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 CWP27/24

Economic and Social Research Council

Policy Choice in Time Series by Empirical Welfare Maximization*

Toru Kitagawa[†] Weining Wang[‡] Mengshan Xu[§]

December 9, 2024

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 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 derive a nonasymptotic upper bound for conditional welfare regret. To illustrate the implementation and uses of T-EWM, we perform simulation studies and apply the method to estimate optimal restriction rules against Covid-19.

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

*We are grateful to Karun Adusumilli, Isaiah Andrews, Joshua Angrist, Giuseppe Cavaliere, Yoosoon Chang, Wei Cui, Whitney Newey, Joon Park, Mikkel Plagborg-Møller, Ashesh Rambachan, Barbara Rossi, Frank Schorfheide, Neil Shephard, Christian Wolf, and the participants at the EEA-ESEM 2021, IAAE Annual Meeting 2022, and seminars at CeMMAP, Glasgow, Harvard/MIT, Indiana, KU Leuven, LMU Munich, Michigan, Notre Dame, OCIS, Oxford, Philadelphia Fed, SFU, TEDS, and UC3M for beneficial comments. We are grateful to Thomas Carr for his excellent research assistance. Financial support from the ESRC through the ESRC CeMMAP (grant number RES-589-28-0001) and the ERC (grant number 715940) is gratefully acknowledged. Weining Wang's research is partially supported by the ESRC (Grant Reference: ES/T01573X/1).

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

[‡]Faculty of economics and business, University of Groningen. Email: weining.wang@rug.nl

[§]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 (White and Lu (2010), 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 structural 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, the outcomes and policies observed in the available data can be statistically and causally dependent in a complex manner, and 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 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 Manski (2004) and 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 including White and Lu (2010), 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 choices based on the historical evidence.

Since the seminal works of Manski (2004) and Dehejia (2005), statistical treatment choice and empirical welfare maximization have been active topics of research, e.g., 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, while to our knowledge, 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), Ida et al. (2024), among others. The problem of dynamic treatment regimes assumes that training data is a short panel in which treatments have been randomized both among cross-sectional units and across time periods. Recently, Adusumilli et al. (2022) consider an optimal policy in a dynamic treatment assignment problem with a budget constraint where the planner allocates treatments to subjects arriving sequentially. 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, and Dimakopoulou et al. (2017), Kock et al. (2020), Kasy and Sautmann (2021), Adusumilli (2021), and Kitagawa and Rowley (2024) for recent works in econometrics. The setting in this paper differs from the standard multi-arm bandit setting in the following three 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 the 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 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 by 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 a case of multi-period welfare functions, how T-EWM is related to the Lucas critique, as well as T-EWM's links with structural vector autoregressive models, Markov Decision Processes, and reinforcement learning. In Section 5 we present an empirical application. Technical proofs, simulation studies, and other details are presented in Appendices.

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). The planner sets $W_T \in \{0,1\}$, $T \geq 1$, making use of 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.

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.[1]

Let

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

collect the observable variables for period t. For $t = 0, 1, 2, \ldots$, let $\mathcal{X}$ denote the sample

[1]We do not allow Z_t to be realised between Y_t and W_t. If there exists some Z that is realised after W_t, before Y_t, and is not causally effected by W_t, the causal link is 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.

space of X_t. Furthermore, 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.[2] 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

[2]Section 4.1 discusses how to extend the single-period policy choice problem to multi-period settings.

welfare, for short) to be[3]

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

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

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

This welfare criterion is conditional on the planner's information set. This contrasts with the unconditional welfare criterion common in the cross-sectional treatment choice setting, where any conditioning variables (observable characteristics of a unit) are averaged out.[4] 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 Appendix B.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 (4). Specifically, we assess the statistical performance of an estimated policy rule $\hat{g}$ in terms of the convergence of conditional welfare regret,

$$\mathcal{W}_T(g^*|X_{0:T-1}=x_{0:T-1}) - \mathcal{W}_T(\hat{g}|X_{0:T-1}=x_{0:T-1}), \quad (5)$$

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

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

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

Section 3.

2.2 An 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 $w_{0:t-2} \neq w'_{0:t-2}$,

$$Y_t(w_{0:t-2}, w_{t-1}, w_t) = Y_t(w'_{0:t-2}, w_{t-1}, w_t) := Y_t(w_{t-1}, w_t) \tag{6}$$

holds with probability one.

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

$$Y_t(w_{0:t}) \perp X_{0:t-1} | W_{t-1}, \tag{7}$$

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

$$W_t \perp X_{0:t-1} | W_{t-1}. \tag{8}$$

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 (6). 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

$W_T \in \{0,1\}$, i.e.,

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

Consequently, welfare conditional on $\mathcal{F}_{T-1}$ can be simplified to

$$\begin{aligned}\mathcal{W}_T(g|W_{T-1}) &= \mathsf{E}[Y_T(W_{T-1},1)g(X_{0:T-1}) + Y_T(W_{T-1},0)(1-g(X_{0:T-1}))|X_{0:T-1}]. \\ &= \mathsf{E}[Y_T(W_{T-1},1)g(W_{T-1}) + Y_T(W_{T-1},0)(1-g(W_{T-1}))|W_{T-1}], \quad (9)\end{aligned}$$

where the first equality follows from the definition of $\mathcal{F}_{T-1}$ and Assumption 2.1(i); the second equality follows from Assumption 2.1(ii).

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.

Example 1. Suppose the planner (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 [5]

$$Y_t = \beta_0 + \beta_1 W_t + \beta_2 W_{t-1} + \varepsilon_t, \quad (10)$$

$$W_t = (1-q) + \lambda W_{t-1} + V_t, \quad \lambda = p + q - 1, \quad (11)$$

$$\varepsilon_t \perp (W_t, X_{0:t-1}) \qquad \forall 1 \le t \le T-1 \qquad \text{and} \qquad \varepsilon_T \perp X_{0:T-1}. \quad (12)$$

$$\text{If } W_{t-1} = 1, \quad \begin{cases} V_t = 1-p & \text{with probability } p \\ V_t = -p & \text{with probability } 1-p, \end{cases} \quad (13)$$

$$\text{if } W_{t-1} = 0, \quad \begin{cases} V_t = -(1-q) & \text{with probability } q \\ V_t = q & \text{with probability } 1-q. \end{cases} \quad (14)$$

Compatibility with Assumption 2.1 can be seen as follows. Assumption 2.1(i) is implied by (10), 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 (10), (12), and the fact that the distribution of V_t depends solely on W_{t-1}, i.e., under (11),(13), and (14), 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 gener-

[5]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.

ating process. First, we impose a strict overlap condition on the propensity score.

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 (Sequential 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. It is worth noting that the above assumption together with Assumption 2.1(ii) implies $Y_t(W_{t-1}, w) \perp W_t | W_{t-1}$.

Under Assumption 2.1, we have that, for any measurable function f of the potential outcome $Y_t(W_{0:t-1})$ and treatment W_t, it holds

$$\mathsf{E}(f(Y_t(W_{0:t}), W_t)|\mathcal{F}_{t-1}) = \mathsf{E}(f(Y_t(W_{t-1}, W_t), W_t)|W_{t-1}) = \mathsf{E}(f(Y_t, W_t)|W_{t-1}). \tag{15}$$

Example 1 continued. Assumption 2.2 is satisfied if $0 < p < 1$ and $0 < q < 1$; Assumption 2.3 is implied by (10) and (12).

Imposing Assumption 2.2 and 2.3 and assuming propensity scores are known, we consider constructing a sample analogue of (9) conditional on $W_{T-1} = w$ based on the historical average of the inverse propensity score weighted outcomes,

$$\widehat{\mathcal{W}}(g|W_{T-1} = w) = \frac{1}{T(w)} \sum_{1 \leq t \leq T-1 : W_{t-1} = w} \left[\frac{Y_t W_t g(W_{t-1})}{e_t(W_{t-1})} + \frac{Y_t(1 - W_t)\{1 - g(W_{t-1})\}}{1 - e_t(W_{t-1})} \right], \tag{16}$$

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 planner'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 $(\cdot|w)$, we define the true optimal policy and its empirical analogue to be,

$$g^*(w) \in \text{argmax}_{g:\{w\}\to\{0,1\}} \mathcal{W}_T(g|w), \quad (17)$$
$$\hat{g}(w) \in \text{argmax}_{g:\{w\}\to\{0,1\}} \widehat{\mathcal{W}}(g|w), \quad (18)$$

where $\hat{g}$ 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 that $\widehat{\mathcal{W}}(g|w)$ 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 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,

$$\bar{\mathcal{W}}(g|w) = T(w)^{-1} \sum_{1\le t\le T-1: W_{t-1}=w} \mathsf{E}\left[Y_t(W_{t-1},1)g(W_{t-1}) + Y_t(W_{t-1},0)\left[1-g(W_{t-1})\right]|W_{t-1}\right]. \quad (19)$$

Under the strict overlap and unconfoundedness assumptions (i.e. Assumptions 2.2 and 2.3), the difference between empirical welfare and (19) is a sum of MDS. A concentration inequality for average MDS then implies that the empirical welfare concentrates around $\bar{\mathcal{W}}(g|w)$. Since $\bar{\mathcal{W}}(g|w)$ is not guaranteed to inform an optimal policy in terms of $\mathcal{W}_T(g|w)$, we impose the assumption:

Assumption 2.4 (Invariance of the welfare ordering)**.** Given $w \in \{0,1\}$, let $g^* = g^*(w)$ defined in (17). There exist a positive constant $c > 0$, such that for any $g \in \{0,1\}$,

$$\mathcal{W}_T(g^*|w) - \mathcal{W}_T(g|w) \le c\big[\bar{\mathcal{W}}(g^*|w) - \bar{\mathcal{W}}(g|w)\big], \quad (20)$$

with probability one, i.e., P_T (inequality (20) holds) $= 1$, where P_T is the probability distribution for $X_{0:T-1}$.

Noting that the left-hand side of (20) 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, the optimality of g^* in terms of the conditional value of welfare at T is maintained in the historical average of the conditional values of welfare. 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 2.1. 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, Assumption 2.4 is implied by the following condition.

A2.4' *The stochastic process*

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

is weakly stationary.

Under A2.4', $\mathsf{E}\left[Y_t(W_{t-1}, 1)g(W_{t-1}) + Y_t(W_{t-1}, 0)\left[1 - g(W_{t-1})\right]|W_{t-1} = w\right]$ is invariant for $2 \leq t \leq T$. Then, Assumption 2.4 will hold naturally.

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.
ε_t remains an i.i.d. noise in the following settings.
(i) By Remark 2.1, 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.

(ii) If we replace (10) 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.,

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

(iii) If we replace (10) 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 (10) by

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

to allow heterogeneous treatment effect. Then

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

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 (17), Assumption 2.4 holds if $\beta_{1,T}$ and $\bar{\beta}_w$ have the same sign and $\frac{|\beta_{1,T}|}{|\bar{\beta}_w|} \le c.$.

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 numbers 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 $\{\mathcal{F}_{t-1} : t = 1, 2, \ldots, T\}$.[6] Specifically,

$$\begin{aligned} &\mathcal{W}_T(g^*|w) - \mathcal{W}_T(\hat{g}|w) \le c\left[\bar{\mathcal{W}}(g^*|w) - \bar{\mathcal{W}}(\hat{g}|w)\right] \\ =&c\left[\bar{\mathcal{W}}(g^*|w) - \widehat{\mathcal{W}}(\hat{g}|w) + \widehat{\mathcal{W}}(\hat{g}|w) - \bar{\mathcal{W}}(\hat{g}|w)\right] \le c\left[\bar{\mathcal{W}}(g^*|w) - \widehat{\mathcal{W}}(g^*|w) + \widehat{\mathcal{W}}(\hat{g}|w) - \bar{\mathcal{W}}(\hat{g}|w)\right] \\ \le& 2c \sup_{g:\{w\}\to\{0,1\}} |\bar{\mathcal{W}}(g|w) - \widehat{\mathcal{W}}(g|w)|, \end{aligned} \tag{21}$$

where the first inequality follows from Assumption 2.4. The second inequality follows from the definition of T-EWM rule $\hat{g}$ in (18).

To bound the right-hand side of (21), define

$$\begin{aligned} \widehat{\mathcal{W}}_t(g|w) =& 1(W_{t-1} = w)\left[\frac{Y_t W_t g(W_{t-1})}{e_t(W_{t-1})} + \frac{Y_t(1 - W_t)\{1 - g(W_{t-1})\}}{1 - e_t(W_{t-1})}\right], \\ \bar{\mathcal{W}}_t(g|w) =& 1(W_{t-1} = w)\, \mathsf{E}\{Y_t(W_{t-1}, 1) g(W_{t-1}) + Y_t(W_{t-1}, 0)(1 - g(W_{t-1}))|W_{t-1} = w\}. \end{aligned}$$

Then, we can express (16) and (19) as

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

[6] Note that by (15), $\mathsf{E}(\cdot|\mathcal{F}_{t-1}) = \mathsf{E}(\cdot|W_{t-1})$.

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. Next, we impose

Assumption 2.5 (Bounded outcomes)**.** There exists $M < \infty$ such that the support of outcome variable Y_t is contained in $[-M/2, M/2]$.

Proposition 2.1. Under Assumptions 2.1-2.3 and 2.5, the sequence $\{\widehat{\mathcal{W}}_t(g|w) - \bar{\mathcal{W}}_t(g|w)\}_{t=0}^{T-1}$ is an MDS.

The proof of this proposition can be found in Appendix A.1. With Proposition 2.1, 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.

THEOREM 2.1. Under Assumptions 2.1 to 2.5, it holds that

$$\sup_{g:\{0,1\}\to\{0,1\}} \mathsf{E}\,|\widehat{\mathcal{W}}(g|w) - \bar{\mathcal{W}}(g|w)| \le \frac{C}{\sqrt{T-1}}, \tag{22}$$

where C is a constant defined in Appendix A.2.

The proof of Theorem 2.1 can be found in Appendix A.2. Combining (21) (an implication of Assumption 2.4) and (22), we can conclude that the convergence rate of expected regret $\mathsf{E}[\mathcal{W}_T(g^*|w) - \mathcal{W}_T(\hat{g}|w)]$ is upper-bounded by $2c \cdot \frac{C}{\sqrt{T-1}}$ uniformly in $w \in \{0,1\}$.

Remark 2.2. [Higher Markov orders] Theorem 2.1 presents our main results of welfare regret upper bounds under the simple first-order Markovian structure outlined in Assumption 2.1. We can extend our analysis to a higher or infinite order Markovian structure as summarized below. We defer the detailed discussion and formal proofs to Appendix A.3.

If current observations can depend causally or statistically on the realized treatment over the preceding q periods (for some $1 < q < T$), we can define $\mathcal{W}_T(g|w_{T-q:T-1})$, $\widehat{\mathcal{W}}(\hat{g}|w_{T-q:T-1})$, $\bar{\mathcal{W}}(g|w_{T-q:T-1})$, g^*, and $\hat{g}$, similar to the definitions in (9), (16), (19), (17), and (18), respectively. Let $w_{T-q:T-1} \in \{0,1\}^q$ be a realization of the treatment path spanning from time $T-q$ to $T-1$. Given the modified Assumptions detailed in Appendix A.3.1, we can apply similar reasoning as in Theorem 2.1 to establish a convergence rate of $\frac{1}{\sqrt{T-q}}$.

Remark 2.3. [Comparison with Bojinov and Shephard (2019)] The major distinction between our work and Bojinov and Shephard (2019) is that we focus on policy decisions and future welfare, while Bojinov and Shephard (2019) study estimation and inference on the retrospective causal effects. An estimand of interest in Bojinov and Shephard (2019) is the temporal (zero-lag) average treatment effect (ATE) defined as

$$\bar{\tau}_0 := \frac{1}{T-1} \sum_{t=1}^{T-1} \left[Y_t \left(W_{0:t-1}, 1 \right) - Y_t \left(W_{0:t-1}, 0 \right) \right]. \tag{23}$$

In contrast, our paper focuses on maximizing

$$\mathcal{W}_T(G|\mathcal{F}_{T-1}) = \mathsf{E}\left[\tau_T(\mathcal{F}_{T-1}) 1\{X_{0:T-1} \in G\} + Y_T(W_{0:T-1}, 0) | \mathcal{F}_{T-1} \right] \tag{24}$$

where $\tau_T(\mathcal{F}_{T-1})$ is the conditional ATE (CATE) at time T,

$$\tau_T(\mathcal{F}_{T-1}) = \mathsf{E}\left[Y_T(W_{0:T-1}, 1) - Y_T(W_{0:T-1}, 0) | \mathcal{F}_{T-1} \right]. \tag{25}$$

$\tau_T(\mathcal{F}_{T-1})$ is the conditional ATE at an upcoming time period of T, while $\bar{\tau}_0$ sums up the causal effects of the past realized time periods from 0 to $T-1$. This distinction necessitates different sets of assumptions between our work and Bojinov and Shephard (2019). Specifically, the sequential unconfoundedness assumption (Assumption 2.3) is sufficient for unbiased estimation for $\bar{\tau}_0$, whereas it falls short for $\tau_T(\mathcal{F}_{T-1})$. Consequently, Assumptions 2.1 (Markov properties) and 2.4 (Invariance of welfare ordering) are not imposed in Bojinov and Shephard (2019).

Analogous to the construction of our empirical welfare criterion, we can consider the following estimator for $\tau_T(\mathcal{F}_{T-1})$:

$$\hat{\tau}_T(w) = T(w)^{-1} \sum_{1 \le t \le T-1 : W_{t-1} = w} \left[\frac{Y_t W_t}{e_t(W_{t-1})} - \frac{Y_t(1 - W_t)}{1 - e_t(W_{t-1})} \right]. \tag{26}$$

To validate $\hat{\tau}_T(w)$ as an estimator for $\tau_T(\mathcal{F}_{T-1})$, the crucial step is to link the past CATEs (or the welfare in our context) from past periods to the future ones. This motivates our introduction of Assumption 2.4, the invariance of welfare ordering.

The construction of $\hat{\tau}_T(w)$ is model-free and selects out a subset of past periods that share the same conditioning with time T. To guarantee the availability of observations sharing the conditioning states, we impose Assumption 2.1, limiting the persistence of carryover effects of treatment.

Remark 2.4. We construct the empirical welfare $\widehat{\mathcal{W}}(g|W_{T-1} = w)$ in (16) by the average with uniform weights. To generalize, we can specify $\widehat{\mathcal{W}}(g|W_{T-1} = w)$ as a weighted average:

$$\widehat{\mathcal{W}}(g|W_{T-1} = w) = \sum_{1 \le t \le T-1 : W_{t-1} = w} a_t \left[\frac{Y_t W_t g(W_{t-1})}{e_t(W_{t-1})} + \frac{Y_t(1 - W_t)\{1 - g(W_{t-1})\}}{1 - e_t(W_{t-1})} \right], \tag{27}$$

where $a_t \geq 0$ is a prespecified weight assigned to period $1 \leq t \leq T-1$. Modifying intermediate welfare $\bar{\mathcal{W}}(g|w)$ accordingly by

$$\bar{\mathcal{W}}(g|w) = \sum_{1 \leq t \leq T-1 : W_{t-1} = w} a_t \, \mathsf{E}\left[Y_t(W_{t-1}, 1) g(W_{t-1}) + Y_t(W_{t-1}, 0)\left[1 - g(W_{t-1})\right] | W_{t-1}\right] \quad (28)$$

and imposing Assumption 2.4 with the modified intermediate welfare, we can study a condition for the weights that generalizes the welfare regret convergence of Theorem 2.1. Specification of nonuniform weights can reflect the planner's belief or knowledge on how the period T potential outcome distribution differs from those of the past periods or which past period observations are more informative for the current period decision making. See, e.g., Ishihara and Kitagawa (2024) for an optimal weighting of pieces of evidence for policy choice when the population that the policy is implemented differs from the populations that pieces of evidence were collected. We leave a formal investigation for the current time-series setting for future research.

2.4 Infinite Markov order

Time-series models commonly used in macroeconomic policy analysis imply an infinite order Markovian structure. For example, we consider the following modification to Example 1.

Example 2. Replace (10) in Example 1 with a structural MA(∞) model:

$$Y_t = \alpha + \sum_{i=0}^{\infty} \beta_i W_{t-i} + \sum_{i=0}^{\infty} \gamma_i \varepsilon_{t-i}, \quad (29)$$

while keeping the conditions (11) to (14) unchanged. This type of MA(∞) process underlies the causal impulse response analysis of structural vector autoregressions; see, e.g., Kilian and Lütkepohl (2017). For the effect of shocks to diminish, both γ_i and β_i must decay in absolute value with $i \to \infty$. In particular, if Y_t is an AR(1) process, then β_i and γ_i are polynomials of the AR coefficient.

Without requiring stationarity or functional form restrictions, we can extend our framework to infinite Markovian order and obtain convergence of welfare regret conditional on a treatment path of infinite length, $\mathcal{W}_T(g^*|w_{-\infty:T-1}) - \mathcal{W}_T(\hat{g}|w_{-\infty:T-1})$. Let $1 < m < T$ and $\hat{g}$ be a policy that maximizes the empirical welfare $\widehat{\mathcal{W}}(g|w_{T-m:T-1})$ with conditioning policy path truncated to $w_{T-m:T-1}$. Appendix A.3.2 shows the following upper bound for the welfare regret of infinite order Markov models:

$$\mathcal{W}_T(g^*|w_{-\infty:T-1}) - \mathcal{W}_T(\hat{g}|w_{-\infty:T-1}) \leq 2c \sup_{g:\{w_{T-m:T-1}\}\to\{0,1\}} |\bar{\mathcal{W}}(g|\mathcal{F}_{t-1}) - \widehat{\mathcal{W}}(g|w_{T-m:T-1})| + 2 \cdot \widetilde{\text{w-bias}}_{\infty}(m),$$

where $\bar{\mathcal{W}}(g|\mathcal{F}_{t-1})$ is the intermediate welfare defined by (A.19) in Appendix A.3.2, around which the empirical welfare is expected to concentrate, and $\widetilde{\text{w-bias}}_{\infty}(m)$ is the welfare bias due to truncation of the empirical welfare defined by (A.21) in Appendix A.3.2. The first term on the right-hand side represents the average of an MDS. Under the regularity conditions presented in Appendix A.3.2, we can show this term to converge at a rate of $\frac{1}{\sqrt{T-m}}$. For the second term on the right-hand side, we shall have $\text{plim}_{m\to\infty}\widetilde{\text{w-bias}}_{\infty}(m) = 0$ under additional conditions that ensure the decay temporal dependence. See Appendix A.3.2 for more details. Furthermore, in Appendix A.3.3, we show that the infinite order T-EWM proposed in Appendix A.3.2 can be applied to the policy choice problem specified in Example 2.

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 incorporate general covariates Z_t into $X_t \in \mathcal{X}$, which can be continuous. Now, $X_t = (Y_t, W_t, Z_t)$. 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*: the same as Assumption 2.1 (i).

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

$$Y_t(w_{0:t}) \perp X_{0:t-1}|X_{t-1}, \tag{30}$$

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

$$W_t \perp X_{0:t-1} | X_{t-1}. \tag{31}$$

Similarly to (9) in the illustrative example, Assumption 3.1 implies that we can reduce the conditioning information of $\mathcal{F}_{t-1}$ to only X_{t-1} and reduce the policy to a binary map of X_{t-1} without any loss of conditional welfare, i.e., we can partition the space of X_{t-1} into G and its complement. Following these reductions and considering the planner's focus on the policy choice in period T, we can formulate the planner's objective function as follows:

$$\mathcal{W}_T(G|X_{T-1}) = \mathsf{E}\left[Y_T(W_{T-1}, 1)1(X_{T-1} \in G) + Y_t(W_{T-1}, 0)1(X_{T-1} \notin G)|X_{T-1}\right]. \tag{32}$$

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 $t = 1, 2, \ldots, T-1$ and $w \in \{0, 1\}$

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

Under Assumptions 3.1 and 3.3, we can generalize (15) by including the set of covariates in the conditioning variables: for any measurable function f,

$$\mathsf{E}(f(Y_t(W_{0:t}), W_t)|\mathcal{F}_{t-1}) = \mathsf{E}(f(Y_t(W_{t-1}, W_t), W_t)|X_{t-1}) = \mathsf{E}(f(Y_t, W_t)|X_{t-1}). \tag{33}$$

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 $K(\cdot)$ and a bandwidth h

$$\widehat{\mathcal{W}}(G|x) = \frac{\sum_{t=1}^{T-1} K(\frac{X_{t-1}-x}{h})\left[\frac{Y_t W_t}{e_t(X_{t-1})} 1(X_{t-1} \in G) + \frac{Y_t(1-W_t)}{1-e_t(X_{t-1})} 1(X_{t-1} \notin G)\right]}{\sum_{t=1}^{T-1} K(\frac{X_{t-1}-x}{h})}, \tag{34}$$

where $(\cdot|x)$ denotes $(\cdot|X_{T-1} = x)$. Theorem B.1 in Appendix B.1 provides a regret bound for (34). 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 the precise characterization of the statistical properties of the regret bound of the kernel approach to Appendix B.1. This slow convergence rate may limit its practicality when the covariates X_{t-1} are multi-dimensional or when we extend the order of Markovian dependence to multi-periods. Therefore, in what follows, we instead pursue a novel approach that estimates an optimal policy rule by maximizing an empirical analogue of *unconditional* welfare over a specified class of decision sets $\mathcal{G}$. We show that under additional assumptions, this approach can lead to the convergence rate of the conditional welfare that is free from the curse of dimensionality.

3.2 Bounding the conditional regret: continuous covariate case

We first clarify how a maximizer of conditional welfare (32) 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. In our setup, the complexity of the functional class needs to be specified by the user.

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 the next assumption.

Assumption 3.4 (Correct specification)**.** Let $\mathcal{W}_T(G|x)$ be the conditional welfare as defined in (32), and

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

be the unconditional welfare under policy G. We have, at every $x \in \mathcal{X}$,

$$\text{argmax}_{G\in\mathcal{G}}\mathcal{W}_T(G) \subset \text{argmax}_{G\in\mathcal{G}}\mathcal{W}_T(G|x).$$

This assumption ensures that maximizing unconditional welfare corresponds to maximizing conditional welfare over $\mathcal{G}$. A sufficient condition for the equivalence of maximizing conditional and unconditional welfare is that the specified class of policy rules, $\mathcal{G}$, includes the first best policy G^*_{FB} for the unconditional problem, where

$$G^*_{FB} := \{x \in \mathcal{X} : \mathsf{E}[Y_T(W_{T-1},1) - Y_T(W_{T-1},0)|X_{T-1} = x] \geq 0\}. \tag{35}$$

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'.[7] The planner can be confident about Assumption 3.4 if, for instance, $\mathsf{E}[Y_T(W_{T-1}, 1) - Y_T(W_{T-1}, 0)|X_{T-1} = x]$ is believed to be monotonic in x (element-wise) and the class $\mathcal{G}$ consists of decision sets with monotonic boundaries (Mbakop and Tabord-Meehan, 2021; Kitagawa et al., 2021).

With Assumption 3.4, we can shift the focus to maximizing unconditional welfare, even when the planner's ultimate objective function is conditional welfare. We also note several side-benefits of considering optimal policy in terms of unconditional welfare. First, an optimal policy in terms of the unconditional welfare can inform an optimal policy action that the planner would choose under a counterfactual history. Such analysis exercise could be of interest in counterfactual policy analysis. Second, the unconditionally optimal policy learned from the time-series data of a particular unit can inform optimal policies for other similar units but with different histories. Third, it can also inform an optimal policy to the setting where the planner implements the same Markovian policy over multiple time periods, assuming the data generating process is stationary.

The following proposition directly results from Assumptions 3.1 (i), Assumption 3.4, and the definition of G^*_{FB}.

Proposition 3.1. Under Assumptions 3.1 (i) and 3.4, an optimal policy rule $G_* \in \mathcal{G}$ in terms of the unconditional welfare

$$G_* \in \text{argmax}_{G \in \mathcal{G}} \mathcal{W}_T(G) \tag{36}$$

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 policy 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 bound the conditional function. For $G \in \mathcal{G}$ and $X_{T-1} = x$, define conditional regret as

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

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

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

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

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

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

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

Assumption 3.5. (Lower bound of conditional density) For $x^{obs} \in \mathcal{X}$ and any $G \in \mathcal{G}$, there exists a positive constant $\underline{p}$ such that

$$p_{T-1}(x^{obs}, G) \geq \underline{p} > 0. \tag{37}$$

Remark 3.1. This assumption is satisfied if X_{T-1} is a discrete random variable taking a finite number of different values. In this case, $p_{T-1}(x^{obs}, G) \geq \min_{x \in \mathcal{X}} \Pr\left(X_{T-1} = x\right) > 0$, so we can set $\underline{p} = \min_{x \in \mathcal{X}} \Pr\left(X_{T-1} = x\right)$. 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,

$$R_T(G|x^{obs}) \leq \frac{1}{\underline{p}} R_T(G). \tag{38}$$

The proof of this lemma can be found in Appendix A.4. Using 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

$$\widehat{\mathcal{W}}(G) = \frac{1}{T-1} \sum_{t=1}^{T-1} \left[\frac{Y_t W_t}{e_t(X_{t-1})} 1(X_{t-1} \in G) + \frac{Y_t(1-W_t)}{1-e_t(X_{t-1})} 1(X_{t-1} \notin G) \right], \tag{39}$$

and we define

$$\hat{G} \in \mathrm{argmax}_{G\in\mathcal{G}}\widehat{\mathcal{W}}(G). \tag{40}$$

In addition, define two intermediate welfare functions,

$$\begin{aligned}
\bar{\mathcal{W}}(G) &= \frac{1}{T-1}\sum_{t=1}^{T-1}\mathsf{E}\left[Y_t(W_{t-1},1)1(X_{t-1}\in G)+Y_t(W_{t-1},0)1(X_{t-1}\notin G)|\mathcal{F}_{t-1}\right], \\
\widetilde{\mathcal{W}}(G) &= \frac{1}{T-1}\sum_{t=1}^{T-1}\mathsf{E}\left[Y_t(W_{t-1},1)1(X_{t-1}\in G)+Y_t(W_{t-1},0)1(X_{t-1}\notin G)\right]. \qquad (41)
\end{aligned}$$

The need for an additional intermediate welfare function, $\widetilde{\mathcal{W}}(G)$, arises because we use unconditional welfare to bound the conditional welfare. After centering the empirical welfare around its conditional mean, $\bar{\mathcal{W}}(G)$, we are left with another difference, $\bar{\mathcal{W}}(G)-\widetilde{\mathcal{W}}(G)$, which we will control below.

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

Assumption 3.6 (Welfare ordering)**.** For any $G\in\mathcal{G}$, there exists some constant c

$$\mathcal{W}_T(G_*)-\mathcal{W}_T(G)\le c\big[\widetilde{\mathcal{W}}(G_*)-\widetilde{\mathcal{W}}(G)\big], \tag{42}$$

with probability one, i.e., P_T (inequality (42) holds) $=1$, 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).

$$\begin{aligned}
\mathcal{W}_T(G_*|x^{obs})-\mathcal{W}_T(\hat{G}|x^{obs}) &\le \frac{1}{\underline{p}}\left[\mathcal{W}_T(G_*)-\mathcal{W}_T(\hat{G})\right] \\
&\le \frac{c}{\underline{p}}\left[\widetilde{\mathcal{W}}(G_*)-\widetilde{\mathcal{W}}(\hat{G})\right] \le \frac{2c}{\underline{p}}\sup_{G\in\mathcal{G}}\left|\widehat{\mathcal{W}}(G)-\widetilde{\mathcal{W}}(G)\right|. \qquad (43)
\end{aligned}$$

The first inequality follows from Lemma 3.1 and Assumption 3.5. The second inequality follows from (42). The last inequality follows from an argument similar to the one below (21).

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

$$\widehat{\mathcal{W}}(G)-\widetilde{\mathcal{W}}(G) \;=\; \bar{\mathcal{W}}(G)-\widetilde{\mathcal{W}}(G)+(\widehat{\mathcal{W}}(G)-\bar{\mathcal{W}}(G)) = I+II, \tag{44}$$

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.

(44) 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 bound for the tail probability of the sum of martingale difference sequences. 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 sample 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.

Similar to our approach with the simple model described in Section 2, we assume that Y_t, which is reintroduced in Section 3.1, has a bounded support.

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

The bound of the conditional regret will be further established in the following two subsections.

3.2.1 Bounding II

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:

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

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

$$\mathcal{H}_t = \{h_t(\cdot; G) = \widehat{\mathcal{W}}_t(G) - \bar{\mathcal{W}}_t(G) : G \in \mathcal{G}\}, \tag{45}$$

where the arguments of the function $h_t(\cdot; G)$ are Y_t, W_t, and X_{t-1}. In the following, we use n to represent the number of summands since the endpoints of samples may vary across different settings in this section, subsequent sections, and appendices. For example, in the case of multi-period welfare functions, the endpoint of a sample is no longer fixed at $T-1$. Given the class of functions $\mathcal{H}_t$, we consider a martingale difference array $\{h_t(Y_t, W_t, X_{t-1}; G)\}_{t=1}^n$, and denote its average by

$$\mathsf{E}_n h \stackrel{\text{def}}{=} \frac{1}{n}\sum_{t=1}^{n} h_t(Y_t, W_t, X_{t-1}, G),$$

where $h \stackrel{\text{def}}{=} \{h_1(\cdot; G), h_2(\cdot; , G), \ldots, h_n(\cdot; G)\}$, and we suppress n and G if there is no confusion in the context.

Since we do not restrict X_t to be stationary, we shall handle a vector of function classes that possibly vary over t. To this end, we define the following set of notations. Let H_t denote the envelope for the function class $\mathcal{H}_t$, and $\overline{H}_n = (H_1, H_2, \cdots, H_n)'$, and $\mathbf{H_n} = \mathcal{H}_1 \times \mathcal{H}_2 \times \cdots \times \mathcal{H}_n$. For a function f supported on $\mathcal{X}$, define $\|f\|_{Q,r} \stackrel{\text{def}}{=} (\int_{x\in\mathcal{X}} |f(x)|^r dQ(x))^{1/r}$, and for an n-dimensional vector $v = \{v_1, \ldots, v_n\}$, its l_2 norm is denoted by $|v|_2 \stackrel{\text{def}}{=} (\sum_{i=1}^n v_i^2)^{1/2}$. The covering number of a function class $\mathcal{H}$ w.r.t. a metric ρ is denoted by $\mathcal{N}(\varepsilon, \mathcal{H}, \rho(.))$. For two series of functions $f = \{f_t\}_{i=1}^n$ and $g = \{g_t\}_{i=1}^n$, define the metrics $\rho_{2,n}(f, g) = (n^{-1}\sum_t |f_t - g_t|^2)^{1/2}$ and $\sigma_n(f, g) = (n^{-1}\sum_t \mathsf{E}[(f_t - g_t)^2|\mathcal{F}_{t-1}])^{1/2}$. Let α_n denote an n-dimensional vector in $\mathbb{R}^n$ and $\circ$ denote the element-wise product. In the next assumption, we want to bound the covering number of $\mathcal{N}(\delta|\overline{H}_n|_2, \mathbf{H_n}, \rho_{2,n})$ by the covering number of all its one-dimensional projection. Now, let $A \lesssim B$ denote that there exists some constant c_0 such that $A \le c_0 \cdot B$.

Assumption 3.8 (Function classes)**.** Let $n = T - 1$. For any discrete measures Q, any $\alpha_n \in \mathbb{R}^n_+$, and all $\delta > 0$, we have

$$\mathcal{N}(\delta|\tilde{\alpha}_n \circ \overline{H}_n|_2, \tilde{\alpha}_{\mathbf{n}} \circ \mathbf{H_n}, |.|_2) \le \max_t \sup_Q \mathcal{N}(\delta\|H_t\|_{Q,2}, \mathcal{H}_n, \|.\|_{Q,2}) \lesssim K(v+1)(4e)^{v+1}(\frac{2}{\delta})^{crv}, \tag{46}$$

where K, v, c, and e are positive constants; r is a positive integer and $\tilde{\alpha}_{n,t} = \frac{\sqrt{\alpha_{n,t}}}{\sqrt{\sum_t \alpha_{n,t}}}$.

Assumption 3.8 restricts the complexity of the function class to be of polynomial discrimination, and the complexity index v appears in the derived regret bounds. See Appendix A.7 for a justification for this assumption.

Assumption 3.9 (Empirical sum)**.** There exists a constant $L > 0$ such that $\Pr(\sigma_n(f, g)/\rho_{2,n}(f, g) > L) \to 0$ 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 equivalent to the sample average of unconditional expectations. 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 of Hall and Heyde (2014).

Now, let $A \lesssim_p B$ denote $A = O_p(B)$. Then, we have for II:

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

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

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

The proof of Theorem 3.1 is presented in Appendix A.5.

3.2.2 Bounding I

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

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

and

$$\overline{S}_t(G) = \mathsf{E}(S_t(G)|\mathcal{F}_{t-1}) - \mathsf{E}(S_t(G)|\mathcal{F}_{t-2}), \tag{47}$$

$$\tilde{S}_t(G) = \mathsf{E}(S_t(G)|\mathcal{F}_{t-2}) - \mathsf{E}(S_t(G)). \tag{48}$$

We can apply a similar technique in Theorem 3.1 (See Lemma A.3 in Appendix A.5) to bound the sum of $\overline{S}_t(G)$. The second term $\tilde{S}_t(G)$ is handled below. Recalling (41) and (44), 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 function classes,

$$\begin{aligned} \overline{\mathcal{S}}_t &= \{f_t = \mathsf{E}(S_t(G)|\mathcal{F}_{t-1}) - \mathsf{E}(S_t(G)|\mathcal{F}_{t-2}) : G \in \mathcal{G}\}, \\ \tilde{\mathcal{S}}_t &= \{f_t = \mathsf{E}(S_t(G)|\mathcal{F}_{t-2}) - \mathsf{E}(S_t(G)) : G \in \mathcal{G}\}. \end{aligned}$$

Note that by Assumption 3.1, we have $\mathsf{E}(S_t(G)|\mathcal{F}_{t-1}) = \mathsf{E}(S_t(G)|X_{t-1})$, and $\mathsf{E}(S_t(G)|\mathcal{F}_{t-2}) = \mathsf{E}(S_t(G)|X_{t-2})$.

DEFINITION 3.1. Let $\{\varepsilon_t\}_{t=-\infty}^{\infty}$ be a sequence of i.i.d. random variables, and $\{g_t\}_{t=-\infty}^{\infty}$ is a sequence of measurable functions of ε's, which might vary with time t. For a process $\xi_. \stackrel{\text{def}}{=} \{\xi_t\}_{t=-\infty}^{\infty}$ with $\xi_t \stackrel{\text{def}}{=} g_t(\varepsilon_t, \varepsilon_{t-1}, \cdots)$ and integers $l, q \geq 0$, we define the dependence adjusted norm for an arbitrary process ξ_t as

$$\theta_{\xi,q} = \sum_{l=0}^{\infty} \max_t \|\xi_t - \xi_{t,l}^*\|_q, \tag{49}$$

where $\|\cdot\|_q$ denotes $(\mathsf{E}|\cdot|^q)^{1/q}$, and $\xi_{t,\ell}^* = g_t(\varepsilon_t, \cdots, \varepsilon'_{t-l}, \cdots)$ is the random variable ξ_t with its l-th lag replaced by ε'_{t-l}, an independent copy of ε_{t-l}. The subexponential/Gaussian dependence adjusted norm is given by:

$$\Phi_{\phi_{\tilde{v}}}(\xi_{.}) = \sup_{q\geq 2}(\theta_{\xi,q}/q^{\tilde{v}}), \tag{50}$$

where $\tilde{v} = 1/2$ (resp. 1) corresponds to the case that the process ξ_i is sub-Gaussian (resp. sub-exponential).

Assumption 3.10 (Data generating process)**.** $X_t = g_t(\varepsilon_t, \varepsilon_{t-1}, \cdots)$, where ε_t is a sequence of i.i.d. random variables, and g_t is measurable functions of ε's.

It shall be noted that Assumption 3.10 implies that $\tilde{S}_t(G) = \tilde{g}_t(\varepsilon_t, \varepsilon_{t-1}, \cdots)$, where $\tilde{g}$ is another measurable function of ε's.

Assumption 3.11. (i) (Markov exogeneity for Z) $Z_t \perp X_{0:t-1}|X_{t-1}$. (ii) (Envelope functions) $\tilde{S}_t(G)$ has an envelope $\tilde{F}_t(\cdot)$, i.e., $\sup_{G\in\mathcal{G}} |\mathsf{E}(S_t(G)|X_{t-2} = x) - \mathsf{E}(S_t(G))| \leq |\tilde{F}_t(x)|$ for every t and $x \in \mathcal{X}$; for every t and any discrete measures Q, it holds that $\max_t \sup_Q \|\tilde{F}_t\|_{Q,2} \leq \infty$.

The next assumption pertains to the tail of the underlying innovations of time series.

Assumption 3.12 (Tail assumption)**.** For $\tilde{v} = 1/2$ or 1, it holds that $\sup_{G\in\mathcal{G}} \Phi_{\phi_{\tilde{v}}}(\tilde{S}_{.}(G)) < \infty$.

Assumption 3.13 (Further function classes)**.** Suppose that F_t (resp. $\tilde{F}_t$) is the envelope of the function class $\overline{\mathcal{S}_t}$ (resp. $\tilde{\mathcal{S}}_t$). Define $\overline{F}_n = (F_1, F_2, \cdots, F_n)$ (resp. $\overline{\tilde{F}}_n = (\tilde{F}_1, \tilde{F}_2, \cdots, \tilde{F}_n)$), and $\mathbf{F}_n = \{\overline{\mathcal{S}}_1, \overline{\mathcal{S}}_2, \cdots, \overline{\mathcal{S}}_n\}$ (resp. $\tilde{\mathbf{F}}_n = \{\tilde{\mathcal{S}}_1, \tilde{\mathcal{S}}_2, \cdots, \tilde{\mathcal{S}}_n\}$). Let Q denote a discrete measure over a finite number of n points. For $n = T - 1$ and all $\delta > 0$, there exists positive constants v, V, and c, such that,

$$\begin{aligned}
\mathcal{N}(\delta|\tilde{\alpha}_n \circ \overline{F}_n|_2, \tilde{\alpha}_{\mathbf{n}} \circ \mathbf{F}_{\mathbf{n}}, |.|_{\mathbf{2}}) &\leq \max_t \sup_Q \mathcal{N}(\delta\|\tilde{F}_t\|_{Q,2}, \overline{\mathcal{S}}_t, \|.\|_{Q,2}) \lesssim (1/\delta)^{cV}, \\
\mathcal{N}(\delta|\tilde{\alpha}_n \circ \overline{\tilde{F}}_n|_2, \tilde{\alpha}_{\mathbf{n}} \circ \tilde{\mathbf{F}}_{\mathbf{n}}, |.|_{\mathbf{2}}) &\leq \max_t \sup_Q \mathcal{N}(\delta\|\tilde{F}_t\|_{Q,2}, \tilde{\mathcal{S}}_t, \|.\|_{Q,2}) \lesssim (1/\delta)^{cv}.
\end{aligned}$$

Assumption 3.10 imposes that the time series $\tilde{S}_t(G)$ and X_t can be expressed as measurable functions of i.i.d. innovations ε_t. Assumption 3.11(i) completes the Markov exogeneity (Assumption 3.1) in the presence of the covariate Z_t, which is introduced at the beginning of Section 3. Assumption 3.11(ii) is a standard envelope assumption on the function class, and it states that the function of interest is enveloped by a function of X_{t-2}. Assumption 3.12 implies that $\Phi_{\phi_{\tilde{v}}}(\tilde{S}_{.}(G)) < \infty$ for $\tilde{v} = 1/2$ or 1. Assumption 3.13 restricts the complexity of the function classes. Based on these assumptions, we have the following rate:

THEOREM 3.2. Under Assumptions 3.1 to 3.3, 3.7, and 3.9 to 3.13,

$$\sup_{P_T\in\mathcal{P}_T(M,\kappa)} \left|\bar{\mathcal{W}}(G)-\widetilde{\mathcal{W}}(G)\right| \lesssim_p \frac{c_T[2V(\log T)e\gamma]^{1/\gamma}\sup_{G\in\mathcal{G}}\Phi_{\phi_{\tilde{v}}}(\tilde{S}_{\cdot}(G))}{\sqrt{T-1}} + C\sqrt{\frac{v}{T-1}},$$

where c_T is a large enough constant; $\tilde{v}=1/2$ or 1, and $\gamma=1/(1+2\tilde{v})$; V and v are the constants defined in Assumption 3.13; and C is the similar constant in Theorem 3.1, which depends only on M and κ.

A proof is presented in Appendix A.6. The bound depends on the complexity of the function class, V and v, and the time-series dependency, $\sup_G\Phi_{\phi_{\tilde{v}}}(\tilde{S}_{\cdot}(G))$. As we consider the sample analogue of unconditional welfare, all the observations are utilized, resulting in a $\frac{1}{\sqrt{T-1}}-$rate of convergence.

3.2.3 The regret bound

Now, we can obtain the overall bound for unconditional welfare. 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 3.1 to 3.7, and E_{P_T} be the expectation taken over different realizations of random samples. Recall that x_{obs} is defined to be the observed value of X_{T-1}.

THEOREM 3.3. Under Assumptions 3.1 to 3.13,

$$\sup_{P_T\in\mathcal{P}_T(M,\kappa)} \mathsf{E}_{P_T}[\mathcal{W}_T(G_*|x_{obs})-\mathcal{W}_T(\hat{G}|x_{obs})] \lesssim \frac{1}{\underline{p}}\left(C\sqrt{\frac{v}{T-1}}+\frac{c_T[2V(\log T)e\gamma]^{1/\gamma}\sup_{G\in\mathcal{G}}\Phi_{\phi_{\tilde{v}}}(\tilde{S}_{\cdot}(G))}{\sqrt{T-1}}\right), \tag{51}$$

where G_* is defined in (36).

This theorem follows from (43), Theorems 3.1 and 3.2, and a similar reasoning in Appendix A.2.

Remark 3.2 (on Assumptions 3.4-3.5)**.** In this theorem, the conditional regret is compared with the conditional welfare achieved under G_*, which is the optimal policy within the class of feasible unconditional decision sets, $\mathcal{G}$. Under Assumption 3.4, G_* can be substituted with G_*^{FB} without altering the regret bound. However, if Assumption 3.4 is violated, the regret bound relative to G_*^{FB} becomes

$$\begin{aligned}&\sup_{P_T\in\mathcal{P}_T(M,\kappa)} \mathsf{E}_{P_T}[\mathcal{W}_T(G_*^{FB}|x_{obs})-\mathcal{W}_T(\hat{G}|x_{obs})]\\ &\lesssim \frac{1}{\underline{p}}\left(C\sqrt{\frac{v}{T-1}}+\frac{c_T[2V(\log T)e\gamma]^{1/\gamma}\sup_{G\in\mathcal{G}}\Phi_{\phi_{\tilde{v}}}(\tilde{S}_{\cdot}(G))}{\sqrt{T-1}}\right)+\left[\mathcal{W}_T(G_*^{FB}|x_{obs})-\mathcal{W}_T(G_*|x_{obs})\right],\end{aligned}$$

where the last term represents the cost to social welfare incurred by adhering to policy restrictions (such as ethical, moral, or legislative considerations) that are implied by $\mathcal{G}$.

The factor $\frac{1}{\underline{p}}$ arises from Assumption 3.5. A violation of this assumption, i.e., setting $\underline{p} = 0$, could result in unbounded conditional regret. Under Assumption 3.5, Theorem 3.3 shows that the regret bound of our proposed policy choice converges at a rate of $\frac{1}{\sqrt{T-1}}$.

4 Extensions and Discussion

In this section, we discuss a few possible extensions. Section 4.1 introduces a multi-period policy-making framework. Section 4.2 concerns the ability of the current methods to handle the Lucas critiques. Section 4.3 discusses a few connections of T-EWM to the literature on optimal policy choice. More extensions are discussed in Appendix B.

4.1 Multi-period welfare

So far we have considered only the case of a one-period welfare function. This subsection discusses how to extend the current 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. Extending this model beyond two periods is straightforward using similar reasoning. 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 planner 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 $g_1(.)$. Furthermore, we assume that period $T+1$ welfare is determined only by the treatment choice in the period T, i.e., $W_T = g_1(W_{T-1})$. This means there is an information update in period T. The two-period welfare function is

$$\begin{aligned} \mathcal{W}_{T:T+1}(g_1(\cdot), g_2(\cdot)|\mathcal{F}_{T-1}) &= \mathcal{W}_{T:T+1}(g_1(\cdot), g_2(\cdot)|W_{T-1}) \\ &= \mathcal{W}_T(g_1(\cdot)|W_{T-1}) + \mathcal{W}_{T+1}\left(g_2(\cdot)|W_T = g_1\left(W_{T-1}\right)\right). \end{aligned} \tag{52}$$

To be more specific with the analytical format, we suppress the $(\cdot)$ in $g_i(\cdot)$, when the meaning is clear from the context.

$$
\begin{aligned}
&\mathcal{W}_{T:T+1}(g_1, g_2|W_{T-1} = w) \\
&= \mathsf{E}\left[Y_T(W_{T-1}, 1)g_1(W_{T-1}) + Y_T(W_{T-1}, 0)\left(1 - g_1(W_{T-1})\right)|W_{T-1} = w\right] \\
&+ \mathsf{E}\left[Y_{T+1}(W_T, 1)g_2(W_T) + Y_{T+1}(W_T, 0)(1 - g_2(W_T))|W_T = g_1(w)\right]. \qquad (53)
\end{aligned}
$$

The second term of (53) follows from

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

where the last equality follows from Assumption 2.1. To estimate the above welfare function, we recall the definition of $T(w) = \#\{1 \leq t \leq T-1 : W_t = w\}$, and we define $T(g_1(w))$ similarly. Then the empirical analogue of (52) can be written as,

$$
\begin{aligned}
\widehat{\mathcal{W}}_{T:T+1}(g_1, g_2|w) =& \frac{1}{T(w)} \sum_{t:W_{t-1}=w} \left\{ \frac{Y_t W_t g_1(W_{t-1})}{e_t(W_{t-1})} + \frac{Y_t\left(1 - W_t\right)\left(1 - g_1(W_{t-1})\right)}{1 - e_t(W_{t-1})} \right\} \\
&+ \frac{1}{T(g_1(w))} \sum_{t:W_{t-1}=g_1(w)} \left\{ \frac{Y_t W_t g_2(W_{t-1})}{e_t(W_{t-1})} + \frac{Y_t\left(1 - W_t\right)\left(1 - g_2(W_{t-1})\right)}{1 - e_t(W_{t-1})} \right\}.
\end{aligned} \qquad (54)
$$

The maximizer of (54), $\hat{g}_1$, $\hat{g}_2$, can be obtained by backward induction, a technique widely applied in the Markov decision process (MDP) and dynamic treatment regime literature. See Section 4.3.1 and Appendix A.10 for more discussion on the relationship between T-EWM and MDP. To derive the theoretical property of the estimator, we also define

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

Similarly to the derivation in the previous sections, $\widehat{\mathcal{W}}_{T:T+1}(g_1, g_2|w) - \bar{\mathcal{W}}_{T:T+1}(g_1, g_2|w)$ is a (weighted) sum of MDS. Its upper bound can be shown by the method of Section 2.3. We show in Appendix B.3 the extension to multi-period welfare with continuous conditioning covariate.

4.2 Accounting for Lucas critique

How can the framework of T-EWM handle the Lucas critique? In this section, we clarify a link between T-EWM and the SVAR approach based on a three-equation New Keynesian model where a choice of a policy regime can take into account economy's policy response.

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.

$$\begin{aligned} \text{Phillips curve:} \quad & \pi_t = \beta \mathsf{E}_t \pi_{t+1} + \kappa x_t + \varepsilon_t, \\ \text{IS curve:} \quad & x_t = \mathsf{E}_t x_{t+1} - \sigma^{-1}(i_t - \mathsf{E}_t \pi_{t+1}), \\ \text{Taylor rule:} \quad & i_t = \delta \pi_t + v_t, \end{aligned} \tag{55}$$

Here v_t is the monetary policy shock (baseline target rate), which can be viewed as a policy variable that the planner can manipulate. In the process of generating the sample, it is often assumed to follow an AR(1) process $v_t = \rho v_{t-1} + e_t$. The AR coefficient of v_t, ρ, represents a monetary policy regime. We also assume that the shocks in the Phillips curve follow AR(1), $\varepsilon_t = \gamma \varepsilon_{t-1} + \delta_t$. 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 d_t can be written as $d_t = F d_{t-1} + \eta_t$.

Define the outcome variables of the system (55) to be $\tilde{Y}_t = \begin{pmatrix} x_t \\ \pi_t \end{pmatrix}$. At the end of time $T-1$, the goal of the planner is minimizing (or maximizing) the expectation of some function of $\tilde{Y}_T$. For example, an objective function is the welfare cost that penalizes 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 A.8 shows that the VAR-reduced form of the system (55) can expressed as:

$$\tilde{Y}_t = M(\rho) d_t, \tag{56}$$

where $M(\rho)$ is a non-random matrix defined in Appendix A.8. If the model (55) is correctly specified, the solution to (56) takes the Lucas critique into account since it solves for a *deep* parameter ρ. A change in ρ incorporates both the direct effect of the policy regime (through ρ in d_t equation) and private agents' anticipation of the policy change (ρ in $M(\rho)$).

Now we show how this is related to the T-EWM framework. The treatment v_t in (55) corresponds to W_t in the previous sections. We can write $M(\rho) = \begin{pmatrix} m_{11}(\rho) & m_{12}(\rho) \\ m_{21}(\rho) & m_{22}(\rho) \end{pmatrix}$. When v_t is a binary variable (e.g., high target rate and low target rate), we can define the potential outcomes by setting $v_t = 1$ or 0 in (56): $\tilde{Y}_t(1) = \begin{pmatrix} m_{11}(\rho) + m_{12}(\rho)\varepsilon_t \\ m_{21}(\rho) + m_{22}(\rho)\varepsilon_t \end{pmatrix}$ 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 can write them as $\tilde{Y}_t(1;\rho)$ and $\tilde{Y}_t(0;\rho)$. Transforming $\tilde{Y}_t = (x_t, \pi_t)'$ into $Y_t = |x_t|^2 + |\pi_t - \pi_0|^2$, we can define the potential outcomes for the welfare $(Y_t(1;\rho), Y_t(0;\rho))$, which also depend on ρ. The policy choice problem for $v_T \in \{0,1\}$ can be set to minimize the expected welfare cost:

$$\mathsf{E}\left[Y_T(1;\rho)g(v_{T-1}) + Y_T(0;\rho)(1 - g(v_{T-1}))|\mathcal{F}_{T-1}\right] \tag{57}$$

in $g(v_{T-1}) \in \{0,1\}$, assuming that the one-time policy choice of v_T does not change the value of ρ governing the outcome generating process (56). Thus, one can view the framework of T-EWM in the previous sections concerns a choice of binary policy shock v_T, assuming the fixed deep parameter of ρ.

In contrast, consider the case where the policy choice of interest concerns regime ρ instead of v_T. We can modify the T-EWM framework as follows in order to handle this case. Assume that the monetary policy shock v_t is continuously distributed and let $f_{v_t|\mathcal{F}_{t-1}}(v;\rho)$ be the conditional density of v_t given the filtration $\mathcal{F}_{t-1}$ evaluated at v and ρ. In the outcome process of $t = T-1$ and earlier, the distribution $f_{v_t|\mathcal{F}_{t-1}}(v;\rho)$ at the value of ρ generates the outcomes up to $T-1$. In the counterfactual policy scenario of period $t = T$ and later, the planner manipulates the distribution of v_T by changing the value of ρ instead of directly setting a particular value of v_T. For such intervention, we specify the planner's problem as a choice of ρ to minimize

$$\mathsf{E}\left[Y_T(v_T;\rho)|\mathcal{F}_{T-1}\right] = \int \mathsf{E}\left[Y_T(v;\rho)|\mathcal{F}_{T-1}\right] f_{v_T|\mathcal{F}_{T-1}}(v;\rho)dv. \tag{58}$$

This objective function differs from the welfare objective function of (57) in the following two aspects. First, in (57), the policy rule for v_T is deterministic, whereas in (58), the planner chooses ρ to change the conditional probability density function of the treatment v_T, i.e., a randomized policy for v_T. Second, we see that in (57), the policy v_T affects the outcome Y_T only through v_T with ρ fixed, whereas in (58), the policy ρ affects the outcome through both the distribution of v_T and the deep policy parameter ρ underlying the potential outcomes.

Despite these differences, we can pursue a T-EWM approach for a statistical policy choice of ρ as far as one can construct a sample analogue of the welfare criterion of (58). For instance, let ρ_t be the policy regime in the sampling period $t = 1, \ldots, T-1$, and assume ρ_t is observable or estimable.[8] For the sake of illustration, we will use a simplified model by assuming that ρ_t takes values in a finite set Ω and $v_t \in \{0,1\}$. For a given $\rho \in \Omega$, we assume that ρ_t is independent of $Y_t(v_t;\rho)$ conditional on $\mathcal{F}_{t-1}$, and we let Assumption 2.1 hold with $W_t = v_t$, such that $\mathsf{E}(\cdot|\mathcal{F}_{t-1}) = \mathsf{E}(\cdot|v_{t-1})$. Then, for a policy function $g_\rho : \{0,1\} \to \Omega$ and

[8] 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.

$v \in \{0,1\}$, we can define the empirical welfare estimating(58) as

$$\widehat{\mathcal{W}}\left(g_\rho|v\right) = T(v)^{-1} \sum_{t:v_{t-1}=v} \left[\sum_{\rho\in\Omega} \frac{Y_t \mathbf{1}(\rho_t = \rho)}{\Pr[\rho_t = \rho|v_{t-1}]} \mathbf{1}[g_\rho(v_{t-1}) = \rho]\right], \tag{59}$$

where $T(v) := \#\{v_{t-1} = v\}$. We then maximize this empirical criterion with respect to $v = v_{T-1}$ to estimate an optimal policy regime.

It is worth noting that as long as ρ_t is observable (or estimable), the empirical analogue of the welfare criterion is free from functional form restrictions of $Y_t(v;\rho)$ or a distributional restriction of $f_{v_t|\mathcal{F}_{t-1}}(v;\rho)$. On the other hand, whether the empirical analogue can estimate the welfare criterion (58) well or not relies on whether the agents in the economy have correct knowledge of ρ_t and behaves in response to the shift of ρ_t. We leave for future research a rigorous characterization for the learnability of an optimal policy regime along this T-EWM approach and its implementability in practice.

4.3 Connections to other policy choice models in the literature

In this section, we discuss T-EWM's relation to the literature on treatment and optimal policy analysis.

4.3.1 Connection to MDP and Reinforcement learning

Markov Decision Process (MDP) is the standard framework for reinforcement learning algorithms commonly applied to decision-making problems in dynamic environments. See, e.g., Kallenberg (2016) for a comprehensive introduction. We can relate the current T-EWM model to a special case of MDP with a finite horizon. The conditioning variables X_{t-1} correspond to the Markov state at time t and the welfare outcome Y_t corresponds to the reward at t. Before the planner intervenes at time T, the Markov state transitions follow the data generating process in which the transitions of the policies are prescribed by the propensity score. After the planner intervenes, the transition of policies is governed by a deterministic rule described by the (estimated) optimal policy function (2). As we show in Appendix A.10, the population conditional welfare of T-EMW with a finite-horizon welfare target can be viewed as the value function of a finite-horizon MDP with a non-stationary solution. See Chapter 2 of Kallenberg (2016) for more details about finite horizon MDPs.

The reinforcement learning (RL) literature is vast and provides a rich toolbox to solve MDPs. There are, however, several main differences between T-EWM and RL approaches. First, T-EWM builds on the framework of potential outcome time series which allows for more flexibility in dynamic causal dependence than the reward generating processes considered in the standard MDPs. Second, the framework of T-EWM can easily accommodate a nonstationary environment in both causal effects and data generating processes without requiring modeling them explicitly. The third difference lies in the estimation algorithms. T-EWM obtains a policy by optimizing the empirical welfare criterion, whereas RL typically employs iteration algorithms to maximize the value function. Fourth, T-EWM performs an optimal policy choice with available data exogenously given to the planner. The RL literature in contrast studies the decision maker's joint strategies of sampling data and learning policies. That is, the T-EWM method can be regarded as an estimation approach for the optimal policy in an offline and off-policy RL problem with a finite horizon, building on the potential outcome time-series framework.

4.3.2 Comparison with impulse response functions

In empirical macroeconomics, a common practice is to measure the causal effects of policies using the impulse response functions (IRF) to the structural policy shocks; see Sims (1980), Ramey (2016), Plagborg-Moller (2016), Stock and Watson (2017), among others. This approach is based on the representation that the endogenous outcome variables are expressed as a weighted sum of the structural shocks, in which the coefficients of the structural shocks correspond to IRFs. With our notation, it can be expressed as

$$Y_t = \sum_{h=0}^{\infty} (\theta_h W_{t-h} + \phi_h \epsilon_{t-h}), \tag{60}$$

where W_t is the policy shock and ϵ_t is a non-policy structural shock that the planner cannot control. Setting the value of structural shocks exogenously to $w_{-\infty:t}$ without intervening the joint distribution of the other shocks $\epsilon_{-\infty:t}$ defines the potential outcome $Y_t(w_{-\infty:t})$. The coefficients (θ_h, ϕ_h), $h = 0, 1, 2, \ldots$, are IRFs of the h-period ahead outcome variable to the unit changes in policy and non-policy shocks, respectively.

Our approach of T-EWM based on the potential outcome time series differs from the approach of causal IRF analysis in several aspects. First, we assume that the policy shocks of interest $\{W_t : t = 0, \ldots, T-1\}$ are directly observable in the available time-series data, while the common framework of IRF analysis such as SVAR considers settings where the policy shocks are not observable and need to be identified through identification of a linear simultaneous equation system. Second, our framework imposes little restrictions on the functional form of how the current and past shocks affect the outcome, while the IRF analysis based on (60) assumes that the structural equation for Y_t is linear in the structural shocks and additive over the horizons. Third, our framework allows unrestricted heterogeneity (i.e., nonstationarity) of the causal effects, whereas the shock-based causal models (60) assume either stationary IRFs that depend only on horizon h or nonstationary IRFs with explicit modeling of how they evolve over time.[9] Fourth, the existing IRF analysis focuses mainly on estimating and inferring the IRFs, whereas the literature has not formulated how to perform statistical policy choice based on the estimated IRFs. Our T-EWM approach, in contrast, explicitly formulates the policy choice problem using potential outcome time series and analyzes the welfare performance of the statistical policy choice.

4.4 Estimation with unknown propensity score

To this point, we have treated the propensity score function as known, but this is infeasible in many applications. Here we consider the case where the propensity score at each time t, $e_t(\cdot)$, is 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 $\hat{e}_t$.

4.4.1 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 one of the rate of convergence of $\hat{e}_t(\cdot)$ and the rate of convergence of the estimated welfare loss (given in Theorem 3.3).

THEOREM 4.1. 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 a 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 above equation (51), we assume that there exists a sequence $\phi_T \to \infty$ such that the series of estimators $\hat{\tau}_t$ satisfy

$$\limsup_{T\to\infty} \sup_{P_T\in\mathcal{P}_T(M,\kappa)} \phi_T \mathsf{E}_{P_T}\left[(T-1)^{-1}\sum_{t=1}^{T-1}|\hat{\tau}_t - \tau_t|\right] < \infty. \tag{61}$$

Then under Assumptions 3.1 to 3.13, we have

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

To ensure that $\hat{e}_t(W_{t-1})$ is a valid estimator of $e_t(W_{t-1})$ in the sense of satisfying (61), we need to restrict the dynamics of W_t. It is, however, important to note that the condition of (61) does not require stationarity of the outcome variable $Y_t(.)$.

[9] See Bojinov and Shephard (2019) and Rambachan and Shepherd (2021) for comparative discussions between the IRFs and the average treatment effects with potential outcome time series.

A proof is presented in Appendix A.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 rate is affected by the estimation accuracy of the propensity score, and it is the maximum of $\frac{1}{\sqrt{T-1}}$ and ϕ_T^{-1}.

Remark 4.1. 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 4.1 can be improved. However, this is not a trivial extension, and we leave it for further research.

4.4.2 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 or nonparametrically. An example of a parametric estimator is the (ordered) probit model, which is employed by Hamilton and Jorda (2002), Scotti (2018), and Angrist et al. (2018). Under Assumption 3.1, e_t can be expressed as a function of X_{t-1}. Then, the structure of the propensity score can be given by a probit model

$$\begin{aligned} e_t(X_{t-1}) &\equiv P(W_t = 1|X_{t-1}) = \Phi(\beta' X_{t-1}), \\ 1 - e_t(X_{t-1}) &\equiv P(W_t = 0|X_{t-1}) = 1 - \Phi(\beta' X_{t-1}). \end{aligned}$$

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

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. Their methods can be employed here. For simplicity, we assume that the propensity score function is invariant across times, i.e., $e_t(\cdot) = e(\cdot)$ for any t, and $e(\cdot)$ is continuous, and we set $\mathcal{X} \subset \mathbb{R}^1$. For a local polynomial of order $p = 1$ and any $x \in \mathcal{X}$, a local likelihood logit model can be specified as

$$\log\left[\frac{e(x)}{1-e(x)}\right] = \alpha_x,$$

for a local parameter α_x. By the continuity of the propensity score function $e(\cdot)$, for some X_{t-1} close to x, we can find a local parameter β_x, such that $\log\left[\frac{e(X_{t-1})}{1-e(X_{t-1})}\right] \approx \alpha_x + \beta_x(X_{t-1} - x)$.

The estimated propensity score evaluated at x, $\hat{e}(x)$, can be obtained by solving

$$(\hat{\alpha}_x, \hat{\beta}_x) = \text{argmax}_{\alpha,\beta} \frac{1}{T-1} \sum_{i=1}^{T-1} \left\{ \left[W_i \log\left(\frac{\exp\left(\alpha + \beta\left(X_{t-1} - x\right)\right)}{1 + \exp\left(\alpha + \beta\left(X_{t-1} - x\right)\right)} \right) + (1 - W_i) \log\left(\frac{1}{1 + \exp\left(\alpha + \beta\left(X_{t-1} - x\right)\right)} \right) \right] K\left(\frac{X_{t-1} - x}{h} \right) \right\}.$$

where $K(\cdot)$ is a kernel function, and h is bandwidth. Then, we have $\hat{e}(x) = \frac{\exp(\hat{\alpha}_x)}{1+\exp(\hat{\alpha}_x)}$.

5 Application

In the previous pandemic of COVID-19, policymakers around the world faced the problem of making effective policies to respond to the Covid-19 pandemic. In this section, we illustrate the usage of T-EWM with an application to choosing the stringency of restrictions imposed in the United States during the pandemic. We let the treatment W_t be a binary indicator for whether the government relaxes restrictions at week t. $W_t = 1$ means the stringency of restrictions is maintained or increased from time $t-1$. $W_t = 0$ means restrictions are relaxed. The stringency of restrictions is measured by the Oxford Stringency Index, which is described in Section 5.1. We assume that a change in the stringency of restrictions at time t will have a lagged effect on deaths, and set the outcome variable Y_t to be $-1\times$*two-week ahead deaths*. The time t information set is,

$$X_t = (\text{cases}_t, \text{deaths}_t, \text{change in cases}_t, \text{change in deaths}_t, \text{restriction stringency}_t, \text{vaccine coverage}_t, \text{economic conditions}_t). \quad (62)$$

The variables in X_t are chosen to include the most important factors considered by policy makers when deciding the stringency of restriction. The inclusion of the economic conditions reflects policymaker's concerns over the economic effect of restrictions. Finally, the propensity score is estimated as a logit model with a linear index,

$$\log\left(\frac{\Pr(\text{keeping or increasing restrictions at week } t)}{\Pr(\text{decreasing restriction at week } t)} \right) = \alpha + \beta X_{t-1}. \quad (63)$$

5.1 Data

The dataset consists of weekly data for the United States. It runs from April 2020 to January 2022 and contains 92 observations. Data on cases, deaths, and vaccine coverage was downloaded from the website of the Centers for Disease Control and Prevention (CDC). These series are plotted in Figures 4 and 6a in Appendix A.12. We use the Oxford Stringency

Index as a measure of the stringency of restrictions. This index is taken from the Oxford COVID-19 Government Response Tracker (OxCGRT), which tracks policy interventions and constructs a suite of composite indices that measure governments' responses.[10] The left panel of Figure 5 in Appendix A.12 plots the time series of the stringency index.

The economic impact of restrictions is a crucial factor that every policymaker has to consider. To measure economic conditions, we use the Lewis-Mertens-Stock weekly economic index (WEI). [11]The right panel of Figure 5 in Appendix A.12 plots the WEI. In general, none of these time series seems stationary over the sample period. They exhibit both seasonal fluctuations and changes with respect to different stages of the pandemic.

5.2 On T-EWM assumptions

Before we proceed with the analysis, we discuss the validity of some of the T-EWM assumptions in this context. The Markov properties (Assumption 3.1) require that (i) the potential outcome at time t depends only on the time t and time $t-1$ treatments; (ii) conditional on X_{t-1}, the potential outcomes and treatment at time t are not affected by the path $X_{0:t-2}$. (i) can be justified here as the treatment is the *direction of change* in the stringency of restrictions. While the *level* before $t-1$ may affect the outcome at t, its directional change may not have such a long-existing impact. (ii) requires that conditional on current cases and deaths, cases and deaths one week ago (as X_t includes the first difference of cases and deaths, it includes the level of cases and deaths from one week ago), current economic conditions, the stringency of restrictions, vaccine coverage, and deaths in two weeks will be independent of lags of these variables if the lag is greater than one. For this assumption to hold, it must be that current infections are unrelated to deaths in more than three weeks. In general, it is unclear whether this is true, although there is some support in the literature. For example, based on data collected during the early stage of the pandemic in China, Verity et al. (2020) calculate that the posterior mean time from infection to death is 17.8 days, with a 95%-confidence interval of [16.9, 19.2].

Strict overlap (Assumption 3.2) requires that, for all $x \in \mathcal{X}$, the propensity score is strictly larger than 0 and smaller than 1. Figure 6b in Appendix A.12 shows the histogram of estimated propensity scores. The strict overlap assumption can be verified by visual inspection: the smallest (estimated) propensity score is larger than 0.3, and the largest is smaller than 0.9. Sequential unconfoundedness (Assumption 3.3) requires that conditional on X_{t-1}, treat-

[10]The index is calculated based on a dataset that is updated by a professional team of over one hundred students, alumni, staff, and project partners. It is a composite index covering different types of restriction, such as school and workplace closures, restrictions on the size of gatherings, internal and international movements, facial covering policies etc. See Hale et al. (2020) for more details.

[11]The WEI is an index of ten indicators of real economic activity. It represents the common component of series covering consumer behavior, the labor market, and production. See Lewis et al. (2021) for more details.

ment assignment at time t is quasi-random. This assumption cannot be tested in general. However, in the past two years, policymakers have had very limited knowledge of Covid-19 beyond the observable data. After controlling for the observables in (62), it seems reasonable that the remaining random factors in two-week-ahead deaths and changes in the stringency of restrictions are independent.

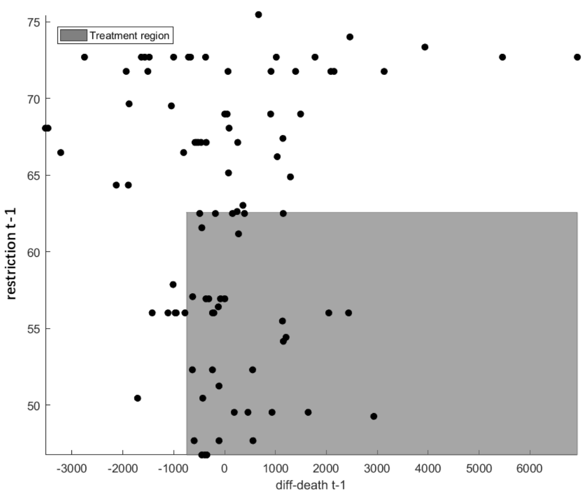
5.3 Estimation results and policy recommendation

In this subsection, we summarize the estimation results and discuss the policy recommendations. We show that T-EWM leads to sensible and robust policy decisions. After observing X_{T-1}, we aim to maximize expected welfare, $\mathsf{E}(-1 \cdot \text{deaths}_{T+1})$, over a set of quadrant policies. A class of quadrant treatment rules with k policy variables, $x = (x_1 \ldots x_k)'$, is defined as

$$\mathcal{G} \equiv \left\{ \begin{array}{c} \{x : s_1(x_1 - b_1) > 0 \ \& \ \ldots \ \& \ s_k(x_k - b_k) > 0\} \, ; \\ s_1, \ldots, s_k \in \{-1, 1\}, b_1, \ldots, b_k \in \mathbb{R} \end{array} \right\}.$$

Let X^P_{T-1} denote a vector of variables for policy choice. This can be any subvector of X_{T-1}. For our first set of results, we use $X^P_{T-1} = \left(\text{change in deaths}_{T-1}, \text{restriction stringency}_{T-1}\right)$. Figure 1 presents the estimated treatment region. The estimated optimal decision rule states that restrictions should not be relaxed (i.e. $W_T = 1$), if the weekly fall in deaths is below 745, and the current level of restrictions is lower than 62.6.

Figure 1. Optimal policy based on $X^P_{T-1} = \left(\text{change in cases}_{T-1}, \text{restriction stringency}_{T-1}\right)$



The x-axis is the change in deaths at $T-1$, and the y-axis is the stringency of restrictions at $T-1$.

We can examine the robustness of this result by expanding the set of variables for policy choice. We first add vaccine coverage: $X^P_{T-1} = (\text{change in deaths}_{T-1}, \text{restriction stringency}_{T-1}, \text{vaccine coverage}_{T-1})$. The optimal estimated quadrant policy is then a 3d-quadrant. Figure 7 in Appendix A.12 shows the projection of this 3d-quadrant onto the 2d-planes of (change in deaths_{T-1}, restriction stringency_{T-1}) and (restriction stringency_{T-1}, vaccine coverage_{T-1}). We further expand the set of variables for policy choice by adding the change in cases, so that X^P_{T-1} = (change of deaths_{T-1}, restriction stringency_{T-1}, vaccine coverage_{T-1}, change in cases_{T-1}). The optimal estimated quadrant policy in this case is a 4d-quadrant. Figure 8 in Appendix A.12 shows projections of this 4d-quadrant onto the 2d-planes of (change in deaths_{T-1}, restriction stringency_{T-1}) and (vaccine coverage_{T-1}, change in cases_{T-1}).

Table 1 summarises the estimation results of Figures 1, 7, and 8. It shows that the thresholds of the T-EWM policies are insensitive to the additions of the variables that the quadrant policies depend on. As the number of variables increases, the threshold of the change in weekly deaths decreases slightly from −745 to −1007.5, while the other thresholds remain stable. During the sample period, the mean of weekly cases is 594344.3. Therefore, the threshold in the last column and the last row corresponds to a 10% fall relative to the average number of cases during the sample period. In sum, T-EWM suggests that the policymaker should not relax restrictions ($W_T = 1$) if there are no significant drops in deaths and cases, current stringency is comparatively low, and vaccine coverage is comparatively low.

Table 1. Estimated optimal quadrant rules

	variables	change in deaths	restriction	vaccine (%)	change in cases
Figure 1	2	> -745	< 62.6	–	–
Figure 7	3	> -1007.5	< 62.6	< 115.2	–
Figure 8	4	> -1007.5	< 62.6	< 115.2	> -50842.5

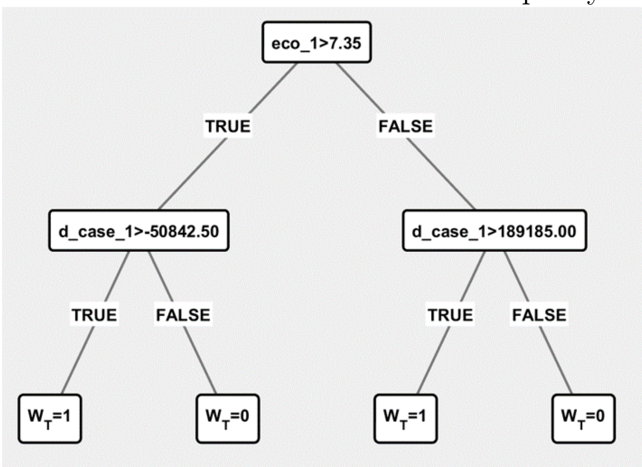
Including more than four policy variables presents a significant computational challenge for the grid-search method we have used to optimize quadrant treatment rules. [12] To overcome this issue, we employ decision trees to search for an optimal policy based on time-series empirical welfare.

The decision tree approach (Breiman et al. (1984)) sequentially searches for policy variables and their threshold-based splits to maximize welfare. Athey and Wager (2021) and Zhou et al. (2023) study the properties and implementation of this approach in maximizing doubly robust empirical welfare criterion with brute force searches for tree partitions. Ida et al. (2022) implement decision tree with a heuristic two-step optimization in the context of estimating a rebate assignment policy in an electricity market. We implement a classification tree algorithm based on Hastie et al. (2009). See Appendix A.12.3 for the details of our implementation.

[12] When five variables are included in X^P_{T-1}, the estimated running time is 12 hours; with six variables, the estimated running time skyrockets to 4459 hours. (These estimates were made on a computer with a 12th Gen Intel(R) Core(TM) i7-1265U 2.7 GHz processor and 32GB of RAM.)

Figure 2 illustrates a policy rule obtained by a T-EWM decision tree, which considers all the seven policy variables within X_{T-1}, i.e., $X^P_{T-1} = X_{T-1} = (\text{cases}_{T-1}, \text{deaths}_{T-1}$, change in cases_{T-1}, change in deaths_{T-1}, restriction stringency_{T-1}, vaccine coverage_{T-1}, economic $\text{conditions}_{T-1})$. Due to the relatively small sample size of 92, we only generate trees with a depth of three.

Figure 2. T-EWM decision tree with seven policy variables



The result from the decision tree indicates that if the economy performs relatively well, policymakers should adopt a more aggressive approach in maintaining or tightening Covid restrictions. (Note that the decision given by the left node of the second level aligns with the last one in the last row of Table 1, which considers four policy variables.) Conversely, if the economy performs relatively poorly, policymakers should refrain from further tightening Covid restrictions unless the increase in cases is significant.

To wrap up, we illustrate the use and implementation of the proposed T-EWM approach in making policy decisions. Furthermore, utilizing decision trees with the T-EWM approach allows for the inclusion of more policy variables while keeping computation time manageable. More results from the T-EWM decision trees including a specification with a higher order of the Markovian structure of $q = 2$ can be found in Appendix A.12.

6 Conclusion

This article proposes T-EWM, a framework and method for choosing optimal policies based on time-series data. We characterise assumptions under which this method can learn an optimal policy. We evaluate its statistical properties by deriving non-asymptotic upper and lower bounds of the conditional welfare. We discuss its connections to the existing literature, including the Markov decision process and impulse response analysis. We present simulation results and empirical applications to illustrate the computational feasibility and applicability of T-EWM. As a benchmark formulation, this paper mainly focuses on a one-period social welfare function as a 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

K. Adusumilli. Risk and optimal policies in bandit experiments. *arXiv preprint arXiv:2112.06363*, 2021.

K. Adusumilli, F. Geiecke, and C. Schilter. Dynamically optimal treatment allocation using reinforcement learning. *arXiv preprint arXiv:1904.01047*, 2022.

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. Policy learning with observational data. *Econometrica*, 89(1): 133–161, 2021.

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.

L. Breiman, J. Friedman, R. Olshen, and C. Stone. *Classification and regression trees*. CRC press, 1984.

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.

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.

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

M. Dimakopoulou, Z. Zhou, S. Athey, and G. Imbens. Estimation considerations in contextual bandits. *arXiv preprint arXiv:1711.07077*, 2017.

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.

T. Hale, N. Angrist, B. Kira, A. Petherick, T. Phillips, and S. Webster. Variation in government responses to covid-19. *Blavatnik School of Government Working Paper*, 2020.

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.

T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer series in statistics. Springer, 2009. ISBN 9780387848846. URL `https://books.google.de/books?id=eBSgoAEACAAJ`.

T. Ida, T. Ishihara, K. Ito, D. Kido, T. Kitagawa, S. Sakaguchi, and S. Sasaki. Choosing who chooses: Selection-driven targeting in energy rebate programs. *NBER working paper*, (30469), 2022.

T. Ida, T. Ishihara, K. Ito, D. Kido, T. Kitagawa, S. Sakaguchi, and S. Sasaki. Dynamic targeting: Experimental evidence from energy rebate programs. *NBER working paper*, (32561), 2024.

T. Ishihara and T. Kitagawa. Evidence aggregation for treatment choice. *arXiv preprint arXiv:2108.06473*, 2024.

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.

M. Kasy and A. Sautmann. Adaptive treatment assignment in experiments for policy choice. *Econometrica*, 89(1):113–132, 2021.

L. Kilian and H. Lütkepohl. *Structural vector autoregressive analysis*. Cambridge University Press, 2017.

T. Kitagawa and J. Rowley. Bandit algorithms for policy learning: methods, implementation, and welfare-performance. *Japanese Economic Review*, forthcoming:1–48, 2024.

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. J. Lewis, K. Mertens, J. H. Stock, and M. Trivedi. Measuring real activity using a weekly economic index 1. *Journal of Applied Econometrics*, 2021.

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 series estimands have nonparametric causal meaning? *working paper*, 2021.

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

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.

R. Verity, L. C. Okell, I. Dorigatti, P. Winskill, C. Whittaker, N. Imai, G. Cuomo-Dannenburg, H. Thompson, P. G. Walker, H. Fu, et al. Estimates of the severity of coronavirus disease 2019: a model-based analysis. *The Lancet infectious diseases*, 20(6): 669–677, 2020.

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.

H. White and X. Lu. Granger causality and dynamic structural systems. *Journal of Financial Econometrics*, 8(2):193–243, 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.

Z. Zhou, S. Athey, and S. Wager. Offline multi-action policy learning: Generalization and optimization. *Operations Research*, 71(1):148–183, 2023.

A Supplemental Appendix A

A.1 Proof of Proposition 2.1

Proof. We have

$$\begin{aligned}
&\mathsf{E}\left[\widehat{\mathcal{W}}_t(g|w) - \bar{\mathcal{W}}_t(g|w)|\mathcal{F}_{t-1}\right] \\
&= \mathsf{E}\left[1(W_{t-1}=w)\left[\frac{Y_t W_t g(W_{t-1})}{e_t(W_{t-1})} + \frac{Y_t(1-W_t)\{1-g(W_{t-1})\}}{1-e_t(W_{t-1})}\right]|\mathcal{F}_{t-1}\right] \\
&- \mathsf{E}\left[1(W_{t-1}=w)\mathsf{E}\{Y_t(W_{0:t-1},1)g(W_{t-1}) + Y_t(W_{0:t-1},0)(1-g(W_{t-1}))|W_{t-1}=w\}|\mathcal{F}_{t-1}\right] \\
&=: I_t - II_t, \qquad \text{(A.1)}
\end{aligned}$$

Under Assumption 2.5, it remains to show $I_t - II_t = 0$. By the Law of iterated expectation,

$$\begin{aligned}
I_t &= \mathsf{E}\left[1(W_{t-1}=w)\left[\frac{Y_t W_t g(W_{t-1})}{e_t(W_{t-1})} + \frac{Y_t(1-W_t)\{1-g(W_{t-1})\}}{1-e_t(W_{t-1})}\right]|\mathcal{F}_{t-1}\right] \\
&= 1(W_{t-1}=w)\left\{\frac{g(W_{t-1})}{e_t(W_{t-1})}\mathsf{E}\left[Y_t W_t|\mathcal{F}_{t-1}\right] + \frac{1-g(W_{t-1})}{1-e_t(W_{t-1})}\mathsf{E}\left[Y_t(1-W_t)|\mathcal{F}_{t-1}\right]\right\}. \qquad \text{(A.2)}
\end{aligned}$$

Note that

$$\begin{aligned}
\mathsf{E}\left[Y_t W_t|\mathcal{F}_{t-1}\right] &= \mathsf{E}\left[Y_t(W_{0:t-1},W_t)W_t|\mathcal{F}_{t-1}\right] = \mathsf{E}\left[Y_t(W_{t-1},W_t)W_t|\mathcal{F}_{t-1}\right] \\
&= \mathsf{E}\left[Y_t(W_{t-1},1)W_t|\mathcal{F}_{t-1}\right] = \mathsf{E}\left[Y_t(W_{t-1},1)|\mathcal{F}_{t-1}\right]\mathsf{E}\left[W_t|\mathcal{F}_{t-1}\right] \\
&= \mathsf{E}\left[Y_t(W_{t-1},1)|\mathcal{F}_{t-1}\right]e_t(W_{t-1}), \qquad \text{(A.3)}
\end{aligned}$$

where the second equality follows from Assumption 2.1(i), the fourth equality follows from Assumption 2.3, and the last equality follows from $\mathsf{E}\left[W_t|\mathcal{F}_{t-1}\right] = \mathsf{E}\left[W_t|W_{t-1}\right] = e_t(W_{t-1})$ by the second statement of Assumption 2.1(ii). Applying the same arguments, we have

$$\mathsf{E}\left[Y_t(1-W_t)|\mathcal{F}_{t-1}\right] = \mathsf{E}\left[Y_t(W_{t-1},0)|\mathcal{F}_{t-1}\right]\left[1-e_t(W_{t-1})\right]. \qquad \text{(A.4)}$$

Combining (A.2), (A.3), (A.4), and the Law of iterated expectation, we obtain

$$I_t = 1(W_{t-1}=w)\mathsf{E}\left[Y_t(W_{t-1},1)g(W_{t-1}) + Y_t(W_{t-1},0)(1-g(W_{t-1}))|\mathcal{F}_{t-1}\right].$$

On the other hand,

$$
\begin{aligned}
II_t &= \mathsf{E}\left[1(W_{t-1}=w)\mathsf{E}\{Y_t(W_{0:t-1},1)g(W_{t-1})+Y_t(W_{0:t-1},0)(1-g(W_{t-1}))|W_{t-1}=w\}|\mathcal{F}_{t-1}\right] \\
&= 1(W_{t-1}=w)\mathsf{E}\left[Y_t g(W_{t-1})+Y_t(1-g(W_{t-1}))|W_{t-1}=w\right] \\
&= 1(W_{t-1}=w)\mathsf{E}\left[Y_t(W_{t-1},1)g(W_{t-1})+Y_t(W_{t-1},0)(1-g(W_{t-1}))|W_{t-1}\right] \\
&= 1(W_{t-1}=w)\mathsf{E}\left[Y_t(W_{t-1},1)g(W_{t-1})+Y_t(W_{t-1},0)(1-g(W_{t-1}))|\mathcal{F}_{t-1}\right],
\end{aligned}
$$

where the second equality follows from $\sigma(W_{t-1}) \subseteq \mathcal{F}_{t-1}$, the third equality follows from the definition of the indicator function, and the last equality follows from (15), which is an implication of Assumption 2.1. Consequently, we have $I_t - II_t = 0$. □

A.2 Proof of Theorem 2.1

A.2.1 Preliminary Lemmas

The following two lemmas will be used in the proofs of the main results.

LEMMA A.1 (Freedman's inequality)**.** Let $\xi_{a,i}$ be a martingale difference sequence indexed by $a \in \mathcal{A}$, $i = 1, \ldots, n$, $\mathcal{F}_i$ 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,

$$\Pr(\max_{a\in\mathcal{A}} |M_a| \geq z) \leq \sum_{i=1}^n \Pr(\max_{a\in\mathcal{A}} \xi_{a,i} \geq A) + 2\Pr(\max_{a\in\mathcal{A}} V_a \geq B) + 2|\mathcal{A}|e^{-z^2/(2zA+2B)}. \quad \text{(A.5)}$$

The proof can be found in Freedman (1975). Next, let $a \lesssim b$ indicate that there is a positive constant C, such that $a \leq C \cdot b$.

LEMMA A.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,

$$\mathsf{E}(\max_{a\in\mathcal{A}} |M_a|) \lesssim A\log(1+|\mathcal{A}|) + \sqrt{B}\sqrt{\log(1+|\mathcal{A}|)}. \quad \text{(A.6)}$$

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

$$
\begin{aligned}
\Pr(\max_a |M_a| \geq z) &\leq 2|\mathcal{A}|\exp(-z/4A) \quad \text{for} \quad z > B/A && \text{(A.7)} \\
&\leq 2|\mathcal{A}|\exp(-z^2/4B) \quad \text{Otherwise.} && \text{(A.8)}
\end{aligned}
$$

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

$$\mathsf{E}(\exp|C_a/4A|) \leq 2, \tag{A.9}$$

and

$$\mathsf{E}(\exp|D_a/\sqrt{4B}|^2) \leq 2. \tag{A.10}$$

For completeness, we show the derivations of these inequalities.

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

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

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

Again, the first inequality follows from Fubini's inequality and the second is due to Lemma A.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. □

A.2.2 Proof of Theorem 2.1

Proof. Define $\tilde{p}_w \stackrel{\text{def}}{=} \frac{1}{T-1}\sum_{t=1}^{T-1}\Pr(W_{t-1} = w|\mathcal{F}_{t-2})$. Note that $\tilde{p}_w$ is only a device for the proof, so it might contain elements that are not observed in the dataset.

$$\sup_{g:\{0,1\}\to\{0,1\}}\left[\widehat{\mathcal{W}}(g|w) - \bar{\mathcal{W}}(g|w)\right] = \sup_{g:\{0,1\}\to\{0,1\}}\frac{1}{T-1}\frac{T-1}{T(w)}\sum_{t=1}^{T-1}\left[\widehat{\mathcal{W}}_t(g|w) - \bar{\mathcal{W}}_t(g|w)\right]$$

$$= \sup_{g:\{0,1\}\to\{0,1\}}(\frac{T(w)}{T-1} - \tilde{p}_w + \tilde{p}_w)^{-1}(T-1)^{-1}\sum_{t=1}^{T-1}\left[\widehat{\mathcal{W}}_t(g|w) - \bar{\mathcal{W}}_t(g|w)\right]. \tag{A.11}$$

And

$$\frac{T(w)}{T-1} - \tilde{p}_w = \frac{1}{T-1}\sum_{t=1}^{T-1}\left[\mathbf{1}\left(W_{t-1} = w\right) - \Pr\left(W_{t-1} = w|\mathcal{F}_{t-2}\right)\right] = \frac{1}{T-1}\sum_{t=1}^{T-1}\xi_t,$$

where $\xi_t \stackrel{\text{def}}{=} \sum_t\left[\mathbf{1}\left(W_{t-1} = w\right) - \Pr\left(W_{t-1} = w|\mathcal{F}_{t-2}\right)\right]$. This is a sum of MDS.

Applying Freedman's inequality with $|\mathcal{A}| = 1$ and

$$V_a = \sum_{t=1}^{T-1} \mathrm{var}\left(\xi_t|\mathcal{F}_{t-2}\right) = \sum_{t=1}^{T-1} e(w|\mathcal{F}_{t-2})\left[1 - e(w|\mathcal{F}_{t-2})\right] < T - 1 \equiv B,$$

where the last inequality is by Assumption 2.2. By the same assumption, we also have $\max_t \xi_t < 1$. Therefore, we can set $A = 1$, then the first term of (A.5) is reduced to $\sum_{t=1}^{T-1} \Pr\left(\max_t \xi_t \geq 1\right) = 0$, and

$$\begin{aligned} \Pr\left(\left|\frac{1}{T-1}\sum_{t=1}^{T-1}\xi_t\right| \geq z\right) &= \Pr\left(\left|\sum_{t=1}^{T-1}\xi_t\right| \geq (T-1)z\right) \\ &\leq 2 \cdot \exp\left[\frac{-z^2(T-1)^2}{2A\,(T-1)\,z + 2B}\right] = 2 \cdot \exp\left[\frac{-z^2(T-1)^2}{2\,(T-1)\,z + 2(T-1)}\right] = o(1). \end{aligned} \tag{A.12}$$

hold for any $z > 0$. Therefore, for sufficient large T, we have $\left|\frac{T(w)}{T-1} - \tilde{p}_w\right| \lesssim_p \kappa/2$, where κ is the constant defined in Assumption 2.2. With probability approaching 1,

$$\left(\frac{T(w)}{T-1} - \tilde{p}_w + \tilde{p}_w\right)^{-1} \leq \left(-\left|\frac{T(w)}{T-1} - \tilde{p}_w\right| + \tilde{p}_w\right)^{-1} \leq (-\kappa/2 + \kappa)^{-1} = (\kappa/2)^{-1}. \tag{A.13}$$

Now we bound $\sum_{t=1}^{T-1}\left[\widehat{\mathcal{W}}_t(g|w) - \bar{\mathcal{W}}(g|w)\right]$, which is also a sum of MDS.

To apply Lemma A.2, we define $\mathcal{G} \equiv \{g : \{0,1\} \to \{0,1\}\}$ and $\xi_{g,t} \equiv \widehat{\mathcal{W}}_t(g|w) - \bar{\mathcal{W}}(g|w)$. (With some abuse of notations, this $\xi_{g,t}$ is different from ξ_t defined above). Note that $|\mathcal{G}|$ is finite, and we can find a constant C_A, such that for any t, $\sup_{g\in\mathcal{G}} \xi_{g,t} \leq (M + \frac{M}{\kappa}) < C_A$, where κ and M are constants defined in Assumption 2.2 and 2.5 respectively. By these two assumptions, we also can find a constant C_B depending only on M and κ, such that $V_g = \sum_{t=1}^{T-1} \mathsf{E}\left(\xi_{g,t}^2|\mathcal{F}_{t-1}\right) < C_B(T-1)$. Then by Lemma A.2,

$$\begin{aligned} \mathsf{E}\left(\sup_{g\in\mathcal{G}}\left|\sum_{t=1}^{T-1}\xi_{g,t}\right|\right) &\lesssim C_A \log\left(1 + |\mathcal{G}|\right) + \sqrt{B}\sqrt{\log\left(1 + |\mathcal{G}|\right)} \\ &\lesssim C_A \log\left(1 + |\mathcal{G}|\right) + \sqrt{C_B(T-1)}\sqrt{\log\left(1 + |\mathcal{G}|\right)} \\ &\lesssim \sqrt{(T-1)}\sqrt{C_B \log\left(1 + |\mathcal{G}|\right)}. \end{aligned} \tag{A.14}$$

By Jensen's inequality $\sup_{g\in\mathcal{G}} \mathsf{E}\left|\sum_{t=1}^{T-1}\xi_{g,t}\right| \leq \mathsf{E}\left(\sup_{g\in\mathcal{G}}\left|\sum_{t=1}^{T-1}\xi_{g,t}\right|\right)$. Then, combining

(A.11), (A.13), and (A.14),

$$\sup_{g\in\mathcal{G}} \mathsf{E}\left(\left|\widehat{\mathcal{W}}(g|w) - \bar{\mathcal{W}}(g|w)\right|\right)$$
$$\lesssim \sup_{g\in\mathcal{G}} \mathsf{E}\left\{\frac{1}{\kappa/2}\frac{1}{T-1}\left|\sum_{t=1}^{T-1}\xi_{g,t}\right|\right\} \lesssim \left\{\frac{1}{\kappa/2}\frac{\sqrt{C_B \log(1+|\mathcal{G}|)}}{\sqrt{(T-1)}}\right\} = \frac{C}{\sqrt{(T-1)}},$$

where C depends only on κ, $|\mathcal{G}|$, and M. The first curved inequality follows from the exponential tail probability of (A.12). □

A.3 Higher/infinite Markov order

A.3.1 The q-th ($q > 1$) Markov order

We modify Assumption 2.1 to:

Assumption 2.1* (q-th order Markov properties). For an integer $q > 1$, 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 $w_{0:t}$,

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

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

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

Under these modified assumptions, the policy function g is now defined on $\{0,1\}^q$, mapping to $\{0,1\}$. For a vector $w \in \{0,1\}^q$, the propensity score can be defined as $e_t(w) = \Pr(W_t = 1 | W_{t-q:t-1} = w)$. Furthermore, we can redefine:

$$\begin{aligned}
\mathcal{W}_T(g|w) &= \mathsf{E}\{Y_T(W_{T-q:T-1},1)g(W_{T-q:T-1}) + Y_T(W_{T-q:T-1},0)(1-g(W_{T-q:T-1}))|W_{T-q:T-1} = w\},\\
\widehat{\mathcal{W}}(g|w) &= \frac{1}{T(w)}\sum_{q\le t\le T-1: W_{t-q:t-1}=w}\left[\frac{Y_t W_t g(W_{t-q:t-1})}{e_t(W_{t-q:t-1})} + \frac{Y_t(1-W_t)\{1-g(W_{t-q:t-1})\}}{1-e_t(W_{t-q:t-1})}\right],\\
\bar{\mathcal{W}}(g|w) &= \frac{1}{T(w)}\sum_{q\le t\le T-1: W_{t-q:t-1}=w}\mathsf{E}\left[Y_t(W_{t-q:t-1},1)g(W_{t-q:t-1}) + Y_t(W_{t-q:t-1},0)\left[1-g(W_{t-q:t-1})\right]|W_{t-q:t-1}\right].
\end{aligned} \tag{A.15}$$

These welfare functions will also used in the subsequent Subsection A.3.2.

Moreover, Assumption 2.3 (sequential unconfoundedness) can be modified to: for any $t = 1, 2, \ldots, T-1$ and $w \in \{0, 1\}$,

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

Then, a convergence rate of $\frac{1}{\sqrt{T-q}}$ can be established by similar arguments used for the proof of Theorem 2.1 (cf. Appendix A.2).

A.3.2 Infinite Markov order

Under the infinite Markov order, the current variables might depend on historical values that precede the first observations of the sample time series. To accommodate this concept, we change the notation from $X_{0:t}$ to $X_{-\infty:t}$ in this subsection.

If we allow all the historical treatments to affect the current outcome, we can drop Assumption 2.1(i) completely. For a concise and consistent discussion of the policy function, we maintain the first statement of Assumption 2.1(ii) and adjust it to accommodate the case of infinite Markov order:

$$Y_t(w_{-\infty:t}) \perp X_{-\infty:t-1} | W_{-\infty:t-1}. \tag{A.16}$$

It means that for simplicity, we continue to exclude the past outcomes, $Y_{-\infty:t-1}$, from the variables influencing the current outcome Y_t. As a result, we can omit $Y_{-\infty:t-1}$ from the arguments of the policy function g, thereby defining g as a function mapping from $\{0,1\}^\infty$ to $\{0, 1\}$.

If the Markovian process has an infinite order, a valid empirical welfare function should take the form of $\widehat{\mathcal{W}}(g|w_{-\infty:T-1})$, which is infeasible. In practice, the planner can only use a truncated treatment path to construct the empirical welfare. Let the truncation be implemented at the period $T - m \in \{0 : T-1\}$, and the policy is given by

$$\hat{g} \in \text{argmax}_{g:\{w_{T-m:T-1}\}\to\{0,1\}} \widehat{\mathcal{W}}(g|w_{T-m:T-1}), \tag{A.17}$$

i.e., the optimizer is based on the truncated Markov order m.

Let us (re)define the optimal policy as $g^* \in \text{argmax}_{g:\{w_{-\infty:T-1}\}\to\{0,1\}} \mathcal{W}_T(g|w_{-\infty:T-1})$. Then, the population-level regret can be decomposed as:

$$\begin{aligned}
&\mathcal{W}_T(g^*|w_{-\infty:T-1}) - \mathcal{W}_T(\hat{g}|w_{-\infty:T-1}) = \mathcal{W}_T(g^*|w_{T-m:T-1}) - \mathcal{W}_T(\hat{g}|w_{T-m:T-1}) \\
&+ \mathcal{W}_T(g^*|w_{-\infty:T-1}) - \mathcal{W}_T(g^*|w_{T-m:T-1}) + \mathcal{W}_T(\hat{g}|w_{T-m:T-1}) - \mathcal{W}_T(\hat{g}|w_{-\infty:T-1}) \\
&\leq \mathcal{W}_T(g^*|w_{T-m:T-1}) - \mathcal{W}_T(\hat{g}|w_{T-m:T-1}) + 2 \sup_{g:\{w_{-\infty:T-1}\}\to\{0,1\}} |\mathcal{W}_T(g|w_{-\infty:T-1}) - \mathcal{W}_T(g|w_{T-m:T-1})| \\
&\leq 2c \sup_{g:\{w_{T-m:T-1}\}\to\{0,1\}} |\bar{\mathcal{W}}(g|w_{T-m:T-1}) - \widehat{\mathcal{W}}(g|w_{T-m:T-1})| + 2 \cdot \text{w-bias}_\infty(m),
\end{aligned} \tag{A.18}$$

where $\text{w-bias}_\infty(m) := \sup_{g:\{w_{-\infty:T-1}\}\to\{0,1\}} |\mathcal{W}_T(g|w_{-\infty:T-1}) - \mathcal{W}_T(g|w_{T-m:T-1})|$. The last inequality follows by imposing an assumption similar to Assumption 2.4. For the second term of the last row, we should have $\lim_{m\to\infty} \text{w-bias}_\infty(m) = 0$ under mild regularity conditions. Thus, we can focus on the conditions required for the convergence of $|\bar{\mathcal{W}}(g|w_{T-m:T-1}) - \widehat{\mathcal{W}}(g|w_{T-m:T-1})|$.

Note that in the case of infinite Markov order, we must drop Assumption 2.1(i) completely. As a result, $\bar{\mathcal{W}}(g|w_{T-m:T-1}) - \widehat{\mathcal{W}}(g|w_{T-m:T-1})$ is no longer an average of an MDS since conditioning on $W_{t-m:t-1}$ is no longer equivalent to conditioning on $\mathcal{F}_{t-1}$ for any $m < \infty$.

To address this issue, we define, for a treatment path $w_{T-m:T-1} \in \{0,1\}^m$,

$$\bar{\mathcal{W}}(g|\mathcal{F}_{t-1}) := \frac{1}{T(w_{T-m:T-1})} \sum_{\substack{m\le t\le T-1: \\ W_{t-m:t-1}=w_{T-m:T-1}}} \mathsf{E}\left[Y_t(W_{-\infty:t-1},1)g(W_{t-m:t-1}) + Y_t(W_{-\infty:t-1},0)\left[1-g(W_{t-m:t-1})\right]|\mathcal{F}_{t-1}\right]. \tag{A.19}$$

Now, we can further decompose the first term of (A.18) into:

$$\begin{aligned}
&\sup_{g:\{w_{T-m:T-1}\}\to\{0,1\}} |\bar{\mathcal{W}}(g|w_{T-m:T-1}) - \widehat{\mathcal{W}}(g|w_{T-m:T-1})| \\
\le& \sup_{g:\{w_{T-m:T-1}\}\to\{0,1\}} |\bar{\mathcal{W}}(g|\mathcal{F}_{t-1}) - \widehat{\mathcal{W}}(g|w_{T-m:T-1})| + \sup_{g:\{w_{T-m:T-1}\}\to\{0,1\}} |\bar{\mathcal{W}}(g|\mathcal{F}_{t-1}) - \bar{\mathcal{W}}(g|w_{T-m:T-1})| \\
=& \sup_{g:\{w_{T-m:T-1}\}\to\{0,1\}} |\bar{\mathcal{W}}(g|\mathcal{F}_{t-1}) - \widehat{\mathcal{W}}(g|w_{T-m:T-1})| + \overline{\text{w-bias}}_\infty(m),
\end{aligned} \tag{A.20}$$

where $\overline{\text{w-bias}}_\infty(m) := \sup_{g:\{w_{T-m:T-1}\}\to\{0,1\}} |\bar{\mathcal{W}}(g|\mathcal{F}_{t-1}) - \bar{\mathcal{W}}(g|w_{T-m:T-1})|$. Summarizing (A.18) and (A.20), we obtain

$$\begin{aligned}
\mathcal{W}_T(g^*|w_{-\infty:T-1}) - \mathcal{W}_T(\hat{g}|w_{-\infty:T-1}) \le 2c &\sup_{g:\{w_{m:T-1}\}\to\{0,1\}} |\bar{\mathcal{W}}(g|\mathcal{F}_{t-1}) - \widehat{\mathcal{W}}(g|w_{T-m:T-1})| \\
&+ 2\cdot \widetilde{\text{w-bias}}_\infty(m),
\end{aligned}$$

where

$$\widetilde{\text{w-bias}}_\infty(m) := c\cdot \overline{\text{w-bias}}_\infty(m) + \text{w-bias}_\infty(m), \tag{A.21}$$

Furthermore, Assumption 2.3 is strengthened to: for $w \in \{0,1\}$,

$$Y_t(W_{-\infty:t-1}, w) \perp W_t | X_{-\infty:t-1}. \tag{A.22}$$

As mentioned above, the first statement of Assumption 2.1(ii) is maintained and modified for simplicity. The second statement of Assumption 2.1(ii) is modified to

$$W_t \perp X_{-\infty:t-1} | W_{t-m:t-1}. \tag{A.23}$$

This requires that conditioning on $W_{t-m:t-1}$, the truncated history, $X_{-\infty:t-m-1}$, does not have direct influences on W_t. In other words, the propensity score has limited overlaps: $e_t(\mathcal{F}_{t-1}) = \Pr(W_t = 1|\mathcal{F}_{t-1}) = \Pr(W_t = 1|W_{t-m:t-1})$.

Let's summarize. First, under conditions (A.22) and (A.23), the difference $\bar{\mathcal{W}}(g|\mathcal{F}_{t-1}) - \widehat{\mathcal{W}}(g|w_{T-m:T-1})$ is an average of an MDS. For a given m, it converges at $\frac{1}{\sqrt{T-m}}$ rate. Second, under additional regularity conditions of decay temporal dependence, we have $\lim_{m\to\infty}$ w-bias$_\infty(m) = 0$. Third, under the condition (A.16), we have

$$\bar{\mathcal{W}}(g|\mathcal{F}_{t-1}) = \frac{1}{T(w_{T-m:T-1})} \sum_{\substack{m\leq t\leq T-1:\\ W_{t-m:t-1}=w_{T-m:T-1}}} \mathsf{E}\left[Y_t(W_{-\infty:t-1},1)g(W_{t-m:t-1}) + Y_t(W_{-\infty:t-1},0)\left[1-g(W_{t-m:t-1})\right]|W_{-\infty:t-1}\right].$$

This equality leads to $\text{plim}_{m\to\infty}\overline{\text{w-bias}}_\infty(m) = 0$, given that T, $T(w_{T-m:T-1})$, and m diverge at appropriate rates, as well as some other additional regularity conditions of decay temporal dependence are met. Consequently, $\text{plim}_{m\to\infty}\widetilde{\text{w-bias}}_\infty(m) = 0$.

A.3.3 Example 2 continued

Now we verify whether the model specified in Example 2 satisfies the adjusted assumptions discussed in Appendix A.3.2. The assumptions adjusted for the case of infinite Markov order are summarized below:

a. The condition (A.22) of sequence unconfoundedness, $Y_t(W_{-\infty:t-1}, w) \perp W_t|X_{-\infty:t-1}$, which represents the adjusted Assumption 2.3.

b. The condition (A.23), $W_t \perp X_{-\infty:t-1}|W_{t-m:t-1}$, which represents the second statement of the adjusted Assumption 2.1(ii).

c. The modified Assumption 2.4 of the invariance of the welfare ordering, $\mathcal{W}_T(g^*|w_{T-m:T-1}) - \mathcal{W}_T(g|w_{T-m:T-1}) \leq c\big[\bar{\mathcal{W}}(g^*|w_{T-m:T-1}) - \bar{\mathcal{W}}(g|w_{T-m:T-1})\big]$.

d. The condition (A.16), $Y_t(w_{-\infty:t}) \perp X_{-\infty:t-1}|W_{-\infty:t-1}$, which represents the first statement of the adjusted Assumption 2.1(ii).

Firstly, the condition (A.22) is guaranteed by the independence between ε_t and W_t, conditional on $X_{-\infty:t-1}$. Secondly, the condition (A.23) is satisfied for any $m \geq 1$, as W_t is an AR(1) process and its distribution depends solely on W_{t-1}, as presented by (11).

Thirdly, regarding the invariance of welfare ordering, it holds that

$$\begin{aligned}
&\mathcal{W}_T(g^*|w_{T-m:T-1}) - \mathcal{W}_T(g|w_{T-m:T-1})\\
&= \mathrm{E}\left[\alpha + \beta_0 g^* + \sum_{i=1}^{\infty}\beta_i W_{T-i} + \sum_{i=0}^{\infty}\gamma_i\varepsilon_{T-i}|W_{T-m:T-1} = w_{T-m:T-1}\right]\\
&- \mathrm{E}\left[\alpha + \beta_0 g + \sum_{i=1}^{\infty}\beta_i W_{T-i} + \sum_{i=0}^{\infty}\gamma_i\varepsilon_{T-i}|W_{T-m:T-1} = w_{T-m:T-1}\right] = \beta_0\left(g^* - g\right),
\end{aligned}$$

and

$$
\begin{aligned}
&\bar{\mathcal{W}}(g^*|w_{T-m:T-1}) - \bar{\mathcal{W}}(g|w_{T-m:T-1}) \\
&= \frac{1}{T(w_{T-m:T-1})} \sum_{m \le t \le T-1: W_{t-m:t-1} = w_{T-m:T-1}} \mathrm{E}\left[\alpha + \beta_0 g^* + \sum_{i=1}^{\infty} \beta_i W_{t-i} + \sum_{i=0}^{\infty} \gamma_i \varepsilon_{t-i} | W_{T-m:T-1} = w_{T-m:T-1}\right] \\
&- \frac{1}{T(w_{T-m:T-1})} \sum_{m \le t \le T-1: W_{t-m:t-1} = w_{T-m:T-1}} \mathrm{E}\left[\alpha + \beta_0 g + \sum_{i=1}^{\infty} \beta_i W_{t-i} + \sum_{i=0}^{\infty} \gamma_i \varepsilon_{t-i} | W_{T-m:T-1} = w_{T-m:T-1}\right] \\
&= \frac{1}{T(w_{T-m:T-1})} \sum_{m \le t \le T-1: W_{t-m:t-1} = w_{T-m:T-1}} \beta_0 \left(g^* - g\right) = \beta_0 \left(g^* - g\right),
\end{aligned}
$$

where the second equality follows from the fact that the sequence $\{Y_t, W_t\}$, which is generated from (29) and (11) to (14), remains stationary. Consequently, Assumption (2.4) holds with $c = 1$.

Finally, as mentioned at the beginning of Appendix A.3.2, the purpose of the condition (A.16), which excludes the effects of past values of Y_t, is to ensure a simple and consistent discussion throughout Section 2. Although it may not be satisfied by the outcome equation (29), the linear structure of (29) ensures that under additional assumptions, our T-EWM approach for the case of infinite Markov order, as presented in Appendix A.3.2, remains valid.

To see this point, note that in Appendix A.3.2, the condition (A.16) is only used for showing $\mathrm{plim}_{m\to\infty} \overline{\text{w-bias}}_{\infty}(m) = 0$. Recall that $\overline{\text{w-bias}}_{\infty}(m) := \sup_{g:\{w_{T-m:T-1}\}\to\{0,1\}} |\bar{\mathcal{W}}(g; m|\mathcal{F}_{t-1}) - \bar{\mathcal{W}}(g|w_{T-m:T-1})|$. In Example 2, this difference becomes

$$
\begin{aligned}
&\bar{\mathcal{W}}(g; m|\mathcal{F}_{t-1}) - \bar{\mathcal{W}}(g|w_{T-m:T-1}) \\
&= \frac{1}{T(w_{T-m:T-1})} \sum_{m \le t \le T-1: W_{t-m:t-1} = w_{T-m:T-1}} \mathsf{E}\left[\alpha + \beta_0 g + \sum_{i=1}^{\infty} \beta_i W_{t-i} + \sum_{i=0}^{\infty} \gamma_i \varepsilon_{t-i} | \mathcal{F}_{t-1}\right] \\
&- \frac{1}{T(w_{T-m:T-1})} \sum_{m \le t \le T-1: W_{t-m:t-1} = w_{T-m:T-1}} \mathsf{E}\left[\alpha + \beta_0 g + \sum_{i=1}^{\infty} \beta_i W_{t-i} + \sum_{i=0}^{\infty} \gamma_i \varepsilon_{t-i} | W_{t-m:t-1} = w_{T-m:T-1}\right] \\
&= \frac{1}{T(w_{T-m:T-1})} \sum_{m \le t \le T-1: W_{t-m:t-1} = w_{T-m:T-1}} \left[\sum_{i=m+1}^{\infty} \beta_i W_{t-i} + \sum_{i=1}^{\infty} \gamma_i \varepsilon_{t-i}\right] \\
&- \frac{1}{T(w_{T-m:T-1})} \sum_{m \le t \le T-1: W_{t-m:t-1} = w_{T-m:T-1}} \mathsf{E}\left(\sum_{i=m+1}^{\infty} \beta_i W_{t-i} + \sum_{i=1}^{\infty} \gamma_i \varepsilon_{t-i} | W_{t-m:t-1} = w_{T-m:T-1}\right) \\
&= \frac{1}{T(w_{T-m:T-1})} \sum_{m \le t \le T-1: W_{t-m:t-1} = w_{T-m:T-1}} \left[\sum_{i=m+1}^{\infty} \beta_i W_{t-i} - \mathsf{E}\left(\sum_{i=m+1}^{\infty} \beta_i W_{t-i} | W_{t-m:t-1} = w_{T-m:T-1}\right)\right] \\
&+ \frac{1}{T(w_{T-m:T-1})} \sum_{m \le t \le T-1: W_{t-m:t-1} = w_{T-m:T-1}} \left[\sum_{i=1}^{\infty} \gamma_i \varepsilon_{t-i} - \mathsf{E}\left(\sum_{i=1}^{\infty} \gamma_i \varepsilon_{t-i} | W_{t-m:t-1} = w_{T-m:T-1}\right)\right] \\
&=: I_m + II_m,
\end{aligned}
$$

where the first term of the second equality follows from that the sum is over only those observations that satisfy $W_{t-m:t-1} = w_{T-m:T-1}$, and thus the path $W_{t-m:t-1}$ is fixed on

$w_{T-m:T-1}$ within both $\bar{\mathcal{W}}(g;m|\mathcal{F}_{t-1})$ and $\bar{\mathcal{W}}(g|w_{T-m:T-1})$.

Given that $\sum_{i=1}^{\infty}|\beta_i| < \infty$ holds and T, $T(w_{T-m:T-1})$, and m diverge at appropriate rates, we shall have $\text{plim}_{m\to\infty} I_m = 0$. For II_m, note that it is an average of an MA(∞) process, centered around its conditional mean given $W_{t-m:t-1} = w_{T-m:T-1}$. This should converge to zero if $\max_{l\geq 1}\sum_{i\geq l}|\gamma_i| \leq k\cdot l^{-\alpha}$ holds for some positive k and α, and the noise ε_t is iid with zero mean and has a finite second moment.

A.4 Proof of Lemma 3.1

Proof. We have

$$\begin{aligned} R_T(G) &= \int R_T(G|x)dF_{X_{T-1}}(x) \\ &= \int_{x\in A(x^{obs},G)} R_T(G|x)dF_{X_{T-1}}(x) + \int_{x\notin A(x^{obs},G)} R_T(G|x)dF_{X_{T-1}}(x) \\ &\geq R_T(G|x^{obs})\cdot p_{T-1}(x^{obs},G) + 0 = R_T(G|x^{obs})\cdot p_{T-1}(x^{obs},G). \end{aligned} \tag{A.24}$$

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

A.5 Proof of Theorem 3.1

We first show the following lemma.

LEMMA A.3. Under Assumptions 3.1 to 3.3, and 3.7 to 3.9, we have

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

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

Proof. h_t denotes a function belonging to the functional class $\mathcal{H}_t$, and $h = \{h_1, h_2, \ldots, h_n\}$. J_k is a cover of the functional class $\mathbf{H_n}$ with radius $2^{-k}M$ with respect to the $\rho_{2,n}(.)$ norm, and $k = 1, \cdots, \overline{K}$. We set $2^{-\overline{K}} \asymp \sqrt{n}^{-1}$, then $\overline{K} \asymp \log(n)$. Recall that M is the constant defined in Assumption 3.7, which implies $\max_t |h_t| \leq M$. We define $h^* = \arg\max_{h\in\mathbf{H_n}} \mathsf{E}_n h_.$. Let $h^{(k)} = \min_{h\in J_k}\rho_{2,n}(h,h^*)$ and $h^{(0)} = (0,\cdots,0) \in \mathbb{R}^n$, then $\rho_{2,n}(h^{(k)},h^*) \leq 2^{-k}M$ holds by the definition of J_k, and

$$\rho_{2,n}(h^{(k-1)}, h^{(k)}) \leq \rho_{2,n}(h^{(k-1)}, h^*) + \rho_{2,n}(h^{(k)}, h^*) \leq 3 \cdot 2^{-k} M. \tag{A.25}$$

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

$$\sum_{t=1}^{n} h_t \leq |\sum_{t=1}^{n} h_t^{(0)}| + |\sum_{k=1}^{\overline{K}} \sum_{t=1}^{n} (h_t^{(k)} - h_t^{(k-1)})| + |\sum_{t=1}^{n} (h_t^{(\overline{K})} - h_t^*)|. \tag{A.26}$$

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 $h^{(\overline{K})}$

$$|\sum_{t=1}^{n} (h_t^{(\overline{K})} - h_t^*)| \leq |(\sum_{t=1}^{n} (h_t^{(\overline{K})} - h_t^*)^2)^{1/2}|\sqrt{n} \leq n 2^{-\overline{K}} M. \tag{A.27}$$

Thus,

$$\mathsf{E}(|\,\mathsf{E}_n h|_{\mathbf{H_n}}) \leq \sum_{k}^{\overline{K}-1} \mathsf{E} \max_{f \in J_k, g \in J_{k-1}, \rho_{2,n}(f,g) \leq 3 \cdot 2^{-k} \cdot M} |\,\mathsf{E}_n(f-g)| + 2^{-\overline{K}} M. \tag{A.28}$$

Apply Lemma A.2 and Assumption 3.9 to (A.28). The maximal inequality (A.6) of Lemma A.2 is reproduced here:

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

where for the first term of (A.28), we have $\mathcal{A} = \{f - g : f \in J_k, g \in J_{k-1}, \rho_{2,n}(f,g) \leq 3 \cdot 2^{-k} \cdot M\}$, $|\mathcal{A}| = |J_k||J_{k-1}| \leq 2\mathcal{N}^2(2^{-k}M, \mathbf{H_n}, \rho_{2,n}(.)) \lesssim_p 2 \max_t \sup_Q \mathcal{N}^2(2^{-k}M, \mathcal{H}_t, \|.\|_{Q,2})$, and $A \leq 3M$. B in (A.6) 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 any pair (f,g) satisfying $f - g \in \mathcal{A}$.

Therefore, by Lemma A.2,

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

By Assumption 3.8, $\max_t \log \sup_Q \mathcal{N}(\varepsilon M, \mathcal{H}_t, \|.\|_{Q,2}) \lesssim \log(K) + \log(v+1) + (v+1)(\log 4 + 1) + (rv)\log(\frac{2}{\varepsilon M})$. Thus, the integral in the last row is finite by a standard argument for bracketing numbers (see, e.g., the comment following Theorem 19.4 in Van der Vaart, 2000). Then, we have $\mathsf{E}(|\mathsf{E}_n h|_{\mathbf{H_n}}) \lesssim M\sqrt{v/n}$. $\square$

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

LEMMA A.4. Under Assumptions 3.1 to 3.3, and 3.7 to 3.9,

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

where c_n is an 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 A.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' such that $c_n' \ll c_n$. We set $x = \mathsf{E}(|\mathsf{E}_n h|_{\mathbf{H_n}}) + c_n\sqrt{v/n}$. The result then follows. $\square$

Finally, Theorem 3.1 follows by combining Lemma A.3, Lemma A.4, and $n = T-1$.

A.6 Proof of Theorem 3.2

In this proof, we will show the concentration of $I = \frac{1}{T-1}\sum_{t=1}^{T-1}\tilde{S}_t(G) + \frac{1}{T-1}\sum_{t=1}^{T-1}\overline{S}_t(G)$. Under Assumptions 3.8 and 3.13 and with similar arguments as those for Lemma A.4, the first term of I satisfies:

$$\sup_{G\in\mathcal{G}}\frac{1}{T-1}\sum_{t=1}^{T-1}\overline{S}_t(G)\lesssim_p M\sqrt{v}/\sqrt{T-1}. \tag{A.29}$$

The concentration rate of the second term of I, $\frac{1}{T-1}\sum_{t=1}^{T-1}\tilde{S}_t(G)$, is shown through the following two steps:

(i) For a finite function class $\mathcal{G}$ with $|\mathcal{G}|=\tilde{M}<\infty$ and under Assumptions 3.10- 3.12,

$$\sup_{G\in\mathcal{G}}\frac{1}{T-1}\sum_{t=1}^{T-1}\tilde{S}_t(G)\lesssim_p \frac{c_T[\log(\tilde{M})2e\gamma]^{1/\gamma}\sup_G\Phi_{\phi_{\tilde{v}}}(\tilde{S}_.(G))}{\sqrt{T-1}} \tag{A.30}$$

holds with probability $1-\exp(-c_T^\gamma)$, where c_T is a large enough constant, and $\Phi_{\phi_{\tilde{v}}}(\tilde{S}_.(G))$ is defined in Assumption 3.12.

(ii) Next, we extend (A.30) to obtain the main result of Theorem 3.2: Let $\mathcal{G}$ be a function class with infinite elements, and its complexity is subject to Assumption 3.13. Under Assumptions 3.10- 3.13, we have

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

Step (i). By Assumption 3.1 and 3.11(i), we have $\mathsf{E}(S_t(G)|\mathcal{F}_{t-2})=\mathsf{E}(S_t(G)|X_{t-2})$. By Assumption 3.10, we have $X_t=g_t(\varepsilon_t,\varepsilon_{t-1},\cdots)$. Define $X^*_{t-2,l}$ as a version of the random variable X_{t-2}, in which ε_{t-2-l} is replaced by ε^*_{t-2-l}: By definition of $\tilde{S}(G)$, we have

$$\Phi_{\phi_{\tilde{v}}}(\tilde{S}_.(G))=\sup_{q\geq 2}(\sum_{l\geq 0}\max_t\|\mathsf{E}(S_t(G)|X_{t-2})-\mathsf{E}(S_t(G)|X^*_{t-2,l})\|_q)/q^{\tilde{v}}.$$

Then, by Assumption 3.12, we have $\sup_G\Phi_{\phi_{\tilde{v}}}(\tilde{S}_.(G))<\infty$. For $\gamma=1/(1+2\tilde{v})$ and a finite $\mathcal{G}$ with $|\mathcal{G}|=\tilde{M}$, the bound (A.30) follows by Theorem 3 of Wu and Wu (2016):

$$\Pr(\sup_{G\in\mathcal{G}}\sum_{t=1}^{T-1}\tilde{S}_t(G)\geq x)\leq\tilde{M}\exp\left[-\left(\frac{x}{\sqrt{T-1}\sup_G\Phi_{\phi_{\tilde{v}}}(\tilde{S}_.(G))}\right)^\gamma\frac{1}{2e\gamma}\right]. \tag{A.31}$$

Specifically, let us set $x=c_T[\log(\tilde{M})2e\gamma]^{1/\gamma}\Phi_{\phi_{\tilde{v}}}(\tilde{S}_.(G))\sqrt{T-1}$, where c_T is a sufficiently large constant, then

$$\sup_{G\in\mathcal{G}}\frac{1}{T-1}\sum_{t=1}^{T-1}\tilde{S}_t(G)\lesssim c_T[\log(\tilde{M})2e\gamma]^{1/\gamma}\sup_{G\in\mathcal{G}}\Phi_{\phi_{\tilde{v}}}(\tilde{S}_.(G))/\sqrt{T-1} \tag{A.32}$$

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

Step (ii). We now consider the case where $\mathcal{G}$ is not finite. We define $\mathcal{G}^{(1)\delta}$ to be a $\delta \max_t \sup_Q \|\tilde{F}_t\|_{Q,2}$-net within $\mathbf{F}_n$ w.r.t. $\mathcal{G}$. We denote $\tilde{S}_t(\pi(G))$ as the closest component to $\tilde{S}_t(G)$ in the net $\mathcal{G}^{(1)\delta}$. Then,

$$\begin{aligned}
\sup_G \frac{1}{T-1}\sum_t \tilde{S}_t(G) &\leq \sup_{G\in\mathcal{G}} \left| \frac{1}{T-1}\sum_{t=1}^{T} \left[\tilde{S}_t(G) - \tilde{S}_t(\pi(G))\right] \right| + \sup_{G\in\mathcal{G}^{(1)\delta}} \left|\frac{1}{T-1}\sum_{t=1}^{T-1}(\tilde{S}_t(G))\right| \\
&\lesssim_p \delta \max_t \sup_Q \|\tilde{F}_t\|_{Q,2} + \sup_{G\in\mathcal{G}^{(1)\delta}} \left|\frac{1}{T-1}\sum_{t=1}^{T-1}(\tilde{S}_t(G))\right| \\
&\lesssim_p \delta \max_t \sup_Q \|\tilde{F}_t\|_{Q,2} + c_T[V\log(1/\delta)2e\gamma]^{1/\gamma} \sup_{G\in\mathcal{G}} \Phi_{\phi_{\tilde{v}}}(\tilde{S}_.(G))/\sqrt{T-1},
\end{aligned}$$

where V and δ following the first $\lesssim_p$ are the constants in the first statement of Assumption 3.13. Recall that $\sup_{G\in\mathcal{G}} \Phi_{\phi_{\tilde{v}}}(\tilde{S}_.(G))$ is finite by Assumption 3.12. By setting $\delta = \frac{1}{T}$, we obtain

$$\sup_{G\in\mathcal{G}} \frac{1}{T-1}\sum_{t=1} \tilde{S}_t(G) \lesssim_p c_T[V\log(T)2e\gamma]^{1/\gamma} \sup_{G\in\mathcal{G}} \Phi_{\phi_{\tilde{v}}}(\tilde{S}_.(G))/\sqrt{T-1}. \qquad \text{(A.33)}$$

A.7 Justifying Assumption 3.8

Here we interpret the entropy condition in Assumption 3.8 in a more general way. We follow the argument of Chapter 11 in Kosorok (2008) for the functional class related to non-i.i.d. data.

First, we illustrate the case that the random stochastic process

$$\{D_t\}_{t=-\infty}^{\infty} \stackrel{\text{def}}{=} \{Y_t, W_t, X_{t-1}\}_{t=-\infty}^{\infty}$$

is assumed to be stationary. In this case, we have for all $t \in \{1, 2, \ldots, n\}$,

$$\begin{aligned}
&h_t(\cdot; G) = h(\cdot; G) \\
&\text{and } \mathcal{H} = \{h(\cdot; G) = \widehat{\mathcal{W}}_t(G) - \bar{\mathcal{W}}_t(G) : G \in \mathcal{G}\}.
\end{aligned}$$

For an n-dimensional non-negative vector $\alpha_n = \{\alpha_{n,1}, \alpha_{n,2}, \ldots, \alpha_{n,n}\}$, define Q_{α_n} as a discrete measure with probability mass $\frac{\alpha_{n,t}}{\sum_{t=1}^n \alpha_{n,t}}$ on the value D_t. Recall for a function f, its $L_r(Q)-$norm is denoted by $||f||_{Q,r} \stackrel{\text{def}}{=} \left[\int_v |f(v)|^r dQ(v)\right]^{1/r}$. Thus, given a sample $\{D_t\}_{t=1}^n$ and $h(\cdot; G)$, we have $||h(\cdot; G)||_{Q_{\alpha_n},2} = \left[\frac{1}{\sum_{t=1}^n \alpha_{n,t}} \sum_t \alpha_{n,t} h(D_t; G)^2\right]^{1/2}$.

In the stationary case, we define the *restricted* function class $\mathbf{H_n} = \{h_{1:n} : h_{1:n} \in \mathcal{H} \times \mathcal{H} \times \cdots \times \mathcal{H}, h_1 = h_2 = \cdots = h_n\}$ and the envelope function $\overline{H}_n = (H_1, H_2, \cdots, H_n)'$. Furthermore, let $\mathcal{Q}$ denote the class of all discrete probability measures on the domain of the random vector $D_t = \{Y_t, W_t, X_{t-1}\}$. Recall for an n-dimensional vector $v = \{v_1, \ldots, v_n\}$, its l_2-norm is denoted by $|v|_2 \stackrel{\text{def}}{=} (\sum_{i=1}^n v_i^2)^{1/2}$. Then, for any $\alpha_n \in \mathbb{R}_+^n$ and $\tilde{\alpha}_{n,t} = \frac{\sqrt{\alpha_{n,t}}}{\sqrt{\sum_t \alpha_{n,t}}}$, we have

$$\mathcal{N}(\delta|\tilde{\alpha}_n \circ \overline{H}_n|_2, \tilde{\alpha}_n \circ \mathbf{H}_n, |.|_2) = \mathcal{N}(\delta||H||_{Q_{\alpha_n},2}, \mathcal{H}, ||.||_{Q_{\alpha_n},2}) \leq \sup_{Q\in\mathcal{Q}} \mathcal{N}(\delta||H||_{Q,2}, \mathcal{H}, ||.||_{Q,2}).$$

In light of this relationship in the stationary case, we generalize the setup. Let $\mathcal{K}$ be a subset of $\{1, \cdots, n\}$, and its dimension is $K = |\mathcal{K}|$. Let $\alpha_{n,K}$ denote the K dimensional subvector of α_n corresponding to the index set $\mathcal{K}$. Recall that H_t denotes the envelope function of $\mathcal{H}_t$, the functional class corresponding to $\{h_t(.,G), G \in \mathcal{G}\}$, and $\mathbf{H_n} = \mathcal{H}_1 \times \mathcal{H}_2 \times \cdots \times \mathcal{H}_n$. Then, for the subset $\mathcal{K}$, we similarly define $\mathbf{H_{n,K}}$ as the corresponding functional class, and the vector of its envelope functions as $\overline{H}_{n,K}$.

Let us consider the following assumption: for any fixed K, the covering number can effectively be reduced to K dimension, i.e.,

$$\begin{aligned}
\mathcal{N}(\delta|\tilde{\alpha}_n \circ \overline{H}_n|_2, \tilde{\alpha}_{\mathbf{n}} \circ \mathbf{H_n}, |.|_2) &\leq \max_{\mathcal{K}\subseteq\{1:n\}} \mathcal{N}(\delta|\tilde{\alpha}_{n,K} \circ \overline{H}_{n,K}|_2, \tilde{\alpha}_{n,K} \circ \mathbf{H_{n,K}}, |.|_2), \\
&\leq \max_{\mathcal{K}\subseteq\{1:n\}} \prod_{t\in\mathcal{K}} \mathcal{N}(\delta|\tilde{\alpha}_{n,t} \cdot H_t|_2, \tilde{\alpha}_{n,t} \cdot \mathcal{H}_t, |.|_2), \\
&\leq \sup_{Q\in\mathcal{Q}} \max_{t\in\mathcal{K}} \max_{\mathcal{K}\subseteq\{1:n\}} \mathcal{N}(\delta\|H_t\|_{Q,2}, \mathcal{H}_t, \|.\|_{Q,2})^K.
\end{aligned}$$

The first inequality of (46) in Assumption 3.8 holds by setting $K = |\mathcal{K}| = 1$. Under this assumption, it suffices to examine the one-dimensional covering number $\mathcal{N}(\delta\|H_t\|_{Q,2}, \mathcal{H}_t, \|.\|_{Q,2})$.

To provide an example for the second inequality of (46), we let $\mathbf{1}(X_{t-1} \in G) = \mathbf{1}(X_{t-1}^\top \theta \leq 0)$ for some $\theta \in \Theta$, where Θ is a compact set in $\mathbb{R}^d$. Without loss of generality, we assume that Θ is the unit ball in $\mathbb{R}^d$, i.e., for all $\theta \in \Theta$: $|\theta|_2 \leq 1$. For $w = 0$ and 1, define $S_{t,w} \stackrel{\text{def}}{=} Y_t(w)\mathbf{1}(W_t = w)/\Pr(W_t = w|X_{t-1})$, where $Y_t(w)$ is the abbreviation of $Y_t(W_{t-1}, w)$. Under Assumption 3.7, $|Y_t(1)|, |Y_t(0)| \leq M/2$, so we have $\|H_t\|_{Q,r} \leq M + M/\kappa \stackrel{\text{def.}}{=} M'$, and

$$\begin{aligned}
h_t^\theta \;=\; & \mathsf{E}(Y_t(1)\mathbf{1}(X_{t-1}^\top\theta \leq 0)|X_{t-1}) + \mathsf{E}(Y_t(0)\mathbf{1}(X_{t-1}^\top\theta > 0)|X_{t-1}) \\
& - S_{t,1}\mathbf{1}(X_{t-1}^\top\theta \leq 0) - S_{t,0}\mathbf{1}(X_{t-1}^\top\theta > 0).
\end{aligned}$$

The corresponding functional class can be written as

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

where $f_{1,1}^\theta$ (resp. $f_{1,0}^\theta, f_{0,1}^\theta$, and $f_{0,0}^\theta$) corresponds to $\mathsf{E}(Y_t(1)\mathbf{1}(X_{t-1}^\top\theta \leq 0)|X_{t-1})$ (resp. $\mathsf{E}(Y_t(0)\mathbf{1}(X_{t-1}^\top\theta > 0)|X_{t-1}), -S_{t,1}\mathbf{1}(X_{t-1}^\top\theta \leq 0)$, and $-S_{t,0}\mathbf{1}(X_{t-1}^\top\theta > 0)$). Let the corresponding functional class be denoted by $\mathcal{F}_{1,1}$ (resp. $\mathcal{F}_{1,0}, \mathcal{F}_{0,1}$, and $\mathcal{F}_{0,0}$). For all finitely discrete norms Q and any positive ε, we know that

$$\sup_Q \mathcal{N}(\varepsilon, \mathcal{H}_t, \|.\|_{Q,r}) \leq \sup_Q \mathcal{N}(\varepsilon/4, \mathcal{F}_{1,1}, \|.\|_{Q,r})\mathcal{N}(\varepsilon/4, \mathcal{F}_{1,0}, \|.\|_{Q,r})\mathcal{N}(\varepsilon/4, \mathcal{F}_{0,1}, \|.\|_{Q,r})\mathcal{N}(\varepsilon/4, \mathcal{F}_{0,0}, \|.\|_{Q,r}). \tag{A.34}$$

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 $d+2$ since the class $\{x \in \mathbb{R}^d, x^\top \theta \leq 0, \theta \in \Theta\}$ is of VC dimension less than $d+2$ (see the proof of Lemma 9.6 of Kosorok (2008)). Therefore, we have $\sup_Q \mathcal{N}(\varepsilon/4, \mathcal{F}_{0,1}, \|.\|_{Q,r}) \vee \mathcal{N}(\varepsilon/4, \mathcal{F}_{0,0}, \|.\|_{Q,r}) \lesssim (4/(\varepsilon M'))^{d+2}$.

Moreover, we impose the following Lipschitz condition on functions $f_{1,1}^\theta$ and $f_{1,0}^\theta$: For any distinct points $\theta, \theta' \in \Theta$ and a positive constant M_d, it holds that $||f_{1,1}^\theta - f_{1,1}^{\theta'}||_{Q,r} \leq M_d|\theta - \theta'|_r$ (a similar equality holds for $f_{1,0}^\theta$). Then, it falls within the type II class defined in Andrews (1994), so according to the derivation of the (A.2) Andrews (1994) we have

$$\sup_Q \mathcal{N}(\varepsilon M', \mathcal{F}_{1,1}, \|.\|_{Q,r}) \leq \sup_Q \mathcal{N}(\varepsilon M'/M_d, \Theta, \|.\|_{Q,r}), \tag{A.35}$$

where the latter is the covering number of an Euclidean ball under the norm $\|.\|_{Q,r}$. Thus, according to Equation (5.9) in Wainwright (2019), $\sup_Q \mathcal{N}(\varepsilon M'/M_d, \Theta, \|.\|_{Q,r}) \lesssim (1+2M_d/\varepsilon M')^d$. Combining the above results, we have $\sup_Q \mathcal{N}(\varepsilon M', \mathcal{H}_t, \|.\|_{Q,r}) \lesssim (4/(\varepsilon M'))^{2(d+2)}(1+2M_d/(\varepsilon M'))^{2d}$. Finally, with some rearrangement and redefinition of constant terms, we can obtain (46).

A.8 Accounting for the Lucas critique

Here we solve the VAR reduced form of the three-equation New Keynesian model discussed in Section 4.2.

Recall that $\tilde{Y}_t := \begin{pmatrix} x_t \\ \pi_t \end{pmatrix}$ and $d_t := \begin{pmatrix} v_t \\ \varepsilon_t \end{pmatrix}$. Rearranging (55) yields

$$\begin{pmatrix} \mathsf{E}_t\, x_{t+1} \\ \mathsf{E}_t\, \pi_{t+1} \end{pmatrix} = \begin{pmatrix} \kappa/\beta + 1 & \delta/\sigma - 1/(\sigma\beta) \\ -\kappa/\beta & 1/\beta \end{pmatrix} \begin{pmatrix} x_t \\ \pi_t \end{pmatrix} + \begin{pmatrix} v_t/\sigma + \varepsilon_t/(\sigma\beta) \\ -\varepsilon_t/\beta \end{pmatrix}. \tag{A.36}$$

Define

$$N = \begin{pmatrix} \kappa/(\sigma\beta) + 1 & \delta/\sigma - 1/(\sigma\beta) \\ -\kappa/\beta & 1/\beta \end{pmatrix}, \quad C = -\begin{pmatrix} 1/\sigma & 1/\sigma\beta \\ 0 & -1/\beta \end{pmatrix}.$$

By assuming that N is invertible, we can define $A = N^{-1}$. Recall that d_t is assumed to be an AR(1) process: $d_t = Fd_{t-1} + \eta_t$, where $F = \begin{pmatrix} \rho & 0 \\ 0 & \gamma \end{pmatrix}$. Then, for $\Gamma_t := ACd_t$, equation (A.36) can be written as $\tilde{Y}_t = A\,\mathsf{E}_t\tilde{Y}_{t+1} + ACd_t$. Solving forward, we obtain $\tilde{Y}_t = \lim_{L\to\infty} A^L\,\mathsf{E}_t(\tilde{Y}_{t+L}) + \sum_{l\geq 0} A^{l+1}C\,\mathsf{E}_t\,(d_{t+l})$.

By assuming $\det(A) < 1$, we have $\lim_{L\to\infty} A^L \mathsf{E}_t(\tilde{Y}_{t+L}) \to_{a.s.} 0$. Thus, $\tilde{Y}_t$ can be solved as

$$\tilde{Y}_t = M(\rho)d_t, \tag{A.37}$$

where $M(\rho) := \sum_{l\geq 0} A^{l+1}CF^l = \sum_{l\geq 0} A^{l+1}C \begin{pmatrix} \rho & 0 \\ 0 & \gamma \end{pmatrix}^l$. Assuming that $M(\rho)$ is invertible, by iterating (A.37), we can solve the VAR reduced form:

$$\tilde{Y}_t = M(\rho) F [M(\rho)]^{-1} \tilde{Y}_{t-1} + M(\rho)\eta_t.$$

A.9 Proof of Theorem 4.1

Under policy G, we use the following notation: $\widetilde{\mathcal{W}}(G)$ is defined in (41); $\widehat{\mathcal{W}}(G)$ represents the estimated welfare defined in (39); $\widehat{\mathcal{W}}^{\hat{e}}(G)$ represents the estimated welfare, with the estimated propensity score $\hat{e}(\cdot)$; and

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

Recall that G_* is the optimal policy defined in (36). Let $\hat{G}^{\hat{e}}$ be the optimal policy estimated using the estimated propensity score $\hat{e}(\cdot)$,

$$\hat{G}^{\hat{e}} \in \operatorname*{argmax}_{G\in\mathcal{G}} \widehat{\mathcal{W}}^{\hat{e}}(G). \tag{A.38}$$

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

$$\begin{aligned}
&\widetilde{\mathcal{W}}(G_*) - \widetilde{\mathcal{W}}(\hat{G}^{\hat{e}}) \\
&= \widetilde{\mathcal{W}}(G_*) - \widetilde{\mathcal{W}}(\hat{G}^{\hat{e}}) + \left[\widehat{\mathcal{W}}^{\hat{e}}(G_*) - \widehat{\mathcal{W}}^{\hat{e}}(G_*)\right] + \left[\widehat{\mathcal{W}}(\hat{G}^{\hat{e}}) - \widehat{\mathcal{W}}(\hat{G}^{\hat{e}})\right] + \left[\widehat{\mathcal{W}}(G_*) - \widehat{\mathcal{W}}(G_*)\right] \\
&\leq \widetilde{\mathcal{W}}(G_*) - \widetilde{\mathcal{W}}(\hat{G}^{\hat{e}}) + \left[\widehat{\mathcal{W}}^{\hat{e}}(\hat{G}^{\hat{e}}) - \widehat{\mathcal{W}}^{\hat{e}}(G_*)\right] + \left[\widehat{\mathcal{W}}(\hat{G}^{\hat{e}}) - \widehat{\mathcal{W}}(\hat{G}^{\hat{e}})\right] + \left[\widehat{\mathcal{W}}(G_*) - \widehat{\mathcal{W}}(G_*)\right] \\
&= \left[\widehat{\mathcal{W}}(G_*) - \widehat{\mathcal{W}}^{\hat{e}}(G_*) - \widehat{\mathcal{W}}(\hat{G}^{\hat{e}}) + \widehat{\mathcal{W}}^{\hat{e}}(\hat{G}^{\hat{e}})\right] + \left[\widetilde{\mathcal{W}}(G_*) - \widetilde{\mathcal{W}}(\hat{G}^{\hat{e}}) - \widehat{\mathcal{W}}(G_*) + \widehat{\mathcal{W}}(\hat{G}^{\hat{e}})\right] \\
&= I^{\hat{e}} + II^{\hat{e}},
\end{aligned} \tag{A.39}$$

where the first inequality comes from $\widehat{\mathcal{W}}^{\hat{e}}(\hat{G}^{\hat{e}}) \geq \widehat{\mathcal{W}}^{\hat{e}}(G_*)$, which is implied by the definition (A.38).

For $II^{\hat{e}}$, we know $II^{\hat{e}} \leq 2\sup_{G\in\mathcal{G}} |\widetilde{\mathcal{W}}(G) - \widehat{\mathcal{W}}(G)|$. Similar arguments to Section 3.2 can then be used to bound it. For $I^{\hat{e}}$, Note that for any $G \in \mathcal{G}$,

$$\widehat{\mathcal{W}}(G) = \frac{1}{T-1}\sum_{t=1}^{T-1}\left[\tau_t 1(X_{t-1} \in G) + \frac{Y_t(1-W_t)}{1-e_t(X_{t-1})}\right]. \tag{A.40}$$

Similarly,

$$\widehat{\mathcal{W}}^{\hat{e}}(G) = \frac{1}{T-1}\sum_{t=1}^{T-1}\left[\hat{\tau}_t 1(X_{t-1} \in G) + \frac{Y_t(1-W_t)}{1-\hat{e}_t(X_{t-1})}\right]. \tag{A.41}$$

Combining (A.40) and (A.41) with $I^{\hat{e}}$,

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

Then,

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

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

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

The statement of Theorem 4.1 follows from (61) and Theorem 3.3.

A.10 Link to Markov decision problems

In this section, we show the connection between our T-EWM setup and models of the Markov Decision Process (MDP). For MDP, we adapt the notation of Kallenberg (2016) (LK hereafter), an online lecture note by Lodewijk Kallenberg. As described in LK, the MDP is the set of models for making decisions for dependent data. An 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 a 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. In this subsection, we show a formal link between our T-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 T-EWM policy in our framework.

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

$$p_{ij}^t(a) \quad = \Pr(W_t = j | W_{t-1} = i; \text{choosing } W_t = a). \tag{A.42}$$

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. We assume that after time $T-1$, the planner implements a deterministic policy, so the probability only takes values in $\{0, 1\}$, i.e.,

$$p_{ij}^t(a) \quad = 1 \text{ if } j = a, \quad p_{ij}^t(a) \quad = 0 \text{ if } j \neq a. \tag{A.43}$$

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

$$r_i^t(a) = \mathsf{E}\left[Y_t(a) | W_{t-1} = i\right]. \tag{A.44}$$

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,

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

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 (of length 2) and the finite-period welfare function

$$\begin{aligned} v_i^{T:T+1}(R) = \mathsf{E}_{i,R}\left[\sum_{k=T}^{T+1} r_i^k(W_k)\right] &= \mathsf{E}\Big[Y_T(1)p_{i1}^T\left(g_1(W_{T-1})\right) + Y_T(0)p_{i0}^T\left(g_1(W_{T-1})\right) \\ &\quad + Y_{T+1}(1)p_{i1}^{T+1}\left(g_2(W_T)\right) + Y_{T+1}(0)p_{i0}^{T+1}\left(g_2(W_T)\right)|W_{T-1} = i\Big] \\ &= \mathsf{E}\left[Y_T(1)g_1(W_{T-1}) + Y_T(0)\left[1 - g_1(W_{T-1})\right]|W_{T-1} = i\right] \\ &\quad + \mathsf{E}\left[Y_{T+1}(1)g_2\left(W_T\right) + Y_{T+1}(0)\left[1 - g_2\left(W_T\right)\right]|W_T = g_1(i)\right]. \end{aligned} \tag{A.45}$$

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 (A.43) and the exclusion condition.

Comparing (A.45) with (53), we can view the population conditional welfare of T-EMW with a finite-period target as the value function of a finite-horizon MDP with a non-stationary solution. According to LK, in this case, the policy is usually obtained by using a backward induction algorithm.

A.11 Simulation

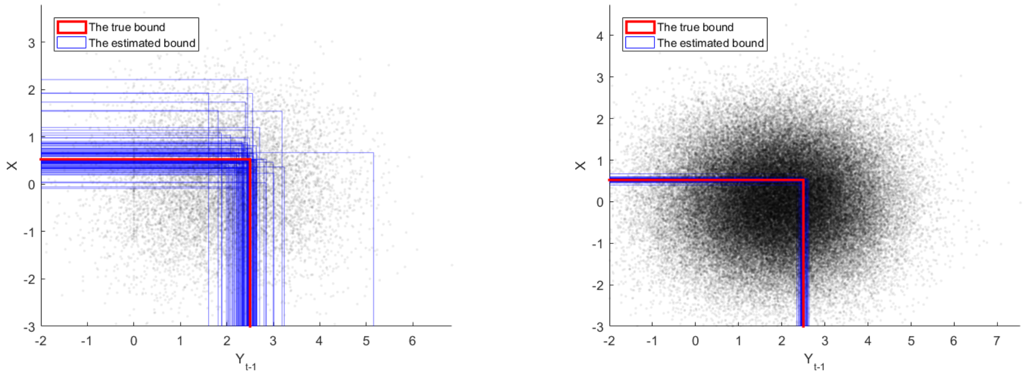
In this subsection, we illustrate the accuracy of our method through a simple simulation exercise. We consider the following model

$$\begin{aligned} Y_t &= W_t \cdot \mu(Y_{t-1}, Z_{t-1}) + \phi Y_{t-1} + \varepsilon_t, \\ \mu(Y_{t-1}, Z_{t-1}) &= 1(Y_{t-1} < B_1) \cdot 1(Z_{t-1} < B_2) - 1(Y_{t-1} > B_1 \vee Z_{t-1} > B_2), \end{aligned} \tag{A.46}$$

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 defined as

$$\mathcal{G} \equiv \left\{ \begin{array}{c} ((y_{t-1}, z_{t-1}) : s_1(y_{t-1} - b_1) > 0 \ \& \ s_2(z_{t-1} - b_2) > 0), \\ s_1, s_2 \in \{-1, 1\}, b_1, b_2 \in \mathbb{R} \end{array} \right\}. \tag{A.47}$$

It is immediate that $G_{\text{FB}}^* \in \mathcal{G}$. Therefore, we can directly estimate the unconditional treatment 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 2 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 multiply the variances and MSEs by the sample size n. The sample sizes are n = 100, 500, 1000, and 2000. The number of Monte Carlo replications is 500.

Table 2. 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 MSEs shrink, even after multiplying by the sample size n, which suggests that the convergence rate in this case is faster than $\frac{1}{\sqrt{n}}$.

A.12 Additional figures and results for the empirical application

A.12.1 Time series plots of the raw data and estimated propensity scores

Figure 4. Weekly cases and deaths from 4-2020 to 1-2022

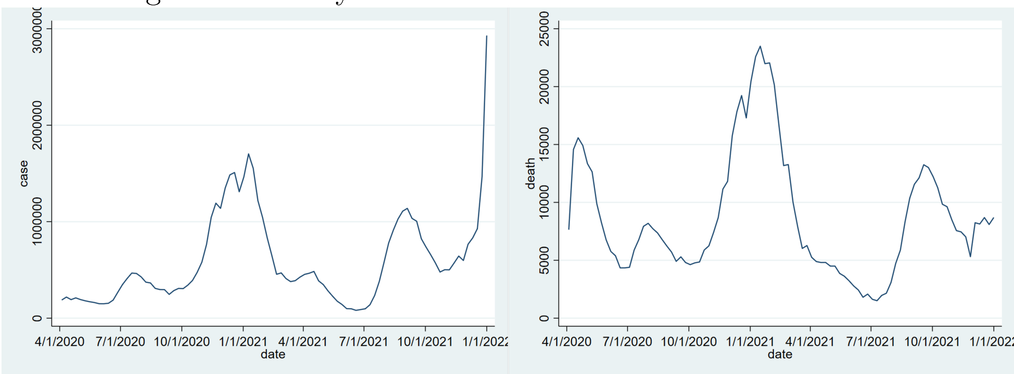


Figure 5. Restriction level and economic condition from 4-2020 to 1-2022

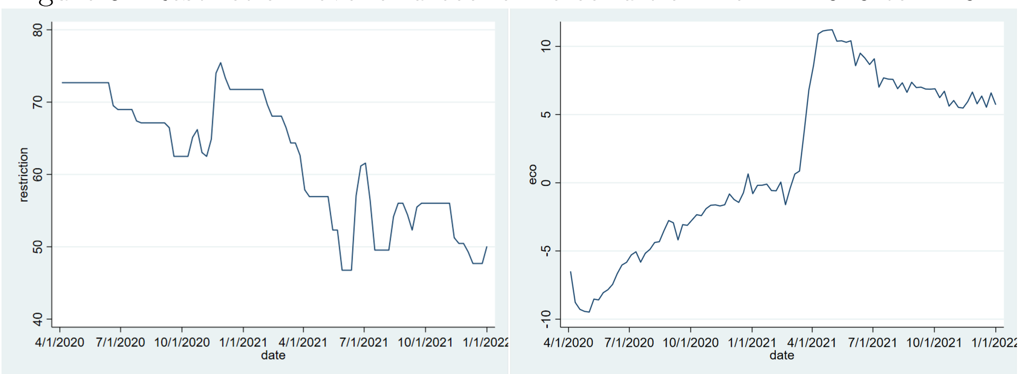
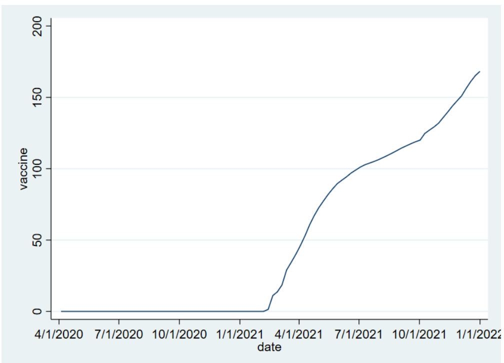
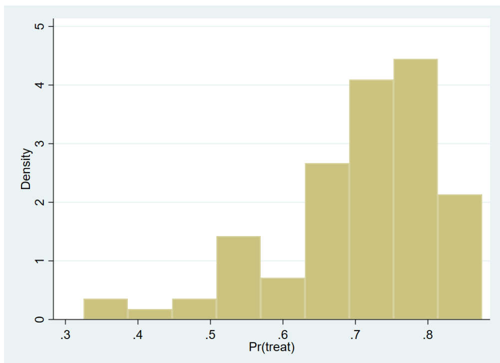


Figure 6b presents the propensity score estimated by (63), in which the set of covariates X_{t-1} is given by (62) (the case of Markov order $q = 1$).



(a) Vaccine coverage from 4-20 to 1-22

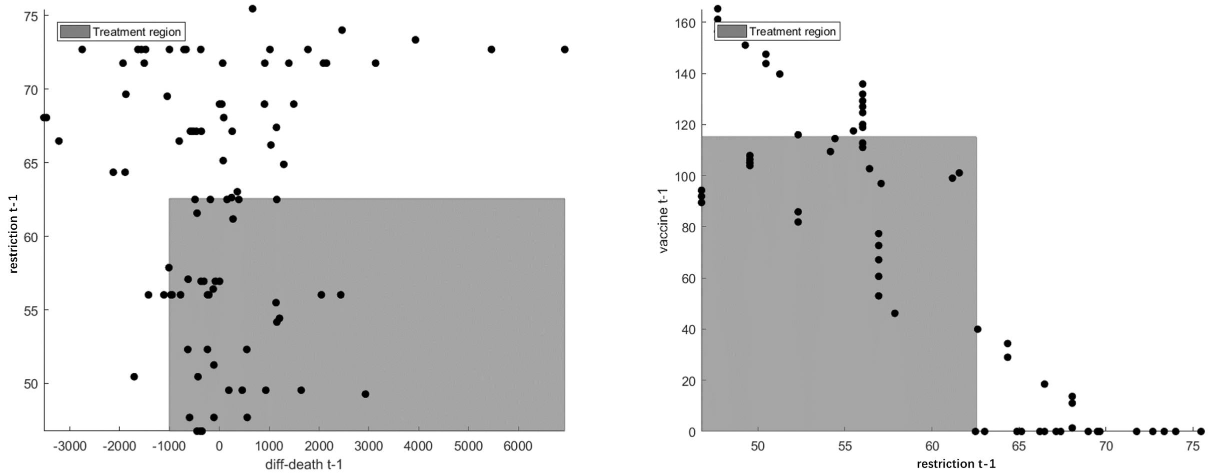


(b) Estimated PS for $q = 1$

Figure 6. Vaccine coverage and estimated propensity scores

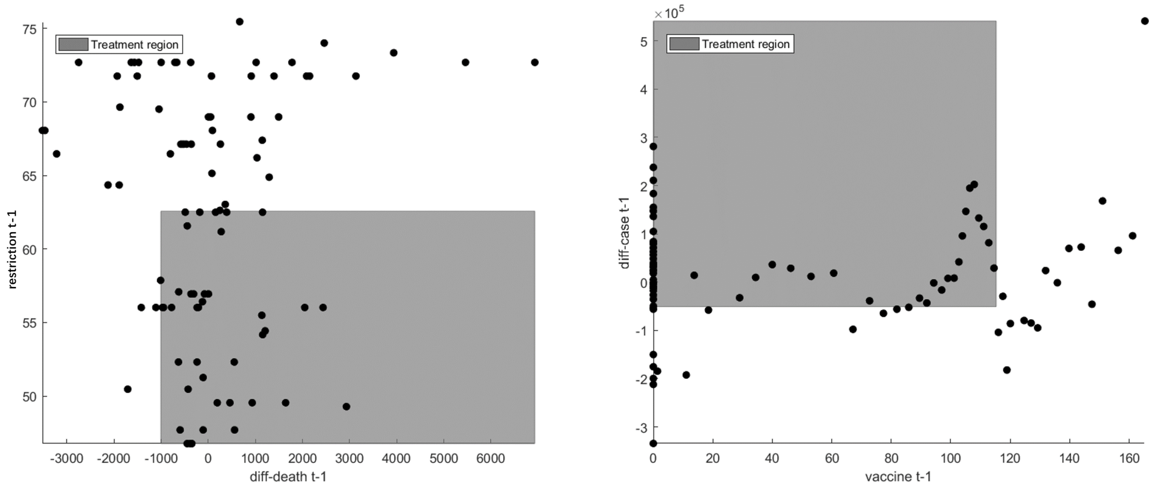
A.12.2 Policy choices based on an increased number of variables

Figure 7. Optimal policies based on
$X^P_{T-1} = \big(\text{change in cases}_{T-1}, \text{restriction stringency}_{T-1}, \text{vaccine coverage}_{T-1}\big)$



In the left panel, the x-axis is the change in deaths at week $T-1$, and the y-axis is the stringency of restrictions at week $T-1$; in the right panel, the x-axis is the stringency of restrictions at week $T-1$, and the y-axis is the vaccine coverage at week $T-1$.

Figure 8. Optimal policy based on
$X^P_{T-1} = (\text{change in deaths}_{T-1}, \text{restriction stringency}_{T-1}, \text{vaccine coverage}_{T-1},$
$\text{change in cases}_{T-1})$



In the left panel, the x-axis is the change in deaths at week $T-1$, and the y-axis is the stringency of restrictions at week $T-1$; in the right panel, the x-axis is the vaccine coverage at week $T-1$, and the y-axis is the change in cases at week $T-1$.

A.12.3 Algorithm details of the T-EWM decision tree

The dataset used in the empirical application contains 92 observations, a relatively small size. If a minimum node size is not set, the decision tree may produce nodes with only one or two observations. In these instances, the policy recommendations derived from the tree, based on empirical welfare, become highly variable and can be sometimes difficult to interpret.

Thus, we set the minimum size for each node at four. During the tree-growing process, if a split determined by the optimal policy variable results in any node having fewer than four observations, the algorithm will disregard this variable. It will then identify a sub-optimal variable from the remaining set of policy variables and execute the split. This process continues until a split is achieved where each resulting node contains at least four observations. If no policy variable can produce a split resulting in nodes with the minimum required size, the algorithm will cease splitting at this node and move to the next. Both trees presented in Figure 2 of the main text and Figure 10 below are generated using this algorithm.

A.12.4 Additional results from the T-EWM decision tree

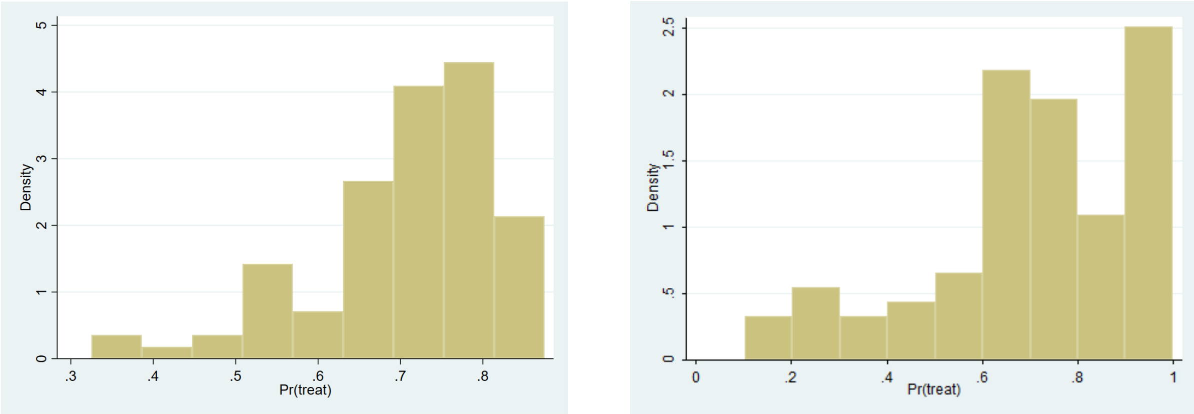
In Remark 2.2 at the end of Section 2.3, we extend our theoretical framework to accommodate higher-order Markovian structures. In this subsection, we revisit the empirical application, setting the Markov order $q = 2$ under the alternative Assumption 2.1*, which is introduced in Appendix A.3.1.

For $q = 2$, we need to re-estimate the propensity score function. Recall that for the case of $q = 1$, the set of covariates of the propensity score is given by (62):

$$
\begin{aligned}
X_t = &(\text{cases}_t, \text{deaths}_t, \text{change in cases}_t, \text{change in deaths}_t, \\
&\text{restriction stringency}_t, \text{vaccine coverage}_t, \text{economic conditions}_t).
\end{aligned}
$$

Now, we shall use both X_{t-1} and X_{t-2} as the covariates for the propensity score. The histograms of the estimated propensity scores for the observed data are presented in Figure 9: panel (a) for $q = 1$ (replicated from Figure 6b), and panel (b) for $q = 2$. For Figure 9b, we have observed that a non-trivial number of estimated propensity scores are close to 1. This is not necessarily evidence of a violation of the overlap assumption. Overfitting may be the main cause of this pattern, as an increased number of regressors can improve the fit of the predicted values to the binary dependent variables. (Currently, the sample size is 92, and the number of regressors is 14.)

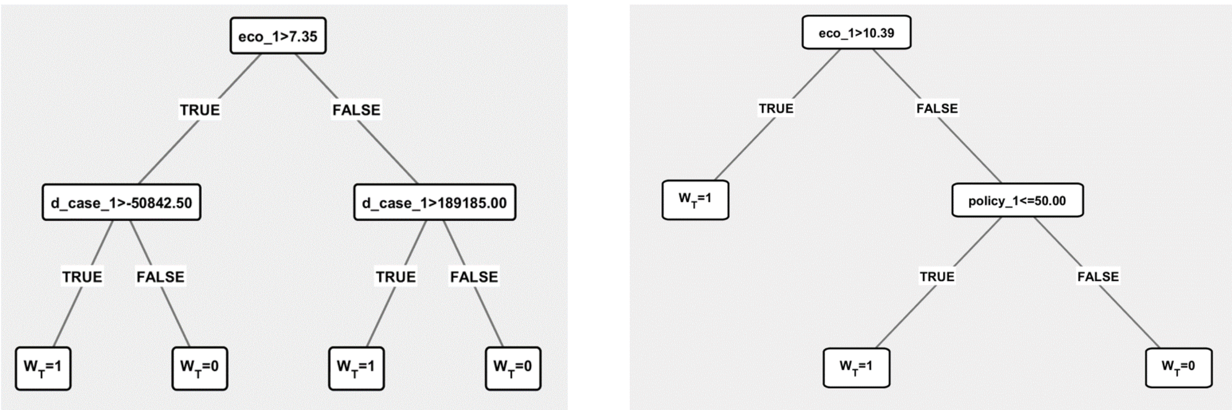
The estimated T-EWM decision tree, as presented in panel (b) of Figure 10, is obtained



(a) $q = 1$ and $e_t = e_t(X_{t-1})$ (b) $q = 2$ and $e_t = e_t(X_{t-1}, X_{t-2})$

Figure 9. The estimated PS for $q = 1$ and $q = 2$.

by censoring the estimated propensity scores at 0.025 and 0.975. The set of policy variables is represented by a 14-dimensional vector, i.e., $X^P_{T-1} = (X_{T-1}, X_{T-2})$, where X_t is given by (62). Figure 2 in the main text is reproduced here in panel (a). We observe that with a higher



(a) $q = 1$ and $X^P_{T-1} = X_{T-1}$ (b) $q = 2$ and $X^P_{T-1} = (X_{T-1}, X_{T-2})$

Figure 10. T-EWM decision trees for $q = 1$ and $q = 2$.

Markov order and an enlarged set of policy variables, the treatment region recommended by the T-EWM tree has changed. Both trees have chosen the economic condition as the variable for the first split, but the thresholds selected are slightly different. In the left branch of the second level of panel (b), the algorithm stops splitting the node since it cannot find a variable that can produce two leaves containing at least four observations each. Nevertheless, both trees provide policy recommendations that are reasonable to interpret.

B Online Appendix: Other Results and Extensions

B.1 Nonparametric method to bound the conditional regret

In Section 3.2, we have shown that we can bound conditional regret by unconditional regret if the unconditional first best policy is feasible. In Appendix B.1.1, we present examples where this method is applicable, and examples where it is not. More examples can be found in Appendix B.2. When Assumption 3.4 does not hold, we proceed to the nonparametric method discussed in Appendix B.1.2.

B.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, 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 3. 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

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

We illustrate how the policy solutions can differ between conditional and unconditional welfare functions. In Figures 11 and 12, 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 (B.1). 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 11). 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 11. $G^*_{\text{FB}} \in \mathcal{G}$

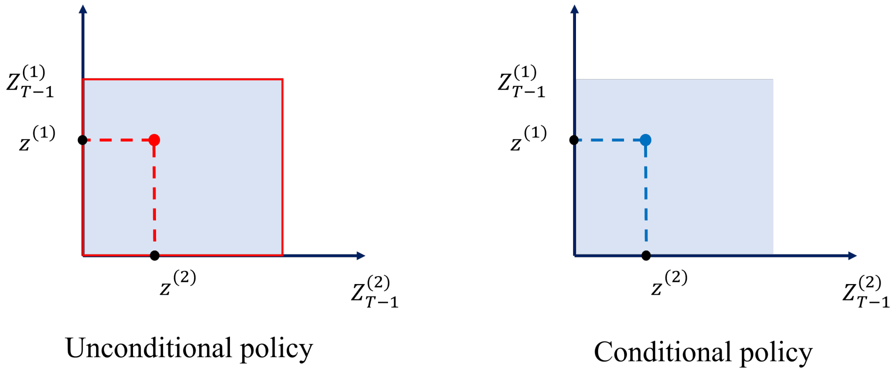


Figure 12. $G^*_{\text{FB}} \notin \mathcal{G}$

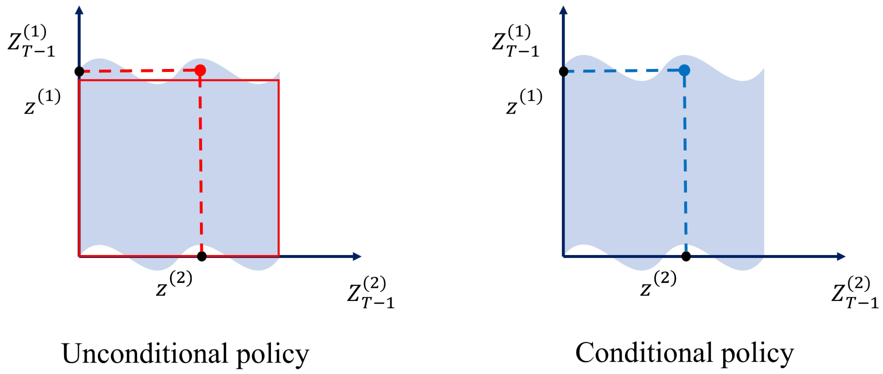


Figure 12 shows a case where the first best unconditional policy is not contained in the class of feasible policies: it is not possible to implement the policy choice that coincides with the shaded area. In this case, the policy chosen by the unconditionally optimal feasible policy (red rectangle in 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 B.2.1, 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.

B.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 planner 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

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

For simplicity, we let $X_{T-1} \in \mathbb{R}$. Then (34) can be rewritten as [13]

$$\widehat{\mathcal{W}}(G|x) = \frac{\sum_{t=1}^{T-1} K_h(X_{t-1}, x)\widehat{\mathcal{W}}_t(G)}{\sum_{t=1}^{T-1} K_h(X_{t-1}, x)}, \tag{B.2}$$

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 $K(\cdot)$ being 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

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

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

$$\bar{\mathcal{W}}_h(G|x) = \frac{\sum_{t=1}^{T-1} K_h(X_{t-1}, x)\mathcal{W}_t(G|x)}{\sum_{t=1}^{T-1} K_h(X_{t-1}, x)}, \tag{B.3}$$

where the second equality follows from Assumption 3.3.

The invariance of welfare ordering assumption is modified to:

Assumption B.1 (Invariance of welfare ordering)**.** For any $G \in \mathcal{G}$ and $x \in \mathcal{X}$, there exists some constant c

$$\mathcal{W}_T(G_x^*|x) - \mathcal{W}_T(G_|x) \le c\big[\bar{\mathcal{W}}_h(G_x^*|x) - \bar{\mathcal{W}}_h(G|x)\big]. \tag{B.4}$$

Similar to Assumption 2.4, (B.4) 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

[13] 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

covariate case (i.e., $X_t \in \mathbb{R}$). This result can be readily extended to the multiple-covariate case.

THEOREM B.1. Under Assumptions B.1 and B.2 specified in Appendix B.1.3, 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 B.1.3.

B.1.3 Proof of Theorem B.1

For simplicity, we maintain Assumption 3.1, and one of its implications is equation (33).

In addition to (B.2) and (B.3), we define

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

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 $G \in \mathcal{G}$, such that

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

where the first inequality follows from Assumption B.1. The second inequality follows from Lemma B.1 below. The third inequality follows from similar arguments to (21). The last equality follows from Lemma B.2 stated below.

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

Assumption B.2. X_t is a 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: $K(x) = 0$ if

$|X| > c_k$. $\mathsf{E}(K^2(\frac{X_t - x}{h})|\mathcal{F}_{t-1}) \le c_m$, 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 that is in the interior point of its support $\mathcal{X}$, which is also bounded; (iii) $\int K(u)^2 du$ is bounded.

LEMMA B.1. Under Assumption B.2 For any $G \in \mathcal{G}$ and $x \in \mathcal{X}$,

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

Proof. Under Assumption B.2,

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

Rearranging the equation we have

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

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 B.2. Under Assumption B.2,

$$\sup_{G\in\mathcal{G}}\sup_{x\in\mathcal{X}} |\widehat{\mathcal{W}}(G|x) - \widetilde{\mathcal{W}}_h(G, x)| \lesssim_p c_w^{-1}(\sqrt{(T-1)h}^{-1}). \tag{B.5}$$

Proof. Note

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

We first 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 $\{X_t\}_{t=1}^{T-1}$ is $X_{(1)}, \cdots, X_{(T-1)}$, and $B_{x,h} = \{t : |(x - x_t)/h| \le$

$c_k\}$.

Because of summation by part, the numerator is bounded by

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

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

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 B.2. Due to the boundedness of $\mathsf{E}(K_h(X_{t-1},x)|\mathcal{F}_{t-2})$ from the Assumption B.2, by following similar steps to the proof of (22), we have

$$\sup_{G\in\mathcal{G}}\sup_{x\in\mathcal{X}}|\widehat{\mathcal{W}}(G|X_{T-1}=x)-\bar{\mathcal{W}}(G|X_{T-1}=x)|\lesssim_p c_w^{-1}(\sqrt{(T-1)h}^{-1}). \tag{B.6}$$

□

B.2 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 3 to show how Assumption 3.4 ensures this equivalence.

B.2.1 $G^*_{\text{FB}}\in\mathcal{G}$ is sufficient: a discrete and a continuous case

Let $X_{t-1}=W_{t-1}\in\{0,1\}$, and $\mathcal{G}$ 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:

$$\begin{aligned}\mathcal{W}_T(G) =& \mathsf{E}\left[Y_T(1)\mathbf{1}(W_{T-1} \in G) + Y_T(0)\mathbf{1}(W_{T-1} \notin G)\right] \\ =& \mathsf{E}\left[\left[Y_T(1) - Y_T(0)\right]\mathbf{1}(W_{T-1} \in G) + Y_T(0)\right] \\ =& \mathsf{E}\left[\tau(W_{T-1})\mathbf{1}(W_{T-1} \in G)\right] + \mathsf{E}\left[Y_T(0)\right],\end{aligned} \tag{B.7}$$

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

$$G^*_{\text{FB}} \equiv \{w_{T-1} \in \{0, 1\} : \tau(w_{T-1}) \geq 0\}. \tag{B.8}$$

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

$$G^*_{\text{FB}} = \text{argmax}_{G \in \mathcal{G}} \mathcal{W}_T(G). \tag{B.9}$$

The planner's conditional objective function can be written as

$$\begin{aligned}\mathcal{W}_T(G|W_{T-1}) &= \mathsf{E}\left[Y_T(1)\mathbf{1}(W_{T-1} \in G) + Y_T(0)\mathbf{1}(W_{T-1} \notin G)|W_{T-1}\right] \\ &= \mathsf{E}\left[\left[Y_T(1) - Y_T(0)\right]\mathbf{1}(W_{T-1} \in G) + Y_T(0)|W_{T-1}\right] \\ &= \mathsf{E}\left[\left[Y_T(1) - Y_T(0)\right]\mathbf{1}(W_{T-1} \in G)|W_{T-1}\right] + \mathsf{E}\left[Y_T(0)|W_{T-1}\right] \\ &= \mathsf{E}\left[\tau(W_{T-1})\mathbf{1}(W_{T-1} \in G)|W_{T-1}\right] + \mathsf{E}\left[Y_T(0)|W_{T-1}\right].\end{aligned} \tag{B.10}$$

To check whether G^*_{FB} is optimal in the conditional problem, we need to study this problem w.r.t. $\mathcal{P}$

$$\begin{aligned}&\max_{G \in \mathcal{P}} \mathcal{W}_T(G|W_{T-1} = w_{T-1}) \\ =&\max_{G \in \mathcal{P}} \mathsf{E}\left[\tau(W_{T-1})\mathbf{1}(W_{T-1} \in G)|W_{T-1} = w_{T-1}\right] \\ =&\max_{G \in \mathcal{P}} \tau(w_{T-1})\mathbf{1}(w_{T-1} \in G) \\ =&\tau(w_{T-1})\mathbf{1}(w_{T-1} \in G^*_{\text{FB}}),\end{aligned} \tag{B.11}$$

where the first equality follows from (B.10), and the last one follows from the definition of G^*_{FB} in (B.8). The equivalence follows by combining (B.9) and (B.11).

We now turn to the continuous conditioning variable case.

The planner'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

$$G^*_{\mathrm{FB}} \equiv \{x_{T-1} \in \mathbb{R}^2 : \tau(x_{T-1}) \geq 0\}. \tag{B.12}$$

(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 $G^*_{\mathrm{FB}} \in \mathcal{G}$,

$$G^*_{\mathrm{FB}} \in \mathrm{argmax}_{G \in \mathcal{G}} \mathcal{W}_T(G). \tag{B.13}$$

We now introduce some notations. 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 planner's conditional objective function can be written as:

$$\begin{aligned}&\mathcal{W}_T(G^{(1)}|X^{(2)}_{T-1})\\ =&\mathsf{E}\left[Y_T(1)\mathbf{1}(X^{(1)}_{T-1} \in G^{(1)}) + Y_T(0)\mathbf{1}(X^{(1)}_{T-1} \notin G^{(1)})|X^{(2)}_{T-1}\right]\\ =&\mathsf{E}\left[\left[Y_T(1) - Y_T(0)\right]\mathbf{1}(X^{(1)}_{T-1} \in G^{(1)}) + Y_T(0)|X^{(2)}_{T-1}\right]\\ =&\mathsf{E}\left[\left[Y_T(1) - Y_T(0)\right]\mathbf{1}(X^{(1)}_{T-1} \in G^{(1)})|X^{(2)}_{T-1}\right] + \mathsf{E}\left[Y_T(0)|X^{(2)}_{T-1}\right]\\ =&\mathsf{E}\left[\mathsf{E}\left[\tau(X_{T-1})\mathbf{1}(X^{(1)}_{T-1} \in G^{(1)})|X_{T-1}\right]|X^{(2)}_{T-1}\right] + \mathsf{E}\left[Y_T(0)|X^{(2)}_{T-1}\right],\end{aligned} \tag{B.14}$$

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., $G^*_{\mathrm{CFB}}(X^{(2)}_{T-1} = x^{(2)}_{T-1}) \in \mathfrak{B}(\mathbb{R})$, for every $x^{(2)}_{T-1} \in \mathbb{R}$. Note that $\tau(x_{t-1}) = \tau(x^{(1)}_{t-1}, x^{(2)}_{t-1})$. Following the last row of (B.14), 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

$$G^*_{\text{CFB}}(X^{(2)}_{T-1}=x^{(2)}_{T-1})=\{x^{(1)}_{T-1}\in\mathbb{R}:\tau(x^{(1)}_{T-1},x^{(2)}_{T-1})\geq 0\}. \tag{B.15}$$

Comparing (B.12) and (B.15), we see $G^*_{\text{CFB}}(X^{(2)}_{T-1}=x^{(2)}_{T-1})$ is given by the intersection of G^*_{FB} with the line $X^{(2)}_{T-1}=x^{(2)}_{T-1}$.

B.2.2 $G^*_{\mathbf{FB}}\in\mathcal{G}$ 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 4. 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 (B.7), 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 $W_{T-1} = 1$.

$$\begin{aligned}
&\max_{G \in \mathcal{P}} \mathcal{W}_T(G | W_{T-1} = 1) \\
=&\max_{G \in \mathcal{P}} \mathsf{E}\left[\tau(W_{T-1})\mathbf{1}(W_{T-1} \in G) | W_{T-1} = 1\right] \\
=&\max_{G \in \mathcal{P}} \tau(1)\mathbf{1}(1 \in G) \\
=&\tau(1)\mathbf{1}(1 \in G_*).
\end{aligned} \tag{B.16}$$

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 5. 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

$$G_* \equiv \operatorname*{argmax}_{G \in \mathcal{G}} \mathcal{W}_T(G) = \{0, 1\} \times \{0, 1, 2, 3, 4, 5\}, \tag{B.17}$$

We can always construct a data-generating process with a certain type of conditional treatment effect τ, such that (B.17) is the best feasible unconditional policy. For example, let $\tau(x_{t-1}) = \tau(w_{t-1}, z_{t-1})$, set $\Pr(Z_{t-1} = i) = 1/10$, $Z_{t-1} \perp W_{t-1}$, 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 $a = 5.5$ includes $\{w, 2\}$ 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

$$\underset{G^z \in \mathcal{G}^z}{\text{argmax}}\ \mathcal{W}_T(G^z | W_{T-1} = w) = \{1, 2, 3, 4, 5\},$$

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

B.2.3 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 3 to show why Assumption 3.4 ensures equivalence between the unconditional and conditional problems.

Example 3, continued.

Figure 13. $G^*_{\text{FB}} \notin \mathcal{G}_0$ but Assumption 3.4 is satisfied

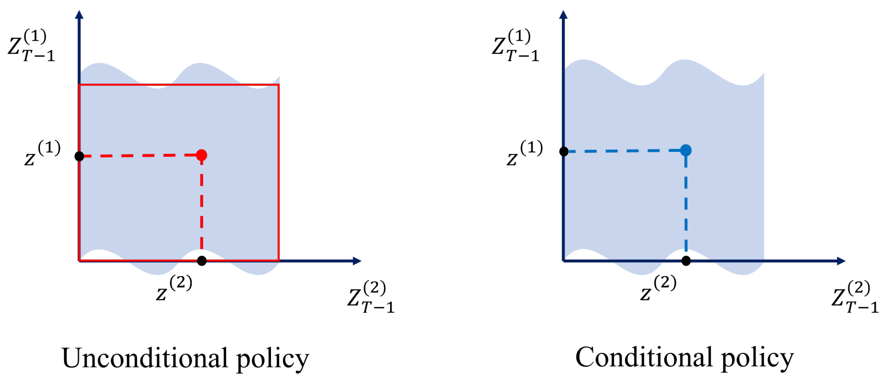


Figure 13 shows a case where the solutions coincide even though the first best unconditional policy is not available. The red square differs from the shaded area (the first best), but the red point is inside the red square, and the blue point is inside the shaded area. The social planner will set $W_T = 1$ 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.3 Multi-period welfare with continuous covariates

In this section, we extend the results in the previous sections to a multi-period setup. We focus on a simple offline decision problem with deterministic treatment rules. Namely, we do not update policy after time T, but we allow the welfare function to include the updated realized observations. A social planner is faced with a finite multi-period welfare target and a set of continuous policy variables is available. Without loss of generality, we focus on the case of a two-period policy assignment.

The planner's decision making procedure can be described as follows. Based on a sample collected from time 0 to $T-1$, the planner chooses (or estimates) a *decision rule* that will be implemented on T and $T+1$. The rule is characterized by two sets defined on $\mathcal{X}$, and we write them as $G_{1:2} = \{G_1, G_2\} \in \mathcal{X}^2$. At the beginning of time T, the planner makes the decision by $W_T = 1(X_{T-1} \in G_1)$, then she/he will observe the outcome Y_T as well as $X_T = (Y_T, W_T, Z_T)$ at the end of time T. At the beginning of time $T+1$, the planner will make the decision by $W_{T+1} = 1(X_T \in G_2)$. Since we focus on an offline problem, G_2 is chosen (estimated) at the end of $T-1$ and implemented at $T+1$ after X_T is revealed. Also, the class of subsets of $\mathcal{X}^2$ where $G_{1:2}$ is chosen from, i.e. $\mathcal{G}_{1:2}$, is assumed to be of polynomial classification and with finite VC dimension.

Recall the notations:

$$\begin{aligned}
S_t(G) &= Y_t(W_{t-1},1)1(X_{t-1} \in G) + Y_t(W_{t-1},0)1(X_{t-1} \notin G), \\
\mathcal{W}_t(G|X_{t-1}) &= \mathsf{E}\left[S_t(G)|X_{t-1}\right], \\
\widehat{\mathcal{W}}_t(G) &= \frac{Y_t W_t}{e_t(X_{t-1})} 1(X_{t-1} \in G) + \frac{Y_t\left(1-W_t\right)}{1-e_t(X_{t-1})} 1(X_{t-1} \notin G).
\end{aligned}$$

Let us define

$$\bar{\mathcal{W}}_t(G) = \mathsf{E}\left[S_t(G)|\mathcal{F}_{t-1}\right].$$

We maintain Assumption 3.1, the Markovian condition of order 1. Then, the conditional two-period welfare function is defined as

$$\begin{aligned}
&\mathcal{W}_{T:T+1}(G_{1:2}|\mathcal{F}_{T-1}) \\
=&\mathcal{W}_T(G_1|X_{T-1}) + \mathsf{E}\left[S_{T+1}(G_2)|\mathcal{F}_{T-1}\right] \\
=&\mathcal{W}_T(G_1|X_{T-1}) + \mathsf{E}\left[\mathsf{E}\left[S_{T+1}(G_2)|\mathcal{F}_T\right]|\mathcal{F}_{T-1}\right] \\
=&\mathcal{W}_T(G_1|X_{T-1}) + \mathsf{E}\left[\mathsf{E}\left[S_{T+1}(G_2)|X_T\right]|X_{T-1}\right] \\
=&\mathcal{W}_T(G_1|X_{T-1}) + \mathsf{E}\left[\mathsf{E}\left[S_{T+1}(G_2)|Y_T(W_{T-1}, 1(X_{T-1} \in G_1)), W_T = 1(X_{T-1} \in G_1)\right]|X_{T-1}\right],
\end{aligned} \tag{B.18}$$

where the third equality follows from Assumption 3.1. It shall be noted that in this case, the welfare function is conditional on the information available up to time $T-1$, and we fixed the conditioning value of X_{T-1} at the observed one.

B.3.1 Direct estimation of the conditional welfare function.

With the defined two-period welfare function on hand, we can discuss how to estimate the welfare function. Similar to that discussed in Section 3.2 for the single-period welfare target, we can either directly estimate the conditional welfare functions or the regret bounds by the unconditional welfare function. In this subsection, we directly estimate the two-period conditional welfare function proposed. Without loss of generality, we start with the case $X_t = (Y_t, W_t) \in \mathbb{R} \times \{0,1\}$, i.e, Y_t is the only continuous variable in X_t. It can be easily extended to the case with other continuous variables $X_t = (Y_t, W_t, Z_t) \in \mathbb{R} \times \{0,1\} \times \mathbb{R}^k$.

Again, we apply the abbreviation: $(\cdot|x) = (\cdot|X_{T-1} = x)$, where $x = (y, w)$. Now, we have

$$\mathcal{W}_{T:T+1}(G_{1:2}|x) = \mathcal{W}_T(G_1|x) + \mathsf{E}\left(\mathcal{W}_{T+1}(G_2|X_T, W_T = 1(X_{T-1} \in G_1))|X_{T-1} = x\right).$$

Let $\mathcal{G}_{1:2}$ be the class of feasible policies, which is a sub-class of the class of all the measurable functions defined on $\mathcal{X} \times \mathcal{X} \to \{0,1\}^2$. Conditional on $X_{T-1} = x$, we define

$$G^*_{1:2} \in \text{argmax}_{G_{1:2} \in \mathcal{G}_{1:2}} \mathcal{W}_{T:T+1}(G_{1:2}|x).$$

Note $G^*_{1:2}$ can depend on x, but we suppress this dependence in the notation.

Similarly to (34), for any policy G_1 and G_2, define

$$\begin{aligned}\widehat{\mathcal{W}}(G_{1:2}|x) = & \frac{\sum_{t=1}^{T-1} \mathbf{1}\left(W_{t-1} = w\right) K_h(Y_{t-1}, y) \widehat{\mathcal{W}}_t(G_1)}{\sum_{t=1}^{T-1} \mathbf{1}\left(W_{t-1} = w\right) K_h(Y_{t-1}, y)} \\ & + \frac{\sum_{t=1}^{T-2} K_h(Y_{t-1}, y) \mathbf{1}\left(W_{t-1} = w\right) \mathbf{1}\left(W_t = 1(X_{t-1} \in G_1)\right) \widehat{\mathcal{W}}_{t+1}(G_2)}{\sum_{t=1}^{T-2} K_h(Y_{t-1}, y) \mathbf{1}\left(W_{t-1} = w\right) \mathbf{1}\left(W_t = 1(X_{t-1} \in G_1)\right)},\end{aligned}$$

where $K_h(a,b) = \frac{1}{h} K(\frac{a-b}{h})$ with $K(\cdot)$ assumed to be a bounded kernel function with a bounded support. Since we have $\mathcal{F}_{t-1} \subset \mathcal{F}_t$, we have the $\mathbf{1}\left(W_t = 1(X_{t-1} \in G_1)\right)$ can be removed from the above sum since W_t is determined by X_{t-1}.

Let $\mathcal{G}^2$ denote the class of feasible unconditional decision sets, which is a class of subsets of $\mathcal{X} \times \mathcal{X}$, and

$$\hat{G}_{1:2} \in \text{argmax}_{G_{1:2} \in \mathcal{G}_{1:2}} \widehat{\mathcal{W}}(G_{1:2}|x).$$

Define $\mathcal{E}\mathcal{W}_{t+1}(G_{1:2}|x) = \mathsf{E}\left[\mathsf{E}\left[S_{T+1}(G_2)|Y_T(W_{T-1}, 1(X_{T-1} \in G_1)), W_T = 1(X_{T-1} \in G_1)\right]|X_{T-1} = x\right]$ and $\mathcal{E}\mathcal{W}_{t+1}(G_{1:2}|X_{T-1}) = \mathsf{E}\left[\mathsf{E}\left[S_{T+1}(G_2)|Y_T(W_{T-1}, 1(X_{T-1} \in G_1)), W_T = 1(X_{T-1} \in G_1)\right]|X_{T-1}\right]$.

First of all, x is a vector of values consisting of w and y. To construct an MDS, we shall define an intermediate counterpart,

$$\bar{\mathcal{W}}_h(G_{1:2}|x) = \frac{\sum_{t=1}^{T-1}\mathbf{1}\left(W_{t-1}=w\right)K_h(Y_{t-1},y)\mathcal{W}_t(G_1|x)}{\sum_{t=1}^{T-1}\mathbf{1}\left(W_{t-1}=w\right)K_h(Y_{t-1},y)} + \frac{\sum_{t=1}^{T-2}K_h(Y_{t-1},y)\mathbf{1}\left(W_{t-1}=w\right)\mathbf{1}\left(W_t=1(X_{t-1}\in G_1)\right)\mathcal{E}\mathcal{W}_{t+1}(G_{1:2}|x)}{\sum_{t=1}^{T-2}K_h(Y_{t-1},y)\mathbf{1}\left(W_{t-1}=w\right)\mathbf{1}\left(W_t=1(X_{t-1}\in G_1)\right)}.$$

Note that

$$\begin{aligned}&\mathsf{E}(K_h(Y_{t-1},y)\mathbf{1}\left(W_{t-1}=w\right)\mathbf{1}\left(W_t=1(X_{t-1}\in G_1)\right)\widehat{\mathcal{W}}_{t+1}(G_2)|\mathcal{F}_{t-1})\\ &= K_h(Y_{t-1},y)\mathbf{1}\left(W_{t-1}=w\right)\mathbf{1}\left(W_t=1(X_{t-1}\in G_1)\right)\mathsf{E}(\widehat{\mathcal{W}}_{t+1}(G_2)|X_{t-1}),\end{aligned}$$

so $K_h(Y_{t-1},y)\mathbf{1}\left(W_{t-1}=w\right)\mathbf{1}\left(W_t=1(X_{t-1}\in G_1)\right)(\widehat{\mathcal{W}}_{t+1}(G_2)-\mathcal{E}\mathcal{W}_{t+1}(G_2|X_{T-1}))$ is an MDS with respect to $\mathcal{F}_{t-1}$. Then with similar steps to those in the proof of Theorem B.1, we can achieve the closeness between $K_h(Y_{t-1},y)\mathbf{1}(W_{t-1}=w)\mathbf{1}(W_t=1(X_{t-1}\in G_1))\mathcal{E}\mathcal{W}_{t+1}(G_2|X_{T-1})$ and $K_h(Y_{t-1},y)\mathbf{1}(W_{t-1}=w)\mathbf{1}(W_t=1(X_{t-1}\in G_1))\mathcal{E}\mathcal{W}_{t+1}(G_2|X_{T-1}=x)$ by a bias term of order $O_p(h^2)$. Thus, with similar assumptions assumed in Theorem B.1, we can achieve the regret bound of the rate $\sqrt{(T-1)h}^{-1}+(T-1)^{-1}+h^2$.

B.3.2 Bounding the conditional regret by the unconditional one

Similar to Section 3.2, under the correct specification assumption, we can bound the conditional regret by unconditional regret. For the multivariate case, the unconditional welfare is defined as

$$\mathcal{W}_{T:T+1}(G_{1:2}) = \mathcal{W}_T(G_1) + \mathsf{E}\{\mathsf{E}\left[S_{T+1}(G_2)|W_T=1(X_{T-1}\in G_1)\right]\}. \tag{B.19}$$

Note that we have
$\mathsf{E}\{\mathsf{E}\left[S_{T+1}(G_2)|W_T=\mathbf{1}(X_{t-1}\in G_1)\right]\} = \mathsf{E}\left[S_{T+1}(G_2)\mathbf{1}(W_T=\mathbf{1}(X_{t-1}\in G_1))\right]/\Pr(W_T=1(X_{t-1}\in G_1))$. Note that slightly different from the one-period welfare function, in this unconditional welfare, the second part is still conditioning on $W_T=1(X_{t-1}\in G_1)$ since the treatment W_T is determined by the planner's policy G_1.

The optimal unconditional policy within the class $\mathcal{G}_{1:2}$ is defined as

$$G^*_{1:2}\in \text{argmax}_{G_{1:2}\in\mathcal{G}_{1:2}}\mathcal{W}_{T:T+1}(G_{1:2}). \tag{B.20}$$

To bound the conditional regret with the unconditional one we also need to impose the following assumption.

Assumption B.3. Let $\mathcal{W}_{T:T+1}(G_{1:2}|x)$ be the conditional welfare defined in (B.18),

$$\text{argsup}_{G_{1:2}\in\mathcal{G}_{1:2}}\mathcal{W}_{T:T+1}(G_{1:2}) \subset \text{argsup}_{G_{1:2}\in\mathcal{G}_{1:2}}\mathcal{W}_{T:T+1}(G_{1:2}|x).$$

Proposition B.1. Under Assumption 3.1 and (B.3),

$$G^*_{1:2} \in \text{argsup}_{G_{1:2}\in\mathcal{G}_{1:2}}\mathcal{W}_{T:T+1}(G_{1:2}|X_{T-1}).$$

The conditional regret is

$$R_{T:T+1}(G_{1:2}|x) := \mathcal{W}_{T:T+1}(G^*_{1:2}|x) - \mathcal{W}_{T:T+1}(G_{1:2}|x).$$

Similarly, unconditional regret can be expressed as an integral of conditional regret. Thus, the unconditional welfare and regret are

$$\begin{aligned}
\mathcal{W}_{T:T+1}(G_{1:2}) &= \int \mathcal{W}_{T:T+1}(G_{1:2}|x)dF_{X_{T-1}}(x),\\
R_{T:T+1}(G_{1:2}) &= \mathcal{W}_{T:T+1}(G^*_{1:2}|x) - \mathcal{W}_{T:T+1}(G_{1:2})\\
&= \int R_{T:T+1}(G_{1:2}|x)dF_{X_{T-1}}(x).
\end{aligned}$$

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

$$\begin{aligned}
A(x', G_{1:2}) &= \{x : x \in \mathcal{X} \text{ and } R_{T:T+1}(G_{1:2}|x) \geq R_{T:T+1}(x', G_{1:2})\},\\
p_{T-1}(x', G_{1:2}) &= \Pr(X_{T-1} \in A(x', G_{1:2})) = \int_{x\in A(x',G_{1:2})} dF_{X_{T-1}}(x).
\end{aligned}$$

Now, we impose the following assumption.

Assumption B.4. For $x^{obs} \in \mathcal{X}$ and any $G_{1:2} \in \mathcal{G}_{1:2}$

$$p_{T-1}(x^{obs}, G_{1:2}) \geq \underline{p} > 0. \tag{B.21}$$

For some positive constant $\underline{p}$.

LEMMA B.3. Under Assumption B.4

$$R_{T:T+1}(G_{1:2}|x^{obs}) \leq \frac{1}{\underline{p}} R_{T:T+1}(G_{1:2}).$$

Now under Assumptions B.4 and Lemma B.3, we can specify the sample analogue

$$\widehat{\mathcal{W}}(G_{1:2}) = \frac{1}{T}\sum_{t=1}^{T-1}\widehat{\mathcal{W}}_t(G_1) + \frac{1}{T(G_1)}\sum_{t=1}^{T-2}\mathbf{1}\left(W_t = \mathbf{1}(X_{t-1}\in G_1)\right)S_{t+1}(G_2),$$

where $T(G_1)$ is a random number defined as $T(G_1) = \#\{1 \le t \le T-1 : W_t = \mathbf{1}(X_{t-1} \in G_1)\}$. We define the estimated welfare policy,

$$\hat{G}_{1:2} \in \text{argmax}_{G_{1:2}\in\mathcal{G}_{1:2}}\widehat{\mathcal{W}}(G_{1:2}).$$

To prove the bound we can proceed with similar steps as in the proof of Theorem 3.1. Namely, we define a conditional welfare function to form an MDS as follows,

$$\begin{aligned}\bar{\mathcal{W}}(G_{1:2}) &= \frac{1}{T}\sum_{t=1}^{T-1}\mathcal{W}_t(G_1|\mathcal{F}_{t-2}) \\ &+ \frac{1}{\mathsf{E}(T(G_1))}\sum_{t=1}^{T-1}\mathsf{E}\{[\mathbf{1}\left(W_t = \mathbf{1}(X_{t-1}\in G_1)\right)S_{t+1}(G_2)|W_t = \mathbf{1}(X_{t-1}\in G_1)]\,|\mathcal{F}_{t-2}\}.\end{aligned}$$

Then the unconditional population counterpart is as follows,

$$\begin{aligned}\widetilde{\mathcal{W}}(G_{1:2}) &= \frac{1}{T}\sum_{t=1}^{T-1}\mathcal{W}_t(G_1) \\ &+ \frac{1}{\mathsf{E}(T(G_1))}\sum_{t=1}^{T-1}\mathsf{E}(\mathbf{1}\left[W_t = \mathbf{1}(X_{t-1}\in G_1)S_{t+1}(G_2)\right]. \qquad \text{(B.22)}\end{aligned}$$

Now, similar to Assumption 3.6, we impose

Assumption 3.6'. For $G^*_{1:2}$ defined in (B.20) and any $G_{1:2} \in \mathcal{G}_{1:2}$, there exists some constant c, such that

$$\mathcal{W}_{T:T+1}(G^*_{1:2}) - \mathcal{W}_{T:T+1}(G_{1:2}) \le c\left(\widetilde{\mathcal{W}}(G^*_{1:2}) - \widetilde{\mathcal{W}}(G_{1:2})\right).$$

Then, the conditional regret can be bounded by

$$\begin{aligned}\mathcal{W}_{T:T+1}(G^*_{1:2}|x^{obs}) - \mathcal{W}_{T:T+1}(G_{1:2}x^{obs}) &\le \frac{1}{\underline{p}}\left[\mathcal{W}_{T:T+1}(G^*_{1:2}) - \mathcal{W}_{T:T+1}(G_{1:2})\right] \\ &\le \frac{c}{\underline{p}}\left[\widetilde{\mathcal{W}}(G^*_{1:2}) - \widetilde{\mathcal{W}}(G_{1:2})\right] \\ &\le \frac{c}{\underline{p}}\sup_{G_{1:2}\in\mathcal{G}_{1:2}}\left[\widehat{\mathcal{W}}(G_{1:2}) - \widetilde{\mathcal{W}}(G_{1:2})\right].\end{aligned}$$

And

$$\begin{aligned}
\widehat{\mathcal{W}}(G_{1:2}) - \widetilde{\mathcal{W}}(G_{1:2}) &= \left[\bar{\mathcal{W}}(G_{1:2}) - \widetilde{\mathcal{W}}(G_{1:2})\right] + \left[\widehat{\mathcal{W}}(G_{1:2}) - \bar{\mathcal{W}}(G_{1:2})\right] \\
&= I + II
\end{aligned}$$

Similar to Section 3, we can drive upper bounds for I and II.