assumptions exclusively for unconditional welfare criteria. These assumptions correspond to $G^*_{FB} \in \mathcal{G}$.

Assumption 3.5. For $x^{obs} \in \mathcal{X}$ and any $G \in \mathcal{G}$, there exists a positive constant $\underline{p}$ such that

$$p_{T-1}(x^{obs}, G) \geq \underline{p} > 0. \tag{31}$$

Remark 2. This assumption is satisfied if X_{T-1} is a discrete random variable taking a finite number of different values. In this case, $p_{T-1}(x^{obs}, G) \geq \min_{x \in \mathcal{X}} \Pr(X_{T-1} = x) > 0$, so we can set $\underline{p} = \min_{x \in \mathcal{X}} \Pr(X_{T-1} = x)$. If X_t is continuous, then we need to exclude a set of points around the maximum of the function $R_T(G|x)$ for the assumption to hold. Namely, we can assume that we focus on x belonging to a compact subset $\tilde{\mathcal{X}} \subset \mathcal{X}$ such that $\arg\max_{x \in \mathcal{X}} R_T(G|x) \notin \tilde{\mathcal{X}}$. If we would like to include the whole support of X_t, we can modify the proof by imposing an additional uniform continuity condition on $R_T(G|\cdot)$.

The following lemma provides a bound for conditional regret $R_T(G|x^{obs})$ using unconditional regret $R_T(G)$.

LEMMA 3.1. Under Assumptions 3.5,

$$R_T(G|x^{obs}) \leq \frac{1}{\underline{p}} R_T(G) \tag{32}$$

Using of Assumption 3.5 and Lemma 3.1 to bound conditional regret, we proceed to construct an empirical analogue of the welfare function and provide theoretical results for the regret bound. The sample analogue of $\mathcal{W}_T(G)$ can be expressed as,

$$\widehat{\mathcal{W}}(G) = \frac{1}{T-1} \sum_{t=1}^{T-1} \left[\frac{Y_t W_t}{e_t(X_{t-1})} 1(X_{t-1} \in G) + \frac{Y_t(1 - W_t)}{1 - e_t(X_{t-1})} 1(X_{t-1} \notin G) \right], \tag{33}$$

and we define

$$\hat{G} \in \operatorname{argmax}_{G^c \in \mathcal{G}} \widehat{\mathcal{W}}(G). \tag{34}$$

Recall that $\mathcal{F}_{t-1} = \sigma(X_{1:t-1})$. In addition, define two intermediate welfare functions,

$$\begin{aligned} \bar{\mathcal{W}}(G) &:= \frac{1}{T-1} \sum_{t=1}^{T-1} \mathsf{E}\left\{ Y_T(W_{T-1}, 1) 1(X_{T-1} \in G) + Y_T(W_{T-1}, 0) 1(X_{T-1} \notin G) | \mathcal{F}_{t-1} \right\}, \\ \widetilde{\mathcal{W}}(G) &:= \frac{1}{T-1} \sum_{t=1}^{T-1} \mathsf{E}\left\{ Y_T(W_{T-1}, 1) 1(X_{T-1} \in G) + Y_T(W_{T-1}, 0) 1(X_{T-1} \notin G) \right\}. \end{aligned} \tag{35}$$

To obtain a regret bound for unconditional welfare, Assumption 2.4 is modified to

Assumption 3.6. For any G_1, $G_2 \in \mathcal{G}$, there exists some constant c

$$\mathcal{W}_T(G_1) - \mathcal{W}_T(G_2) \leq c\big[\widetilde{\mathcal{W}}(G_1) - \widetilde{\mathcal{W}}(G_2)\big], \tag{36}$$

with probability approaching one in the sense that P_T (inequality (36) holds) $\to 1$ as $T \to \infty$, where P_T is the probability distribution for $X_{0:T-1}$.

Below, we bound the regret for conditional welfare $\mathcal{W}_T(G_*|x^{obs}) - \mathcal{W}_T(\hat{G}|x^{obs})$ by regret for unconditional welfare $\left[\mathcal{W}_T(G_*) - \mathcal{W}_T(\hat{G})\right]$, and further by $\left[\widetilde{\mathcal{W}}(G) - \widehat{\mathcal{W}}(G)\right]$ (up to constant factors).

$$\begin{aligned}
\mathcal{W}_T(G_*|x^{obs}) - \mathcal{W}_T(\hat{G}|x^{obs}) &\leq \frac{1}{\underline{p}}\left[\mathcal{W}_T(G_*) - \mathcal{W}_T(\hat{G})\right], \\
&\leq \frac{c}{\underline{p}}\left[\widetilde{\mathcal{W}}(G_*) - \widetilde{\mathcal{W}}(\hat{G})\right], \\
&\leq \frac{2c}{\underline{p}} \sup_{G\in\mathcal{G}} \left|\widehat{\mathcal{W}}(G) - \widetilde{\mathcal{W}}(G)\right|, \qquad (37)
\end{aligned}$$

The first inequality follows from Lemma 3.1 and Assumption 3.5. The second inequality follows from (36). The last inequality follows from a similar argument to (22).

Note $\widehat{\mathcal{W}}(G) - \widetilde{\mathcal{W}}(G)$ is *not* a sum of MDS. Instead, it can be decomposed as

$$\begin{aligned}
\widehat{\mathcal{W}}(G) - \widetilde{\mathcal{W}}(G) &= \bar{\mathcal{W}}(G) - \widetilde{\mathcal{W}}(G) + (\widehat{\mathcal{W}}(G) - \bar{\mathcal{W}}(G)), \\
&= I + II, \qquad (38)
\end{aligned}$$

where $I := \bar{\mathcal{W}}(G) - \widetilde{\mathcal{W}}(G)$ and $II := \widehat{\mathcal{W}}(G) - \bar{\mathcal{W}}(G)$. Subject to assumptions specified later, Theorem 3.1 below shows that II, which is a sum of MDS, converges at $\frac{1}{\sqrt{T-1}}$-rate, and Theorem 3.2 below shows that I converges at the same rate.

(38) reveals that our proof strategy is considerably more complicated than the proof for the EWM model with i.i.d. observations of Kitagawa and Tetenov (2018), although the rates are similar. Specifically, we need to derive a novel bound for the tail probability of the sum of martingale difference sequences. Moreover, to achieve uniform bounds, we need to utilize a chaining technique without applying symmetrization. In addition, we need to handle complex functional classes induced by non-stationary processes. For the EWM model, the main task is to show the convergence rate of a simpler analogue of II, which can be achieved with standard empirical process theory for i.i.d. samples. In comparison, we not only have to treat our II more carefully due to time-series dependence, but we also have to deal with I.

Assumption 3.7 is the first of several further assumptions we must impose to proceed.

Assumption 3.7. There exists $M < \infty$ such that the support of outcome variable Y_t is contained in $[-M/2, M/2]$.

This implies that the welfare functions are also bounded.

3.2.1 Bounding II by MDS

Define empirical welfare at time t and its population conditional expectation as follows,

$$\widehat{\mathcal{W}}_t(G) = \frac{Y_t W_t}{e_t(X_{t-1})} 1(X_{t-1} \in G) + \frac{Y_t(1-W_t)}{1-e_t(X_{t-1})} 1(X_{t-1} \notin G),$$
$$\bar{\mathcal{W}}_t(G) = \mathsf{E}\left\{Y_t(W_{t-1},1)1(X_{t-1} \in G) + Y_t(W_{t-1},0)1(X_{t-1} \notin G)|\mathcal{F}_{t-1}\right\}.$$

Thus, we examine two summations

$$\widehat{\mathcal{W}}(G) = \frac{1}{T-1}\sum_{t=1}^{T-1} \widehat{\mathcal{W}}_t(G),$$
$$\bar{\mathcal{W}}(G) = \frac{1}{T-1}\sum_{t=1}^{T-1} \bar{\mathcal{W}}_t(G).$$

For each $t = 1, \ldots, T-1$, define a functional class indexed by $G \in \mathcal{G}$,

$$\mathcal{H}_t = \{h_t(\cdot; G) = \widehat{\mathcal{W}}_t(G) - \bar{\mathcal{W}}_t(G) : G \in \mathcal{G}\}. \tag{39}$$

The arguments of the function $h_t(\cdot; G)$ are Y_t, W_t, and X_{t-1}. Given the class of functions $\mathcal{H}_t$, we consider a martingale difference array $\{h_t(y_t, w_t, x_{t-1}; G)\}_{t=1}^n$,[10] and denote the average by

$$\mathsf{E}_n h := \frac{1}{n}\sum_{t=1}^{n} h_t(y_t, w_t, , x_{t-1}, G).$$

Denote d_t as the set of the random variable $\{Y_t, W_t, X_{t-1}\}$, and $X_t := \{Y_t, W_t, Z_t\}$. Since we do not restrict d_t to be stationary, we shall handle a vector of functional classes which possibly varies over t. Later we shall assume that the covering number for the functional classes over all t can be bounded the covering number of all one-dimensional projection. We define the covering number for $\overline{h}_n \stackrel{\text{def}}{=} \{h_1(d_1, G), h_2(d_2, G), \cdots, h_n(d_n, G)\}$, with the function classes respectively $\mathbf{H_n} = \{\mathcal{H}_1, \mathcal{H}_2, \cdots, \mathcal{H}_n\}$. Let α_n be a n dimensional vector in $\mathbb{R}^n$. Let $\circ$ denote the element-wise product. In addition, the envelope for each function class is denoted by $\overline{H}_n = \{H_1, H_2, \cdots, H_n\}$. Furthermore, we denote the one-dimensional covering number for time t as $\mathcal{N}(\varepsilon, \mathcal{H}_t, \rho(.))$, the minimum number of balls with $\{g, h \in \mathbf{H_n} : \rho(g,h) < \varepsilon\}$ covering the classes $\mathbf{H_n}$ for a metric ρ. For the following we try to obtain maximum inequality for $\mathsf{E}[|\,\mathsf{E}_n\, h|_{\mathbf{H_n}}] \stackrel{\text{def}}{=} \mathsf{E}[\sup_{\overline{h}_n \in \mathbf{H_n}} |\,\mathsf{E}_n\, h(X)|]$. H_t is an envelope function of $\mathcal{H}_t$, and $\|H_t\|_{Q,r} \stackrel{\text{def}}{=} (\int_x |H_t(x)|^r dQ(x))^{1/r}$. And we choose ρ to be the $\|.\|_{Q,r}$ norm.

[10] We use n as the number of summands since for different theorems in this and following sections (such as those for multi-period and multi-lag welfare functions), the numbers of summands are different.

We define the covering number for the function class $\mathbf{H_n}$ to be,

$$\mathcal{N}(\delta|\alpha_n \circ \overline{H}_n|_2, \alpha_n \circ \mathbf{H_n}, |.|_2).$$

Assumption 3.8. For a finite discrete measures Q, positive constants K, v, e, and an integer r, we assume, uniformly over all $\tilde{\alpha}_n$, $\tilde{\alpha}_n = \alpha_n/|\alpha_n|_2$,

$$\mathcal{N}(\delta|\tilde{\alpha}_n \circ \overline{H}_n|_2, \tilde{\alpha}_\mathbf{n} \circ \mathbf{H_n}, |.|_\mathbf{2}) \leq \max_t \sup_Q \mathcal{N}(\varepsilon \|H_t\|_{Q,r}, \mathcal{H}_t, \|.\|_{Q,2}) \lesssim K(v+1)(4e)^{v+1}(\frac{2}{\varepsilon})^{rv}. \tag{40}$$

The above assumption restricts the complexity of the functional class to be of polynomial discrimination. v appears in the derived regret bounds. See Appendix D.4 for a justification of Assumption 3.8. For a series of functions f_t and g_t, we define the metric $\rho_{2,n}(f,g) = (n^{-1}\sum_t |f_t - g_t|^2)^{1/2}$ and $\sigma_n(f,g) = (n^{-1}\sum_t \mathsf{E}[(f_t - g_t)^2|\mathcal{F}_{t-1}])^{1/2}$.

Assumption 3.9. There exist some constant $L > 0$, $\Pr(\sigma_n(f,g)/\rho_{2,n}(f,g) > L) \to 0$ as $n \to \infty$. Also, $\Pr((n^{-1}\sum_t \mathsf{E}[(f_t - g_t)^2|\mathcal{F}_{t-2}])^{1/2}/\rho_{2,n}(f,g) > L) \to 0$ as $n \to \infty$.

$\rho_{2,n}(f,g)^2$ is the quadratic variation difference and $\sigma_n(f,g)^2$ is its conditional equivalent. It is evident that $\rho_{2,n}(f,g)^2 - \sigma_n(f,g)^2$ involves martingale difference sequences. In the special case of i.i.d. observations, $\sigma_n(f,g)^2$ is the unconditional expectation. Assumption 3.9 can thus be viewed as specifying that $\rho_{2,n}(f,g)^2$ and $\sigma_n(f,g)^2$ are asymptotically equivalent in a probability sense. A similar condition can be seen, for example, in Theorem 2.23 in Hall and Heyde (2014).

Then we have for II:

THEOREM 3.1. Under Assumption 3.1 to 3.3, and 3.7 to 3.9,

$$\sup_{G\in\mathcal{G}} |\widehat{\mathcal{W}}(G) - \bar{\mathcal{W}}(G)| \lesssim_p C\frac{M}{\kappa}\sqrt{\frac{v}{T-1}},$$

where C is a constant that depends only on M and κ.

The proof of Theorem 3.1 is presented in Appendix D.3.

3.2.2 Bounding I

Here, we complete the process of bounding unconditional regret. Let

$$S_t(G) := \; Y_t(W_{t-1}, 1)1(X_{t-1} \in G) + Y_t(W_{t-1}, 0)1(X_{t-1} \notin G),$$

and

$$\overline{S}_t(G) = \mathsf{E}(S_t(G)|\mathcal{F}_{t-1}) - \mathsf{E}(S_t(G)|\mathcal{F}_{t-2}). \tag{41}$$

$$\tilde{S}_t(G) := \mathsf{E}(S_t(G)|\mathcal{F}_{t-2}) - \mathsf{E}(S_t(G)). \tag{42}$$

We apply a tail bound to the summ of $\mathsf{E}(S_t(G)|\mathcal{F}_{t-1}) - \mathsf{E}(S_t(G)|\mathcal{F}_{t-2})$. The summand $\mathsf{E}(S_t(G)|\mathcal{F}_{t-2}) - \mathsf{E}(S_t(G))$ is handled below.

Recalling (35) and (38), we have $I = \frac{1}{T-1}\sum_{t=1}^{T-1}\tilde{S}_t(G) + \frac{1}{T-1}\sum_{t=1}^{T-1}\overline{S}_t(G)$.

Define the functional class,

$$\mathcal{S}_t = \{f_t : f_t = \mathsf{E}(S_t(G)|\mathcal{F}_{t-2}) - \mathsf{E}(S_t(G)) : G \in \mathcal{G}\}, \tag{43}$$

$$\tilde{\mathcal{S}}_t = \{f_t : f_t = \mathsf{E}(S_t(G)|\mathcal{F}_{t-1}) - \mathsf{E}(S_t(G)|\mathcal{F}_{t-2}) : G \in \mathcal{G}\}. \tag{44}$$

We assume that a process X_t can be expressed as the following form of structural equation, namely, $X_t = g_t(\varepsilon_t, \varepsilon_{t-1}, \cdots, \varepsilon_{-\infty})$, where ε_ts are i.i.d. random variables. We now define the dependence adjusted norm for an arbitrary process X_t, (where $\ell, q \geq 0$ are integers) as

$$\theta_{x,q} \stackrel{\text{def}}{=} \max_t \sum_{\ell \geq 0}^{\infty} \|X_t - X^*_{t,\ell}\|_q, \tag{45}$$

where $\|X\|_q$ is defined as $(\mathsf{E}\,|X|^q)^{1/1}$ for a random variable X, $X^*_{t,\ell}$ is the random variable $X_{t,\ell}$ with the ℓth lag replaced by an independent copy of $X_{t,\ell}$, and the subexponential/Gaussian dependence adjusted norm is denoted as,

$$\Phi_{\phi_v} = \sup_{q \geq 2}(\theta_{x,q}/q^{\tilde{v}}), \tag{46}$$

where $\tilde{v} = 1$ for subexponential, and $\tilde{v} = 1/2$ in the subGaussian case.

We impose the following:

Assumption 3.10. $S_t(G) = g_t(\varepsilon_t, \cdots, \varepsilon_{-\infty})$, where ε_t are i.i.d. random variables.

Assumption 3.11. $Z_t \perp X_{0:t-1}|X_{t-1}$. $\partial\,\mathsf{E}(S_t(G)|\mathcal{F}_{t-2})/\partial\varepsilon_{t-l-1}$ has envelope $F_{t,l}(X_{t-2})$. $\sup_G |\partial\,\mathsf{E}(S_t(G)|\mathcal{F}_{t-2})/\partial\varepsilon_{t-l-1}| \leq F_{t,l}(X_{t-2})$ for every t, l.

Assumption 3.12. $\sup_{q\geq 2}(\sum_{l\geq 0}\max_t \|F_{t,l}(X_{t-2})\varepsilon_{t-l-1}\|_q)/q^{\tilde{\nu}} \stackrel{\text{def}}{=} \Phi_{\phi_v,F} < \infty$. $\Phi_{\phi_v} < \infty$. $\gamma = 1/(1+2\tilde{v})$ ($\tilde{v} = 1/2, \quad \text{or} \quad 1$).

Assumption 3.13. Suppose that $F_t(\tilde{F}_t)$ is the envelope of the functional class $\tilde{S}_t(G)(\overline{\tilde{S}}_t(G))$. Define $\overline{F}_T = (F_1, F_2, \cdots, F_{T-1})$ ($\overline{\tilde{F}}_T = (\tilde{F}_1, \tilde{F}_2, \cdots, \tilde{F}_{T-1})$) , and $\mathbf{F}_T = \{\mathcal{S}_1, \mathcal{S}_2, \cdots, \mathcal{S}_{T-1}\}(\tilde{\mathbf{F}}_{T-2} = \{\tilde{\mathcal{S}}_1, \tilde{\mathcal{S}}_2, \cdots, \tilde{\mathcal{S}}_{T-1}\})$. Let Q denote a discrete measure over a finite number of points, and V

be a positive integer,

$$\mathcal{N}(\delta|\tilde{\alpha}_T \circ \overline{F}_T|_2, \tilde{\alpha}_{\mathbf{T}} \circ \mathbf{F}_{\mathbf{T}}, |.|_{\mathbf{2}}) \leq \max_t \sup_Q \mathcal{N}(\delta|F_{p,t}|_{2,Q}, \mathcal{S}_t, |.|_{Q,2}) \lesssim (1/\delta)^V. \tag{47}$$

$$\mathcal{N}(\delta|\tilde{\alpha}_T \circ \tilde{\overline{F}}_T|_2, \tilde{\alpha}_{\mathbf{T}} \circ \tilde{\mathbf{F}}_{\mathbf{T}}, |.|_{\mathbf{2}}) \leq \max_t \sup_Q \mathcal{N}(\delta|\tilde{F}_{p,t}|_{2,Q}, \tilde{\mathcal{S}}_t, |.|_{Q,2}) \lesssim (1/\delta)^v. \tag{48}$$

Assumption 3.10 imposes that the time series $S_t(G)$ as the structural equation of a sequence of i.i.d. innovations ε_t. Assumption 3.11 is a standard envelope assumption on the functional class. It states that the partial derivative is enveloped with a function of X_{t-2}. This is partly due to Assumption 3.1. Assumption 3.12 concerns moments, and is also related to the dependency of time series over time. Moreover, it implicitly imposes a tail assumption on the underlying time series. Assumption 3.13 restricts the complexity of the the functional class.

Then, we have the following tail probability bound.

THEOREM 3.2. Assuming Assumptions 3.1 to 3.3 and 3.9- 3.13,

$$\sup_{P_T \in \mathcal{P}_T(M,\kappa)} \left| \bar{\mathcal{W}}(G) - \widetilde{\mathcal{W}}(G) \right| \lesssim_p \frac{c_T [V \log T]^{1/\gamma} 2e\gamma \Phi_{\phi_v,F}}{\sqrt{T-1}} + CM\sqrt{v/(T-1)}. \tag{49}$$

A proof is presented in Appendix D.5. The bound depends on the complexity of the functional class, V, and the time series dependency, $\Phi_{\phi_v,F}$. As we consider the sample analogue of unconditional welfare, we have the $\frac{1}{\sqrt{T-1}}$ rate of convergence.

3.2.3 The regret bound

The overall bound for unconditional welfare is obtained by combining the bounds of Theorem 3.1 and 3.2 with (37). Let P_T be a joint probability distribution of a sample path of length $(T-1)$, $\mathcal{P}_T(M,\kappa)$ be the class of P_T, which satisfies Assumptions 3.2 and 3.7, and E_{P_T} be the expectation taken over different realizations of random samples. Then

THEOREM 3.3. Under Assumption 3.1 to 3.13,

$$\sup_{P_T \in \mathcal{P}_T(M,\kappa)} \mathsf{E}_{P_T}[\mathcal{W}_T(G_*) - \mathcal{W}_T(\hat{G})] \lesssim CM\sqrt{\frac{v}{T-1}} + \frac{c_T [V \log T]^{1/\gamma} 2e\gamma \Phi_{\phi_v,F}}{\sqrt{T-1}}, \tag{50}$$

where G_* is defined in (29).

The proof follows directly from Theorems 3.1 and 3.2.

Remark. We presented our main results of welfare regret upper bounds under the first-order Markovian structure of Assumption 3.1. This is for ease of exposition, and it is

straightforward to relax Assumption 3.1 by introducing a higher order Markovian structure. For example, current observations can depend causally or statistically on the realized treatment and covariates over the last q periods, $q \geq 1$. Assumption 3.1 corresponds to $q = 1$.

Assumption 3.1* [q-th order Markov properties] For an integer $q \geq 0$, the time-series of potential outcomes and observable variables satisfy the following conditions:

(i) *q-th order Markovian exclusion*: for $t = q+1, q+2, \ldots, T$ and for arbitrary treatment paths $(w_{0:t-q-1}, w_{t-q:t})$ and $(w'_{0:t-q-1}, w_{t-q:t})$, where $w_{0:t-q-1} \neq w'_{0:t-q-1}$,

$$Y_t(w_{0:t-q-1}, w_{t-q:t}) = Y_t(w'_{0:t-q-1}, w_{t-q:t}) := Y_t(w_{t-q:t})$$

holds with probability one.

(ii) *q-th order Markovian exogeneity*: for $t = q, q+1, \ldots, T$ and any treatment path $w_{0:t}$,

$$Y_t(w_{0:t}) \perp X_{0:t-1} | X_{t-q:t-1},$$

and for $t = q, q+1, \ldots, T-1$,

$$W_t \perp X_{0:t-1} | X_{t-q:t-1}.$$

When the Markov properties are of order q, the propensity score can be defined as $e_t(x) = \Pr(W_t = 1 | X_{t-q:t-1} = x)$, where the vector x is of the the same dimension as the random vector $X_{t-q:t-1}$. Assumption 3.3 can be modified to: for any t and $w \in \{0, 1\}$, $Y_t(W_{t-q:t-1}, w) \perp W_t | X_{t-q:t-1}$.

In addition, a policy G becomes a region defined on the space of the random vector $X_{T-q:T-1}$, and

$$\begin{aligned}
\mathcal{W}_T(G) &= \mathsf{E}\left\{Y_T(W_{T-q:T-1}, 1)1(X_{T-q:T-1} \in G) + Y_T(W_{T-q:T-1}, 0)1(X_{T-q:T-1} \notin G)\right\}, \\
\widehat{\mathcal{W}}(G) &= \frac{1}{T-q} \sum_{t=q}^{T-1} \left[\frac{Y_t W_t}{e_t(X_{t-q:t-1})} 1(X_{t-q:t-1} \in G) + \frac{Y_t(1-W_t)}{1-e_t(X_{t-q:t-1})} 1(X_{t-q:t-1} \notin G) \right].
\end{aligned}$$

$\bar{\mathcal{W}}(G)$, $\widetilde{\mathcal{W}}(G)$, and all associated function classes can be similarly redefined.

4 Extensions

In this section, we discuss a few possible extensions. Section 4.1 describes a non-parametric kernel method which directly bounds the conditional regret, and discusses its statistical properties. To motivate adopting a conditional welfare function, we present in Section 4.1 an example where the optimal policies in terms of unconditional and conditional welfare do not coincide. In the Appendix we present further examples where we find the unconditional solution corresponds to the conditional solution, which turns out to depend on the feasibility of the first best solution. Section 4.2 introduces a multi-period policy making framework. Section 4.3 concerns the ability of the current methods to handle the Lucas critiques. Section 4.4 discusses a few connections to the literature.

4.1 Nonparametric method to bound the conditional regret

In Section 3.2, we showed that we can boundconditional regret by unconditional regret if the unconditional first best policy is feasible. In Section 4.1.1 and Appendix A, we present examples where this method is applicable, and examples where it is not. When there is no direct correspondence between the solutions for the unconditional and conditional welfare functions, we proceed to the nonparametric method of Subsection 4.1.2.

4.1.1 Motivation

In this subsection, we discuss the relationship between the optimal policy solutions in terms of conditional and unconditional welfare. This relationship is not straightforward. In some cases, we do find a trivial correspondence between the two solutions, which implies that unconditional welfare does bound conditional welfare, and the methods and results in Section 3.2 directly apply. However, in other cases, when the first best policy is not feasible, we shall use the kernel estimator. This motivates us to present the kernel estimator as an important alternative to the method of Section 3.2.

Example 4.1. Observing $X_{T-1} = (Y_{T-1}, W_{T-1}, Z_{T-1})' \in \mathbb{R} \times \{0,1\} \times \mathbb{R}^2$ at time $T-1$, the planner chooses W_T based on the last two continuous variables. The feasible policy class is rectangles in $\mathbb{R}^2$:

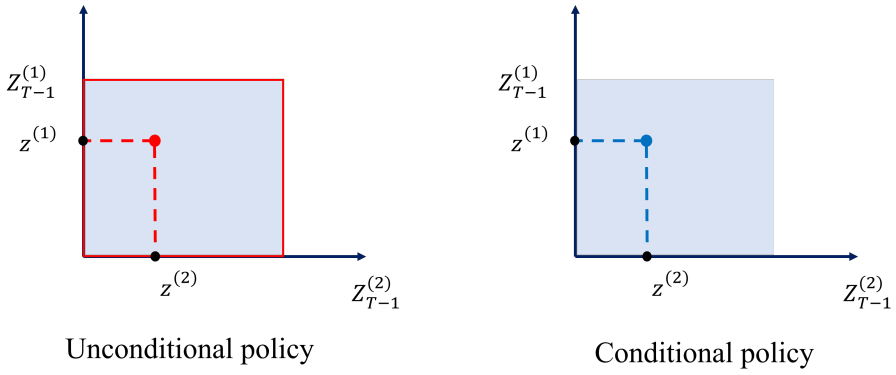
$$\mathcal{G} = \{z \in [a_1, a_2] \times [b_1, b_2] : a_1, a_2, b_1, b_2 \in \mathbb{R}\}.$$

The corresponding unconditional problem is $\max_{G \in \mathcal{G}} \mathcal{W}_T(G)$. Suppose the planner is interested in maximizing the welfare conditional on $Z_{T-1} = z := (z^{(1)}, z^{(2)})$. The conditional problem is to find

$$\arg\max_{G \in \mathcal{G}} \mathcal{W}_T(G|Z_{T-1} = z), \tag{51}$$

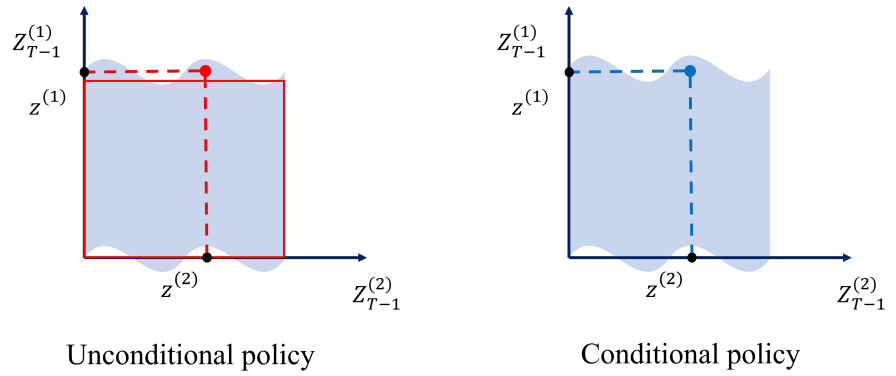
We illustrate how the policy solutions can differ between conditional and unconditional welfare functions. In Figures 1 and 2, the *shaded area* represents the region where the conditional average treatment effect is positive. The first best unconditional policy assigns $W_T = 1$ to any value of Z_{T-1} inside this region, and $W_T = 0$ to any point outside it. This policy is also the solution to (51). The *red rectangle* is the best feasible (i.e. rectangular) unconditional policy. This policy assigns $W_T = 1$ to any realization of Z_{T-1} inside the rectangle. The conditional policy concerns what policy to assign only at a particular value of Z_{T-1} corresponding to its realized value in the data (blue point in the right-hand side panel of Figure 1). If the best feasible unconditional policy (red rectangle) agrees with the first best unconditional policy (shared area), then the policy choice informed by the unconditional policy is guaranteed to be optimal in terms of the conditional policy at any conditioning value of Z_{T-1}.

Figure 1. $G^*_{\text{FB}} \in \mathcal{G}$



The dashed red line is the same as the dashed blue line.

Figure 2. $G^*_{\text{FB}} \notin \mathcal{G}$



The dashed red line is different to the dashed blue line.

Figure 2 shows a case where the first best unconditional policy is not contained in the class of feasible policies: it is not possible to implement the policy choice that coincides with the shaded area. In this case, the policy chosen by the unconditionally optimal feasible policy

(red rectangle in the left-hand side panel) does not coincide with the optimal policy choice of the conditional policy. The highlighted point $(z^{(1)}, z^{(2)})$ lies outside the red rectangle, so the best feasible unconditional policy would set $W_t = 0$. However, it lies within the shaded region, so the conditional average treatment effect given $Z_{T-1} = (z^{(1)}, z^{(2)})$ is positive, and the optimal policy conditional on $Z_{T-1} = (z^{(1)}, z^{(2)})$ is to set $W_t = 1$.

These two examples show the importance of the first best unconditional policy: when it is included in the set of feasible unconditional policies, the solution to the conditional problem corresponds to the solution to the unconditional problem. When it is not included, we do not have this correspondence. In Appendix A.1, we show that the feasibility of the first best solution to the unconditional problem is a sufficient condition. A sufficient and necessary condition is given by Assumption 3.4.

4.1.2 Nonparametric estimator of the optimal conditional policy

If the solution to the conditional problem cannot be obtained from the unconditional problem, the conditional problem must be solved directly. That is, the SP should estimate an optimal policy from the empirical analogue of the conditional welfare function. If the conditioning variables are continuous, some type of nonparametric smoothing is unavoidable. Here we use a kernel-based method to estimate the optimal conditional policy. Recall that welfare conditional on $X_{T-1} = x$ can be written as

$$\mathcal{W}_T(G|x) = E\left\{Y_T(W_{T-1}, 1)1(X_{T-1} \in G) + Y_t(W_{T-1}, 0)1(X_{T-1} \notin G)|X_{T-1} = x\right\}.$$

For simplicity, we let $X_{T-1} \in \mathbb{R}$. Then (28) can be rewritten as [11]

$$\widehat{\mathcal{W}}(G|x) = \frac{\sum_{t=1}^{T-1} K_h(X_{t-1}, x)\widehat{\mathcal{W}}_t(G)}{\sum_{t=1}^{T-1} K_h(X_{t-1}, x)}, \tag{52}$$

where $\widehat{\mathcal{W}}_t(G) := \frac{Y_t W_t}{e_t(X_{t-1})}1(X_{t-1} \in G) + \frac{Y_t(1-W_t)}{1-e_t(X_{t-1})}1(X_{t-1} \notin G)$, and $K_h(a, b) := \frac{1}{h}K(\frac{a-b}{h})$ with $K(\cdot)$assumed to be a bounded kernel function with a bounded support.

Recall that $\mathcal{G}$ is the set of feasible policies conditional on $X_{T-1} = x$. We define

$$\begin{aligned} G_x^* &\in \text{argmax}_{G\in\mathcal{G}}\mathcal{W}(G|x) \\ \hat{G}_x &\in \text{argmax}_{G\in\mathcal{G}}\widehat{\mathcal{W}}(G|x) \end{aligned}$$

[11] If the set of conditioning variables X_{T-1} contains both continuous and discrete components, we can adapt a hyper method to construct a valid sample analogue combing kernel-smoothing (for continuous variables) and subsamples (for discrete variables). In this section, we focus on the case where the target welfare function is conditional on a univariate continuous variable

to be the maximizers of $\mathcal{W}(G|x)$ and $\widehat{\mathcal{W}}(G|x)$, and

$$\bar{\mathcal{W}}_h(G|x) = \frac{\sum_{t=1}^{T-1} K_h(X_{t-1}, x)\mathcal{W}_t(G|x)}{\sum_{t=1}^{T-1} K_h(X_{t-1}, x)}, \tag{53}$$

where the second equality follows from Assumption 3.3.

The invariance of welfare ordering assumption is modified to,

Assumption 4.1 (Invariance of welfare ordering)**.** For any G_1, $G_2 \in \mathcal{G}$ and $x \in \mathcal{X}$, there exists some constant c

$$\mathcal{W}_T(G_1|x) - \mathcal{W}_T(G_2|x) \leq c\big[\bar{\mathcal{W}}_h(G_1|x) - \bar{\mathcal{W}}_h(G_2|x)\big]. \tag{54}$$

Similar to Assumption 2.4, (54) holds if the stochastic process $S_t(x) := Y_T(W_{T-1}, 1)1(X_{T-1} \in G) + Y_t(W_{T-1}, 0)1(X_{T-1} \notin G)|_{X_{T-1}=x}$ is weakly stationary.

The following theorem shows an upper bound for conditional regret in the one dimensional covariate case (i.e., $X_t \in \mathbb{R}$). This result can be readily extended to the multiple covariate case.

THEOREM 4.1. Under the assumptions specified in Appendix D.6, we will have

$$\sup_{P_T \in \mathcal{P}_T(M,\kappa)} \sup_{G \in \mathcal{G}} \sup_{x \in \mathcal{X}} \mathsf{E}_{P_T}[\mathcal{W}_T(G|x) - \mathcal{W}_T(\hat{G}_x|x)] \leq c_1(\sqrt{(T-1)h}^{-1} + (T-1)^{-1} + h^2).$$

Setting $h = O(T^{-1/5})$, the right hand side bound is $O(T^{-2/5})$.

A proof is presented in Appendix D.6.

4.2 Multi-period welfare

So far we have considered only the case of one period and one lag. This subsection extends the setting to a multiple period policy framework. In the interest of space, we focus on cases with discrete covariates, and extend the simple model in Section 2.2 to a two-period welfare function. Extension to cases with more than two periods is trivial. Recall Assumption 2.1, which imposed Markov properties on the data generating processes,

$$\begin{aligned} Y_t(w_{0:t}) &= Y_t(w_{t-1}, w_t), \\ Y_t(w_{0:t}) &\perp X_{0:t-1}|W_{t-1}, \\ W_t &\perp X_{0:t-1}|W_{t-1}. \end{aligned}$$

The SP chooses policy rules for two periods, $g_1(\cdot)$ and $g_2(\cdot) : \{0,1\} \to \{0,1\}$, to maximise

aggregate welfare over periods T and $T+1$. The decision in the second period, $g_2(.)$, is not contingent on the functional form of $g_1(.)$. However, we assume that period $T+1$ welfare is conditional on the treatment choice in the period T i.e. $g_1(W_{T-1})$. This means there is an information update in period T.

$$\begin{aligned}\mathcal{W}_{T:T+1}(g_1(.), g_2(.)|\mathcal{F}_{T-1}) &= \mathcal{W}_{T:T+1}(g_1(.), g_2(.)|W_{T-1}), \\ &= \mathcal{W}_T(g_1(.)|W_{T-1}) + \mathcal{W}_{T+1}\left(g_2(.)|W_T = g_1\left(W_{T-1}\right)\right), \end{aligned} \tag{55}$$

Note that there are only two feasible maps in this case: for $i \in 1, 2$, $g_i(\cdot)$ maps 0 to 0 and 1 to 1, or $g_i(\cdot)$ maps 0 to 1 and 1 to 0. In the following, we suppress the (.) in $g_i(.)$, when the meaning is clear from context.

$$\begin{aligned}&\mathcal{W}_{T:T+1}(g_1, g_2|W_{T-1} = w) \\ &= \mathsf{E}\left\{Y_T(W_{T-1}, 1)g_1(W_{T-1}) + Y_T(W_{T-1}, 0)\left(1 - g_1(W_{T-1})\right)|W_{T-1} = w\right\} \\ &+ \mathsf{E}\left\{Y_{T+1}(W_T, 1)g_2(W_T) + Y_{T+1}(W_T, 0)(1 - g_2(W_T))|W_T = g_1(w)\right\}, \end{aligned} \tag{56}$$

where the last term follows from the exclusion condition:

$$\begin{aligned}&\mathsf{E}\left\{Y_{T+1}(g_1(W_{T-1}), 1)g_2(W_T) + Y_{T+1}(g_1(W_{T-1}), 0)(1 - g_2(W_T))|W_{T-1} = w\right\} \\ &= \mathsf{E}\left\{Y_{T+1}(W_T, 1)g_2(W_T) + Y_{T+1}(W_T, 0)(1 - g_2(W_T))|W_T = g_1(W_{T-1}), W_{T-1} = w\right\} \\ &= \mathsf{E}\left\{Y_{T+1}(W_T, 1)g_2(W_T) + Y_{T+1}(W_T, 0)(1 - g_2(W_T))|W_T = g_1(w)\right\}.\end{aligned}$$

Recall the definition $T(w) = \#\{1 \leq t \leq T-1 : W_t = w\}$. We define $T(g_1(w))$ similarly. Then the empirical counterpart of (55) can be written,

$$\begin{aligned}\widehat{\mathcal{W}}_{T:T+1}(g_1, g_2|w) =& \frac{1}{T(w)} \sum_{t:W_{t-1}=w} \left\{\frac{Y_t W_t g_1(W_{t-1})}{e_t(W_{t-1})} + \frac{Y_t\left(1 - W_t\right)\left(1 - g_1(W_{t-1})\right)}{1 - e_t(W_{t-1})}\right\} \\ &+ \frac{1}{T(g_1(w))} \sum_{t:W_{t-1}=g_1(w)} \left\{\frac{Y_t W_t g_2(W_{t-1})}{e_t(W_{t-1})} + \frac{Y_t\left(1 - W_t\right)\left(1 - g_2(W_{t-1})\right)}{1 - e_t(W_{t-1})}\right\}. \end{aligned} \tag{57}$$

The maximizer of (57), $\hat{g}_1$, $\hat{g}_2$, can be obtained by backward induction, a technique widely applied in the Markov decision process (MDP) literature. See Section 4.4.1 and Appendix C for more discussion on the relationship between T-EWM and MDP.

We also define

$$\bar{\mathcal{W}}_{T:T+1}(g_1, g_2|w) = \frac{1}{T(w)} \sum_{t:W_{t-1}=w} \mathsf{E}\left\{Y_t(1)g_1(W_{t-1}) + Y_t(0)\left[1 - g_1(W_{t-1})\right] | W_{t-1} = w\right\}$$
$$+ \frac{1}{T(g_1(w))} \sum_{t:W_{t-1}=g_1(w)} \mathsf{E}\left\{Y_t(1)g_2(W_{t-1}) + Y_t(0)\left[1 - g_2(W_{t-1})\right] | W_{t-1} = g_1(w)\right\}.$$

The regret $\widehat{\mathcal{W}}_{T:T+1}(g_1, g_2|w) - \bar{\mathcal{W}}_{T:T+1}(g_1, g_2|w)$ is a (weighted) sum of MDS. Its upper bound can be shown by the method of Section 2.3.

4.3 Accounting for Lucas critique

An interesting question is the extent to which T-EWM can deal with the Lucas critique. In this section, we attempt to answer the question by showing the link between T-EWM and the conventional SVAR approach. In VAR analysis, economists address the Lucas critique by explicitly modeling the conditional expectations of economic variables. In this way, the optimal policy implied by a structural VAR takes into account changes in the data generating process in response to a policy decision. In most cases, the policy function and dynamics of the state variables share a(some) common deep parameter(s). In comparison, our approach has less structure. Nevertheless, we discuss the possibility of linking our framework with the structural approach. The key to extending our method is to make the agent's response function dependent on the policy parameter(s).

We start with a three-equation new Keynesian model. (See, e.g., Chapter 8 of Walsh (2010).) At time t, let π_t denote inflation, x_t the output gap, and i_t the interest rate.

$$\begin{aligned} \text{Phillips curve:} \quad & \pi_t = \beta \mathsf{E}_t \pi_{t+1} + \kappa x_t + \varepsilon_t, \\ \text{IS curve:} \quad & x_t = \mathsf{E}_t x_{t+1} - \sigma^{-1}(i_t - \mathsf{E}_t \pi_{t+1}), \\ \text{Taylor rule:} \quad & i_t = \delta \pi_t + v_t, \end{aligned} \tag{58}$$

Here v_t is the treatment variable. It represents the baseline target rate, and is assumed to follow an AR(1) process $v_t = \rho v_{t-1} + e_t$. We also assume that $\varepsilon_t = \gamma \varepsilon_{t-1} + \delta_t$. Define $d_t = \begin{pmatrix} v_t \\ \varepsilon_t \end{pmatrix}$, $F = \begin{pmatrix} \rho & 0 \\ 0 & \gamma \end{pmatrix}$, and a vector of noises $\eta_t = \begin{pmatrix} e_t \\ \delta_t \end{pmatrix}$. Then the process for $\begin{pmatrix} v_t \\ \varepsilon_t \end{pmatrix}$ can be written as $d_t = F d_{t-1} + \eta_t$.

Define the outcome variables of the system (58) to be $\tilde{Y}_t := \begin{pmatrix} x_t \\ \pi_t \end{pmatrix}$. At the end of time $T-1$, the goal of the SP is minimizing (or maximizing) the expectation of some function of $\tilde{Y}_T$. For example, an objective function that balances the time T output gap and inflation,

$Y_T := |x_T|^2 + |\pi_T - \pi_0|^2$, where the π_0 is the inflation target.

Appendix B shows that the VAR reduced form of the system (58) can be solved to obtain

$$\tilde{Y}_t = AC(I - F)^{-1} d_t, \tag{59}$$

where A and C are non-random matrices defined in Appendix B. If the model (58) is correctly specified, the solution to (59) takes the Lucas critique into account since it solves for a *deep* parameter ρ, which reflects both the direct effect of the treatment (ρ in d_t) and private agent's anticipation of the policy change (ρ in $AC(I - F)^{-1}$).

Now we show how this is related to the T-EWM framework. The treatment v_t in (58) corresponds to W_t in the previous sections. Since F contains ρ, we can write $AC(I-F)^{-1} = M(\rho) = \begin{pmatrix} m_{11}(\rho) & m_{12}(\rho) \\ m_{21}(\rho) & m_{22}(\rho) \end{pmatrix}$. When v_t is a binary variable (e.g., high target rate and low target rate), we have the counterfactual outcomes: $\tilde{Y}_t(1) = \begin{pmatrix} |m_{11}(\rho)| + |m_{12}(\rho)\varepsilon_t| \\ |m_{21}(\rho)| + |m_{22}(\rho)\varepsilon_t| \end{pmatrix}$ and $\tilde{Y}_t(0) = \begin{pmatrix} |m_{12}(\rho)\varepsilon_t| \\ |m_{22}(\rho)\varepsilon_t| \end{pmatrix}$. Both $\tilde{Y}_t(1)$ and $\tilde{Y}_t(0)$ depend on ρ, so we write them as $\tilde{Y}_t(1;\rho)$ and $\tilde{Y}_t(0;\rho)$. The dependence of $\tilde{Y}_t(1)$ and $\tilde{Y}_t(0)$ on ρ remains when we move back to the continuous v_t case, as Y_t is a transformation of $\tilde{Y}_t$, the counterfactual outcomes with respect to Y_t should also depend on ρ. Denote $f_{v_T|\mathcal{F}_{T-1}}(v;\rho)$ (resp. $f_{W_T|\mathcal{F}_{T-1}}(w)$) as the conditional density of v_t (resp. W_T) on the filtration $\mathcal{F}_{T-1}$ evaluated at v, ρ (resp. w). Therefore, the SP's problem can is to choose ρ to minimize the following welfare function,

$$\mathsf{E}\left[Y_T(v_T;\rho)|\mathcal{F}_{T-1}\right] = \int \mathsf{E}\left(Y_T(v;\rho)|\mathcal{F}_{T-1}\right) f_{v_T|\mathcal{F}_{T-1}}(v;\rho)dv. \tag{60}$$

This formula is similar to (10). To make the link clear, recall the welfare function (10):

$$\mathcal{W}_T(g|\mathcal{F}_{T-1}) = \mathsf{E}\left[Y_T(W_T)|\mathcal{F}_{T-1}\right] = \mathsf{E}\left[Y_T(1)g(W_{T-1}) + Y_T(0)\left(1 - g(W_{T-1})\right)|\mathcal{F}_{T-1}\right],$$

where $\mathcal{F}_{T-1} = \sigma(W_{T-1})$ by Assumption 2.1. Since $g(\cdot)$ is a deterministic policy function, by definition $g(W_{T-1}) = \Pr(W_T = 1|W_{T-1}) = \Pr(W_T = 1|\mathcal{F}_{T-1}) \in \{0, 1\}$. (10) can then be written as

$$\begin{aligned} \mathsf{E}\left[Y_T(W_T)|\mathcal{F}_{t-1}\right] &= \sum_{w\in\{0,1\}} \mathsf{E}\left[Y_T(w)|\mathcal{F}_{t-1}\right] \Pr(W_T = w|\mathcal{F}_{T-1}), \\ &= \int_{w\in\{0,1\}} \mathsf{E}\left(Y_T(w)|\mathcal{F}_{T-1}\right) f_{W_T|\mathcal{F}_{T-1}}(w)dw. \end{aligned} \tag{61}$$

The treatment variable is w in (61), and v in (60). There are two differences between the

two formulas. First, in (61) we have a deterministic treatment rule, while in (60), the SP chooses ρ to change the conditional probability density function of the treatment. They are essentially similar since a deterministic treatment rule can be regarded as a degenerate probability distribution. Second, we see that in (61), the policy only affects the outcome Y_T through the value of the treatment value w, while in (60), policy affects the outcome through both the treatment value v and the policy parameter ρ. Note that in (60), the same parameter ρ appears in both $Y_T(\cdot;\rho)$ and $f_{v_T|\mathcal{F}_{T-1}}(v;\rho)$, which corresponds to the case of the SVAR model. Thus, the deep policy parameter ρ must be solved for, taking into account of the change of data generating process $Y_T(\cdot;\rho)$ in response to a change in policy.

The link between (10) and (60) reveals how the Lucas critique can be accounted for within the T-EWM framework. We extend the analytical form of the welfare function to allow counterfactual outcomes to depend on policy through channels other than the implemented treatment. In (60), this dependence is represented by the policy parameter ρ, which appears not only in the conditional policy distribution function $f_{v_T|\mathcal{F}_{T-1}}(\cdot;\rho)$, but also in the counterfactual outcomes $Y_T(\cdot;\rho)$. In the same spirit, (10) can be modified to

$$\mathcal{W}_T(g|\mathcal{F}_{T-1}) = \mathsf{E}[Y_T(1;\rho)g(v_{T-1};\rho) + Y_T(0;\rho)(1-g(v_{T-1};\rho))|\mathcal{F}_{T-1}]. \tag{62}$$

If we have a valid sample counterpart to (62), T-EWM allows us to obtain an optimal policy from it. Let ρ_t be a random process over time. Assume ρ_t is either observable or estimable. For simplicity, we assume that ρ_t is discrete random process taking finite many values. Recall from (10) that we have already $\mathcal{F}_{T-1} = \sigma(W_{T-1})$. Conditional on $W_{T-1} = w$, the sample counterpart to (62) is:

$$\begin{aligned}&\widehat{\mathcal{W}}\left(g(\cdot;\rho)|w\right)\\ =&T(w,\rho)^{-1}\sum_{t:W_{t-1}=w \text{ and } \rho_t=\rho}\left[\frac{Y_t(v_{t-1},1;\rho_t)v_t}{e_t(W_{t-1})}g(v_{t-1};\rho_t) + \frac{Y_t(v_{t-1},1;\rho_t)\left(1-v_t\right)}{1-e_t(W_{t-1})}\big(1-g(v_{t-1};\rho_t)\big)\right],\end{aligned}$$

where $T(w,\rho) := \#\{W_{t-1} = w \text{ and } \rho_t = \rho\}$.

In addition, we note the following,

(i) As long as ρ_t is an observable or estimable random process, the functional forms of $Y_t(v_{t-1},1;\rho_t)$ and $g(\cdot;\rho_t)$ w.r.t ρ_t can remain unspecified. The framework remains mostly model-free.

(ii) ρ_t can be estimated by the method proposed in Schorfheide (2005), assuming that monetary policy follows a nominal interest rate rule that is subject to regime shifts.

(iii) Under the assumption that private agents have knowledge of ρ_t only, other components of the functional form $g(.;.)$ are not known by general society. This may be because

private agents are myopic. In this framework we can estimate $g(.;.)$ using the usual EWM framework.

Remark 3 (Continuous treatment)**.** Equation (60) is a welfare function that combines a continuous treatment dose ρ with randomized treatment outcomes. In this paper, we do not focus on continuous treatment choice, but there is a sizable literature concerning this topic, see, e.g., Hirano and Imbens (2004), Kennedy et al. (2017), Kallus and Zhou (2018), and Colangelo and Lee (2021) for discussion of the i.i.d. case. We briefly comment on how T-EWM can be applied to this problem.

Let W_t be a continuous treatment variable at time t taking values in [0,1], and $g : \mathcal{X} \to [0, 1]$ be a continuous policy function. For simplicity, we assume $g(\cdot)$ is a deterministic policy (treatment dose) function. For simplicity, let X_t be a univariate discrete random variable taking finite many values, and $\mathcal{F}_t = \sigma(X_t)$. The SP's problem is to maximize

$$E\left[Y_T\left(g(X_{T-1})\right)|X_{T-1}\right],$$

where $Y_t(w)$ is the counterfactual outcome under the treatment dose w. Based on the set up for the i.i.d. case in Kallus and Zhou (2018), a sample analogue of $E\left[Y_T\left(g(X_{T-1})\right)|X_{T-1} = x\right]$ is

$$\frac{1}{T(x)} \sum_{t:X_{t-1}=x} \frac{Y_t}{f_{W|X_t}(g(X_t))} K_h\left(g(X_{t-1}), W_t\right), \tag{63}$$

where $K_h(a, b) = \frac{1}{h}K(\frac{a-b}{h})$ is a kernel function. The conditional density $f_{W|X_t}(g(X_t))$ also needs to be estimated. See Colangelo and Lee (2021) for different estimation methods. The best policy function can be obtained by maximizing (63) at each point of $\mathcal{X}$. We expect that the rate of convergence for regret will be determined by the rate of convergence of the nonparametric estimator.

4.4 Connections to other policy choice model in the literature

In this section we discuss T-EWM's relation to several literatures on treatment and optimal policy analysis.

4.4.1 Connection to MDP

Markov Decision processes (MDP) are popular models used for decision making. See, e.g., Kallenberg (2016) for a comprehensive introduction. The current T-EWM model can be viewed as a special case of a Markov decision processes (MDP) with a finite horizon. Briefly, the conditioning variables X_{t-1} at each time t correspond to the Markov state at time t, and the welfare outcome Y_t corresponds to the reward. Before the SP intervenes (at time $T-1$),

the transition probability of the Markov state is described by a rule P_T, and the transition of policies is governed by the propensity score. After the SP intervenes, the transition of policies is governed by a deterministic rule described by the (estimated) optimal policy function (2). In this MDP, the expected reward Y_T under policy W_T and state X_{T-1} is unknown, and optimizing the conditional empirical welfare function (e.g., (17)) implicitly estimates the expected reward at state X_{T-1} (in the language of T-EWM, the expected potential outcomes conditioning on X_{T-1}). If the Markov transition probability (the propensity score) before time $T-1$ is unknown, then it needs to be estimated by the methods described in Section 5.3.2. See Appendix C for a detailed description of the link between T-EWM and MDP.

Note that, for multi-period welfare functions, if the welfare function has a finite horizon (as in the case discussed in here), the optimal policy is often nonstationary. See Chapter 2 of Kallenberg (2016) for more details about finite horizon MDPs.

4.4.2 Connection to IRF

Sims (1980) propose analysing "causal effects" using a structural equation framework. In particular, the vector autoregressive model (VAR) circumvents the endogeneity issues of ordinary least squares regression. It is common to measure "causal effects" using the impulse response function (IRF) induced by the structural equations; see Ramey (2016), Plagborg-Moller (2016), Stock and Watson (2017), among others. Hinging causal effect on a specific structural equation has an advantage in terms of interpretability. However, it depends crucially on the belief that the structure model is not misspecified. Our approach is to optimize treatment choices using a potential outcome framework, which is not linked to a specific structure model, Bojinov and Shephard (2019) show the connection of the defined treatment with the impulse response function (IRF) within a structural Vector autoregression (VAR) framework. In particular, the lag s IRF at time $t+s$ with shocks $(w_{0:t+s})$ and observations $(Y_{t+s}(.))$ is can be written as,

$$IRF_{t,s} = \mathsf{E}\left\{Y_{t+s}\left(w_{0:t+s}\right) \mid W_{0:t+s} = w_{0:t+s}, Y_{0:t-1}\right\} - \mathsf{E}\left\{Y_{t+s}\left(w'_{0:t+s}\right) \mid W_{0:t+s} = w'_{0:t+s}, Y_{0:t-1}\right\},$$

where $w_{0:t-1} = w'_{0:t-1} = 0, w_{t+1:t+s} = w'_{t+1:t+s} = 0, w_t = 1, w'_t = 0$ and the expectation is with respect to the treatment path. Therefore, the treatment effects do not require specifying a structural equation, and lead to a flexible optimal treatment analysis. In addition, it provides an alternative robustness check for structural models. Since assuming linear Vector Auto regressive structure might be too strong for decision making, people rarely make policy decisions based on IRF. Moreover, the usual IRF implied from a linear VAR would not allow us to model heterogenous causal effects of multiple shocks.

4.4.3 Connection to reinforcement learning

Reinforcement learning tries to understand and automate goal-directed learning and decision making. The decision making agents (who correspond to the social planner) are allowed to interact with the environment by formalizing decision based on reward (which corresponds to the social welfare function). The decision makers base their actions on the state and their actions influence the environment, which then changes the state. In our learning framework, we do not update the policies over time. The environment of RL is often formulated in the form of an MDP. In problems of complete knowledge, the agent possesses a complete and accurate model of the environment's dynamics. In problems of incomplete knowledge, such a model is not available. In our corresponding MDP framework we do not have a complete knowledge of the environment and our estimate framework implicitly fulfills the task of policy making by estimating the targeted reward function without parameterizing an underlying dynamic model. As our model can be directly linked with MDP, Our decision making framework can be regarded as a class of off-line RL problems.

5 Other theoretical results

In this section, we extend other results in Kitagawa and Tetenov (2018) to the time series settings. Here we maintain the setting and method of Section 3.2. In particular, we assume that the first best policy is feasible, and we bound conditional regret with unconditional regret.

5.1 Improved rates with margin assumption

In this section, we show special cases where we can improve the rate by imposing additional margin assumptions. For simplicity, we assume that $X_t = \{Y_t, W_t\}$ are strictly stationary. We also use the compact notation $Y_t(W_{1:t-1}, 1) = Y_t(1)$ and $Y_t(W_{1:t-1}, 0) = Y_t(0)$. Define

$$\tau(W_{t-1}) = \mathsf{E}((Y_t(1) - Y_t(0))|W_{t-1}).$$

If we are willing to make stronger assumption on observations to have a very small probability to fall into a small region near the margin of the above optimal classification region, we can potentially achieve a better rate. For detailed insight in the i.i.d. case, see, for example, Boucheron et al. (2005). Let

$$\begin{aligned} & \bar{\mathcal{W}}(g, W_{t-2}) \\ \stackrel{\text{def}}{=} \; & (T-1)^{-1} \sum_{1 \le t \le T-1} \mathsf{E}\left\{ Y_t(W_{t-1}, 1) g(W_{t-1}) + Y_t(W_{t-1}, 0) \left[1 - g(W_{t-1})\right] | W_{t-2} \right\}. \end{aligned} \tag{64}$$

$$\begin{aligned} & \bar{\mathcal{W}}(g) \\ \stackrel{\text{def}}{=} \; & (T-1)^{-1} \sum_{1 \le t \le T-1} \mathsf{E}\left\{ Y_t(W_{t-1},1) g(W_{t-1}) + Y_t(W_{t-1},0)\left[1 - g(W_{t-1})\right] \right\}. \end{aligned} \tag{65}$$

$$\begin{aligned} & \widehat{\mathcal{W}}(g) \\ \stackrel{\text{def}}{=} \; & (T-1)^{-1} \sum_{1 \le t \le T-1} \left\{ Y_t(W_{t-1},1) g(W_{t-1}) + Y_t(W_{t-1},0)\left[1 - g(W_{t-1})\right] \right\}. \end{aligned} \tag{66}$$

We shall assume that

Assumption 5.1. The first best treatment rule g^*_{FB} defined in (30) belongs to the class of candidate treatment rules g, i.e., $g^*_{\text{FB}} = g^*$.

Assumption 5.2. We assume that the class of treatment rules is finite and countable with $|\mathcal{G}| = 2 \lesssim n^V$, where V is a fixed integer.

Assumption 5.3. $Y_t(0), Y_t(1)$ belong to $[-C, C]$, where C is a constant.

Assumption 5.4. The following margin condition is assumed. Let $\Pr_{t-1}$ be $\Pr(.|\mathcal{F}_{t-2})$. There exists a constant $0 \le \eta \le C$, $0 < \alpha < \infty$ and $0 \le u \le \eta$, such that

$$\max_t \Pr{}_{t-1}(|\tau(W_{t-1})| \le u) \le (u/\eta)^\alpha, \qquad \forall 0 \le u \le \eta, \tag{67}$$

the above assumption can be implied in the discrete case by the condition that there exists a positive constant c such that $\tau(0), \tau(1) > c$. Pick $\eta = c$ then, we have $\alpha = \infty$, and $\mathrm{P}_{t-1}(|\tau(W_{t-1})| \le u) = 0, \quad \forall 0 \le u \le \eta$.

THEOREM 5.1. Under Assumption 5.2-5.4 and Assumption 3.3 , for a small enough positive $\delta > 0$, we have the follow bound, with probability $1 - \delta$,

$$\bar{\mathcal{W}}(g^*) - \bar{\mathcal{W}}(\hat{g}) \lesssim_p C_{\alpha,\eta,V} \left(\frac{\sqrt{\log(2T/\delta)}}{\sqrt{T-1}} \right)^{2(\alpha+1)/(\alpha+2)},$$

where $C_{\alpha,\eta,V}$ is a constant depending only on α, η and V.

We can see that when $\alpha = 0$, the rate becomes $\frac{1}{\sqrt{T-1}}\sqrt{\log 2T/\delta}$, which means it has almost no constraint. When $\alpha \to \infty$, it is approaching $\frac{1}{T-1}\sqrt{\log 2T/\delta}$ (the best rate). Then, with the arguments in Section 3.2 we can bound the conditional regret. A proof is in Appendix D.7.

5.2 Minimax lower bound

We now prove the lower bound of the risk of the method proposed in Section. We follow the proof strategy of Theorem 3 in Devroye and Lugosi (1995), which is based on Hellinger distances.

THEOREM 5.2. Assume Assumptions 2.3, 2.2, 3.2 and 3.8 , we have the following lower bound,

$$\sup_{P_T \in \mathcal{P}(M,\kappa)} \mathsf{E}_{P_T}\{\mathcal{W}(G_*) - \mathcal{W}(\hat{G})\} \geq \sqrt{v/(T-1)}\exp(-1). \tag{68}$$

A proof is presented in Appendix D.8. This theorem and Theorem 3.3 together show the minimax optimality of the proposed T-EWM method.

We shall now bound the conditional welfare below by the unconditional welfare function. Assume we have bounded support of $X^{(2)}_{T-1}$, B. Note we have for constant C_B

$$\begin{aligned} &\int_B R_T(x,G) f_{X^{(2)}_{T-1}}(x)dx \\ &\leq C_B R_T(x_{\max},G) f_{X^{(2)}_{T-1}}(x_{\max}), \end{aligned} \tag{69}$$

where $x_{\max}$ is the value at which attached maximum for the function $R_T(x,G)f_{X^{(2)}_{T-1}}(x)$. Then we have the above results translated to the (worst-case) conditional welfare function.

5.3 Estimation with unknown propensity score

To this point, we have treated the propensity score function as known, but this is infeasible in practice. Here we consider the case where the propensity score at each time t, $e_t(\cdot)$, is an estimated unknown. Estimation can be either parametric or non-parametric. Let $\hat{e}_t(\cdot)$ denote the estimator of the propensity score function, and $\hat{G}_e$ denote the optimal policy obtained using $\hat{e}_t$.

5.3.1 The convergence rate with estimated propensity scores

In this subsection, we adapt Theorem 2.5 of Kitagawa and Tetenov (2018) to our setting and obtain a new regret bound with estimated propensity scores. We show that, with estimated propensity scores, the convergence rate is determined by the slower of the rate of convergence of $\hat{e}_t(\cdot)$ and the rate of convergence of the estimated welfare loss (given in Theorem 3.1).

THEOREM 5.3. Let $\hat{e}_t(\cdot)$ be an estimated propensity score of $e_t(\cdot)$, and $\hat{\tau}_t = \frac{Y_t W_t}{\hat{e}_t(W_{t-1})} - \frac{Y_t(1-W_t)}{1-\hat{e}_t(W_{t-1})}$ be an feasible estimator for $\tau_t = \frac{Y_t W_t}{e_t(W_{t-1})} - \frac{Y_t(1-W_t)}{1-e_t(W_{t-1})}$. Given a class of data generating processes $\mathcal{P}_T(M,\kappa)$ defined under equation (50), we assume that there exists a

sequence $\phi_T \to \infty$ such that the series of estimators $\hat{\tau}_t$ satisfy

$$\limsup_{T\to\infty} \sup_{P_T \in \mathcal{P}_T(M,\kappa)} \phi_T \mathsf{E}_{P_T}\left[(T-1)^{-1}\sum_{t=1}^{T-1}|\hat{\tau}_t - \tau_t|\right] < \infty. \tag{70}$$

Then we have

$$\sup_{P_T \in \mathcal{P}_T(M,\kappa)} \mathsf{E}_{\mathcal{P}_T}[\mathcal{W}(G_*) - \mathcal{W}(\hat{G}_e)] \lesssim (\phi_T^{-1} \vee \frac{1}{\sqrt{T-1}}),$$

A proof is presented in Appendix D.9. This theorem shows that if the propensity score is estimated with sufficient accuracy (a rate of $\phi_T^{-1} \lesssim \sqrt{T}^{-1}$), we obtain a similar regret bound to the previous sections. It is not surprising to see that the rule is affected by the estimation accuracy of the propensity score, which is the maximum of $\frac{1}{\sqrt{T-1}}$ and ϕ_T^{-1}.

Remark 4. In the cross-sectional setting, Athey and Wager (2021) show an improved rate of welfare convergence when propensity scores are unknown and estimated. It is possible to extend their analysis to our time-series setting and assess whether or not the rate shown in Theorem 5.3 can be improved. However this is not a trivial extension, and we leave it for further research.

5.3.2 Estimation of propensity scores

In this subsection, we briefly review various methods of estimating propensity score functions. The propensity score function $e_t(\cdot)$ can be estimated parametrically models or nonparametrically:

1. Parametric estimator

   An example of a parametric estimator is the (ordered) probit model, which is employed in Hamilton and Jorda (2002), Scotti (2018), and Angrist et al. (2018). e_t can be expressed as function of X_{t-1}, with the structure of the propensity score given by the probit model

   $$e_t(1) = P(W_t = 1|\mathcal{F}_{T,t-1}) := \Phi(\beta' X_{t-1}),$$
   $$e_t(0) = P(W_t = 0|\mathcal{F}_{T,t-1}) := 1 - \Phi(\beta' X_{t-1}).$$

   A more complicated structure, such as the dynamic probit model (Eichengreen et al. (1985); Davutyan and Parke (1995)), can also be employed.

2. Nonparametric estimator

We can also use a nonparametric estimator to estimate $e_t(\cdot)$. For example, Frölich (2006) and Park et al. (2017) extend the local polynomial regression of Fan and Gijbels (1995) to a dynamic setting. Let $p = 0$, a local likelihood logit model can be specified. Since $e_t(1)$ is a function of X_{t-1}, we let $e_t(1) := e(X_{t-1})$. Assuming that $e(\cdot)$ is a continuous function of X_{t-1}, we have the logit function

$$\log\left[\frac{e(X_{t-1})}{1-e(X_{t-1})}\right] := \alpha_t. \tag{71}$$

By the continuity of $e(\cdot)$ and stationarity of Z_t, for x close to X_{t-1}, we can find β_t, such that $\log\left[\frac{e(x)}{1-e(x)}\right] \approx \alpha_t + \beta_t(X_{t-1} - x)$. $\hat{\alpha}_t$ and $\hat{\beta}_t$ can be obtained by solving

$$\begin{aligned}(\hat{\alpha}_t, \hat{\beta}_t)' = \underset{\alpha,\beta}{\operatorname{argmax}} \frac{1}{T}\sum_{i=2}^{T} \Big\{ & W_i \left[\alpha_t + \beta_t(X_{t-1} - X_{i-1})\right] \\ & - \log\left[1 + \exp(\alpha_t + \beta_t(X_{t-1} - X_{i-1}))\right] \Big\} K_H(X_{t-1} - X_{i-1}).\end{aligned}$$

where K_H is a kernel function and H is a vector of bandwidths.

6 Simulation

In this subsection, we illustrate the accuracy of our method through a simple simulation exercise. We consider the following model

$$\begin{aligned} Y_t(W_{0:t}) &= W_t \cdot \mu(Y_{t-1}, Z_{t-1}) + \phi Y_{t-1}(W_{0:t-1}) + \varepsilon_t \\ \mu(Y_{t-1}, Z_{t-1}) &= 1(Y_{t-1} < B_1) \cdot 1(Z_{t-1} < B_2) - 1(Y_{t-1} > B_1 \vee Z_{t-1} > B_2), \end{aligned} \tag{72}$$

where μ is a function determining the direction of the treatment effect. The treatment effect at time t is positive if both $Y_{t-1} < B_1$ and $Z_{t-1} < B_2$ and is negative otherwise. The optimal treatment rule is therefore $G_* = \{(Y_{t-1}, Z_{t-1}) : Y_{t-1} < B_1 \text{ and } Z_{t-1} < B_2\}$.

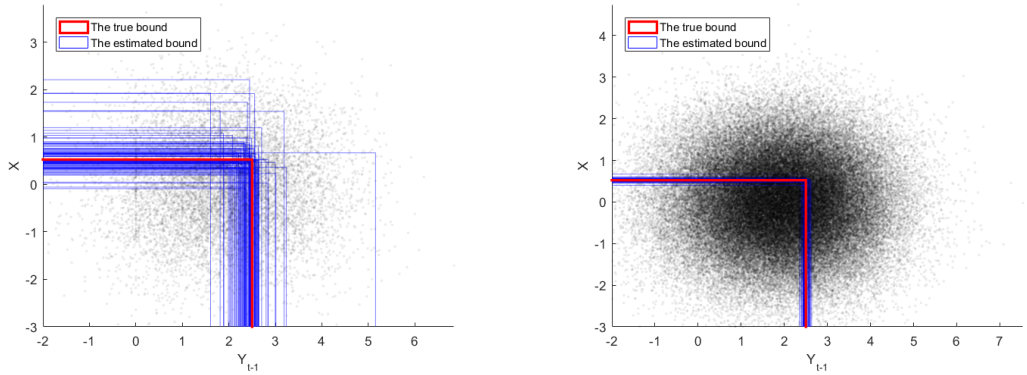
We set $\varepsilon_t \overset{\text{i.i.d.}}{\sim} N(0,1)$, $Z_{t-1} \overset{\text{i.i.d.}}{\sim} N(0,1)$, $\phi = 0.5$, $B_1 = 2.5$,and $B_2 = 0.52$ (approximately the 70% quantile of the standard normal distribution). The propensity score $e_t(Y_{t-1}, Z_{t-1})$ is set to 0.5. Our goal is to estimate G_*. We consider the quadrant treatment rules:

$$\mathcal{G} \equiv \left\{ \begin{array}{c} ((y_{t-1}, x_t) : s_1(y_{t-1} - b_1) > 0 \;\&\; s_2(z_{t-1} - b_2) > 0), \\ s_1, s_2 \in \{-1, 1\}, b_1, b_2 \in R \end{array} \right\}. \tag{73}$$

It is immediate that $G^*_{\text{FB}} \in \mathcal{G}$. Therefore, we can directly estimate the unconditional treat-

ment rule as described in Section 3.2.

Figure 3 illustrates the estimated bound and true bounds for sample sizes $n = 100$ and $n = 1000$. In each case we draw 100 Monte Carlo samples.



(a) $n = 100$ (b) $n = 1000$

Figure 3. The estimated bound for $n = 100$ and $n = 1000$.

The blue lines are estimated bounds, and the red lines are the true bounds. For n=100, the majority of the blue lines are close to the red line. As the sample size increases from 100 to 1000, the blue lines become tightly concentrated around the red line. The results in Table 1 confirm this. This table presents the Monte Carlo averages $(\hat{\mu}_{B_1}, \hat{\mu}_{B_2})$, variances $(\hat{\sigma}^2_{B_1}, \hat{\sigma}^2_{B_2})$ and MSE of estimated B_1 and B_2. We mutliply the variances and MSE by the sample size n. The sample sizes are $n = 100$, 500, 1000 and 2000. The number of Monte Carlo replications is 500.

Table 1. Simulation results for B_1 and B_2

	B_1			B_2		
n	$\hat{\mu}_{B_1}$	$n \cdot \hat{\sigma}^2_{B_1}$	n·MSE$_{B_1}$	$\hat{\mu}_{B_2}$	$n \cdot \hat{\sigma}^2_{B_2}$	n·MSE$_{B_2}$
100	2.4688	14.4331	14.5016	0.6589	18.9382	20.7080
500	2.4919	2.3881	2.4162	0.5433	3.3775	3.5490
1000	2.4924	1.4676	1.5224	0.5310	1.2620	1.3027
2000	2.4958	0.5981	0.6327	0.5267	0.7672	0.7767

As the sample size increases, both the $\hat{\mu}_{B_1}$ and $\hat{\mu}_{B_2}$ converge to their true values, 2.5 and 0.52, respectively. The variances and MSE's shrink, even after multiplying by the sample size n, which suggests that the convergence rate in this case is above $\frac{1}{\sqrt{n}}$.

7 Application

In this section, we illustrate the usefulness of T-EWM with an empirical application. Maintaining low unemployment rate, stable inflation rate, and rapid economic growth are the ultimate goals of a central bank. How monetary policy, such as changes in the FOMC target rate, can affect macroeconomic targets, has been extensively studied. See, e.g., Romer and Romer (1989), Bagliano and Favero (1998), Christiano et al. (1999), and many others. On the other hand, the finance literature documents an increase in the equity premium following Federal Open Market Committee (FOMC) meetings. In particular,Lucca and Moench (2015) find that excess returns tend to be high on the day of FOMC announcements. Cieslak et al. (2019) show that the excess returns in even weeks after the FOMC announcement are significantly higher than those in odd weeks.

In this section, we study the optimal policy region for a fictional monetary policy target: excess stock returns. Excess returns instantly reflect market confidence and are highly correlated with economic indicators. We attempt to address the following questions: if the objective of a fictional decision maker is to increase stock market returns, when should they decrease the target interest rate?

We take treatment variables and covariates that predict the treatment from the monthly dataset of Angrist et al. (2018)(AJK hereafter).[12] Our treatment variable is the change in the Federal target rate. We define the treatment $W_t = 1$ to correspond to a decrease in the federal target rate, and set $W_t = 0$ otherwise. We use the same specification for the propensity score function as AJK, which is constructed to be a Taylor-type monetary policy rule. The covariates for the propensity score function include market expectations, which are given by an adjusted difference between the futures contract price and the current target rate (See AJK and their supplemental data Appendix for more details of this measure), inflation, changes of unemployement rates, lags of inflation, lagged changes in unemployement rates, the current target rate, the last change of FOMC, a dummy variable indicating months with a scheduled FOMC meeting, an interaction term of the last change of FOMC and the dummy for FOMC, a scale factor indicating the part of a month in which a FOMC meeting is scheduled, a set of monthly dummies, and dummies for Y2K and the September 11 attacks. The data runs from August 1989 to July 2005 and contains 128 FOMC meetings.

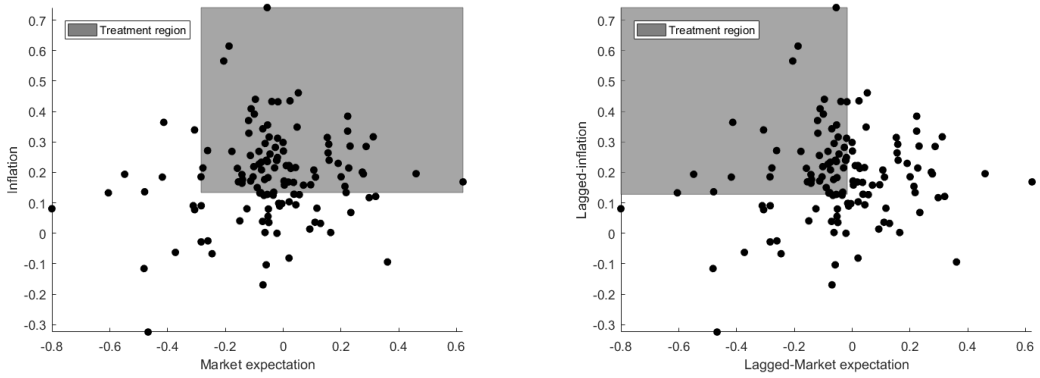
We use the five-day cumulative excess stock returns in the week of FOMC meetings (Week-0) as our outcome variable. This data is taken from the dataset of Cieslak et al. (2019).[13] The five-day cumulative excess return is defined as the difference between the cumulative stock return and the five-day cumulative return on 30-day Treasury bills. We select the

[12]We are grateful to Joshua Angrist, who shared the data and the replication codes. It is now also available at his website https://economics.mit.edu/faculty/angrist/data1/data

[13]We grateful to Anna Cieslak, Adair Morse and Annette Vissing-Jorgensen for posting the data and the replication codes at https://onlinelibrary.wiley.com/doi/abs/10.1111/jofi.12818

excess returns on the days of the 128 FOMC meetings.

We consider the quadrant treatment rules described in (73). In this case, we assume that the decision-maker makes decisions based on whether the inflation rate and market expectations exceed certain bounds.



(a) Five-day cumulative excess return

(b) Lagged Five-days cumulative excess return

Figure 4. Treatment rules from the quadrant class conditioning on market expectation and inflation rate.

Figure 4 shows the quadrant treatment rules selected by T-EWM. The shaded region corresponds to the area in which the policy maker should set $W_t = 1$ (i.e. cut the interest rate). The left panel, (a), shows that a decision maker aiming to maximize the week-0 cumulative excess return should decrease the target rate if *market expectation* $\geq$ -0.2845 and *inflation* $\geq$ 0.1337. A cut in the interest rate is more likely to increase stock market values if inflation is high, and the market expects an increase in the target rate.

The right panel of Figure 4 illustrates the treatment rule for the lag-1 case. The shaded region is consistent with the following mechanism. Suppose that the inflation rate reflects the expectation of the general public, while market expectations reflect the expectations of professional traders who collectively determine the prices of futures. Panel (b) shows that experienced traders react to monetary policy faster than the general public: excess stock returns at the beginning of the next FOMC cycle are more likely to increase following a cut in the target rate in the current period if the inflation rate is high but experienced traders anticipated the decrease in the target rate.

8 Conclusion

This article proposed T-EWM, a framework and method for choosing optimal policies based on time-series data. We characterised assumptions under which this method can learn an optimal policy. We evaluated its statistical properties by deriving non-asymptotic upper

and lower bounds of the conditional welfare. We discussed its connections to the existing literature, including Markov decision process and reinforcement learning. We presented simulation results and empirical applications to illustrate the computational feasibility and applicability of T-EWM. As a benchmark formulation, this paper mainly focused on a one-period social welfare as planner's objective. Extensions of the analysis to policy choices for the middle-run and long-run cumulative social welfare are left for future research.

References

A. Ananth. Optimal treatment assignment rules on networked populations. *working paper*, 2020.

D. W. Andrews. Empirical process methods in econometrics. *Handbook of econometrics*, 4: 2247–2294, 1994.

J. D. Angrist, Ò. Jordà, and G. M. Kuersteiner. Semiparametric estimates of monetary policy effects: string theory revisited. *Journal of Business & Economic Statistics*, 36(3): 371–387, 2018.

S. Athey and S. Wager. Efficient policy learning with observational data. *Econometrica*, 89 (1):133–161, 2021.

F. C. Bagliano and C. A. Favero. Measuring monetary policy with var models: An evaluation. *European Economic Review*, 42(6):1069–1112, 1998.

D. Bhattacharya and P. Dupas. Inferring welfare maximizing treatment assignment under budget constraints. *Journal of Econometrics*, 167(1):168–196, 2012. ISSN 0304-4076.

I. Bojinov and N. Shephard. Time series experiments and causal estimands: exact randomization tests and trading. *Journal of the American Statistical Association*, pages 1–36, 2019.

S. Boucheron, O. Bousquet, and G. Lugosi. Theory of classification: A survey of some recent advances. *ESAIM: probability and statistics*, 9:323–375, 2005.

C. Brownlees and G. Gudmundsson. Performance of empirical risk minimization for linear regression with dependent data. *working paper*, 2021.

C. Brownlees and J. Llorens-Terrazas. Empirical risk minimization for time series: nonparametric performance bounds for prediction. *working paper*, 2021.

L. J. Christiano, M. Eichenbaum, and C. L. Evans. Monetary policy shocks: What have we learned and to what end? *Handbook of macroeconomics*, 1:65–148, 1999.

A. Cieslak, A. Morse, and A. Vissing-Jorgensen. Stock returns over the fomc cycle. *The Journal of Finance*, 74(5):2201–2248, 2019.

K. Colangelo and Y.-Y. Lee. Double debiased machine learning nonparametric inference with continuous treatments, 2021.

N. Davutyan and W. R. Parke. The operations of the bank of england, 1890-1908: a dynamic probit approach. *Journal of Money, Credit and Banking*, 27(4):1099–1112, 1995.

Dehejia. Program evaluation as a decision problem. *Journal of Econometrics*, 125:141–173, 2005.

L. Devroye and G. Lugosi. Lower bounds in pattern recognition and learning. *Pattern recognition*, 28(7):1011–1018, 1995.

B. Eichengreen, M. W. Watson, and R. S. Grossman. Bank rate policy under the interwar gold standard: a dynamic probit model. *The Economic Journal*, 95(379):725–745, 1985.

J. Fan and I. Gijbels. Data-driven bandwidth selection in local polynomial fitting: variable bandwidth and spatial adaptation. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(2):371–394, 1995.

D. A. Freedman. On tail probabilities for martingales. *the Annals of Probability*, pages 100–118, 1975.

M. Frölich. Non-parametric regression for binary dependent variables. *The Econometrics Journal*, 9(3):511–540, 2006.

P. Hall and C. C. Heyde. *Martingale limit theory and its application*. Academic press, 2014.

J. D. Hamilton. A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica: Journal of the econometric society*, pages 357–384, 1989.

J. D. Hamilton and O. Jorda. A model of the federal funds rate target. *Journal of Political Economy*, 110(5):1135–1167, 2002.

S. Han. Optimal dynamic treatment regimes and partial welfare ordering. *working paper*, 2021.

K. Hirano and G. W. Imbens. The propensity score with continuous treatments. *Applied Bayesian modeling and causal inference from incomplete-data perspectives*, 226164:73–84, 2004.

W. Jiang and M. A. Tanner. Risk minimization for time series binary choice with variable selection. *Econometric Theory*, 26:1437–1452, 2010.

L. Kallenberg. Markov decision processes. *Lecture Notes. University of Leiden*, 2016.

N. Kallus. More efficient policy learning via optimal retargeting. *Journal of the American Statistical Association*, 116(534):646–658, 2021.

N. Kallus and A. Zhou. Confounding-robust policy improvement. *arXiv preprint arXiv:1805.08593*, 2018.

M. Kasy and A. Sautmann. Adaptive treatment assignment in experiments for policy choice. *CESifo Working Paper*, 2019.

E. H. Kennedy, Z. Ma, M. D. McHugh, and D. S. Small. Non-parametric methods for doubly robust estimation of continuous treatment effects. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(4):1229–1245, 2017.

T. Kitagawa and A. Tetenov. Who should be treated? empirical welfare maximization methods for treatment choice. *Econometrica*, 86(2):591–616, 2018.

T. Kitagawa and A. Tetenov. Equality-minded treatment choice. *Journal of Business Economics and Statistics*, 39(2):561–574, 2021.

T. Kitagawa, S. Sakaguchi, and A. Tetenov. Constrained classification and policy learning. *arXiv preprint*, 2021.

A. B. Kock, D. Preinerstorfer, and B. Veliyev. Functional sequential treatment allocation. *Journal of the American Statistical Association*, pages 1–36, 2020.

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

T. Lattimore and C. Szepesvári. *Bandit Algorithms*. Cambridge University Press, 2020.

D. O. Lucca and E. Moench. The pre-fomc announcement drift. *The Journal of Finance*, 70 (1):329–371, 2015.

G. Lugosi. Pattern classification and learning theory. In L. Györfi, editor, *Principles of Nonparametric Learning*, pages 1–56, Vienna, 2002. Springer.

C. F. Manski. Statistical treatment rules for heterogeneous populations. *Econometrica*, 72 (4):1221–1246, 2004.

E. Mbakop and M. Tabord-Meehan. Model selection for treatment choice: Penalized welfare maximization. *Econometrica*, 89(2):825–848, 2021. arXiv 1609.03167.

S. A. Murphy. Optimal dynamic treatment regimes. *Journal of the Royal Statistical Society. Series B*, 65(2):331–366, 2003.

B. U. Park, L. Simar, and V. Zelenyuk. Nonparametric estimation of dynamic discrete choice models for time series data. *Computational Statistics & Data Analysis*, 108:97–120, 2017.

M. Plagborg-Moller. *Essays in macroeconometrics*. PhD thesis, 2016.

M. Qian and S. A. Murphy. Performance guarantees for individualized treatment rules. *The Annals of Statistics*, 39(2):1180–1210, 04 2011.

A. Rambachan and N. Shepherd. When do common time seriesestimands have nonparametric causal meaning? *working paper*, 2021.

V. A. Ramey. Macroeconomic shocks and their propagation. *Handbook of macroeconomics*, 2:71–162, 2016.

C. D. Romer and D. H. Romer. Does monetary policy matter? a new test in the spirit of friedman and schwartz. *NBER macroeconomics annual*, 4:121–170, 1989.

S. Sakaguchi. Estimation of optimal dynamic treatment assignment rules under policy constraint. *arXiv preprint*, 2021.

F. Schorfheide. Learning and monetary policy shifts. *Review of Economic dynamics*, 8(2): 392–419, 2005.

C. Scotti. A bivariate model of federal reserve and ecb main policy rates. *26th issue (September 2011) of the International Journal of Central Banking*, 2018.

C. A. Sims. Macroeconomics and reality. *Econometrica: journal of the Econometric Society*, pages 1–48, 1980.

J. H. Stock and M. W. Watson. Twenty years of time series econometrics in ten pictures. *Journal of Economic Perspectives*, 31(2):59–86, 2017.

J. Stoye. Minimax regret treatment choice with finite samples. *Journal of Econometrics*, 151(1):70–81, 2009.

J. Stoye. Minimax regret treatment choice with covariates or with limited validity of experiments. *Journal of Econometrics*, 166(1):138–156, 2012.

A. Tetenov. Statistical treatment choice based on asymmetric minimax regret criteria. *Journal of Econometrics*, 166(1):157–165, 2012.

A. W. Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

V. N. Vapnik. *Statistical Learning Theory*. John Wiley & Sons, 1998.

D. Viviano. Policy targeting under network interference. *arXiv preprint*, 2021.

M. J. Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

C. E. Walsh. Monetary theory and policy. 2010.

W.-B. Wu and Y. N. Wu. Performance bounds for parameter estimates of high-dimensional linear models with correlated errors. *Electronic Journal of Statistics*, 10(1):352–379, 2016.

Y. Zhao, D. Zeng, A. J. Rush, and M. R. Kosorok. Estimating individualized treatment rules using outcome weighted learning. *Journal of the American Statistical Association*, 107(499):1106–1118, 2012.

Y.-Q. Zhao, D. Zeng, E. B. Laber, and M. R. Kosorok. New statistical learning methods for estimating optimal dynamic treatment regimes. *Journal of the American Statistical Association*, 110(510):583–598, 2015.

A The relationship between the conditional and unconditional cases

In this section, we present further examples to illustrate that $G^*_{\text{FB}} \in \mathcal{G}$ is a sufficient but not necessary condition for equivalence between the unconditional and conditional problems. Finally, we extend Example 4.1 to show how Assumption 3.4 ensures this equivalence.

A.1 $G^*_{\text{FB}} \in \mathcal{G}$ is sufficient: a discrete and a continuous case

Let $X_{t-1} = W_{t-1} \in \{0, 1\}$. $\mathcal{G}$ be a subclass of the power set of $\{0, 1\}$, $\mathcal{P} = \{\emptyset, \{0\}, \{1\}, \{0, 1\}\}$.

For compactness, suppress W_{T-1} in $Y_T(W_{T-1}, 1)$. Unconditional welfare can be written as:

$$\begin{aligned}\mathcal{W}_T(G) =& \mathsf{E}\left\{Y_T(1)\mathbf{1}(W_{T-1} \in G) + Y_T(0)\mathbf{1}(W_{T-1} \notin G)\right\} \\ =& \mathsf{E}\left\{\left[Y_T(1) - Y_T(0)\right]\mathbf{1}(W_{T-1} \in G) + Y_T(0)\right\} \\ =& \mathsf{E}\left[\tau(W_{T-1})\mathbf{1}(W_{T-1} \in G)\right] + \mathsf{E}\left[Y_T(0)\right], \end{aligned} \tag{74}$$

where $\tau(w_{T-1}) = \mathsf{E}\left[Y_T(1) - Y_T(0)|W_{T-1} = w_{T-1}\right]$, and the last equality follows from the law of iterated expectations.

The first best unconditional policy is

$$G^*_{\text{FB}} \equiv \{w_{T-1} \in \{0, 1\} : \tau(w_{T-1}) \geq 0\}. \tag{75}$$

By the assumption that $G^*_{\text{FB}} \in \mathcal{G}$

$$G^*_{\text{FB}} = \text{argmax}_{G \in \mathcal{G}} \mathcal{W}_T(G). \tag{76}$$

The SP's conditional objective function can be written as:

$$\begin{aligned}\mathcal{W}_T(G|W_{T-1}) &= \mathsf{E}\left\{Y_T(1)\mathbf{1}(W_{T-1} \in G) + Y_T(0)\mathbf{1}(W_{T-1} \notin G)|W_{T-1}\right\} \\ &= \mathsf{E}\left\{\left[Y_T(1) - Y_T(0)\right]\mathbf{1}(W_{T-1} \in G) + Y_T(0)|W_{T-1}\right\} \\ &= \mathsf{E}\left\{\left[Y_T(1) - Y_T(0)\right]\mathbf{1}(W_{T-1} \in G)|W_{T-1}\right\} + \mathsf{E}\left[Y_T(0)|W_{T-1}\right] \\ &= \mathsf{E}\left[\tau(W_{T-1})\mathbf{1}(W_{T-1} \in G)|W_{T-1}\right] + \mathsf{E}\left[Y_T(0)|W_{T-1}\right]. \end{aligned} \tag{77}$$

To check whether G^*_{FB} is optimal in the conditional problem, we need to study this problem

w.r.t. $\mathcal{P}$

$$\begin{aligned}
&\max_{G\in\mathcal{P}} \mathcal{W}_T(G|W_{T-1}=w_{T-1}) \\
=&\max_{G\in\mathcal{P}} \mathsf{E}\left[\tau(W_{T-1})\mathbf{1}(W_{T-1}\in G)|W_{T-1}=w_{T-1}\right] \\
=&\max_{G\in\mathcal{P}} \tau(w_{T-1})\mathbf{1}(w_{T-1}\in G) \\
=&\tau(w_{T-1})\mathbf{1}(w_{T-1}\in G^*_{\mathrm{FB}}), \qquad (78)
\end{aligned}$$

The first equality follows from (77). The last follows from the definition of G^*_{FB} in (75). Equivalence follows by combining (76) and (78).

We now turn the continuous conditioning variable case.

The SP's unconditional welfare function can be rewritten as (suppressing W_{T-1} in $Y_T(W_{T-1},1)$.)

$$\begin{aligned}
\mathcal{W}_T(G) =&\mathsf{E}\left\{Y_T(1)\mathbf{1}(X_{T-1}\in G)+Y_T(0)\mathbf{1}(X_{T-1}\notin G)\right\} \\
=&\mathsf{E}\left\{\left[Y_T(1)-Y_T(0)\right]\mathbf{1}(X_{T-1}\in G)+Y_T(0)\right\} \\
=&\mathsf{E}\left[\tau(X_{T-1})\mathbf{1}(X_{T-1}\in G)\right]+\mathsf{E}\left[Y_T(0)\right],
\end{aligned}$$

where $\tau(x_{T-1})=\mathsf{E}\left[Y_T(1)-Y_t(0)|X_{T-1}=x_{T-1}\right]$, and the last equality follows from the law of iterated expectations.

The first best policy is

$$G^*_{\mathrm{FB}}\equiv\{x_{T-1}\in\mathbb{R}^2:\tau(x_{T-1})\geq 0\}. \qquad (79)$$

(For simplicity, we assume that G^*_{FB} is measurable, i.e., $G^*_{\mathrm{FB}}\in\mathfrak{B}(\mathbb{R}^2)\subset\mathcal{P}(\mathbb{R}^2)$. This means that we don't have to deal with the outer probability and expectation.) By assumption $G^*_{\mathrm{FB}}\in\mathcal{G}$,

$$G^*_{\mathrm{FB}}\in \mathrm{argmax}_{G\in\mathcal{G}}\mathcal{W}_T(G). \qquad (80)$$

We now introduce some notation. Let $X_{t-1}=[X^{(1)}_{t-1},X^{(2)}_{t-1}]'\in\mathbb{R}^2$. $X^{(1)}_{t-1}$ and $X^{(2)}_{t-1}$ can be continuous or discrete, e.g., one of them can be the treatment W_{t-1}. Let $\mathcal{G}$ be a class of subsets of $\mathbb{R}^2$, and $\mathcal{G}^{(1)}$ a class of subsets of $\mathbb{R}$. The SP's conditional objective function can

be written as:

$$
\begin{aligned}
&\mathcal{W}_T(G^{(1)}|X^{(2)}_{T-1}) \\
=&\mathsf{E}\left\{Y_T(1)\mathbf{1}(X^{(1)}_{T-1}\in G^{(1)}) + Y_T(0)\mathbf{1}(X^{(1)}_{T-1}\notin G^{(1)})|X^{(2)}_{T-1}\right\} \\
=&\mathsf{E}\left\{[Y_T(1)-Y_T(0)]\,\mathbf{1}(X^{(1)}_{T-1}\in G^{(1)}) + Y_T(0)|X^{(2)}_{T-1}\right\} \\
=&\mathsf{E}\left\{[Y_T(1)-Y_T(0)]\,\mathbf{1}(X^{(1)}_{T-1}\in G^{(1)})|X^{(2)}_{T-1}\right\} + \mathsf{E}\left[Y_T(0)|X^{(2)}_{T-1}\right] \\
=&\mathsf{E}\left\{\mathsf{E}\left[\tau(X_{T-1})\mathbf{1}(X^{(1)}_{T-1}\in G^{(1)})|X_{T-1}\right]|X^{(2)}_{T-1}\right\} + \mathsf{E}\left[Y_T(0)|X^{(2)}_{T-1}\right], \qquad (81)
\end{aligned}
$$

where the last equality follows from the law of iterated expectations.

We also assume that the conditional first-best policy G^*_{CFB} is measurable, i.e., $G^*_{\text{CFB}}(X^{(2)}_{T-1} = x^{(2)}_{T-1}) \in \mathfrak{B}(\mathbb{R})$, for every $x^{(2)}_{T-1} \in \mathbb{R}$. Note $\tau(x_{t-1}) = \tau(x^{(1)}_{t-1}, x^{(2)}_{t-1})$. Following the last row of (81), the optimal conditional policy is defined to be

$$
\begin{aligned}
&\text{argmax}_{G^{(1)}\in\mathfrak{B}(\mathbb{R})}\mathcal{W}_T(G^{(1)}|X^{(2)}_{T-1}=x^{(2)}_{T-1}) \\
=&\text{argmax}_{G^{(1)}\in\mathfrak{B}(\mathbb{R})}\mathsf{E}\left\{\mathsf{E}\left[\tau(X_{T-1})\mathbf{1}(X^{(1)}_{T-1}\in G^{(1)})|X_{T-1}\right]|X^{(2)}_{T-1}=x^{(2)}_{T-1})\right\} \\
=&\text{argmax}_{G^{(1)}\in\mathfrak{B}(\mathbb{R})}\mathsf{E}\left\{\mathsf{E}\left[\tau(X^{(1)}_{T-1},x^{(2)}_{T-1})\mathbf{1}(X^{(1)}_{T-1}\in G^{(1)})|X^{(1)}_{T-1}\right]\right\} \\
=&\text{argmax}_{G^{(1)}\in\mathfrak{B}(\mathbb{R})}\mathsf{E}\left[\tau(X^{(1)}_{T-1},x^{(2)}_{T-1})\mathbf{1}(X^{(1)}_{T-1}\in G^{(1)})\right].
\end{aligned}
$$

The optimal policy conditional on $X^{(2)}_{T-1} = x^{(2)}_{T-1}$ is

$$
G^*_{\text{CFB}}(X^{(2)}_{T-1}=x^{(2)}_{T-1}) = \{x^{(1)}_{T-1}\in\mathbb{R} : \tau(x^{(1)}_{T-1}, x^{(2)}_{T-1}) \geq 0\}. \qquad (82)
$$

Comparing (79) and (82), we see $G^*_{\text{CFB}}(X^{(2)}_{T-1} = x^{(2)}_{T-1})$ is given by the intersection of G^*_{FB} with the line $X^{(2)}_{T-1} = x^{(2)}_{T-1}$.

A.2 $G^*_{\text{FB}} \in \mathcal{G}$ is not a necessary condition

Here we show that $G^*_{\text{FB}} \in \mathcal{G}$ is sufficient but not necessary for correspondence between the optimal conditional policy and the first-best unconditional policy.

Example A.1. Univariate covariates

We set $X_{t-1} = W_{t-1} \in \{0,1\}$ and $\mathcal{G} = \{\emptyset, \{1\}\} \subset \mathcal{P} = \{\emptyset, \{0\}, \{1\}, \{0,1\}\}$. Suppose

$$
\tau(1) = \tau(0) = 1 > 0.
$$

For the unconditional problem (74), the first best policy is then

$$G^*_{\text{FB}} = \{0, 1\}.$$

Note that $G^*_{\text{FB}} \notin \mathcal{G}$, so the solution to the unconditional problem is

$$G_* \equiv \underset{G \in \mathcal{G}}{\text{argmax}}\ \mathcal{W}_T(G) = \{1\}.$$

Consider the conditional problem for $W_{T-1} = 1$.

$$\begin{aligned}
&\max_{G \in \mathcal{P}} \mathcal{W}_T(G|W_{T-1} = 1) \\
=&\max_{G \in \mathcal{P}} \mathsf{E}\left[\tau(W_{T-1})\mathbf{1}(W_{T-1} \in G)|W_{T-1} = 1\right] \\
=&\max_{G \in \mathcal{P}} \tau(1)\mathbf{1}(1 \in G) \\
=&\tau(1)\mathbf{1}(1 \in G_*). \qquad (83)
\end{aligned}$$

Therefore, the solution to the unconditional problem is also the solution to the conditional problem. (This is only the case for $W_{T-1} = 1$.)

Example A.2. Two dimensional discrete covariates

Set $X_{t-1} = (W_{t-1}, Z_{t-1})' \in \{0, 1\} \times \{i\}_{i=0}^{10}$. Suppose

$$G^*_{\text{FB}} = \{0, 1\} \times \{1, 3, 5, 7, 9\}.$$

and

$$\mathcal{G} = \left\{ \begin{array}{c} \{(w, z) : w \in \{0, 1\}, z \in \{i\}_{i=0}^{10}, \text{ and } z \in [0, a)\}, \\ a \in \mathbb{R}^+ \end{array} \right\}.$$

For example, if $G \in \mathcal{G}$ and $a = 4.5$, $(0, 4) \in G$, so $(0, 0)$, (0,1), (0,2), and (0,3) are also in G.

Note that $G^*_{\text{FB}} \notin \mathcal{G}$. Suppose the best feasible unconditional policy is

$$G_* \equiv \underset{G \in \mathcal{G}}{\text{argmax}}\ \mathcal{W}_T(G) = \{0, 1\} \times \{0, 1, 2, 3, 4, 5\},$$

We can always construct a data generating process with a certain type of conditional treatment effect τ to achieve that. For example, let $\tau(x_{t-1}) = \tau(w_{t-1}, z_{t-1})$, set $P(Z_{t-1} = i) = 1/10$, $Z_{t-1} \perp W_{t-1}$, and for any $w \in \{0, 1\}$ assume

$$\tau(w,z) = \begin{cases} 2 & z \in \{1,3,5\} \\ 0.1 & z \in \{7,9\} \\ -1.5 & z \text{ is even} \end{cases}$$

For example, a policy G_a with $a = 5.5$ will include $\{w, 2\}$ and $\{w, 4\}$, which has a welfare cost of 2×-1.5.

The conditional problem is

$$\max_{G^z \in \mathcal{G}^z} \mathcal{W}_T(G^z | W_{T-1} = w)$$

where $\mathcal{G}^z = \{z : z \in \{i\}_{i=1}^{10}, \text{ and } z \in [0, a), a \in \mathbb{R}^+\}$. Here

$$\begin{aligned} &\operatorname*{argmax}_{G^z \in \mathcal{G}^z} \mathcal{W}_T(G^z | W_{T-1} = w) \\ =&\{1,2,3,4,5\}, \end{aligned}$$

which is the intersection of G_* with $\text{Supp}(Z_{T-1})$. We have the conclusion.

A.3 An illustration of Assumption 3.4

Recall Assumption 3.4

$$\arg\sup_{G \in \mathcal{G}} \mathcal{W}_T(G) \subset \arg\sup_{G \in \mathcal{G}} \mathcal{W}_T(G|x).$$

We extend Example 4.1 to show why Assumption 3.4 ensures equivalence between the unconditional and conditional problems.

Example 4.1, continued.

Figure 5. $G^*_{\text{FB}} \notin \mathcal{G}_0$ but Assumption 3.4 is satisfied

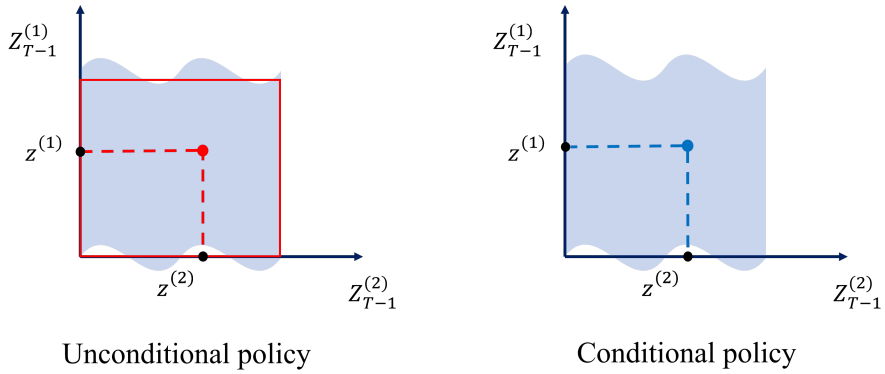


Figure 5 shows a case where the solutions coincide even though the first best unconditional

policy is not available. The red square is differs from the shaded area (the first best), but the red point is inside the red square, and the blue point is inside the shaded area. The social planner will set $W_T = 1$ in both cases. Hence, the feasibility of the first best solution is sufficient but not necessary for conditional and unconditional welfare to coincide.

Thus, there exist situations where the first best solution is not feasible, but we can still achieve correspondence. This example confirms the validity of Assumption 3.4.

B Accounting for the Lucas critique

Here we solve the VAR reduced form of the three-equation New Keynesian model discussed in Section 4.3.

We seek the solution to

$$\begin{pmatrix} \mathsf{E}_t\, x_{t+1} \\ \mathsf{E}_t\, \pi_{t+1} \end{pmatrix} = \begin{pmatrix} \kappa/\beta + 1 & \delta/\sigma - 1/(\sigma\beta) \\ -\kappa/\beta & 1/\beta \end{pmatrix} \begin{pmatrix} x_t \\ \pi_t \end{pmatrix} + \begin{pmatrix} v_t/\sigma \\ \varepsilon_t/\beta \end{pmatrix}.$$

Denote

$$N = \begin{pmatrix} \kappa/\beta + 1 & \delta/\sigma - 1/(\sigma\beta) \\ -\kappa/\beta & 1/\beta \end{pmatrix},$$

$$\tilde{Y}_t = \begin{pmatrix} x_t \\ \pi_t \end{pmatrix}.$$

Let

$$A = N^{-1}.$$

We let $\Gamma_t = N^{-1}\begin{pmatrix} v_t/\sigma \\ \varepsilon_t/\beta \end{pmatrix} = ACd_t$, with $C = \text{diag}[\sigma^{-1}, \beta^{-1}]$. In addition, define

$$d_t = \begin{pmatrix} v_t \\ \varepsilon_t \end{pmatrix}.$$

and

$$d_t = Fd_{t-1} + \eta_t. \tag{84}$$

$$F = \begin{bmatrix} \rho & 0 \\ 0 & \gamma \end{bmatrix} \text{ for } (\rho, \gamma \neq 0).$$

Suppose N is invertible, then

$$\tilde{Y}_t = A\,\mathsf{E}_t\tilde{Y}_{t+1} + \Gamma_t.$$

Solving forward,

$$\tilde{Y}_t = \lim_{L\to\infty}(A^L \,\mathsf{E}_t(\tilde{Y}_{t+L}) + \sum_{l\geq 0} A^l \,\mathsf{E}_t\Gamma_{t+l}).$$

Let $\rho(A) < 1$, $\rho(F) < 1$,

$$\lim_{L\to\infty} A^L \,\mathsf{E}_t(\tilde{Y}_{t+L}) \to_{a.s.} 0.$$

$$\mathsf{E}_t\Gamma_{t+l} = ACF^l d_t.$$

Thus,

$$\tilde{Y}_t = AC(I-F)^{-1}d_t. \tag{85}$$

Combining (84) and (85), we can solve the VAR reduced form for $\tilde{Y}_t := (x_t, \pi_t)'$

$$\tilde{Y}_{t+1} = AC(I-F)^{-1}F(AC(I-F)^{-1})^{-1}\tilde{Y}_t + AC(I-F)^{-1}\eta_{t+1}.$$

C Link to Markov decision problems

In this section, we show the connection between our T-EWM setup and models of Markov Decision Processes (MDP). For MDP, we adapt the notation of Kallenberg (2016), an online lecture notes by Lodewijk Kallenberg (LK hereafter). As described in LK, MDP are models for making decisions for dependent data. A MDP typically has components $\{[p_{ij}(a)]_{i,j}, r_i^t(a), W_{t-1}\}$. In period t, W_{t-1} is the state and W_t is the action. The agent chooses their decision according to a policy (a map from the state W_{t-1} to an action a). They then receive a reward $r_i^t(a)$. The reward function depends on the transition probabilities of a Markov process, which are determined by the action a. Thus their action affects the reward via its effect on the transition probability matrix. The optimal policy is estimated by optimizing an aggregated reward function. We show a formal link between our EWM framework and an MDP. In particular, we show that the MDP's reward function corresponds to our welfare function, and the optimal mapping between states and actions corresponds to the EWM policy in our framework.

In the following equations, the left-hand sides are the notation for MDP in LK, and the right-hand sides are notation for T-EWM from this paper. We consider the model of Section 2.2. For $i, j, a \in \{0, 1\}$, at time t: First, we link the transition probability with the propensity score.

$$p_{ij}^t(a) = P(W_t = j | W_{t-1} = i; \text{choosing } W_t = a),$$

The left-hand side is the Markov transition probability between states i and j under policy a.

The right-hand side is a propensity score under policy a: the probability $W_t = j$, conditional on $W_{t-1} = i$, given $W_t = a$. Note that, in this simple model, the state at time t is the previous treatment W_{t-1}, and the current policy and the next-period state are both W_t. In our setting, after time $T-1$, the SP implements a deterministic policy, so the probability only takes values in $\{0, 1\}$, i.e.,

$$\begin{aligned} p_{ij}^t(a) &= 1 \text{ if } j = a. \\ p_{ij}^t(a) &= 0 \text{ if } j \neq a. \end{aligned} \tag{86}$$

Secondly, we connect the reward function with the expected conditional counterfactual outcome. We denote the reward associated with action a for state i at time t as

$$r_i^t(a) = \mathsf{E}\left[Y_t(a) | W_{t-1} = i\right], \tag{87}$$

The left-hand side is the reward in state i under action a. The right-hand side is the conditional expected counterfactual outcome of $Y_t(a)$, (recall $a \in \{0, 1\}$) conditional on $W_{t-1} = i$.

Thirdly, we link the expected reward function and the expected unconditional counterfactual outcome.

$$\left\{ \sum_i \beta_i r_i^t(a) = r^t(a) \right\} = \mathsf{E}\left[Y_t(a)\right],$$

The left-hand side is the expected reward for action a, with β_i as the initial probability of state i. The right-hand side is the unconditional expected counterfactual outcome.

Finally, we show the link between the total expected reward over a finite horizon and the finite-period welfare function

$$\begin{aligned} &\left\{ v_i^{T:T+1}(R) := \mathsf{E}_{i,R}\left[\sum_{k=T}^{T+1} r_{X_k}^k(Y_k) \right] \right\} \\ &= \mathsf{E}\Big\{ Y_T(1) p_{i1}^T\left(g_1(W_{T-1})\right) + Y_T(0) p_{i0}^T\left(g_1(W_{T-1})\right) \\ &+ Y_{T+1}(1) p_{i1}^{T+1}\left(g_2(W_T)\right) + Y_{T+1}(0) p_{i0}^{T+1}\left(g_2(W_T)\right) | W_{T-1} = i \Big\} \\ &= \mathsf{E}\left\{ Y_T(1) g_1(W_{T-1}) + Y_T(0)\left[1 - g_1(W_{T-1})\right] | W_{T-1} = i \right\} \\ &+ \mathsf{E}\left\{ Y_{T+1}(1) g_2\left(W_T\right) + Y_{T+1}(0)\left[1 - g_2\left(W_T\right)\right] | W_T = g_1(i) \right\}, \end{aligned} \tag{88}$$

The left-hand side is the total expected reward over the planning horizon from T to $T+1$ under the policy $R = (g_1, g_2)$, with the initial state i. The last equality follows from (86)

and the exclusion condition.

With these connections established, we can regard EMW as an MDP with a finite period reward and a non-stationary solution. According to LK, in this case the policy R is usually obtained using a dynamic programming algorithm.

D Proof of lemmas and theorems

D.1 Proof of (23)

Proof. Define $\tilde{p}_w \stackrel{\text{def}}{=} \frac{1}{T-1}\sum_t \Pr(W_{t-1}=w|\mathcal{F}_{t-2})$.

$$\begin{aligned}
&\sup_{g:\{0,1\}\to\{0,1\}} \left\{\widehat{\mathcal{W}}(g|w) - \bar{\mathcal{W}}(g|w)\right\} \\
= &\sup_{g:\{0,1\}\to\{0,1\}} \frac{1}{T-1}\frac{T-1}{T(w)}\sum_t [\widehat{\mathcal{W}}_t(g|w) - \bar{\mathcal{W}}_t(g|w)] \\
= &\sup_{g:\{0,1\}\to\{0,1\}} (\frac{T(w)}{T-1} - \tilde{p}_w + \tilde{p}_w)^{-1}(T-1)^{-1}\sum_t (\widehat{\mathcal{W}}_t(g|w) - \bar{\mathcal{W}}_t(g|w)).
\end{aligned}$$

For large positive constants c_T and C_p, we have with probability $1-2\exp(-c_T)$, by Lemma D.1

$$\frac{T(w)}{T-1} - \tilde{p}_w \lesssim_p c_T\frac{2C_p}{T-1} + \sqrt{2(T-1)C_p}\frac{\sqrt{c_T}}{T-1} = r_{1T}.$$

Since for sufficient large T, we have $r_{1T} < \kappa/2$. Thus with probability approaching 1, by Assumption 2.2,

$$(\frac{T(w)}{T-1} - \tilde{p}_w + \tilde{p}_w)^{-1} \le (-|\frac{T(w)}{T-1} - \tilde{p}_w| + \tilde{p}_w)^{-1} \le (-\kappa/2+\kappa)^{-1} = (\kappa/2)^{-1}. \tag{89}$$

Similarly, for large positive constant C_b, by Lemma D.1, with probability $1-2\exp(-c_T)$

$$\sup_g|\frac{1}{T-1}\sum_t(\widehat{\mathcal{W}}_t(g|w) - \bar{\mathcal{W}}_t(g|w))| < c_T 2C_b\frac{\log 2}{T-1} + \sqrt{2TC_b}\frac{\sqrt{c_T\log 2}}{(T-1)} = r_{2T}. \tag{90}$$

Finally we have,

$$\sup_{g:\{0,1\}\to\{0,1\}} \left\{\widehat{\mathcal{W}}_T(g|w) - \bar{\mathcal{W}}(g|w)\right\} \lesssim_p (\kappa/2)^{-1} r_{2T}. \tag{91}$$

□

D.2 Proof of Lemma 3.1

Proof.

$$\begin{aligned} R_T(G) &= \int R_T(G|x) dF_{X_{T-1}}(x), \\ &= \int_{x \in A(x^{obs}, G)} R_T(G|x) dF_{X_{T-1}}(x) + \int_{x \notin A(x^{obs}, G)} R_T(G|x) dF_{X_{T-1}}(x), \\ &\geq R_T(G|x^{obs}) \cdot p_{T-1}(x^{obs}, G) + 0, \\ &= R_T(G|x^{obs}) \cdot p_{T-1}(x^{obs}, G). \end{aligned} \tag{92}$$

The first inequality follows from the definition of $A(x', G)$ the first best policy being feasible, and $R_T(G|x)$ being non-negative. Then, assumption 3.5 yields $R_T(G|x^{obs}) \leq \frac{1}{\underline{p}} R_T(G)$. □

D.3 Proof of Theorem 3.1

We first show several lemmas.

LEMMA D.1 (Freedman's inequality)**.** Let $\xi_{a,i}$ be a martingale difference sequence indexed by $a \in \mathcal{A}$ and $i = 1, \ldots, n$, $\mathcal{F}_i$ be the filtration, $V_a = \sum_{i=1}^n \mathsf{E}(\xi_{a,i}^2 | \mathcal{F}_{i-1})$, and $M_a = \sum_{i=1}^n \xi_{a,i}$. For positive numbers A and B, we have,

$$\Pr(\max_{a \in \mathcal{A}} |M_a| \geq z) \leq \sum_{i=1}^n \Pr(\max_{a \in \mathcal{A}} \xi_{a,i} \geq A) + 2 \Pr(\max_{a \in \mathcal{A}} V_a \geq B) + 2|\mathcal{A}| e^{-z^2/(2zA+2B)}. \tag{93}$$

The proof can be found in Freedman (1975).

LEMMA D.2 (Maximal inequality based on Freedman's inequality)**.** Let $\xi_{a,i}$ be a martingale difference sequence indexed by $a \in \mathcal{A}$ and $i = 1, \ldots, n$. If, for some positive constants A and B, $\max_{a \in \mathcal{A}} \xi_{a,i} \leq A$, $V_a = \sum_{i=1}^n \mathsf{E}(\xi_{a,i}^2 | \mathcal{F}_{i-1}) \leq B$, and $M_a = \sum_{i=1}^n \xi_{a,i}$ we have,

$$\mathsf{E}(\max_{a \in \mathcal{A}} |M_a|) \lesssim A \log(1 + |\mathcal{A}|) + \sqrt{B} \sqrt{\log(1 + |\mathcal{A}|)} \tag{94}$$

Proof. This follows from Lemma 19.33 of Van der Vaart (2000) , and Lemma D.1. From Freedman's inequality we have

$$\begin{aligned} \Pr(\max_a |M_a| \geq z) &\leq 2|\mathcal{A}| \exp(-z/4A) \quad \text{for} \quad z > B/A & (95) \\ &\leq 2|\mathcal{A}| \exp(-z^2/4B) \quad \text{Otherwise.} & (96) \end{aligned}$$

We truncate M_a into $C_a = M_a \mathbf{1}\{M_a > B/A\}$ and $D_a = M_a \mathbf{1}\{M_a \leq B/A\}$.

Transforming C_a and D_a with $\phi_p(x) = \exp(x^p) - 1$ $(p = 1, 2)$ and applying Fubini's

Theorem, we have

$$\mathsf{E} \exp |C_a/4A| \leq 2 \tag{97}$$

and

$$\mathsf{E}(\exp |D_a/\sqrt{4B}|^2) \leq 2. \tag{98}$$

For completeness, we show the derivations of these inequalities.

$$\begin{aligned} \mathsf{E} \exp |C_a/4A| &\leq \int_0^\infty P(|C_a/4A| \geq x) dx \\ &\leq 2 \int_0^\infty \exp(-4Ax/4A) dx \leq 2, \end{aligned}$$

the first equality follows from Fubini's inequality and the second is due to Lemma D.1.

$$\begin{aligned} \mathsf{E} \exp |D_a^2/4B| &\leq \int_0^\infty P(|D_a^2/4B| \geq x) dx \\ &\leq 2 \int_0^\infty \exp(-4Bx/4B) dx \leq 2. \end{aligned}$$

Again, the first inequality follows from Fubini's inequality and the second is due to Lemma D.1. Now we show that due to Jensen's inequality,

$$\phi_1(\mathsf{E}(\max_a |C_a|/4A)) \leq \sum_a \mathsf{E}\, \phi_1(|C_a|/4A)) \leq |\mathcal{A}|.$$

and similarly,

$$\phi_2(\mathsf{E}(\max_a |D_a|/\sqrt{4B})) \leq \sum_a \mathsf{E}\, \phi_2(|D_a|/\sqrt{4B})) \leq |\mathcal{A}|.$$

Then we have $\mathsf{E}(\max_a |C_a|/4A) \leq \log(|\mathcal{A}| + 1)$ and $\mathsf{E}(\max_a |D_a|/\sqrt{4B}) \leq \sqrt{\log(|\mathcal{A}| + 1)}$. The result thus follows. □

LEMMA D.3. Under Assumption 3.2, 3.3, and 3.7 to 3.9, we have

$$\mathsf{E}[|\, \mathsf{E}_n h|_{\mathbf{H_T}}] \lesssim M\sqrt{v/n}.$$

It shall be noted that the above lemma is of the maximal inequality type and has a standard $\sqrt{n}^{-1}$ rate. The complexity of the function class v also plays role. This is in line with other results in the literature, such as Kitagawa and Tetenov (2018).

Proof. Let h_t denote function belonging to the functional class $\mathcal{H}_t$. Let $h_t^{(0)} = (0, \cdots, 0)$, and the set $J_K : k = 1, \cdots, \overline{K}$. We let $2^{-\overline{K}} \asymp \sqrt{n}^{-1}$, therefore $\overline{K} \asymp \log(n)$. As we assume $\max_t |h_t| \leq M$ for a constant M. J_k is a cover of the functional class $\mathbf{H_T}$ with

radius $2^{-k}M$ with respect to the $\rho_{2,n}(.)$ norm. We denote $\overline{h}_T^* = \arg\max_{\overline{h}_T \in \mathbf{H_T}} \mathsf{E}_n h_.$. Let $h^{(k)} = \min_{h \in J_k} \rho_{2,n}(h, h^*) \leq 2^{-k}M$. Then $\rho_{2,n}(h^{(k)}, h^*) \leq 2^{-k}M$ holds for any $h \in J_k$. Moreover we have

$$\rho_{2,n}(h^{(k-1)}, h^{(k)}) \leq \rho_{2,n}(h^{(k-1)}, h^*) + \rho_{2,n}(h^{(k)}, h^*) \leq 3 \cdot 2^{-k}M. \tag{99}$$

By a standard chaining argument, we express any partial sum of $h \in \mathbf{H_T}$ as a telescoping sum,

$$\sum_{t=1}^{n} h_t \leq |\sum_{t=1}^{n} h_t^{(0)}| + |\sum_{k=1}^{\overline{K}} \sum_{t=1}^{n} (h_t^{(k)} - h_t^{(k-1)})| + |\sum_{t=1}^{n} (h_t^{(\overline{K})} - h_t^*)|. \tag{100}$$

The inequality $|\sum_t a_t| \leq \sum_t |a_t| \leq |\sum_t a_t^2|^{1/2}\sqrt{n}$ can be applied to the third term. Notice that, by the definition of the $h_t^{(\overline{K})}$m

$$|\sum_{t=1}^{n} (h_t^{(\overline{K})} - h_t^*)| \leq |(\sum_{t=1}^{n} (h_t^{(\overline{K})} - h_t^*)^2)^{1/2}|\sqrt{n} \leq n2^{-\overline{K}}M, \tag{101}$$

Thus,

$$\mathsf{E}(|\mathsf{E}_n h|_{\mathbf{H_T}}) \leq \sum_{k}^{\overline{K}-1} \mathsf{E} \max_{f \in J_k, g \in J_{k-1}, \rho_{2,n}(f,g) \leq 3 \cdot 2^{-k} \cdot M} |\mathsf{E}_n (f - g)| + 2^{-\overline{K}}M. \tag{102}$$

Apply Lemma D.2 and Assumption 3.9 to (102). The maximal inequality (94) of Lemma D.2 is reproduced here:

$$\mathsf{E}(\max_{a \in \mathcal{A}} |M_a|) \lesssim A \log(1 + |\mathcal{A}|) + \sqrt{B}\sqrt{\log(1 + |\mathcal{A}|)}$$

For the first term of (102), we have $\mathcal{A} = \{f - g : f \in J_k, g \in J_{k-1}, \rho_{2,n}(f, g) \leq 3 \cdot 2^{-k} \cdot M\}$, $|\mathcal{A}| = |J_k||J_{k-1}| \leq 2\mathcal{N}^2(2^{-k}M, \mathbf{H_T}, \rho_{2,n}(.)) \lesssim_p 2 \max_t \sup_Q \mathcal{N}^2(2^{-k}M, \mathcal{H}_t, \|.\|_{Q,2})$, and $A \leq 3M$. B in (94) is an upper bound of the sum of conditional variances of an MDS. By Assumption 3.9, we have $B = \sum_t \mathsf{E}[(f_t - g_t)^2 | \mathcal{F}_{t-1}] \leq nL^2 \rho_{2,n}(f, g)^2 \leq nL^2(3 \cdot 2^{-k}M)^2$ for $f - g \in \mathcal{A}$.

Therefore, by Lemma D.2, we have the following

$$\begin{aligned} n\,\mathsf{E}(|\mathsf{E}_n h|_{\mathbf{H_T}}) \quad &\lesssim \sum_{k=1}^{\overline{K}} (L * 3 * 2^{-k}M\sqrt{n}) \sqrt{\log(1 + 2 \max_t \sup_Q \mathcal{N}^2(2^{-k}M, \mathcal{H}_t, \|.\|_{Q,2}))} \\ &\quad + 3 * M \sum_{k=1}^{\overline{K}} \log(1 + 2 * \max_t \sup_Q \mathcal{N}^2(2^{-k}M, \mathcal{H}_t, \|.\|_{Q,2})) + o_p(\sqrt{n}) \\ &\lesssim 6\sqrt{n} \int_0^1 M \sqrt{\log(2^{1/2} \max_t \sup_Q \mathcal{N}(2^{-k}M, \mathcal{H}_t, \|.\|_{Q,2}))} d\varepsilon. \end{aligned}$$

The second inequality is because $\overline{K} \lesssim \log(n)$ and because, by Assumption 3.8, $\max_t \log\sup_Q \mathcal{N}^2(2^{-k}M, \mathcal{H}_t, \|.\|_{Q,2}) \leq \log(K)+\log(v+1)+(v+1)(\log 4+1)+(2v)\log(\frac{2}{2^{-k}M})$. Then from equation (40) in Assumption 3.8, we have

$$\mathsf{E}(|\,\mathsf{E}_n h|_{\mathbf{H_T}}) \lesssim M\sqrt{v/n}.$$

□

The next lemma concerns the tail probability bound. It states that, under certain regularity conditions, $|\,\mathsf{E}_n h|_{\mathbf{H_T}}$ is very close to $\mathsf{E}(|\,\mathsf{E}_n h|_{\mathbf{H_T}})$.

LEMMA D.4. Under Assumption 3.2, 3.3, and 3.7 to 3.9,

$$|\,\mathsf{E}_n h|_{\mathbf{H_T}} - \mathsf{E}(|\,\mathsf{E}_n h|_{\mathbf{H_T}}) \lesssim_p M c_n\sqrt{v/n},$$

where c_n is arbitrarily slowly growing sequence.

Proof. Similar to the above derivation, for a positive constant η_k, with $\sum_k \eta_k \leq 1$,

$$\begin{aligned}
& \Pr(n^{-1}\sum_{t=1}^{n} h_t \geq x) \\
\leq\ & \Pr(n^{-1}|\sum_{k=1}^{\overline{K}}\sum_{t=1}^{n} h_t^{(k)} - h_t^{(k-1)}| \geq x - \sqrt{n}^{-1}2^{-\overline{K}}M) \\
\leq\ & \sum_{k=1}^{\overline{K}} \Pr(|n^{-1}\sum_{t=1}^{n} h_t^{(k)} - h_t^{(k-1)}| \geq \eta_k(x - \sqrt{n}^{-1}2^{-\overline{K}}M)) \\
\leq\ & \sum_{k=1}^{\overline{K}} \exp\{\log\max_t\sup_Q \mathcal{N}^2(2^{-k}M, \mathcal{H}_t, \|.\|_{Q,2}) - \eta_k^2(nx - \sqrt{n}2^{-\overline{K}}M)^2/[2\{(nx - \sqrt{n}2^{-\overline{K}}M) \\
& +2((3\cdot 2^{-k}\cdot M)^2 n)\}]\} \\
\leq\ & \sum_{k=1}^{\overline{K}} \exp(\log(K) + \log(v+1) + (v+1)(\log 4+1) + (2v)\log(\frac{2}{2^{-k}M}) \\
& -\eta_k^2(nx - \sqrt{n}2^{-\overline{K}}M))^2/(2((nx - \sqrt{n}2^{-\overline{K}}M) + 2((3\cdot 2^{-k}\cdot M)^2 n))),
\end{aligned}$$

where the above derivation is due to the tail probability in Lemma D.1. We pick η_k and x to ensure the right hand side converges to zero and $\sum_k \eta_k \leq 1$.

We take $b_k = \log(\overline{K}) + \log(v+1) + (v+1)(\log 4+1) + (2v)\log(\frac{2}{2^{-k}M})$, $a_k = 2^{-1}(nx - \sqrt{n}2^{-\overline{K}}M))^2/((nx-\sqrt{n}2^{-\overline{K}}M)+2((3*2^{-k}*M)^2 n))$. We pick $\eta_k \geq \sqrt{a_k/b_k}$, so that $b_k \leq \eta_k^2 a_k$. We also need to choose x to ensure that $\sum_k \eta_k \leq 1$ and $\sum_k \exp(b_k - \eta_k^2 a_k) \to 0$. We pick $c_n\sqrt{v/n} \lesssim x$, and $\eta_k = c_n'\sqrt{b_k/a_k}$, with two slowly growing functions c_n and c_n' and $c_n' \ll c_n$.

We set $x = \mathsf{E}(|\,\mathsf{E}_n h|_{\mathbf{H_T}}) + c_n\sqrt{v/n}$. The result then follows. Finally, Theorem 3.1 follows from Assumption 3.7, Lemma D.3, and Lemma D.4, with $n = T - 1$.

□

D.4 Justifying Assumption 3.8

Here we now intepret the entropy condition in Assumption 3.8. We follow the argument of Chapter 11 in Kosorok (2008) for the functional class related to non i.i.d. data.

We define $|.|_2$ to be the L_2 norm of a vector. Note that, if the function is stationary, then

$$h_1(.,.) = h_2(.,.) = \cdots = h_{T-1}(.,.).$$

In the stationary case, define $\mathcal{H} = h(.,.)$, and the envelope function to be $H(.,.)$. Define the finite discrete measure to be $Q_{\alpha_T} = \{|\alpha_{T-1}|_2\}^{-1}\sum_t \alpha_{T,t}\delta_{d_t}$($\alpha_{T,t}$ is the tth element of α_T), where δ_{d_t} is the Dirac measure at point d_t. Let $\tilde{\alpha}_{T,t} = \alpha_{T,t}/|\alpha_T|_2$. Define $|h(.,.)|_{Q,\alpha,2} = (\{|\alpha_T|_2^2\}^{-1}\sum_t \alpha_{T,t}^2 h(d_t, G)^2)^{1/2}$. Thus, in the stationary case, for any α_T, there exists a Q_{α_T}, such that,

$$\mathcal{N}(\delta|\tilde{\alpha}_T \circ \overline{H}_T|_2, \tilde{\alpha}_T \circ \mathbf{H}_T, |.|_2) \leq \sup_Q \mathcal{N}(\delta|H|_{Q,\alpha,2}, \mathcal{H}, |.|_{Q,\alpha,2}).$$

This is because that, there exist a Q , such that,

$$|\tilde{\alpha}_T \circ H_T|_2 = |\alpha_T|_2^{-1}(\sum_t \alpha_{T,t}^2 H_t^2)^{1/2} = |H_T|_{Q,\alpha,2}.$$

In light of this relationship in the stationary case, we generalize the assumption to some extent. Assume a subset $\mathcal{K}$ of $\{1, \cdot, T-1\}$ of $K = |\mathcal{K}|$ dimension. Let $\alpha_{T,K}$ denote the sub-vector of K dimension corresponding to the index set $\mathcal{K}$. Moreover, we similarly define the vector of the envelope function $\overline{H}_{T,K}$, and $\mathbf{H}_{T,K}$ as the corresponding functional class. In addition, let H_t denote the envelope function and $\mathcal{H}_t$ the functional class corresponding to $\{h_t(., G), G \in \mathcal{G}\}$.

We assume the following: for any fixed K, the covering number can effectively be reduced to K dimension,

$$\begin{aligned}\mathcal{N}(\delta|\tilde{\alpha}_T \circ \overline{H}_T|_2, \tilde{\alpha}_{\mathbf{T}} \circ \mathbf{H_T}, |.|_{\mathbf{2}}) &\leq \max_{\mathbf{H_{T,K}} \in \mathbf{H_T}} \mathcal{N}(\delta|\tilde{\alpha}_{T,K} \circ \overline{H}_{T,K}|_2, \tilde{\alpha}_{\mathbf{T,K}} \circ \mathbf{H_{T,K}}, |.|_{\mathbf{2}}), \\ &\leq \max_{\mathbf{H_{T,K}} \in \mathbf{H_T}} \prod_{t \in \mathcal{K}} \mathcal{N}(\delta|\tilde{\alpha}_{T,t} \circ \overline{H}_t|_2, \tilde{\alpha}_{T,t} \circ \mathbf{H}_t, |.|_2), \\ &\leq \max_{t \in \mathcal{K}} \sup_Q \mathcal{N}(\delta|H_t|_{Q,\alpha,2}, \mathcal{H}_t, |.|_{Q,\alpha,2})^K, \\ &\leq \max_{t \in \mathcal{K}} (\sup_Q \mathcal{N}(\delta\|H_t\|_{Q,2}, \mathcal{H}_t, \|.\|_{Q,2}))^K.\end{aligned}$$

Assumption 3.8 sets $K = 1$. Then it suffices to examine the one-dimensional covering number,i.e., $\mathcal{N}(\delta\|H_t\|_{Q,2}, \mathcal{H}_t, \|.\|_{Q,2})$. In the following, we study $\mathcal{N}(\delta\|H_t\|_{Q,2}, \mathcal{H}_t, \|.\|_{Q,2})$ for a single t.

We let $\mathbf{1}(X_{t-1}^\top \theta \leq 0) = \mathbf{1}(X_{t-1} \in G)$ for some $\theta \in \Theta$, where Θ is a compact set in $\mathbb{R}^d$. Without loss of generality we assume that it is an L_2 ball of $|.|_2 \leq 1$. Define $S_{t,1} \stackrel{\text{def}}{=} Y_t(1)\mathbf{1}(W_t = 1)/P(W_t = 1|\mathcal{F}_{t-1})$. Let $\mathcal{F}_{t-1} = \sigma(X_{0:t-1}, Y_{0:t-1}(.), W_{0:t-1})$. $P^1_{\theta,t-1} \stackrel{\text{def}}{=} \mathsf{E}(Y_t(1)\mathbf{1}(X_{t-1}^\top \theta \leq 0)|\mathcal{F}_{t-1})$, and $P^0_{\theta,t-1} \stackrel{\text{def}}{=} \mathsf{E}(Y_t(0)\mathbf{1}(X_{t-1}^\top \theta > 0)|\mathcal{F}_{t-1})$. We abbreviate $Y_t(W_{0:t-1}, 1)$ to $Y_t(1)$. Recall that we assume $|Y_t(1)|, |Y_t(0)| \leq M$. $\|H_t\|_{Q,r} \leq 2M + 2M/\kappa \stackrel{\text{def.}}{=} M'$.

$$h_{\theta,t-1} \;=\; P^1_{\theta,t-1} + P^0_{\theta,t-1} - S_{t,1}\mathbf{1}(X_{t-1}^\top \theta \leq 0) - S_{t,0}\mathbf{1}(X_{t-1}^\top \theta > 0).$$

and that, by definition, $\mathsf{E}(h_{\theta,t-1}|\mathcal{F}_{t-1}) = 0$. Thus, the corresponding functional class is

$$\mathcal{H}_t = \{h : (y_t, x_{t-1}) \to f^\theta_{1,1} + f^\theta_{1,0} + f^\theta_{0,1} + f^\theta_{0,0}, \theta \in \Theta\},$$

where $f^\theta_{1,1}$(resp. $f^\theta_{1,0}, f^\theta_{0,1}, f^\theta_{0,0}$) corresponds to $Y_t(1)\mathbf{1}(X_{t-1}^\top \theta \leq 0)$ (resp. $Y_t(0)\mathbf{1}(X_{t-1}^\top \theta > 0), -S_{t,1}\mathbf{1}(X_{t-1}^\top \theta \leq 0), -S_{t,0}\mathbf{1}(X_{t-1}^\top \theta > 0)$). For all finitely discrete norms Q. We know that

$$\sup_Q \mathcal{N}(\varepsilon, \mathcal{H}_t, \|.\|_{Q,r}) \leq \sup_Q \mathcal{N}(\varepsilon/4, \mathcal{F}_{1,1}, \|.\|_{Q,r})\mathcal{N}(\varepsilon/4, \mathcal{F}_{1,0}, \|.\|_{Q,r})\mathcal{N}(\varepsilon/4, \mathcal{F}_{0,1}, \|.\|_{Q,r})\mathcal{N}(\varepsilon/4, \mathcal{F}_{0,0}, \|.\|_{Q,r}). \tag{103}$$

We look at the covering number of the respective functional class. According to Lemma 9.8 of Kosorok (2008), the subgraph of the function $\mathbf{1}(X_{t-1}^\top \theta \leq 0)$ is of VC dimension less than $d+2$ because the class $\{x \in \mathbb{R}^d, x^\top \theta \leq 0, \theta \in \Theta\}$ is VC of dimension less than $d+2$ (see the proof of Lemma 9.6 of Kosorok (2008)).

Therefore $\sup_Q \mathcal{N}(\varepsilon/4, \mathcal{F}_{0,1}, \|.\|_{Q,r}) \vee \mathcal{N}(\varepsilon/4, \mathcal{F}_{0,1}, \|.\|_{Q,r}) \lesssim (4/(\varepsilon M'))^{d+2}$. Moreover, we assume the following Lipschitz condition on the function $P^1_{\theta,t-1}$,$P^0_{\theta,t-1}$: $|P^1_{\theta,t-1} - P^1_{\theta',t-1}|_{Q,r} \leq M_d|\theta - \theta'|_r$ (for any distinct points $\theta, \theta' \in \Theta$) and a positive constant M_d. Therefore it falls within the type II class defined in Andrews (1994), and according to the derivation of the

(A.2) Andrews (1994) we have

$$\sup_Q \mathcal{N}(\varepsilon M', \mathcal{F}_{1,1}, \|.\|_{Q,r}) \leq \sup_Q \mathcal{N}(\varepsilon M'/M_d, \Theta, \|.\|_{Q,r}), \tag{104}$$

where the latter is the covering number of a Euclidean ball with $\|.\|_{Q,r}$. Thus, according Equation (5.9) in Wainwright (2019), $\sup_Q \mathcal{N}(\varepsilon M'/M_d, \Theta, \|.\|_{Q,r}) \lesssim (1 + 2M_d/\varepsilon M')^d$. Finally, we can show that the covering number satisfying $\sup_Q \mathcal{N}(\varepsilon M', \mathcal{H}, \|.\|_{Q,r}) \lesssim (4/(\varepsilon M'))^{2(d+2)}(1 + 2M_d/(\varepsilon M'))^{2d}$.

D.5 Proof of Theorem 3.2

By Lemma D.4, based on Assumptions 3.8, and Assumptions 3.13, we have

$$\sup_{G\in\mathcal{G}}(T-1)^{-1}\sum_{t=1}^{T-1}\overline{S}_t(G) \lesssim_p M\sqrt{v}/\sqrt{T-1}. \tag{105}$$

We first show
(i) Let $\mathcal{G}$ be a function class of finite size. Assume Assumptions 3.10- 3.12 and $|\mathcal{G}| = \tilde{M} < \infty$. For a large enough constant c_T,

$$\sup_{G\in\mathcal{G}}\frac{1}{T-1}\sum_{t=1}^{T-1}\tilde{S}_t(G) \lesssim_p \frac{c_T[\log\tilde{M}]^{1/\gamma}4er\Phi^2_{\phi_v}}{\sqrt{T-2}}, \tag{106}$$

with probability $1 - \exp(-c_T^\gamma)$.

Then we extend the above result to obtain Theorem 3.2:
(ii) Let $\mathcal{G}$ be a function class of infinite whose complexity is subject to Assumption [D.4]. Assume Assumptions 3.10- 3.13, we have

$$\sup_{G\in\mathcal{G}}\frac{1}{T-1}\sum_{t=1}^{T-1}\tilde{S}_t(G) \lesssim_p \frac{c_T[V\log T]^{1/\gamma}2e\gamma\Phi_{\phi_v,F}}{\sqrt{T-1}}.$$

Proof. And the object $\mathsf{E}(S_t(G)|\mathcal{F}_{t-2})$ are continuously differentiable with respect to the underlying i.i.d. innovations. We have, for $\tilde{S}_t(G)$,

$$\begin{aligned}
&\tilde{S}_t(G) - \tilde{S}^*_{t,l}(G)\\
=\ &\mathsf{E}(S_t(G)|\mathcal{F}_{t-2}) - \mathsf{E}(S_t(G)|\mathcal{F}^*_{t-2,l})\\
=\ &\partial\,\mathsf{E}(S_t(G)|\mathcal{F}_{t-2})/\partial\varepsilon_{t-l-2}(\varepsilon_{t-l-2} - \varepsilon^*_{t-l-2})\\
\leq\ &|F_{t,l}(\varepsilon_{t-l-2} - \varepsilon^*_{t-l-2})|.
\end{aligned}$$

Thus the dependence adjusted norm satisfies $\theta_{x,q} \leq \sum_{l\geq 0}\max_t\|F_{t,l}(X_{t-2})\|_q$.

And

$$\Phi_{\phi_v} \leq \sup_{q\geq 2}(\sum_{l\geq 0} \max_t \|F_{t,l}(X_{t-2})(\varepsilon_{t-l-2} - \varepsilon^*_{t-l-2})\|_q)/q^{\tilde{v}} \stackrel{\text{def}}{=} \Phi_{\phi_v,F}, \tag{107}$$

$\tilde{v} = 1$ for subexponential, and $\tilde{v} = 1/2$ for subGaussian. Following Assumption 3.12, we assume that $\Phi_{\phi_v,F}$ is finite. $\Phi_{\phi_v} < \infty$. Define $\gamma = 1/(1 + 2\tilde{v})$. Recall that we assume $\mathcal{G}$ is finite, and $|\mathcal{G}| = \tilde{M}$.

According to Theorem 3 of Wu and Wu (2016). The exponential bound is stated as follows

$$\Pr(\sup_{G\in\mathcal{G}} \sum_{t=1}^{T-1} \tilde{S}_t(G) \geq x) \leq \tilde{M} \exp[-x^\gamma/(2e\gamma)(\sqrt{T-2}\Phi_{\phi_v})^\gamma]. \tag{108}$$

We now verify (ii), where $\mathcal{G}$ is not finite. Set $x = c_T[\log \tilde{M}]^{1/\gamma} 2e\gamma\Phi_{\phi_v}/\sqrt{T-2}$, where α_T is a sufficiently large constant. So

$$\sup_{G\in\mathcal{G}} \frac{1}{T-1} \sum_{t=1}^{T-1} \tilde{S}_t(G) \lesssim_p c_T[\log \tilde{M}]^{1/\gamma} 4er\Phi^2_{\phi_v}/\sqrt{T-1}, \tag{109}$$

with probability $1 - \exp(-c_T^\gamma)$.

We define $\mathcal{G}^{(1)\delta}$ to be the $\delta\text{sup}_Q\|F_p\|_{Q,2}$ net of $\mathcal{G}$. We denote $\pi(G)$ as closest component of G in the net $\mathcal{G}^{(1)\delta}$. Then,

$$\begin{aligned}
& \sup_G \frac{1}{T-1} \sum_t \tilde{S}_t(G) \\
\leq\ & \sup_{G\in\mathcal{G}} \frac{1}{T-1} \sum_{t=3}^{T} (\tilde{S}_t(G) - \tilde{S}_t(\pi(G))) + \sup_{G\in\mathcal{G}^{(1)\delta}} |\frac{1}{T-1} \sum_{t=1}^{T-1} (\tilde{S}_t(G)| \\
\leq\ & \delta\text{sup}_Q\|F_{p,t}\|_{Q,2} + \sup_{G\in\mathcal{G}^{(1)\delta}} |\frac{1}{T-1} \sum_{t=1}^{T-1} (\tilde{S}_t(G)| \\
\lesssim_p\ & \delta\text{sup}_Q\|F_{p,t}\|_{Q,2} + c_T[V\log(1/\delta)]^{1/\gamma} 4e\gamma\Phi^2_{\phi_v}/\sqrt{T-1}.
\end{aligned}$$

Finally, by setting $\delta \ll \sqrt{T}^{-1}$,

$$\sup_G \frac{1}{T-1} \sum_{t=3} \tilde{S}_t(G) \lesssim_p c_T[V\log T]^{1/\gamma} 2e\gamma\Phi_{\phi_v,F}/\sqrt{T-1}. \tag{110}$$

Recall that V, $\tilde{v}$ are constants related to the VC class. $\Phi_{\phi_v,F}$ is a subexponential dependence adjusted norm, and $\gamma = 1/(1 + 2\tilde{v})$ is a constant related to the moment and dependency conditions. □

D.6 Proof of Theorem 4.1

For simplicity, we maintain Assumption 3.1, which implies (27)

In addition to (52) and (53), we define

$$\begin{aligned}\widetilde{\mathcal{W}}_h(G,x) &= \frac{\sum_{t=1}^{T-1} K_h(X_{t-1},x)\,\mathsf{E}\left[\frac{Y_tW_t}{e_t(X_{t-1})}1(X_{t-1}\in G)+\frac{Y_t(1-W_t)}{1-e_t(X_{t-1})}1(X_{t-1}\notin G)|\mathcal{F}_{t-1}\right]}{\sum_{t=1}^{T-1}K_h(X_{t-1},x)}\\ &= \frac{\sum_{t=1}^{T-1}K_h(X_{t-1},x)\mathcal{W}_t(G|\mathcal{F}_{t-1})}{\sum_{t=1}^{T-1}K_h(X_{t-1},x)},\end{aligned}$$

where $K(\cdot)$ is a bounded kernel with a bounded support, $K_h(a,b) := \frac{1}{h}K(\frac{a-b}{h})$. The second equality follows from Assumption 3.3.

Our strategy is to show for any $x \in \mathcal{X}$ and any $G \in \mathcal{G}$:

$$\begin{aligned}&\mathcal{W}_T(G|x) - \mathcal{W}_T(\hat{G}|x)\\ &\le c\big[\bar{\mathcal{W}}_h(G|x) - \bar{\mathcal{W}}_h(\hat{G}|x)\big]\\ &\le c\big[\widetilde{\mathcal{W}}_h(G,x) - \widetilde{\mathcal{W}}_h(\hat{G},x)\big] + \mathcal{O}_p(h^2) + \mathcal{O}_p(c_w^{-1}(\sqrt{(T-1)h})^{-1} + (T-1)^{-1})\\ &\le \sup_{G\in\mathcal{G}} 2c|\widetilde{\mathcal{W}}_h(G,x) - \widehat{\mathcal{W}}(G|x)| + \mathcal{O}_p(h^2) + \mathcal{O}_p(c_w^{-1}(\sqrt{(T-1)h})^{-1} + (T-1)^{-1})\\ &= \mathcal{O}_p(h^2) + \mathcal{O}_p(c_w^{-1}(\sqrt{(T-1)h})^{-1} + (T-1)^{-1})\end{aligned}$$

where the first inequality follows from Assumption 4.1. The second inequality follows from Lemma D.5 below. The third inequality follows from similar arguments to (22). The last equality follows from Lemma D.6 stated below.

We present these two lemmas and their proofs. First, let us assume:

Assumption D.1. X_t is one dimensional covariate. Let c_m, c_k, and c_w be positive constants. (i) The kernel function $K(x)$ is bounded and has bounded support $\{|x| \le c_k\}$. $\mathsf{E}(K^2(\frac{X_t-x}{h})|\mathcal{F}_{t-1}) \le c_m$, and $\min_{t,x}\mathsf{E}(K((X_t-x)/h)|\mathcal{F}_{t-1}) \ge c_w$; (ii) $\mathcal{W}_t(G|x)$ is second order differentiable w.r.t x. x is in the interior point of its support $\mathcal{X}$, which is also bounded; (iii) $\int K(u)^2du$ is bounded.

LEMMA D.5. Under Assumption D.1 For any $G \in \mathcal{G}$ and $x \in \mathcal{X}$,

$$\widetilde{\mathcal{W}}_h(G,x) - \bar{\mathcal{W}}_h(G|x) = \mathcal{O}_p(h^2) + \mathcal{O}_p(c_w^{-1}(\sqrt{(T-1)h})^{-1} + ((T-1))^{-1}).$$

Proof. Under Assumption D.1,

$$\begin{aligned}
&\left[\widetilde{\mathcal{W}}_h(G,x) - \bar{\mathcal{W}}_h(G|x)\right]\left[\frac{1}{T-1}\sum_{t=1}^{T-1} K_h(X_{t-1},x)\right] \\
&= \frac{1}{T-1}\sum_{t=1}^{T-1} K_h(X_{t-1},x)(\mathcal{W}_t(G|X_{t-1}) - \mathcal{W}_t(G|x)) \\
&= \frac{1}{T-1}\sum_{t=1}^{T-1}\Big\{ K_h(X_{t-1},x)(\mathcal{W}_t(G|X_{t-1}) - \mathcal{W}_t(G|x)) \\
&- \mathsf{E}[K_h(X_{t-1},x)(\mathcal{W}_t(G|X_{t-1}) - \mathcal{W}_t(G|x))|\mathcal{F}_{t-2}]\Big\} \\
&+ \frac{1}{T-1}\sum_{t=1}^{T-1}\Big\{ \mathsf{E}[K_h(X_{t-1},x)(\mathcal{W}_t(G|X_{t-1}) - \mathcal{W}_t(G|x))|\mathcal{F}_{t-2}]\Big\}.
\end{aligned}$$

Rearranging the equation we have

$$\widetilde{\mathcal{W}}_h(G,x) - \bar{\mathcal{W}}_h(G|x) = \mathcal{O}_p(c_1^{-1}(\sqrt{(T-1)h})^{-1} + (T-1)^{-1}) + \mathcal{O}_p(h^2),$$

where the first term on the right hand side follows from Theorem 3.1 and the second term follows from the standard result concerning the bias of the kernel estimator. □

LEMMA D.6. Under Assumption D.1,

$$\sup_{G\in\mathcal{G}}\sup_{x\in\mathcal{X}} |\widehat{\mathcal{W}}(G|x) - \widetilde{\mathcal{W}}_h(G,x)| \lesssim_p c_w^{-1}(\sqrt{(T-1)h}^{-1} + ((T-1))^{-1}). \tag{111}$$

Proof. Note

$$\begin{aligned}
&\sup_{G\in\mathcal{G}}\sup_{x\in\mathcal{X}} |\widehat{\mathcal{W}}(G|x) - \widetilde{\mathcal{W}}_h(G,x)| \\
\leq\ &\sup_{G\in\mathcal{G}}\sup_{x\in\mathcal{X}} \sum_{t=1}^{T-1} K_h(X_{t-1},x)(\widehat{\mathcal{W}}_t(G) - \mathcal{W}_t(G|\mathcal{F}_{t-1}))/\sum_{t=1}^{T-1} K_h(X_{t-1},x)
\end{aligned}$$

We just look at the numerator, $\sup_{G\in\mathcal{G}}\sup_{x\in\mathcal{X}}\sum_{t=1}^{T-1} K_h(X_{t-1},x)(\widehat{\mathcal{W}}_t(G) - \mathcal{W}_t(G|\mathcal{F}_{t-1}))$. Suppose the order statistics of X_t is $X_{(1)},\cdots,X_{(T-1)}$. $B_{x,h} = \{t : |(x-x_t)/h| \leq c_k\}$.

Now because of summation by part, the above object is bounded by,

$$\begin{aligned}
&\sup_{G\in\mathcal{G}}\sup_{x\in\mathcal{X}}|\sum_{t}^{T-1}K_h(X_{(t)},x)-K_h(X_{(t-1)},x)|\\
&\max_{1\le l\le T-1,t\in B_{x,h}}|\sum_{t=1,t\in B_{x,h}}^{l}(\widehat{\mathcal{W}}_t(G)-\mathcal{W}_t(G|\mathcal{F}_{t-1}))|\\
&+\sup_{G\in\mathcal{G}}\sup_{x\in\mathcal{X}}|K_h(X_{(T-1)},x)||\sum_{t\in B_{x,h}}(\widehat{\mathcal{W}}_t(G)-\mathcal{W}_t(G|\mathcal{F}_{t-1}))|\\
&\lesssim_p h^{-1}\sup_{G}|\sum_{t\in B_{x,h}}(\widehat{\mathcal{W}}_t(G)-\mathcal{W}_t(G|\mathcal{F}_{t-1}))|.
\end{aligned}$$

$|\sum_t^{T-1}[K_h(X_{(t)},x)-K_h(X_{(t-1)},x)]|\lesssim_p h^{-1}$ if the total variation of the function $hK_h(.,x)$ is bounded. $\sup_{G\in\mathcal{G}}(T-1)^{-1}|\sum_{t\in B_{x,h}}(\widehat{\mathcal{W}}_t(G)-\mathcal{W}_t(G|\mathcal{F}_{t-1}))|\lesssim_p(\sqrt{h}/\sqrt{T})$, where $\lesssim_p$ follows from $|\widehat{\mathcal{W}}_t(G)|\le M$.

For the denominator, $(T-1)^{-1}\sum_{t=1}^{T-1}\{K_h(X_{t-1},x)-\mathsf{E}(K_h(X_{t-1},x)|X_{t-2})\}\lesssim_p 1/\sqrt{h(T-1)}+(T-1)^{-1}$, where $\lesssim_p$ follows from c_m, $\int K(u)^2du$ and the bound on the kernel function as assumed in Assumption D.1. Due to the boundedness of $\mathsf{E}(K_h(X_{t-1},x)|\mathcal{F}_{t-2})$ from the Assumption D.1, by following similar steps to the Proof of (23), we have

$$\sup_{G\in\mathcal{G}}\sup_{x\in\mathcal{X}}|\widehat{\mathcal{W}}(G|X_{T-1}=x)-\bar{\mathcal{W}}(G|X_{T-1}=x)|\lesssim_p c_w^{-1}(\sqrt{(T-1)h}^{-1}). \tag{112}$$

□

D.7 Proof of Theorem 5.1

Recall the definitions,

$$\begin{aligned}
&\bar{\mathcal{W}}(g|w) \qquad (113)\\
=\;& T(w)^{-1}\sum_{1\le t\le T-1:W_{t-1}=w}\mathsf{E}\left\{Y_t(W_{t-1},1)g(W_{t-1})+Y_t(W_{t-1},0)\left[1-g(W_{t-1})\right]|W_{t-1}\right\}.
\end{aligned}$$

and

$$g^*(w)=\arg\max_{g(w)\to 0,1}\bar{\mathcal{W}}(g|w). \tag{114}$$

$$\begin{aligned} & \bar{\mathcal{W}}(g, W_{t-2}) \\ \stackrel{\text{def}}{=} \; & (T-1)^{-1} \sum_{1 \le t \le T-1} \mathsf{E}\left\{ Y_t(W_{t-1}, 1) g(W_{t-1}) + Y_t(W_{t-1}, 0) \left[1 - g(W_{t-1})\right] | W_{t-2} \right\}. \end{aligned} \tag{115}$$

$$\begin{aligned} & \widehat{\mathcal{W}}(g) \\ \stackrel{\text{def}}{=} \; & (T-1)^{-1} \sum_{1 \le t \le T-1} \left\{ Y_t(W_{t-1}, 1) g(W_{t-1}) + Y_t(W_{t-1}, 0) \left[1 - g(W_{t-1})\right] \right\}. \end{aligned} \tag{116}$$

$$\begin{aligned} & \bar{\mathcal{W}}(g) \\ \stackrel{\text{def}}{=} \; & (T-1)^{-1} \sum_{1 \le t \le T-1} \mathsf{E}\left\{ Y_t(W_{t-1}, 1) g(W_{t-1}) + Y_t(W_{t-1}, 0) \left[1 - g(W_{t-1})\right] \right\}. \end{aligned} \tag{117}$$

Proof. It is simple to show that,

$$g^*(.) = \arg \max_{g(w) \to 0,1} \bar{\mathcal{W}}(g, W_{t-2}).$$

By a similar derivation to (22), there exists a positive constant c_w, c'_w such that

$$\begin{aligned} & \bar{\mathcal{W}}(g^*|w) - \bar{\mathcal{W}}(g|w) \le c_w (\bar{\mathcal{W}}(g^*, W_{t-2}) - \bar{\mathcal{W}}(g, W_{t-2})), \\ & (\bar{\mathcal{W}}(g^*, W_{t-2}) - \bar{\mathcal{W}}(g, W_{t-2})) \le c'_w (\bar{\mathcal{W}}(g^*) - \bar{\mathcal{W}}(g)). \end{aligned}$$

Therefore it suffices to look at $\hat{g} = \arg\max_{g(w) \to 0,1} \widehat{\mathcal{W}}(g)$. Throughout this section we take $\mathsf{E}_{t-1}(\cdot) = \mathsf{E}(\cdot|\mathcal{F}_{t-2})$, and $\mathrm{P}_{t-1}(\cdot) = \Pr(\cdot|\mathcal{F}_{t-2})$. $\hat{g}$ is defined as in Section 2.2.

The first best rule is $\{x : \tau(x) \ge 0\}$. Assume that there exists a positive constant $\overline{P}$ such that $\frac{1}{T-1} \sum_t \mathrm{P}_{t-1}(g^*(W_{t-1}) \ne g(W_{t-1})) \ge \overline{P}$ happens with probability 1, where P_{t-1} is the probability conditional on $\mathcal{F}_{t-2}$. We define the set of events $\{A : \min_t |\tau(W_{t-1})| > u\}$. Then,

$$\begin{aligned} & \bar{\mathcal{W}}(g^*, W_{t-2}) - \bar{\mathcal{W}}(g, W_{t-2}) = \sum_{1 \le t \le T-1} \frac{1}{T-1} \mathsf{E}_{t-1}\{|\tau(W_{t-1})| |\mathbf{1}(g^*(W_{t-1}) \ne g(W_{t-1}))|\}, \\ \ge \; & u \frac{1}{T-1} \sum_{1 \le t \le T-1} \mathrm{P}_{t-1}(g^*(W_{t-1}) \ne g(W_{t-1}) \cap A), \\ \ge \; & u \{ \sum_{1 \le t \le T-1} \frac{1}{T-1} [\mathrm{P}_{t-1}(g^*(W_{t-1}) \ne g(W_{t-1})) - (u/\eta)^\alpha] \}. \end{aligned}$$

Set $u = \eta(1+\alpha)^{-1/\alpha}\overline{P}^{1/\alpha} \leq \eta$, then we have

$$\bar{\mathcal{W}}(g^*, W_{t-2}) - \bar{\mathcal{W}}(g, W_{t-2}) \geq \overline{P}^{(1+\alpha)/\alpha}(1+\alpha)^{-(1+\alpha)/\alpha}\alpha\eta. \tag{118}$$

Let $\mathsf{Var}_{t-1}(.) = \mathsf{E}_{t-1}(.^2) - (\mathsf{E}_{t-1}(.))^2$.

From this definition,

$$\begin{aligned}
& \frac{1}{T-1}\sum_t \mathsf{Var}_{t-1}\{(Y_t(1) - Y_t(0))\mathbf{1}(g^*(W_{t-1}) \neq g(W_{t-1})\}, \\
= \; & \frac{1}{T-1}2\sum_t \mathsf{E}_{t-1}\{(Y_t^2(1) + Y_t^2(0))\mathbf{1}(g^*(W_{t-1}) \neq g(W_{t-1})\} \\
& -\frac{1}{T-1}\sum_t [\mathsf{E}_{t-1}\{(Y_t(1) - Y_t(0))\mathbf{1}(g^*(W_{t-1}) \neq g(W_{t-1})\}]^2, \\
\leq \; & \frac{1}{T-1}\sum_t 2\,\mathsf{E}_{t-1}\{(Y_t^2(1) + Y_t^2(0))\mathbf{1}(g^*(W_{t-1}) \neq g(W_{t-1})\}, \\
\leq \; & (2C^2)(\overline{P}), \\
\leq \; & (2C^2)(1/\eta\alpha)^{\alpha/(1+\alpha)}(1+\alpha)(\bar{\mathcal{W}}(g^*, W_{t-2}) - \bar{\mathcal{W}}(g, W_{t-2}))^{\alpha/(1+\alpha)},
\end{aligned} \tag{119}$$

where the second to last line follows from the bounds imposed on $Y_t(0), Y_t(1)$ in Assumption 5.3, and Assumption 2.2. The last line is due to Assumption 5.4.

Next from the Freedman inequality we have, with probability $1-\delta$, for a positive constant $C > 2$,

$$\begin{aligned}
& \bar{\mathcal{W}}(g^*, W_{t-2}) - \bar{\mathcal{W}}(\hat{g}, W_{t-2}) \\
\leq \; & \max_g(\widehat{\mathcal{W}}(g^*) - \widehat{\mathcal{W}}(g)) + \max_g |\bar{\mathcal{W}}(g^*, W_{t-2}) - \bar{\mathcal{W}}(g, W_{t-2}) - (\widehat{\mathcal{W}}(g^*) - \widehat{\mathcal{W}}(g))| \\
\leq \; & \frac{2\max_g \sqrt{\frac{1}{T-1}\sum_t C\,\mathsf{Var}_{t-1}\{(Y_t(1) - Y_t(0))\mathbf{1}(g^*(W_{t-1}) \neq g(W_{t-1})\}}}{\sqrt{T-1}}\sqrt{\log(2M/\delta)} \\
& + \log(2M/\delta)2C/(T-1), \\
\leq \; & \sqrt{(2C^2/\kappa))(1/\eta\alpha)^{\alpha/(1+\alpha)}(1+\alpha)\max_{G\in\mathcal{G}}(\bar{\mathcal{W}}(g^*, W_{t-2}) - \bar{\mathcal{W}}(g, W_{t-2}))^{\alpha/(1+\alpha)}}\sqrt{\log(2M/\delta)}/\sqrt{T-1} \\
& + \log(2M/\delta)4C/(\kappa(T-1)),
\end{aligned}$$

The first inequality is due to $\widehat{\mathcal{W}}(g^*) - \widehat{\mathcal{W}}(\hat{g}) \leq 0$, which follows from the definition of $\hat{g}$. The second line is based on (119).

Solving the inequality on both side with respect to $\bar{\mathcal{W}}(g^*, W_{t-2}) - \bar{\mathcal{W}}(\hat{g}, W_{t-2})$, we have with probability $1-\delta$, and there exists a positive constant c'_w such that,

$$\bar{\mathcal{W}}(g^*) - \bar{\mathcal{W}}(\hat{g}) \leq c'_w(\bar{\mathcal{W}}(g^*, W_{t-2}) - \bar{\mathcal{W}}(\hat{g}, W_{t-2})) \lesssim_p C_{\alpha,\eta,V}(\frac{\sqrt{\log 2/\delta)}}{\sqrt{T-1}})^{2(\alpha+1)/(\alpha+2)},$$

Under the stationarity of $X_t = \{Y_t, W_t\}$, $C_{\alpha,\eta,V}$ is a constant depending only on α, η, C and V. □

D.8 Proof of Theorem 5.2

Proof. We can assume a specific data generating process. Let $Y_t(W_{0:t}) = Y_t(W_t)$ with W_t i.i.d. over time with $\mathrm{P}(W_t = 1) = 1/2$ and $\mathrm{P}(W_t = 0) = 1/2$. Let v be an integer. Let $\mathbf{b} = (b_1, b_2, \cdots, b_v)$ be the class of models indexed by a v dimensional $0, 1$ vector, which will have 2^v elements. $j = 1, \cdots, v$ corresponds to j different partitions of X_t related to the set G_j. Next, we assume that $Y_t(1)$ has a binary distribution, with a positive constant $0 < r < 1/2$,

$$Y_t(1) = \begin{cases} 1/2, & \text{with probability } 1/2 + r \quad \text{if} \quad b_j = 1, X_{t-1} \in G_j, \\ -1/2, & \text{with probability } 1/2 - r \quad \text{if} \quad b_j = 1, X_{t-1} \in G_j. \end{cases} \tag{120}$$

$$Y_t(1) = \begin{cases} 1/2, & \text{with probability } 1/2 \quad \text{if} \quad b_j = 1, X_{t-1} \notin G_j, \\ -1/2, & \text{with probability } 1/2 \quad \text{if} \quad b_j = 1, X_{t-1} \notin G_j. \end{cases} \tag{121}$$

Otherwise,

$$Y_t(1) = \begin{cases} 1/2, & \text{with probability } 1/2 \quad \text{if} \quad b_j = 0, X_{t-1} \notin G_j, \\ -1/2, & \text{with probability } 1/2 \quad \text{if} \quad b_j = 0, X_{t-1} \notin G_j. \end{cases} \tag{122}$$

$$Y_t(1) = \begin{cases} 1/2, & \text{with probability } 1/2 - r \quad \text{if} \quad b_j = 0, X_{t-1} \in G_j, \\ -1/2, & \text{with probability } 1/2 + r \quad \text{if} \quad b_j = 0, X_{t-1} \in G_j. \end{cases} \tag{123}$$

and $Y_t(0) = 0$ if $W_t = 0$. Let X_t be a stationary Markovian process of order one. There exists G_j $(j = 1, \cdots, v)$, such that $X_t \in G_j$ corresponds to b_j and $\Pr(X_t \in G_j) > 0$ for all j. $\cup_j G_j = G$ is contained in the compact set from which X_t takes its value. Thus $\Pr(Y_t(1) = 1/2 | X_{t-1} \in G_j, b_j = 1) = 1/2 + r$. We also assume that X_t only causes $Y_t(.)$ via X_{t-1}, not directly. There exists a compact support of $x_{t-1:t}$ containing G_j and G, denoted as B_x. We abbreviate $Y_{0:T}(.)$ to $Y_{0:T}$. As $\hat{G}$ is estimated based on the whole sample $(Y_{0:T}, W_{0:T})$, we denote $g_T(i, X_{0:T}, Y_{0:T}) = \mathbf{1}(X_{t-1} \in \hat{G}, X_{t-1} \in G_i)$ and $f_{\mathbf{b}}(i) = \mathbf{1}(X_{t-1} \in G_i, b_i = 1)$, $\mathbf{1}(X_{t-1} \in \hat{G}\Delta G^*(\mathbf{b})) = \sum_i \mathbf{1}(g_T(i, X_{0:T}, Y_{0:T}) = 1 - f_{\mathbf{b}}(i))$. We assume $\mathbf{b}$ is uniformly distributed over 2^v elements. Moreover, $\mathsf{E}(Y_t(1) | b_j = 1, X_{t-1} \in G_j) = 1/2((1/2 + r) - (1/2 -$

$r)) = r$. Then,

$$\begin{aligned}
& \mathcal{W}_T(G^*(\mathbf{b})) - \mathcal{W}_T(\hat{G}), \\
\geq\ & r\,\mathsf{E}(\mathbf{1}(X_{t-1} \in \hat{G}\Delta G^*(\mathbf{b}))), \qquad \text{where } \widehat{G}\Delta G^*(\mathbf{b}) = (\hat{G} \cap G^*(\mathbf{b})^c) \cup (G^*(\mathbf{b}) \cap \hat{G}^c), \\
\geq\ & \frac{r}{v}\sum_{i=1}^{v} \mathsf{E}\{\mathbf{1}(g_T(i, X_{0:T}, Y_{0:T}) = 1 - f_{\mathbf{b}}(i))\}.
\end{aligned}$$

We let $y_{0:T}$ denote any particular realized path of $Y_{0:T}(w_{0:T})$. The likelihood of the model given one $\mathbf{b}$ can be written as

$$\mathrm{P}_{\mathbf{b}}(x_0, x_1, x_2, \cdots, x_T, y_0, y_1, y_2, \cdots, y_T, w_{0:T}) = \Pi_{t=1}^{T-1}\{p(x_t|x_{t-1})\mathrm{P}_{\mathbf{b}}(y_t|x_{t-1}, w_t)\mathrm{P}(w_t)\}p(x_0, y_0, w_0), \tag{124}$$

where $\mathrm{P}_{\mathbf{b}}(y_t|x_{t-1}, w_t)$ is the conditional density of y_t on x_{t-1}, w_t and $\mathbf{b}$. We let $\mathsf{E}\{\mathbf{1}(g_T(i, X_{0:T}, Y_{0:T}) = 1 - f_{\mathbf{b}}(i))\} = E_i$.

We proceed to bound the value of E_i. Let $\mathrm{P}_{b_{i+}}(y_t|x_{t-1})$ denote the probability conditional on $b_i = 1$, and $\mathrm{P}_{b_{i-}}(y_t|x_{t-1})$ denote the probability conditonal on $b_i = 0$. We let

$$p_+(x_{0:T}, y_{0:T}) \stackrel{\text{def}}{=} [\Pi_{t=1}^{T-1}\{p(x_t|x_{t-1})\mathrm{P}_{b_{i+}}(y_t|x_{t-1}, w_t)\mathrm{P}(w_t)\}, p(x_0, y_0, w_0)],$$

and

$$p_-(x_{0:T}, y_{0:T}) \stackrel{\text{def}}{=} [\Pi_{t=1}^{T-1}\{p(x_t|x_{t-1})\mathrm{P}_{b_{i-}}(y_t|x_{t-1}, w_t)\mathrm{P}(w_t)\}p(x_0, y_0, w_0)].$$

Then we have,

$$\begin{aligned}
& E_i = 2^{-v}\sum_{y_{0:T}, w_{0:T}}\sum_{\mathbf{b}}\int_{x_{0:T}} \mathbf{1}(g_T(i, x_{0:T}, y_{0:T}) = 1 - f_{\mathbf{b}}(i)) \\
& \Pi_{t=0}^{T-1}\{p(x_t|x_{t-1})\mathrm{P}_{\mathbf{b}}(y_t|x_{t-1}, w_t)\mathrm{P}(w_t)\}p(x_0, y_0, w_0)dx_{0:T}, \\
\geq\ & 2^{-v}\sum_{y_{0:T}, w_{0:T}}\sum_{\mathbf{b}}\int_{x_{0:T}} 1/2[\mathbf{1}(g_T(i, x_{0:T}, y_{0:T}) = 1)p_+(x_{0:T}, y_{0:T}) \\
& +\mathbf{1}(g_T(i, x_{0:T}, y_{0:T}) = 0)p_-(x_{0:T}, y_{0:T})]dx_{0:T}, \\
\geq\ & 2^{-v}\sum_{y_{0:T}, w_{0:T}}\sum_{\mathbf{b}}\int_{x_{0:T}} (1/2\{\mathbf{1}(g_T(i, x_{0:T}, y_{0:T}) = 1) + \mathbf{1}(g_T(i, x_{0:T}, y_{0:T}) = 0)\} \\
& \min[p_+(x_{0:T}, y_{0:T}), p_-(x_{0:T}, y_{0:T})])dx_{0:T}.
\end{aligned}$$

There exists a sequence of $x_{t,min,i} \in G_i$ $(t = 1 : T-1)$ such that the following lower bound

holds,

$$
\begin{aligned}
&\geq 2^{-(v+1)} \sum_{y_{0:T},w_{0:T}} \sum_{\mathbf{b}} \int_{x_{0:T}} \min[p_+(x_{0:T},y_{0:T}),p_-(x_{0:T},y_{0:T})]dx_{0:T},\\
&\geq 2^{-(v+1)} \sum_{\mathbf{b}} (\sum_{y_{0:T},w_{0:T}} \int_{x_{0:T}} [p_+(x_{0:T},y_{0:T})p_-(x_{0:T},y_{0:T})]^{1/2} dx_{0:T})^2,\\
&\geq 2^{-(v+2)} \sum_{\mathbf{b}} p_0 (\sum_{y_t,w_t} \int_{x_t} \mathrm{P}(w_t)p(x_t|x_{t-1,min,i}) \sqrt{\mathrm{P}_{b_{i-}}(y_t|x_{t-1,min,i},w_t)\mathrm{P}_{b_{i+}}(y_t|x_{t-1,min,i},w_t)} dx_t)^{2(T-1)},
\end{aligned}
$$

where the second to last line is due to the LeCam's Inequality: for any positive sequence of a_i and b_i,

$$\sum_i \min(a_i,b_i) \geq 1/2 (\sum_i \sqrt{a_i b_i})^2. \tag{125}$$

The last line follows from $p(x_0,y_0,w_0) \geq p_0$ for any $x_{t-1} \in G$, which holds as $\sqrt{\mathrm{P}_{b_{i-}}(y_t|x_{t-1})\mathrm{P}_{b_{i+}}(y_t|x_{t-1})}$ takes the value $\sqrt{1/4-r^2}$ only for $x_{t-1} \in G_i$ and has value $1/2$ otherwise.

For $r \geq 0$,

$$
\begin{aligned}
&\sum_{y_t \in \{1/2,-1/2,0\}, w_t \in \{0,1\}} \int_{x_t} \mathrm{P}(w_t)p(x_t|x_{t-1,min,i}) \sqrt{\mathrm{P}_{b_{i-}}(y_t|x_{t-1,min,i})\mathrm{P}_{b_{i+}}(y_t|x_{t-1,min,i})} dx_t\\
&\geq \min\{1/2 \Big(\int_{x_t} p(x_t|x_{t-1,min,i})dx_t \{1/2+1/2\} + 2 \int_{x_t} p(x_t|x_{t-1,min,i})dx_t \sqrt{1/4-r^2} \Big),\\
&\quad 1/2 \Big(\int_{x_t} p(x_t|x_{t-1,min,i})dx_t + \int_{x_t} p(x_t|x_{t-1,min,i})dx_t \Big) \}\\
&\geq (1/2 + 1/2\sqrt{1-4r^2}).
\end{aligned}
$$

Thus, as $\sqrt{1-4r^2} \geq -4r^2+1$, we have

$$E_i \geq p_0(1-2r^2)^{2(T-1)}.$$

So

$$W_T(G^*(\mathbf{b})) - W_T(\hat{G}) \geq (r/v)p_0(1-2r^2)^{2(T-1)}.$$

We maximize the right hand side with respect to r. Take logs to obtain, $\log(r) + \log(p_0/v) + 2(T-1)\log(1-2r^2)$. The first derivative is $r^{-1} + 2(T-1)(1-2r^2)^{-1}(-4r) = 0$. r is thus of order $\sqrt{T-1}^{-1}$.

We take $r = \sqrt{v/(2(T-1))}$, then $E_i \geq p_0(1-v/(T-1))^{2(T-1)} \geq p_0\{\exp\{-(v/(T-1))/(1-\frac{v}{T-1})\}\}^{2(T-1)}$, so $r\sum_{i=1}^v E_i \geq \sqrt{v/(2(T-1))}\exp(-1)$. □

D.9 Proof of Theorem 5.3

Proof. Under policy G, we use (restate) the following notation:

$\mathcal{W}(G)$ is defined in (35).

$\hat{\mathcal{W}}(G)$ represents the estimated welfare defined in (33).

$\hat{\mathcal{W}}^{\hat{e}}(G)$ represents the estimated welfare, with the estimated propensity score $\hat{e}(\cdot)$,

$$\widehat{\mathcal{W}}(G) = \frac{1}{T-1}\sum_{t=1}^{T-1}[\frac{Y_t W_t}{\hat{e}_t(X_{t-1})}1(X_{t-1}\in G) + \frac{Y_t(1-W_t)}{1-\hat{e}_t(X_{t-1})}1(X_{t-1}\notin G)]$$

We let $\tilde{G}$ denote an arbitrary element in $\mathcal{G}$. Let $\hat{G}^{\hat{e}}$ be the optimal policy estimated using the estimated propensity score $\hat{e}(\cdot)$,

$$\hat{G}^{\hat{e}} \stackrel{\text{def}}{=} \underset{G\in\mathcal{G}}{\text{argmax}}\ \hat{\mathcal{W}}^{\hat{e}}(G). \tag{126}$$

Recall $\tau_t = \frac{Y_t W_t}{e_t(W_{t-1})} - \frac{Y_t(1-W_t)}{1-e_t(W_{t-1})}$ and $\hat{\tau}_t = \frac{Y_t W_t}{\hat{e}_t(W_{t-1})} - \frac{Y_t(1-W_t)}{1-\hat{e}_t(W_{t-1})}$. Similar to (A.29) in the supplementary material for Kitagawa and Tetenov (2018)

$$\begin{aligned}
&\tilde{\mathcal{W}}(\tilde{G}) - \tilde{\mathcal{W}}(\hat{G}^{\hat{e}}),\\
&= \tilde{\mathcal{W}}(\tilde{G}) - \tilde{\mathcal{W}}(\hat{G}^{\hat{e}}) + \left[\hat{\mathcal{W}}^{\hat{e}}(\tilde{G}) - \hat{\mathcal{W}}^{\hat{e}}(\tilde{G})\right] + \left[\hat{\mathcal{W}}(\hat{G}^{\hat{e}}) - \hat{\mathcal{W}}(\hat{G}^{\hat{e}})\right] + \left[\hat{\mathcal{W}}(\tilde{G}) - \hat{\mathcal{W}}(\tilde{G})\right]\\
&\le \tilde{\mathcal{W}}(\tilde{G}) - \tilde{\mathcal{W}}(\hat{G}^{\hat{e}}) + \left[\hat{\mathcal{W}}^{\hat{e}}(\hat{G}^{\hat{e}}) - \hat{\mathcal{W}}^{\hat{e}}(\tilde{G})\right] + \left[\hat{\mathcal{W}}(\hat{G}^{\hat{e}}) - \hat{\mathcal{W}}(\hat{G}^{\hat{e}})\right] + \left[\hat{\mathcal{W}}(\tilde{G}) - \hat{\mathcal{W}}(\tilde{G})\right]\\
&= \left[\hat{\mathcal{W}}(\tilde{G}) - \hat{\mathcal{W}}^{\hat{e}}(\tilde{G}) - \hat{\mathcal{W}}(\hat{G}^{\hat{e}}) + \hat{\mathcal{W}}^{\hat{e}}(\hat{G}^{\hat{e}})\right] + \left[\tilde{\mathcal{W}}(\tilde{G}) - \tilde{\mathcal{W}}(\hat{G}^{\hat{e}}) - \hat{\mathcal{W}}(\tilde{G}) + \hat{\mathcal{W}}(\hat{G}^{\hat{e}})\right]\\
&= I^{\hat{e}} + II^{\hat{e}},
\end{aligned} \tag{127}$$

where the first inequality comes from $\hat{\mathcal{W}}^{\hat{e}}(\hat{G}^{\hat{e}}) \ge \hat{\mathcal{W}}^{\hat{e}}(\tilde{G})$, which is implied by the definition (126).

For $II^{\hat{e}}$, we know $II^{\hat{e}} \le 2\sup_{G\in\mathcal{G}}|\tilde{\mathcal{W}}(G) - \hat{\mathcal{W}}(G)|$. Similar arguments to Section 3.2 can then be used to bound it.

For $I^{\hat{e}}$, Note that for any $G \in \mathcal{G}$

$$\begin{aligned}
\widehat{\mathcal{W}}(G) \quad &= \frac{1}{T-1}\sum_{t=1}^{T-1}[\frac{Y_t W_t}{e_t(X_{t-1})}1(X_{t-1}\in G) + \frac{Y_t(1-W_t)}{1-e_t(X_{t-1})}1(X_{t-1}\notin G)]\\
&= \frac{1}{T-1}\sum_{t=1}^{T-1}[\tau_t 1(X_{t-1}\in G) + \frac{Y_t(1-W_t)}{1-e_t(X_{t-1})}]
\end{aligned} \tag{128}$$

Similarly,

$$\widehat{\mathcal{W}}^{\hat{e}}(G) = \frac{1}{T-1}\sum_{t=1}^{T-1}[\hat{\tau}_t 1(X_{t-1} \in G) + \frac{Y_t(1-W_t)}{1-\hat{e}_t(X_{t-1})}] \tag{129}$$

Combining (128) and (129) with $I^{\hat{e}}$,

$$\begin{aligned}
\hat{\mathcal{W}}(\tilde{G}) - \hat{\mathcal{W}}(\hat{G}^{\hat{e}}) &= \frac{1}{T-1}\sum_{t=1}^{T-1}\tau_t\left[\mathbf{1}\{X_{t-p-1} \in \tilde{G}\} - \mathbf{1}\{X_{t-p-1} \in \hat{G}^{\hat{e}}\}\right] \\
\hat{\mathcal{W}}^{\hat{e}}(\tilde{G}) - \hat{\mathcal{W}}^{\hat{e}}(\hat{G}^{\hat{e}}) &= \frac{1}{T-1}\sum_{t=1}^{T-1}\hat{\tau}_t\left[\mathbf{1}\{X_{t-p-1} \in \tilde{G}\} - \mathbf{1}\{X_{t-p-1} \in \hat{G}^{\hat{e}}\}\right].
\end{aligned}$$

Then,

$$\begin{aligned}
I^{\hat{e}} &= \hat{\mathcal{W}}(\tilde{G}) - \hat{\mathcal{W}}(\hat{G}^{\hat{e}}) - [\hat{\mathcal{W}}^{\hat{e}}(\tilde{G}) - \hat{\mathcal{W}}^{\hat{e}}(\hat{G}^{\hat{e}})], \\
&= \frac{1}{T-1}\sum_{t=1}^{T-1}\left[(\tau_t - \hat{\tau}_t)\cdot 1\{X_i \in \tilde{G}\} - (\tau_t - \hat{\tau}_t)\cdot 1\{X_i \in \hat{G}^{\hat{e}}\}\right], \\
&= \frac{1}{T-1}\sum_{t=1}^{T-1}\left[(\tau_t - \hat{\tau}_t)\left(1\{X_i \in \tilde{G}\} - 1\{X_i \in \hat{G}^{\hat{e}}\}\right)\right], \\
&\leq \frac{1}{T-1}\sum_{t=1}^{T-1}|\tau_t - \hat{\tau}_t|.
\end{aligned}$$

Finally, we have that the rate of convergence is bounded by the accuracy of propensity score estimation and the bound with known propensity scores,

$$\begin{aligned}
\mathsf{E}_{P_T}[\mathcal{W}(G_*) - \mathcal{W}(\hat{G}^{\hat{e}})] \leq \mathsf{E}_{P_T}&\left[\frac{1}{T-1}\sum_{t=1}^{T-1}|\tau_t - \hat{\tau}_t|\right] \\
&+ 2\,\mathsf{E}_{P_T}\left[\sup_{G\in\mathcal{G}}|\mathcal{W}(G) - \hat{\mathcal{W}}(G)|\right].
\end{aligned}$$

The statement of the proof follows by (70) and Theorem 3.3.

□