Research Article

Nathan Kallus and Michele Santacatterina*

Optimal balancing of time-dependent confounders for marginal structural models

<https://doi.org/10.1515/jci-2020-0033>

received November 23, 2020; accepted December 14, 2021

Abstract: Marginal structural models (MSMs) can be used to estimate the causal effect of a potentially time-varying treatment in the presence of time-dependent confounding via weighted regression. The standard approach of using inverse probability of treatment weighting (IPTW) can be sensitive to model misspecification and lead to high-variance estimates due to extreme weights. Various methods have been proposed to partially address this, including covariate balancing propensity score (CBPS) to mitigate treatment model misspecification, and truncation and stabilized-IPTW (sIPTW) to temper extreme weights. In this article, we present kernel optimal weighting (KOW), a convex-optimization-based approach that finds weights for fitting the MSMs that flexibly balance time-dependent confounders while simultaneously penalizing extreme weights, directly addressing the above limitations. We further extend KOW to control for informative censoring. We evaluate the performance of KOW in a simulation study, comparing it with IPTW, sIPTW, and CBPS. We demonstrate the use of KOW in studying the effect of treatment initiation on time-to-death among people living with human immunodeficiency virus and the effect of negative advertising on elections in the United States.

Keywords: causal inference, optimization, covariate balance, time-dependent treatments, marginal structural models

MSC 2020: 92B15, 92C60, 62N99, 62P10

1 Introduction

Marginal structural models (MSMs) offer a successful way to estimate the causal effect of a time-varying treatment on an outcome of interest from longitudinal data in observational studies [1,2]. For example, they have been used to estimate the optimal timing of human immunodeficiency virus (HIV) treatment initiation [3], to evaluate the effect of hormone therapy on cardiovascular outcomes [4], and to evaluate the impact of negative advertising on election outcomes [5]. The increasing popularity of MSMs among applied researchers derives from their ability to control for time-dependent confounders, which are confounders that are affected by previous treatments and affect future ones. In particular, as shown by Robins et al. [2] and Blackwell [5], standard methods, such as regression or matching, fail to control for time-dependent confounding, introducing post-treatment bias. In contrast, MSMs consistently estimate the causal effect of a time-varying treatment via inverse probability of treatment weighting (IPTW), which controls for time-dependent confounding by weighting each subject under study by the inverse of their probability of being treated given covariates, i.e., the propensity score [6], mimicking a sequential randomized trial. In other

* **Corresponding author: Michele Santacatterina**, Department of Population Health, New York University Grossman School of Medicine, New York 10016, New York, USA, e-mail: michele.santacatterina@nyulangone.org

Nathan Kallus: Department of Operations Research and Information Engineering and Cornell Tech, Cornell University, New York 10044, New York, USA, e-mail: kallus@cornell.edu

words, IPTW creates a hypothetical pseudo-population where time-dependent confounders are balanced over time.

Despite their wide range of applications, the use of these methods in observational studies may be jeopardized by their considerable dependence on positivity. This assumption requires that, at each time period, the probability of being assigned to the treatment, conditional on the history of treatment and confounders, is not 0 or 1 [1]. Even if positivity holds theoretically in the population, when propensities are close to 0 or 1, it can be easily *practically* violated, i.e., when some combinations of confounders and treatment are rare in the observed data [7]. Practical positivity violations lead to extreme and unstable weights, which in turn yield very low precision and misleading inferences [8–10]. This is a particular concern with longitudinal data, since IPTW requires conditioning on past treatment sequences in addition to time-dependent confounders thus having greater risk of running into near-zero denominators. In addition, MSMs using IPTW are highly sensitive to *misspecification* of the treatment assignment model, which can lead to biased estimates [8,11,12].

Various statistical methods have been proposed in an attempt to overcome these challenges. To deal with extreme weights, several authors [12,13] have suggested truncation, whereby outlying weights are replaced with less extreme ones. Santacatterina et al. [14] proposed to use shrinkage instead of truncation as a more direct way to control the bias-variance trade-off. Robins et al. [2] recommended the use of stabilized-IPTW (sIPTW) where inverse probability weights are normalized by the marginal probability of treatment. To mitigate misspecification of the treatment assignment model, Imai and Ratkovic [15] proposed to use the covariate balance propensity score (CBPS), which instead of plugging in a logistic regression estimate of propensity into IPTW finds the logistic model that balances covariates via the generalized method of moments. The method tries to balance the first, and possibly higher, moment of each covariate even if a logistic model is misspecified [16]. The authors also provided balancing conditions for longitudinal studies. Yiu and Su [17] proposed a joint calibration approach to covariate balancing weight estimation for MSMs. The method aims to compute weights that jointly eliminate covariate associations with both treatment assignment and censoring processes. Other methods have been proposed in the literature to manage the trade-off between balance and precision [18–20, among others], and that discuss reproducing kernel Hilbert spaces (RKHSs) [21–23, among others]. These methods, however, do not directly deal with time-dependent confounders and informative censoring.

In this article, we present and apply kernel optimal weighting (KOW), which provides weights for fitting an MSM that balance time-dependent confounders while controlling for precision. Specifically, by solving a quadratic optimization problem over weights, the proposed method directly minimizes *imbalance*, defined as the sum of discrepancies between the weighted observed data and the counterfactual of interest over all treatment regimes, while penalizing extreme weights.

This extends the kernel optimal matching method of Kallus [24] and Kallus et al. [25] to the longitudinal setting and to dealing with time-dependent confounders, where, similar to regression and matching, it cannot be applied without introducing post-treatment bias.

The proposed method has several attractive characteristics. First, KOW can balance non-additive covariate relationships by using kernels, which generalize the structure of conditional expectation functions, and does not restrict weights to follow a fixed logistic (or other parametric) form. By doing so, KOW mitigates the effects of possible misspecification of the treatment model. In the simulation study presented in Section 5, we show that KOW is more robust to model misspecification compared with the other methods. In Section 5, we also show how KOW compares favorably with the aforementioned methods in all nonlinear scenarios, and in Section 7.2 we use KOW to balance non-additive covariate relationships estimating the effect of negative advertising on election outcomes. Second, by balancing time-dependent confounders using kernels while penalizing extreme weights, KOW leads to better accuracy, precision, and total error. In Section 5, we show that the mean squared error (MSE) of the estimated effect of a time-varying treatment obtained by using KOW is lower than that obtained by using IPTW, sIPTW, and CBPS in all considered simulated scenarios. Third, differently from ref. [15], where the number of covariate-balancing conditions grows exponentially in the number of time periods, KOW only needs to minimize a number of discrepancies that grows linearly in the number of time periods. This feature leads to a lower computational time of KOW

compared with CBPS when the total number of time periods increases, as shown in our simulation study in Section 5.5 and in our study on the effect of negative advertising on election outcomes in Section 7.2. Fourth, KOW can be easily generalized to other settings, such as informative censoring. We do just that in Section 6, and in Section 7.1, we use this extension to study the effect of HIV treatment on time to death among people living with HIV (PLWH). Finally, KOW can be solved by using off-the-shelf solvers for quadratic optimization.

In the next section, we briefly introduce the literature of MSMs (Section 2). In Section 3, we develop and define KOW. We then discuss some practical guidelines on the use of KOW (Section 4). In Section 5, we report the results of a simulation study aimed at comparing KOW with IPTW, sIPTW, and CBPS. In Section 6, we extend KOW to control for informative censoring. We then present two empirical applications of KOW in medicine and political science (Section 7). We offer some concluding remarks in Section 8.

2 MSMs for longitudinal data

In this section, we briefly review MSMs [1,2]. Suppose we have a simple random sample with replacement of size n from a population. For each unit $i = 1, \dots, n$ and time period $t = 1, \dots, T$, we denote the binary time-varying treatment variable by A_{it} , with $A_{it} = 0$ meaning not being treated at time t and $A_{it} = 1$ being treated at time t , and time-dependent confounders X_{it} . We denote by $\bar{A}_{it} = \{A_{i1}, \dots, A_{it}\}$ the treatment history up to time t and by $\bar{X}_{it} = \{X_{i1}, \dots, X_{it}\}$ the history of confounders up to time t . X_{i1} represents the time-invariant confounders, i.e., confounders that do not depend on past treatments. We denote by $\bar{a}_t$ and $\bar{x}_t$ possible realizations of the treatment history $\bar{A}_{it}$ and the confounder history $\bar{X}_{it}$, respectively. We use $\mathbb{I}[\cdot]$ to denote the indicator so that $\mathbb{I}[\bar{A}_{it} = \bar{a}_t]$ is the variable that is 1 if $\bar{A}_{it} = \bar{a}_t$ and 0 otherwise. To streamline notation, we will refer to $\bar{A}_{iT}$ as $\bar{A}_i$, $\bar{a}_T$ as $\bar{a}$, $\bar{X}_{iT}$ as $\bar{X}_i$, and to $\bar{x}_T$ as $\bar{x}$. For each unit $i = 1, \dots, n$, we denote by Y_i the outcome variable observed at the end of the study. Using the potential outcome framework [26], we denote by $Y_i(\bar{a})$ the potential outcome we would see if we were to apply the treatment regime $\bar{a} \in \mathcal{A}$ to the i th unit, where $\mathcal{A} = \{0, 1\}^T$ is the space of treatment regimes. Throughout, we drop the subscripts i on these variables to refer to a generic unit.

We impose the assumptions of consistency, non-interference, positivity, and sequential ignorability [26,27]. Consistency and non-interference [also known as SUTVA; 28] can be encapsulated in that the potential outcomes are well-defined and the observed outcome corresponds to the potential outcome of the treatment regime applied to that unit, i.e., $Y = Y(\bar{A})$. As previously introduced, positivity states that, for each time $t = 1, \dots, T$, the probability of being treated at time t conditioned on the treatment history up to time $t - 1$ and the confounder history up to time, t , is not 0 or 1, i.e.,

$$0 < \mathbb{P}(A_t = 1 | \bar{A}_{t-1}, \bar{X}_t) < 1 \quad \forall t \in \{1, \dots, T\}. \quad (1)$$

Sequential ignorability states that the potential outcome $Y(\bar{a})$ is independent of treatment assignment at time t , given the treatment history up to time $t - 1$ and the confounder history up to time t . Formally, sequential ignorability is defined as

$$Y(\bar{a}) \perp\!\!\!\perp A_t | \bar{A}_{t-1}, \bar{X}_t \quad \forall t \in \{1, \dots, T\}. \quad (2)$$

An MSM is a model for the marginal causal effect of a time-varying treatment regime on the mean of Y , that is,

$$\mathbb{E}[Y(\bar{a})] = g(\bar{a}, \boldsymbol{\beta}), \quad (3)$$

where $g(\bar{a}, \boldsymbol{\beta})$ is some known function class parameterized by $\boldsymbol{\beta}$. For example, a commonly used MSM is based on additive effects with a common coefficient: $g(\bar{a}, \boldsymbol{\beta}) = \beta_1 + \beta_2 \sum_{t=1}^T a_t$, where the parameter β_2 is the causal parameter of interest. Usually, $\boldsymbol{\beta}$ is computed by a weighted regression of the outcome on the treatment regime alone using weighted least squares (WLSs), i.e., $\min_{\boldsymbol{\beta}} \sum_{i=1}^n W_i (Y_i - g(\bar{A}_i, \boldsymbol{\beta}))^2$, and Wald confidence intervals are constructed using robust (sandwich) standard errors [1,29,30]. In order to consistently estimate

β , the weights $W_{1:n} = (W_1, \dots, W_n)$, must account for the non-randomness of the treatment assignment mechanism, i.e., the confounding. Robins [1] showed that the set of inverse probability weights and stabilized inverse probability weights achieve this objective. These weights are defined as follows:

$$W_i^{\text{IPTW}} = w(\bar{A}_i, \bar{X}_i), \quad w(\bar{a}, \bar{x}) = \prod_{t=1}^T \frac{h_t(\bar{a}_t)}{\mathbb{P}(A_t = a_t | \bar{A}_{t-1} = \bar{a}_{t-1}, \bar{X}_t = \bar{x}_t)}, \quad (4)$$

where $h_t(\bar{a}_t)$ is a known function of treatment history. The set of inverse probability weights is obtained by setting $h_t(\bar{a}_t) = 1$, while the set of stabilized inverse probability weights is obtained by setting $h_t(\bar{a}_t) = \mathbb{P}(A_t = a_t | \bar{A}_{t-1} = \bar{a}_{t-1})$. To estimate weights of the form of equation (4), one first estimates the conditional probability models using either parametric methods such as logistic regression or other machine learning methods [31–33] and then these estimates are plugged in directly into equation (4) to derive weights, which are then plugged into the WLS. Stabilized weights seek to attenuate the variability of inverse probability weights by normalizing them by the marginal probability of treatment. Since the additional factor is a function of treatment regime alone, it does not affect the consistency of the WLS if the MSM is well specified. Both sets of weights, however, rely on plugging in an estimate of a probability into the denominator, meaning that when the true probability is even modestly close to 0, any small error in estimating it can translate to very large errors in estimating the weights and to estimated weights that are extremely variable. Furthermore, both sets of weights rely on the correct specification of the conditional probability models used to estimate the weights in equation (4).

To overcome this issue, Imai and Ratkovic [15] proposed to estimate weights of the form of equation (4) that improve balance of confounders by generalizing the covariate balancing propensity score (CBPS) methodology. Instead of plugging in probability estimates based on logistic regression, CBPS uses the generalized method of moments to find the logistic regression model that if plugged in would lead to weights, W_i^{CBPS} , that approximately solve a subset of the moment conditions that the true inverse probability weights, equation (4), satisfy.

Differently than IPTW, sIPTW, and CBPS, in the next section, we characterize imbalance as the discrepancies in observed average outcomes due to confounding, consider their worst case values, and use quadratic optimization to obtain weights that directly optimize the balance of time-invariant and time-dependent confounders over all possible weights while controlling precision.

3 KOW

In this section, we present a convex-optimization-based approach that obtains weights that minimize the imbalance due to time-dependent confounding (i.e., maximize balance thereof) while controlling precision. Toward that end, in Section 3.1, we provide a definition of imbalance. Specifically, we define imbalance as the sum of discrepancies between the weighted *observed* data and the *unobserved* counterfactual of interest over all treatment regimes. Since this imbalance depends on unknown functions, in Section 3.2 we consider the worst case imbalance, which guards against all possible realizations of the unknown functions. We also show that the worst case imbalance has the attractive characteristic that the number of discrepancies considered grows *linearly* in the number of time periods and not *exponentially* like the number of treatment regimes. We finally show how to minimize this quantity while controlling precision using kernels, RKHS and off-the-shelf solvers for quadratic optimization (Sections 3.3 and 3.4).

3.1 Defining imbalance

Consider any population weights $W = w(\bar{A}, \bar{X})$, where $w(\cdot)$ is a function that depends on the treatment and confounder histories. In this section, we will show that, under consistency and Assumptions (1)–(2), we can decompose the difference between the weighted average outcome among the $\bar{a}$ -treated units, $\mathbb{E}[W \mathbb{1}[\bar{A} = \bar{a}]Y]$,

and the average potential outcome of $\bar{a}$, $\mathbb{E}[Y(\bar{a})]$, into a sum over time points t of discrepancies involving the values of treatment and confounder histories up to time t .

To build intuition we start by explaining this decomposition in the case of two time periods $T = 2$. Assuming consistency and Assumptions (1)–(2), for each $\bar{a} = (a_1, a_2) \in \mathcal{A}$, we can decompose the weighted average outcome among the $\bar{a}$ -treated units as follows:

$$\begin{aligned}
 \mathbb{E}[W\mathbb{1}[\bar{A} = \bar{a}]Y] &= \mathbb{E}[W\mathbb{1}[A_1 = a_1]\mathbb{1}[A_2 = a_2]\mathbb{E}[Y(\bar{a})|A_1, A_2, X_1, X_2]] \\
 &= \mathbb{E}[W\mathbb{1}[A_1 = a_1]\mathbb{1}[A_2 = a_2]\mathbb{E}[Y(\bar{a})|A_1, X_1, X_2]] \\
 &= \mathbb{E}[W\mathbb{1}[A_1 = a_1]\mathbb{E}[Y(\bar{a})|A_1, X_1, X_2]] + \delta_{a_2}^{(2)}(W, g_{\bar{a}}^{(2)}) \\
 &= \mathbb{E}[W\mathbb{1}[A_1 = a_1]\mathbb{E}[Y(\bar{a})|X_1]] + \delta_{a_2}^{(2)}(W, g_{\bar{a}}^{(2)}) \\
 &= \mathbb{E}[Y(\bar{a})] + \delta_{a_1}^{(1)}(W, g_{\bar{a}}^{(1)}) + \delta_{a_2}^{(2)}(W, g_{\bar{a}}^{(2)}) \\
 &= \mathbb{E}[Y(\bar{a})] + \sum_{t=1}^2 \delta_{a_t}^{(t)}(W, g_{\bar{a}}^{(t)}),
 \end{aligned} \tag{5}$$

where the first equality follows from iterated expectation, the second from sequential ignorability, the fourth from iterated expectation and sequential ignorability, and the third and fifth from the following definitions, which exactly capture the difference between the two sides of the third and fifth equalities,

$$\begin{aligned}
 \delta_{a_2}^{(2)}(W, h^{(2)}) &= \mathbb{E}[W\mathbb{1}[A_2 = a_2]h^{(2)}(A_1, X_1, X_2)] - \mathbb{E}[Wh^{(2)}(A_1, X_1, X_2)] \\
 g_{\bar{a}}^{(2)}(A_1, X_1, X_2) &= \mathbb{1}[A_1 = a_1]\mathbb{E}[Y(\bar{a})|A_1, X_1, X_2] \\
 \delta_{a_1}^{(1)}(W, h^{(1)}) &= \mathbb{E}[W\mathbb{1}[A_1 = a_1]h^{(1)}(X_1)] - \mathbb{E}[h^{(1)}(X_1)] \\
 g_{\bar{a}}^{(1)}(X_1) &= \mathbb{E}[Y(\bar{a})|X_1].
 \end{aligned} \tag{6}$$

Note our use of $h^{(t)}$ as a generic dummy function and $g_{\bar{a}}^{(t)}$ as a *specific* function that depends on the particular (unknown) distribution of $\bar{X}_t, \bar{A}_{t-1}, Y(\bar{a})$.

This gives a definition of discrepancy, $\delta_{a_t}^{(t)}(W, h^{(t)})$, where the subscript $a_t \in \{0, 1\}$ refers to the treatment assigned at time t , $W = w(\bar{A}, \bar{X})$ is a population weight, and $h^{(t)}$ is a given function of interest of the treatment and confounder history up to t , $\bar{A}_{t-1}, \bar{X}_t$. The function $g_{\bar{a}}^{(t)}$ is one such function. In particular, for every $a_1 \in \{0, 1\}$, the quantity $\delta_{a_1}^{(1)}(W, h^{(1)})$ is the discrepancy between the $h^{(1)}$ -moments of the baseline confounder distribution in the weighted a_1 -treated population and of the distribution in the whole population. Similarly, for every $a_2 \in \{0, 1\}$, $\delta_{a_2}^{(2)}(W, h^{(2)})$ is a discrepancy in the $h^{(2)}$ -moment of treatment and confounder histories at the start of time step 2. What we have shown above is how these discrepancies directly relate to the difference between weighted averages of observed outcomes and true averages of unknown counterfactuals of interest. Specifically, we have shown that when we measure these discrepancies with respect to the specific function $g_{\bar{a}}^{(t)}$, then their sum gives that difference.

We can extend this decomposition to general horizons $T \geq 1$. Let us define the same discrepancies for any time $t \geq 3$ as

$$\begin{aligned}
 \delta_{a_t}^{(t)}(W, h^{(t)}) &= \mathbb{E}[W\mathbb{1}[A_t = a_t]h^{(t)}(\bar{A}_{t-1}, \bar{X}_t)] - \mathbb{E}[Wh^{(t)}(\bar{A}_{t-1}, \bar{X}_t)], \\
 g_{\bar{a}}^{(t)}(\bar{A}_{t-1}, \bar{X}_t) &= \mathbb{1}[\bar{A}_{t-1} = \bar{a}_{t-1}]\mathbb{E}[Y(\bar{a})|\bar{A}_{t-1}, \bar{X}_t].
 \end{aligned}$$

The following result gives the general decomposition of the difference between weighted average of observed outcomes and true average of counterfactuals as the sum of T discrepancies, one for every time step:

Theorem 1. Under Assumptions (1)–(2), for each $\bar{a} \in \mathcal{A} = \{0, 1\}^T$,

$$\mathbb{E}[W\mathbb{1}[\bar{A} = \bar{a}]Y] - \mathbb{E}[Y(\bar{a})] = \sum_{t=1}^T \delta_{a_t}^{(t)}(W, g_{\bar{a}}^{(t)}).$$

Based on the results of Theorem 1, it is clear that if we want the difference between average counterfactual outcomes and average weighted factual outcomes to be small for all treatment regimes $\bar{a}$, then we should seek weights W that make

$$\bar{\delta}_{\bar{a}}(W, \bar{g}_{\bar{a}}) = \sum_{t=1}^T \delta_{a_t}^{(t)}(W, g_{\bar{a}}^{(t)})$$

small for all $\bar{a}$, where we write $\bar{h} = (h^{(1)}, \dots, h^{(T)})$ for any set of T functions.

The empirical counterparts to $\delta_{a_t}^{(t)}(W, h^{(t)})$ are the sample moment discrepancies for a given set of sample weights $W_{1:n}$:

$$\begin{aligned} \hat{\delta}_{a_t}^{(t)}(W_{1:n}, h^{(t)}) &= \frac{1}{n} \sum_{i=1}^n (W_i \mathbb{1}[A_{it} = a_t] - W_i) h^{(t)}(\bar{A}_{i,t-1}, \bar{X}_{it}), \quad \forall t \geq 2, \\ \hat{\delta}_{a_1}^{(1)}(W_{1:n}, h^{(1)}) &= \frac{1}{n} \sum_{i=1}^n W_i \mathbb{1}[A_{i1} = a_1] h^{(1)}(X_{i1}) - \frac{1}{n} \sum_{i=1}^n h^{(1)}(X_{i1}), \\ \hat{\delta}_{\bar{a}}(W_{1:n}, \bar{h}) &= \sum_{t=1}^T \hat{\delta}_{a_t}^{(t)}(W_{1:n}, h^{(t)}). \end{aligned} \quad (7)$$

Thus, we will seek sample weights $W_{1:n}$ that make $\hat{\delta}_{\bar{a}}(W_{1:n}, \bar{g}_{\bar{a}})$ small for all treatment regimes $\bar{a}$. Toward that end, for any set of given functions $(\bar{h}_{\bar{a}})_{\bar{a} \in \mathcal{A}}$, we define *imbalance* of a set of weights $W_{1:n}$ as the average squared discrepancy over treatment regimes:

$$\text{IMB}(W_{1:n}; (\bar{h}_{\bar{a}})_{\bar{a} \in \mathcal{A}}) = \frac{1}{|\mathcal{A}|} \sum_{\bar{a} \in \mathcal{A}} \hat{\delta}_{\bar{a}}^2(W_{1:n}, \bar{h}_{\bar{a}}). \quad (8)$$

The particular imbalance of interest is given when we consider $\bar{h}_{\bar{a}} = \bar{g}_{\bar{a}}$. One way to control this imbalance, $\text{IMB}(W_{1:n}; (\bar{g}_{\bar{a}})_{\bar{a} \in \mathcal{A}})$, and consequently control the empirical discrepancies of interest, $\hat{\delta}_{\bar{a}}(W_{1:n}, \bar{g}_{\bar{a}})$, is by using inverse probability weights. If known, these weights make this quantity a sample average of mean-zero variables and thus close to zero for large n . However, the difficulties are that (a) even mild practical violations of positivity can lead to large variance of each of these terms and (b) we need to correctly estimate the sequential propensities.

Differently, we will seek to find weights that directly *minimize* imbalance. There are two main challenges in this task. The first challenge is that the imbalance of interest depends on some unknown functions $\bar{g}_{\bar{a}}$. The second is that the number of treatment regimes grows exponentially in the number of time periods. In the next section, we show how the proposed methodology overcomes these two challenges.

3.2 Worst case imbalance

To overcome the fact that we do not actually know the functions $\bar{g}_{\bar{a}}$ on which imbalance $\text{IMB}(W_{1:n}; (\bar{g}_{\bar{a}})_{\bar{a} \in \mathcal{A}})$ depends, we will guard against all possible realizations of the unknown functions. Specifically, since $\hat{\delta}_{\bar{a}}(W_{1:n}, \bar{g}_{\bar{a}})$ scales linearly with $\bar{g}_{\bar{a}}$, we will consider its magnitude relative to that of $\bar{g}_{\bar{a}}$. We therefore need to define a magnitude. In particular, let us define

$$\|\bar{h}\| = \sqrt{\|h^{(1)}\|_{(1)}^2 + \dots + \|h^{(T)}\|_{(T)}^2},$$

where $\|\cdot\|_{(t)}$ are some given extended seminorms on functions from the space of time-dependent confounders and treatment histories up to time t to the space of outcomes. Compared to a norm, an extended seminorm may also assign the values of 0 and ∞ to nonzero elements but must still satisfy triangle inequality and absolute homogeneity. We will discuss specific choices of such seminorms $\|\cdot\|_{(t)}$ in Section 3.4.

Given these, we can define the *worst case discrepancies*,

$$\Delta_{a_t}^{(t)}(W_{1:n}) = \sup_{h^{(t)}} \frac{\hat{\delta}_{a_t}^{(t)}(W_{1:n}, h^{(t)})}{\|h^{(t)}\|_{(t)}} = \sup_{\|h^{(t)}\|_{(t)} \leq 1} \hat{\delta}_{a_t}^{(t)}(W_{1:n}, h^{(t)}).$$

Note that $\Delta_{a_t}^{(t)}(W_{1:n})$ depends *only* on the treatment at time t , a_t , and *not* the whole treatment regime, $\bar{a}$.

Then the *worst case imbalance* is given by

$$\begin{aligned} \mathcal{B}^2(W_{1:n}) &= \sup_{\|\bar{h}_{\bar{a}}\| \leq 1 \quad \forall \bar{a} \in \mathcal{A}} \text{IMB}(W_{1:n}; (\bar{h}_{\bar{a}}^{(t)})_{\bar{a} \in \mathcal{A}}) \\ &= \sup_{\bar{h}_{\bar{a}}, \bar{a} \in \mathcal{A}} \frac{1}{|\mathcal{A}|} \sum_{\bar{a} \in \mathcal{A}} \frac{\hat{\delta}_{\bar{a}}^2(W_{1:n}, \bar{h}_{\bar{a}})}{\|\bar{h}_{\bar{a}}\|^2} \\ &= \frac{1}{2} \sum_{t=1}^T (\Delta_0^{(t)}(W_{1:n})^2 + \Delta_1^{(t)}(W_{1:n})^2). \end{aligned} \quad (9)$$

What is important to note is that this shows that the discrepancies of interest are essentially the same regardless of the particular treatment regime trajectory $\bar{a}$. That is, to control the discrepancies for *all* trajectories $\bar{a}$ for *all* possible realizations of $\bar{g}_{\bar{a}}$, at any time point t , we are only concerned with the discrepancies of histories $\bar{A}_{t-1}, \bar{X}_t$ for those units treated at time t , $A_t = 1$, and for those not, $A_t = 0$. So, while the number of treatment regimes grows *exponentially* in the number of periods, we need only to keep track of and minimize a number of discrepancies growing *linearly* in the number of periods T . By eliminating each of these linearly-many imbalances, any time-dependent confounding would necessarily be removed, as shown by Theorem 1. In Section 5.5, we show how this feature also translates to favorable computational time when dealing with many time periods.

3.3 Minimizing imbalance while controlling precision

We can obtain minimal imbalance by minimizing $\mathcal{B}^2(W)$. However, to control for extreme weights we propose to regularize the weight variables $W_{1:n}$. We therefore wish to find weights that minimize $\mathcal{B}^2(W_{1:n})$ plus a penalty for deviations from uniform weighting. Formally, we want to solve

$$\min_{W_{1:n} \in \mathcal{W}} \mathcal{B}^2(W_{1:n}) + \lambda \|W_{1:n} - e\|_2^2, \quad (10)$$

where e is the vector of ones and $\mathcal{W} = \{W_{1:n} : W_i \geq 0 \forall i\}$ is the space of nonnegative weights $W_{1:n}$. The squared distance of the weights from uniform weights here serves as a convex surrogate for the variance of the resulting MSM (assuming homoskedasticity or bounded residual variances) and λ in equation (10) can be interpreted as a penalization parameter that controls the trade off between imbalance and precision. When λ is equal to zero, the obtained weights provide minimal imbalance. When $\lambda \rightarrow \infty$, the weights become uniformly distributed leading to an ordinary least squares estimator for the MSM.

In the next section, we discuss a specific choice of the norm that specified the worst case discrepancies $\Delta_{a_t}^{(t)}(W_{1:n})$, presented in Section 3.2. Specifically, we show that by choosing an RKHS to specify the norm, we can express the optimization problem in equation (10) as a convex-quadratic function in $W_{1:n}$, which can be easily solved by using off-the-shelf solvers for quadratic optimization.

3.4 RKHS and quadratic optimization to balance time-dependent confounders

An RKHS is a Hilbert space of functions which is associated with a kernel (the reproducing kernel). Specifically, any positive semi-definite kernel $\mathcal{K} : \mathcal{Z} \times \mathcal{Z} \rightarrow \mathbb{R}$ on a ground space $\mathcal{Z}$ defines a Hilbert space given by (the unique completion of) the span of all functions $\mathcal{K}(z, \cdot)$ for $z \in \mathcal{Z}$, endowed with the

inner product $\langle \mathcal{K}(z, \cdot), \mathcal{K}(z', \cdot) \rangle = \mathcal{K}(z, z')$. Kernels are widely used in machine learning to generalize the structure of conditional expectation functions with many applications in statistics [24, 34–36]. Commonly used kernels are the polynomial, Gaussian, and Matérn kernels [34].

The following theorem shows that if $\|\cdot\|_{(t)}$, the norm that specified the worst case discrepancies $\Delta_{a_t}^{(t)}(W_{1:n})$, is an RKHS norm given by the kernel $\mathcal{K}_t$, then we can express it as a convex-quadratic function in $W_{1:n}$.

Theorem 2. Define the matrix $K_t \in \mathbb{R}^{n \times n}$ as

$$K_{tij} = \mathcal{K}_t(\bar{A}_{i,t-1}, \bar{X}_{it}), (\bar{A}_{j,t-1}, \bar{X}_{jt})$$

and note that it is positive semidefinite by definition. Then, if the norm $\|\cdot\|_{(t)}$ is the RKHS norm given by the kernel $\mathcal{K}_t$, the squared worst case discrepancies are

$$\begin{aligned} \Delta_{a_t}^{(1)}(W_{1:n})^2 &= \frac{1}{n^2} W_{1:n}^T I_{a_t}^{(1)} K_t I_{a_t}^{(1)} W_{1:n} - 2e^T K_t I_{a_t}^{(1)} W_{1:n} + e^T K_t e, \\ \Delta_{a_t}^{(t)}(W_{1:n})^2 &= \frac{1}{n^2} W_{1:n}^T (I - I_{a_t}^{(t)}) K_t (I - I_{a_t}^{(t)}) W_{1:n}, \end{aligned}$$

where I is the identity matrix and $I_{a_t}^{(t)}$ is the diagonal matrix with $\mathbb{I}[A_{it} = a_t]$ in its i th diagonal entry.

As an aside, we note that, when $\|\cdot\|_t$ is given by an RKHS norm, our worst case squared discrepancies can in fact equivalently also be written as *average* square discrepancies over random function following a Gaussian process with covariance operator given by the same kernel as reproduces the RKHS (see ref. [24], Proposition 16). This is a consequence of the norm being given by an inner product in a Hilbert space.

Based on Theorem 2, we can now express the worst case imbalance, $\mathcal{B}^2(W_{1:n})$, defined in equation (9), as a convex-quadratic function. Specifically, let $K_t^\circ = I_0^{(t)} K_t I_0^{(t)} + I_1^{(t)} K_t I_1^{(t)}$, which is given by setting every entry i, j of K_t to 0 whenever $A_{it} \neq A_{jt}$, and $K^\circ = \sum_{t=1}^T K_t^\circ$. We then get that

$$\mathcal{B}^2(W_{1:n}) = \frac{1}{2} \sum_{t=1}^T (\Delta_0^{(t)}(W_{1:n})^2 + \Delta_1^{(t)}(W_{1:n})^2) = \frac{1}{n^2} \left(\frac{1}{2} W_{1:n}^T K^\circ W_{1:n} - e^T K_t W_{1:n} + e^T K_t e \right). \quad (11)$$

Finally, to obtain weights that balance covariates to control for time-dependent confounding while controlling precision we solve the quadratic optimization problem,

$$\min_{W_{1:n} \in \mathcal{W}} \frac{1}{2} W_{1:n}^T K_\lambda^\circ W_{1:n} - e^T K_\lambda W_{1:n} \quad (12)$$

where $K_\lambda^\circ = K^\circ + 2\mathcal{M}$, $K_\lambda = K_t + 2\mathcal{M}$. We call this proposed methodology and the result of equation (12), KOW.

4 Practical guidelines

Solutions to the quadratic optimization problem (12) depend on several factors. First, they depend on the choice of the kernel and its hyperparameters. There are some existing practical guidelines on these choices [34, 37], on which we rely as explained below. Second, they depend on the penalization parameter λ . Finally, solutions to equation (12) depend on the chosen set of lagged covariates to include in each kernel. In this section, we introduce some practical guidelines on how to apply KOW in consideration of these factors.

For each t , the unknown function $g_{\bar{a}}^{(t)}(\bar{A}_{t-1}, \bar{X}_t)$ has two distinct inputs: the treatment history and the confounder history. To reflect this structure, we suggest to specify the kernel $\mathcal{K}_t$ as a *product kernel*, i.e., $\mathcal{K}_t(\bar{a}_{t-1}, \bar{x}_t), (\bar{a}'_{t-1}, \bar{x}'_t) = \mathcal{K}_t^{(1)}(\bar{a}_{t-1}, \bar{a}'_{t-1}) \mathcal{K}_t^{(2)}(\bar{x}_t, \bar{x}'_t)$ given a treatment history kernel $\mathcal{K}_t^{(1)}$ and a confounder history kernel $\mathcal{K}_t^{(2)}$. This simplifies the process of specifying the kernels without placing strong restrictions. We further suggest that for the treatment history to use a linear kernel involving ℓ lagged

treatments, $\mathcal{K}_t^{(1)}(\bar{a}_{t-1}, \bar{a}'_{t-1}) = \sum_{s=\max(1, t-\ell)}^{t-1} a_s a'_s$ and for the confounder history to use a polynomial kernel involving the time-invariant confounders and ℓ lagged time-dependent confounders, $\mathcal{K}_t^{(d)}(\bar{x}_t, \bar{x}'_t) = (1 + \theta x_1^T x'_1 + \theta \sum_{s=\max(2, t-\ell+1)}^t x_s^T x'_s)^d$, where $\theta > 0$ and $d \in \mathbb{N}$ are hyperparameters. The linear kernel in treatments includes linear functions in the treatment, and the product specification for the whole kernel covers forms that include interactions between the ℓ lagged treatments and up-to-order- d terms in the covariates. We discuss the choice of the number of lags and the hyperparameters below. In our simulation study in Section 5, we show that the MSE of the MSM-estimated effect using KOW with a product of linear kernel and a quadratic kernel ($d = 2$) outperforms estimates using weights obtained by IPTW, siPTW, and CBPS in all considered simulated scenarios. We again use this choice of kernels in our empirical applications of KOW to real datasets in Section 7. Many other choices of kernel are also possible and may be more appropriate in a particular application, but we suggest the above combination as a generic and successful recipe.

When using kernels, preprocessing the data is an important step. In particular, normalization is employed to avoid unit dependence and covariates with high variance dominating those with smaller ones. Consequently, we suggest, beforehand, to scale the covariates related to the treatment and confounder histories to have mean 0 and variance 1.

To tune the kernels' hyperparameters and the penalization parameter λ , we follow ref. [24] and use the empirical Bayes approach of marginal likelihood [37]. We postulate a Gaussian process prior to $g^{(t)} \sim \mathcal{GP}(c_t \mathbf{1}, \mathcal{K}_t(\theta))$, where $c_t \mathbf{1}$ is a constant function and $\mathcal{K}_t(\theta)$ is a kernel that depends on some set of hyperparameters θ_t . For each t , we then maximize the marginal likelihood of seeing the data $Y \sim \mathcal{N}(g^{(t)}(\bar{X}_t, \bar{A}_{t-1}), \lambda_t)$ over θ_t, λ_t, c_t and let $\lambda = \sum_{t=1}^T \lambda_t$. It would be more correct to consider the marginal likelihood of observing the partial means of outcomes, but we find that this much simpler approach suffices for learning the right representation of the data (θ_t) and the right penalization parameter (λ) and it enables the use of existing packages such as GPML [38]. We demonstrate this in the simulations presented in Section 5, and in particular in Figures 4 and 5 we see that this approach leads to a value of the penalization parameter that is near that which minimizes the resulting MSE of the MSM over possible parameters.

Another practical concern is how many lagged covariates to include in each of the kernels $\mathcal{K}_t$. When deriving inverse probability weights, it is common to model the denominator in equation (4) by fitting a pooled logistic model [39] including only the time-invariant confounders, X_1 , the time-dependent confounders at time t , X_t , and the one-time lagged treatment history, A_{t-1} , rather than the entire histories, i.e., logit $P(A_t = a_t | \bar{A}_{t-1} = \bar{a}_{t-1}, \bar{X}_t = \bar{x}_t) = \alpha_t + \beta_1 A_{t-1} + \beta_2 X_1 + \beta_3 X_t$ [30,40]. This can be understood as a certain Markovian assumption about the data generating process which simplifies the modeling when T is large. The same can be done in the case of KOW, where we may assume that $g_a^{(t)}$ is only a function of the one-time lagged treatment, the time-dependent confounders at time t , and the time-invariant confounders, i.e., $g_a^{(t)}(\bar{A}_{t-1}, \bar{X}_t) = g_a^{(t)}(A_{t-1}, X_1, X_t)$, and correspondingly let the kernel K_t only depend on A_{t-1} , X_1 , and X_t . More generally, we can consider including any amount of lagged variables, as represented by the parameter ℓ in the above specification of the linear and polynomial kernels. In Section 7.2, we consider an empirical setting where T is small and specify the kernels using the whole treatment and confounder histories ($\ell = T$). However, in Section 7.1 we consider a setting where T is large and, following previous approaches studying the same dataset using IPTW with a logistic model of only the one-time lags [30,40,41], we keep only the baseline and one-time-lagged data in each kernel specification ($\ell = 1$).

Certain datasets, such as the one we study in Section 7.1, have repeated observations of outcomes at each time $t = 1, \dots, T$. Thus, for each subject, we have T observations to be used to fit the MSM. Correspondingly, each observation should be weighted appropriately. This can be seen as T instances of the weighting problem. For siPTW, this boils down to restricting the products in the numerator and denominator of equation (4) to be only up to t for each $t = 1, \dots, T$. Similarly, in the case of KOW, we propose to solve equation (12) for each value of $t = 1, \dots, T$, producing $n \times T$ weights, one for each of the outcome observations, to be used in fitting the MSM. This is demonstrated in Section 7.1.

In the case of a single, final observation of outcome, normalizing the weights, whether IPTW or KOW, does not affect the fitted MSM as it amounts to multiplying the least-squares loss by a constant factor. But in the repeated observation setting described above, normalizing each set of weights for each time period

separately can help. Correspondingly, we can add a constraint to equation (12) that the mean of the weights must equal one for each time period separately, which we demonstrate in Section 7.1.

Similar to refs [1,29,30] we suggest using Wald confidence intervals constructed using robust (sandwich) standard errors.

5 Simulations

In this section, we show the results of a simulation study aimed at comparing the bias and MSE of estimating the cumulative effect of a time-varying treatment on a continuous outcome by using an MSM with weights obtained by each of KOW, IPTW, sIPTW, and CBPS.

5.1 Setup

We considered two different simulated scenarios with $T = 3$ time periods, (1) linear, where the treatment was modeled linearly, and (2) nonlinear, where it was modeled quadratically. In both scenarios, we modeled the outcomes nonlinearly so as not to favor our method unfairly. We tuned the kernel's hyperparameters and the penalization parameter as presented in Section 4 and computed bias and MSE over 1,000 replications for each of the varying sample sizes, $n = 100, \dots, 1,000$. In addition, to study the impact of the penalization parameter λ on bias and MSE, in both scenarios, we fixed the sample size to $n = 500$ and considered a grid of 25 values for λ .

For the linear scenario, we drew the data from the following model:

$$Y_i = -1.91 + 0.8 \sum_{t=1}^T A_{i,t} + 0.5 \sum_{k=1}^3 Z_{i,k} + 0.05 \sum_{k \neq m} Z_{i,k} Z_{i,m} + N(0, 5),$$

where $Z_{i,k} = \sum_{t=1}^T X_{i,t,k}$, $A_{i,t} \sim \text{binom}(\pi_{i,t}^{(L)})$, $X_{i,t,k} \sim N(X_{i,t-1,k} + 0.1, 1)$, $k = 1, 2, 3$, and

$$\pi_{i,t}^{(L)} = \left(1 + \exp \left(- \left(0.5 + 0.5A_{i,t-1} + 0.05X_{i,t,1} + 0.08X_{i,t,2} - 0.03X_{i,t,3} + 0.2A_{i,t-1} \sum_{k=1}^3 X_{i,t,k} \right) \right) \right)^{-1}.$$

For the nonlinear scenario, we drew the data from the following model:

$$Y_i = -21.46 + 0.8 \sum_{t=1}^T A_{i,t} + 0.5 \sum_{k=1}^3 Z_{i,k} + 0.1 \left(\sum_{k \neq m} Z_{i,k} Z_{i,m} \right) + N(0, 5),$$

where $Z_{i,k} = \sum_{t=1}^T X_{i,t,k}^2$, $A_{i,t} \sim \text{binom}(\pi_{i,t}^{(NL)})$, $X_{i,t,k} \sim N(X_{i,t-1,k} + 0.1, 1)$, $k = 1, 2, 3$, and

$$\begin{aligned} \pi_{i,t}^{(NL)} = & \left(1 + \exp \left(- \left(0.5 + 0.5A_{i,t-1} + 0.05X_{i,t,1} + 0.08X_{i,t,2} - 0.03X_{i,t,3} \right. \right. \right. \\ & + 0.025X_{i,t,1}^2 + 0.04X_{i,t,2}^2 - 0.015X_{i,t,3}^2 + 0.3 \sum_{k \neq m} X_{i,t,k} X_{i,t,m} \\ & \left. \left. \left. + 0.2A_{i,t-1} \sum_{k=1}^3 X_{i,t,k} + 0.1A_{i,t-1} \sum_{k=1}^3 X_{i,t,k}^2 + 0.05A_{i,t-1} \sum_{k \neq m} X_{i,t,k} X_{i,t,m} \right) \right) \right)^{-1} \end{aligned}$$

The intercepts -1.91 and -21.46 are chosen so the marginal mean of Y_i is 0.

In each scenario and for each replication, we computed two sets of KOW weights. We obtain the first by using the product of two linear kernels ($\mathcal{K}_1$), one for the treatment history and one for the confounder

history, and the second by using the product of a linear kernel for the treatment history and a quadratic kernel for the confounder history ($\mathcal{K}_2$). As presented in Section 4, we rescaled the variables before inputting them to the kernel and, for each replication, we tuned λ and the kernels' hyperparameters by using Gaussian-process marginal likelihood. We also computed two sets of IPTW and sIPTW weights. We obtained the first by fitting for each $t = 1, 2, 3$ a logistic regression model for the treatment $\bar{A}_{i,t}$ conditioned on $\bar{A}_{i,t-1}$, $\bar{X}_{i,t}$ and their interactions, which is well-specified for $\pi_{i,t}^{(L)}$ (we term this the linear specification) and the second by adding all quadratic confounder terms and their interactions with $\bar{A}_{i,t-1}$ which is well-specified for $\pi_{i,t}^{(NL)}$ (we term this the non-linear specification). The numerator of sIPTW in either case was obtained by fitting a logistic regression on the treatment history alone. We obtain the final set of IPTW and sIPTW weights by multiplying the weights over t as shown in equation (4). Finally, we computed two sets of weights using CBPS: one using the covariates as they are (linear CBPS) and one by augmenting the covariates with all quadratic monomials (non-linear CBPS). We used the full (non-approximate) version of CBPS, which outperformed the approximate version in all of our experiments.

We computed the causal parameter of interest by using WLS, regressing the outcome on the cumulative treatment and using weights computed by each of the methods. Specifically, in the linear scenario, we computed weights using (1) $\mathcal{K}_1$ for KOW, the linear specification for IPTW and sIPTW, and linear CBPS, which we refer to as the *correct* case, and (2) $\mathcal{K}_2$ for KOW, the nonlinear specification for IPTW and sIPTW, and the nonlinear CBPS, which we refer to as the *overspecified* case. In the nonlinear scenario, we again used each of the above, but refer to the first as the *misspecified* case and the second as the *correct* case. We highlight that these terms may only reflect the model specification for IPTW and sIPTW, as CBPS does not require a particular specification and the function $g_a^{(t)}$ need not necessarily be in the RKHS that either kernel specify.

We used Gurobi and its R interface [42] to solve equation (10) and optimize the KOW weights, the MatLab package GPML [38] to perform the marginal likelihood estimation of hyperparameters, the R package R.matlab to call MatLab from within R, the R command glm to fit treatment models for IPTW and sIPTW, the R package CBMSM for CBPS, and the R command lm to fit the MSM.

5.2 Results

In this section, we discuss the results obtained in the simulation study across sample sizes and across values of the penalization parameter, λ . In summary, the proposed KOW outperformed IPTW, sIPTW, and CBPS with respect to MSE across all sample sizes and simulation scenarios. An important result is that, in the misspecified case, KOW showed a lower bias and MSE than that of IPTW, sIPTW, and CBPS across all considered sample sizes.

5.3 Across sample sizes

Figure 1 shows bias and MSE of the estimated time-varying treatment effect using KOW (solid), IPTW (dashed), sIPTW (dotted), and CBPS (dashed-dotted) when increasing the sample size from $n = 100$ to $n = 1,000$. In the linear-correct scenario, IPTW had a lower bias compared with sIPTW, CBPS, and KOW in small samples (top-left panel of Figure 1). However, for larger samples, KOW had a smaller bias compared with IPTW, sIPTW and CBPS. KOW outperformed IPTW, sIPTW, and CBPS in terms of MSE across samples sizes (top-right panel of Figure 1). KOW outperformed the other methods with regard to MSE (bottom-right panel of Figure 1) across all sample sizes, in the linear-overspecified scenario. KOW and sIPTW performed similarly with respect to bias in the nonlinear-misspecified scenario (top-left panel of Figure 2), while KOW outperformed IPTW, sIPTW, and CBPS with respect to MSE in all sample sizes (top-right panel of Figure 2). KOW, IPTW, and sIPTW had similar bias in the nonlinear-correct scenario (bottom-left panel of Figure 2), with KOW outperforming the other methods, with respect to MSE, across all sample

Figure 1: Bias and MSE of the estimated time-varying treatment effect using KOW (solid), IPTW (dashed), sIPTW (dotted), and CBPS (dashed-dotted) when increasing the sample size from $n = 100$ to $n = 1,000$ in the linear-correct scenario (top panels) and in the linear-overspecified scenario (bottom panels).

sizes (bottom-right panel of Figure 2). In summary, the MSE obtained by using KOW was lower than that of IPTW, sIPTW, and CBPS across all considered sample sizes. As the next section shows, the larger biases in some of the cases are driven by the choice of penalization parameter λ . Here we choose λ with an eye toward minimizing MSE. A smaller λ , it is shown next, can lead to KOW having *both* smaller bias and MSE than other methods, but the total benefit in MSE is smaller. Figure 3 shows coverage of the 95% Wald confidence interval using KOW across sample sizes and across simulation scenarios. Coverage was close to the nominal level under the linear-correct (solid), linear-overspecified (dot-dashed), and nonlinear-correct (dotted). Lower levels of coverage were obtained under the nonlinear-misspecified scenario (dashed).

Figure 2: Bias and MSE of the estimated time-varying treatment effect using KOW (solid), IPTW (dashed), sIPTW (dotted), and CBPS (dashed-dotted) when increasing the sample size from $n = 100$ to $n = 1,000$, in the nonlinear-misspecified scenario (top panels) and in the nonlinear-correct scenario (bottom panels).

5.4 Across values of the penalization parameter, λ

Figures 4 and 5 show the ratios of squared biases (left panels) and of MSEs (right panels) when comparing KOW (denominator) with IPTW (solid), sIPTW (dashed), and CBPS (dotted) (numerators) across different values of λ and $n = 500$ in the linear and nonlinear scenarios, respectively. Since λ affects only KOW, squared biases and of MSEs obtained using IPTW, sIPTW, and CBPS were the same across values of λ . Values above 1 means that KOW had a lower bias or MSE. For zero or small λ , KOW significantly outperformed IPTW, sIPTW, and CBPS with respect to bias. In many cases, the MSE was also smaller for zero λ . But, the biggest benefit in MSE was seen for larger λ . The peaks of the right panels represent the points for which λ is optimal, i.e., the MSE of KOW is minimized. The solid vertical lines on the right panels show the mean values across replications of the λ value obtained by the procedure described in Sections 4 and 5.1 as

Figure 3: Coverage of the 95% Wald confidence interval using KOW across sample sizes and across simulation scenarios: linear-correct (solid), linear-overspecified (dot-dashed), nonlinear-correct (dotted) and nonlinear-misspecified (dashed).

done in the previous section. It can be seen that these are very near the points at which the MSE is minimized. The benefit in MSE both at and around this point was significant across all scenarios.

5.5 Computational time of KOW

In this section, we present the results of a simulation study aimed at comparing the mean computational time of KOW and CBPS. Compared to SIPTW based on pooled logistic regression, which is generally very fast, both KOW and CBPS have a nontrivial computational time that can grow with both the total number time periods T and the number of covariates (which, for KOW, manifests as the complexity of the kernel functions). For KOW, the most time-consuming tasks are tuning λ by marginal likelihood and computing the matrices that define problem (12), which are affected by these two factors, while solving problem (12) is fast and does not depend on those factors. CBPS computational time is dominated by inverting a covariance matrix with dimensions increasing exponentially in T and linearly in the number of covariates. Imai and Ratkovic [15] also propose using an approximate low-rank matrix that ignores certain covariance terms to make the matrix inversion faster.

Here we compare KOW, CBPS with full covariance matrix (CBPS-full), and CBPS with its low-rank approximation (CBPS-approx) when increasing the number of time periods and the number of covariates. Specifically, following the linear-correct scenario presented in Section 5.1, we fixed the sample size equal to $n = 100$ and randomly generated 100 samples considering $T = 3, \dots, 10$, and $p = 3, \dots, 8$, where p is the total number of covariates X_t for each t . We fixed the number of covariates to be equal to $p = 3$ when evaluating the mean computational times over time periods, while we fixed the number of time periods to be equal to $T = 5$ when analyzing over the number of covariates. For each sample, we computed the KOW weights by solving equation (12) using kernel K_t . We used Gaussian process marginal likelihood to tune the kernels' hyperparameters and penalization parameter. We computed CBPS weights using the linear CBPS as in

Figure 4: Ratios of squared biases and MSEs comparing KOW with IPTW (solid), sIPTW (dashed), and CBPS (dotted) across values of $\lambda = 0, \dots, 1,500$ in the linear-correct scenario (top panels) and in the linear-overspecified scenario (bottom panels). Ratios above 1 means that KOW had a lower bias or MSE. Vertical bars show the mean value of λ , across simulations, obtained as described in Section 5.3.

Section 5.1. We used the R package `rbenchmark` to compute the computational time on a PC with an i7-3770 processor, 3.4 GHz, 8GB RAM, and a Linux Ubuntu 16.04 operating system.

Solid lines of Figure 6 represent mean computational times for KOW, dashed for CBPS-full, and dotted for CBPS-approx. When the number of time periods was relatively small, the mean computational time of KOW was higher compared with both CBPS methods (left panel of Figure 6). However, the mean computation time of KOW over time periods increased linearly while that of both CBPS methods increased exponentially. This is due to the fact that, as presented in Section 3.1, the number of imbalances that we need to minimize grows linearly in the number of time periods. The mean computational time required by KOW when increasing the number of covariates remained constant, while it increased for both CBPS-full and CBPS-approx, with CBPS-full increasing more rapidly. In summary, KOW was less affected by the total number of time periods and covariates compared with CBPS with full and low-rank approximation matrix.

Figure 5: Ratios of squared biases and MSEs comparing KOW with IPTW (solid), sIPTW (dashed), and CBPS (dotted) across values of $\lambda = 0, \dots, 3,000$ in the nonlinear-misspecified scenario (top panels) and in the nonlinear-correct scenario (bottom panels). Ratios above 1 means that KOW had a lower bias or MSE. Vertical bars show the mean value of $\lambda = 0, \dots, 3,000$, across simulations, obtained as described in Section 5.3.

Computing KOW required three steps: tuning the parameters, constructing the matrices for problem (12), and solving problem (12). On average, for $T = 3$, the first step required 21% of the total computational time, the second 78.8%, and the last 0.2%. Thus, solving the optimization problem itself is very fast and is not the bottleneck.

6 KOW with informative censoring

In longitudinal studies, participants may drop out the study before the end of the follow-up time and their outcomes are, naturally, missing observations. When this missingness is due to reasons related to the study (i.e., related to the potential outcomes), selection bias is introduced. This phenomenon is referred to as

Figure 6: Mean computational time in seconds of KOW (solid), CBPS with full covariate matrix (dashed), and CBPS with the low-rank approximation of the full matrix (dotted) over time periods when $n = 100$, $p = 3$ and $T = 2, \dots, 10$ (left panel) and over number of covariates, when $n = 100$, $T = 5$ and $p = 3, \dots, 8$ (right panel).

informative censoring and it is common in the context of survival analysis where the interest is on analyzing time-to-event outcomes. Under the assumptions of consistency, positivity, and sequential ignorability of both treatment and censoring, Robins et al. [43] showed that a consistent estimate of the causal effect of a time-varying treatment can be obtained by weighting each subject $i = 1, \dots, n$ at each time period by the product of inverse probability of treatment and censoring weights. Inverse probability of treatment weights is obtained as presented in Section 2, while inverse probability of censoring weights is usually obtained by inverting the probability of being uncensored at time t , given the treatment and confounder history up to time t [30].

In this section, we extend KOW to similarly handle informative censoring. We demonstrate that under sequentially ignorable censoring, minimizing the very same discrepancies as before at each time period, restricted to the units for which data are available, actually controls for both time-dependent confounding and informative censoring. Thus, KOW naturally extends to the setting with informative censoring.

Let $C_{it} \in \{0, 1\}$ for $t = 1, \dots, T$ indicate whether unit i is censored in time period t and let $C_{i0} = 0$. Note that $C_{it} = 1$ implies that $C_{i,t+1} = 1$ and that $C_{it} = 0$ implies that $C_{i,t-1} = 0$. All we require is that we (at least) observe outcomes Y_i whenever $C_{iT} = 0$, X_{it} whenever $C_{i,t-1} = 0$, and A_{it} whenever $C_{it} = 0$. Note we might observe more, such as the treatment at time t for a unit with $C_{i,t-1} = 0$, or perhaps only some of the data after censoring is corrupted, but that is not required. We summarize the assumption of sequentially ignorable censoring as

$$Y(\bar{a}) \perp\!\!\!\perp \bar{C}_t | \bar{A}_t, \bar{X}_t. \quad (13)$$

Let us redefine

$$\begin{aligned} \delta_{a_1}^{(1)}(W, h^{(1)}) &= \mathbb{E}[W \mathbb{1}[A_1 = a_1] \mathbb{1}[C_1 = 0] h^{(1)}(X_1)] - \mathbb{E}[h^{(1)}(X_1)] \\ g_{\bar{a}}^{(1)}(X_1) &= \mathbb{E}[Y(\bar{a}) | X_1], \\ \delta_{a_t}^{(t)}(W, h^{(t)}) &= \mathbb{E}[W \mathbb{1}[A_t = a_t] \mathbb{1}[C_t = 0] h^{(t)}(\bar{A}_{t-1}, \bar{X}_t)] - \mathbb{E}[W \mathbb{1}[C_{t-1} = 0] h^{(t)}(\bar{A}_{t-1}, \bar{X}_t)], \quad \forall t \geq 2, \\ g_{\bar{a}}^{(t)}(\bar{A}_{t-1}, \bar{X}_t) &= \mathbb{1}[\bar{A}_{t-1} = \bar{a}_{t-1}] \mathbb{E}[Y(\bar{a}) | \bar{A}_{t-1}, \bar{X}_t], \quad \forall t \geq 2. \end{aligned} \quad (14)$$

Similar to Theorem 1, the following theorem shows that we can write the difference between the weighted average outcome among the *uncensored* $\bar{a}$ -treated units, $\mathbb{E}[W \mathbb{1}[\bar{A} = \bar{a}] \mathbb{1}[C_T = 0] Y]$, and the true average potential outcome of $\bar{a}$, $\mathbb{E}[Y(\bar{a})]$, as the sum over time points t of discrepancies involving the values of treatment and confounder histories up to time t .

Theorem 3. Under Assumptions (1)–(2) and (13),

$$\mathbb{E}[W\mathbb{I}[\bar{A} = \bar{a}]\mathbb{I}[C_T = 0]Y] - \mathbb{E}[Y(\bar{a})] = \sum_{t=1}^T \delta_{a_t}^{(t)}(W, g_{\bar{a}}^{(t)}). \quad (15)$$

We then define the empirical counterparts to $\delta_{a_t}^{(t)}(W, h^{(t)})$ as before in equation (7) but limit ourselves to *uncensored* units, as in equation (14). We similarly define imbalance, $\text{IMB}(W_{1:n}; (\bar{g}_{\bar{a}}^{(t)})_{\bar{a} \in \mathcal{A}})$, and the worst case imbalance $\mathcal{B}^2(W_{1:n})$, as before in equations (8) and (9). Finally, again using kernels to specify norms, we obtain weights that balance covariates to control for time-dependent confounding and account for informative censoring while controlling precision by solving the quadratic optimization problem,

$$\min_{W_{1:n} \in \mathcal{W}} \frac{1}{2} W_{1:n}^T K_{\lambda}^{\circ} W_{1:n} - e^T K_{\lambda} W_{1:n}, \quad (16)$$

where $K_{\lambda}^{\circ} = K^{\circ} + 2\lambda I$, $K_{\lambda} = K + 2\lambda I$, $K^{\circ} = \sum_{t=1}^T K_t^{\circ}$, $K_t^{\circ} = \sum_{a_t \in \{0,1\}} I_{a_t}^{(t)} K_t I_{a_t}^{(t)}$, $K_t \in \mathbb{R}^{n \times n}$ is a semidefinite positive matrix defined as $K_{tij} = \mathcal{K}_t((\bar{A}_{i,t-1}, \bar{X}_{it}), (\bar{A}_{j,t-1}, \bar{X}_{jt}))$, $I_{a_t}^{(t)}$ is the diagonal matrix with $\mathbb{I}[A_{it} = a_t]\mathbb{I}[C_{it} = 0] - \mathbb{I}[C_{i,t-1} = 0]$ in its i th diagonal entry (recall $C_{i,0} = 0$ for all i), and e is the vector of all ones.

7 Applications

In this section, we present two empirical applications of KOW. In the first, we estimate the effect of treatment initiation on time to death among PLWH. In the second, we evaluate the impact of negative advertising on election outcomes.

7.1 Effect of HIV treatment on time to death

In this section, we analyze data from the Multicenter AIDS Cohort Study (MACS) to study the effect of the initiation time of treatment on time to death among PLWH. Indeed, due to the longitudinal nature of HIV treatment and the presence of time-dependent confounding, MSMs have been widely used to study causal effects in this domain [3,30,41,44,45, among others]. As an example of time-dependent confounding, CD4 cell count, a measurement used to monitor immune defenses in PLWH and to make clinical decisions, is a predictor of both treatment initiation and survival, as well as being itself influenced by prior treatments. Recognizing the censoring in the MACS data, Hernán et al. [41] showed how to estimate the parameters of the MSM by inverse probability of treatment and censoring weighting (IPTCW).

Here, we apply KOW as proposed in Section 6 to handle both time-dependent confounding and informative censoring while controlling precision. We considered the following potential time-dependent confounders associated with the effect of treatment initiation and the risk of death: CD4 cell count, white blood cell count, red blood cell count, and platelets. We also identified the age at baseline as a potential time-invariant confounding factor. We considered only recently developed HIV treatments, thus, including in the analysis only PLWH that started treatment after 2001. The final sample was composed of a total of $n = 344$ people and 760 visits, with a maximum of $T = 8$ visits per person. We considered two sets of KOW weights, either obtained by using a product of (1) two linear kernels, one for the treatment history and one for the confounder history ($\mathcal{K}_1$) or (2) a linear kernel for the treatment history and a polynomial kernel of degree 2 for the confounder history ($\mathcal{K}_2$). We scaled the covariates related to the treatment and confounder history, and tuned the kernels' hyperparameters and the penalization parameter by using Gaussian processes marginal likelihood as presented in Section 4. Following previous approaches studying the HIV treatment using IPTCW that modeled treatment and censoring using single time lags [30,40,41], we included in each kernel the time-invariant confounders, the previous treatment, A_{t-1} , and the time-dependent confounders

at time t , X_t , instead of the entire histories. As described in Section 4, since we have repeated observations of outcomes, we compute a set of KOW weights by solving the optimization problem (16) for each horizon up to T . In addition, as described in Section 4, we constrained the mean of the weights to be equal to one.

We compared the results obtained by KOW with those from IPTCW and sIPTCW. The latter sets of weights were obtained by using a logistic regression on the treatment history and the aforementioned time-invariant and time-dependent confounders and using only one time lag for each of the treatment and time-dependent confounders as done in previous approaches studying the HIV treatment using IPTCW [30,40,41]. The numerator of sIPTCW was computed by modeling $h(\bar{A}_t)$ in equation (4) with a logistic regression on the treatment history only using one time lag. We modeled the inverse probability of censoring weights similarly. The final sets of IPTCW and sIPTCW weights were obtained by multiplying inverse probability of treatment and censoring weights. We did not compare the results with those of CBPS because it does not handle informative censoring. In particular, CBPS requires a complete $n \times T$ matrix of observed time-dependent confounders, while in the MACS dataset many entries are missing.

We estimated the hazard ratio of the risk of death by using a weighted Cox regression model [41] weighted by KOW, IPTCW, or sIPTCW and using robust standard errors [29]. We used Gurobi and its R interface to solve equation (16) and obtain the KOW weights, the MatLab package GPML to perform the marginal likelihood estimation of hyperparameters, the R package R.matlab to call MatLab from within R, the R package ipw [46] to fit the treatment models for IPTCW and sIPTCW, and the R command coxph (with robust variance estimation) to fit the outcome model. It took 13.5 s to obtain a solution for KOW. Table 1 summarizes the result of our analysis. Both KOW ($\mathcal{K}_1$) and ($\mathcal{K}_2$) showed a significant protective effect of HIV treatment on time to death among PLWH. IPTCW showed a similar effect but with lower precision, resulting in a non-significant effect. With similar precision obtained by KOW, sIPTCW showed a non-significant effect of HIV treatment on time to death. Whereas analyses based on IPTCW and sIPTCW lead to non-significant and inconsistent conclusions, the results we obtained by using KOW show that PLWH can benefit from HIV treatment, as shown in independent randomized placebo-controlled trials [47,48].

7.2 Impact of negative advertising on election outcomes

In this section, we analyze a subset of the dataset from the study by Blackwell [5] to estimate the impact of negative advertising on election outcomes. Because of the dynamic and longitudinal nature of the problem and presence of time-dependent confounders, MSMs have been previously used to study the question [5]. Specifically, poll numbers are time-dependent confounders as they might both be affected by negative advertising and might also affect future poll numbers. We constructed the subset of the data from the study by Blackwell [5] by considering the 5 weeks leading up to each of 114 elections held 2000–2006 (58 US Senate, 56 US gubernatorial). Differently from Section 7.1 in which the outcome was observed at each time

Table 1: Effect of HIV treatment on time to death

	KOW		Logistic	
	$\mathcal{K}_1$	$\mathcal{K}_2$	IPTCW	sIPTCW
H $\hat{R}$	0.40*	0.48*	0.14	1.25
SE	(0.30)	(0.28)	(1.15)	(0.30)

Note: H $\hat{R}$ is the estimated hazard ratio of the effect of HIV treatment initiation on time to death. SE is the estimated robust standard error. Weights were obtained by using KOW ($\mathcal{K}_1$): a product of two linear kernels, one for the treatment history and one for the confounder history; KOW ($\mathcal{K}_2$): a product between a linear kernel for the treatment history and a polynomial kernel of degree 2 for the confounder history; IPTCW: a logistic regression on the treatment history and the time-invariant and time-dependent confounders (using only one time lag for each of the treatment and time-dependent confounders); sIPTCW: stabilized IPTCW. * indicates statistical significance at the 0.05 level.

period, in this analysis, the binary election outcome was observed only at the end of each 5-week trajectory. In addition, all units were uncensored.

We estimated the parameters of two MSMs, the first having separate coefficients for negative advertising in each time period and the second having one coefficient for the cumulative effect of negative advertising. Each MSM was fit using weights given by each of KOW, IPTW, sIPTW, and CBPS (both full and approximate). We used the following time-dependent confounders: Democratic share of the polls, proportion of undecided voters, and campaign length. We also used the following time-invariant confounders: baseline Democratic vote share, proportion of undecided voters, status of incumbency, and election year and type of office. We obtained two sets of KOW weights by using a product of (1) two linear kernels, one for the history of negative advertising and one for the confounder history ($\mathcal{K}_1$) and (2) a linear kernel for the history of negative advertising and a polynomial kernel of degree 2 for the confounder history ($\mathcal{K}_2$). The kernels were over the complete confounder history up to time t , $\bar{X}_t$, and two time-lags of treatment history, A_{t-1} , A_{t-2} . We scaled the covariates and tuned the kernels' hyperparameters and the penalization parameter by using Gaussian processes marginal likelihood. We obtained the final set of KOW weights by solving equation (12). We compared the results obtained by KOW with those from IPTW, sIPTW, CBPS-full, and CBPS-approx. To obtain the sets of IPTW, sIPTW, and CBPS weights, we used logistic models conditioned on the confounder history and two time-lags from the treatment history. To compute the numerator of sIPTW weights, we used a logistic regression conditioned only on two time-lags from the treatment history. We used Gurobi and its R interface to solve equation (16) and obtain the KOW weights, the MatLab package GPML to perform the marginal likelihood estimation of hyperparameters, the R package R.matlab to call MatLab from within R, the R command glm to fit the treatment models for IPTW and sIPTW, the R package CBMSM for CBPS, the R command lm to fit the outcome model, and the R package sandwich to estimate robust standard errors. The computational time to obtain a solution was equal to 12.6 s for KOW, while it was equal to 104 s for CBPS-full and 3.8 s for CBPS-approx.

Table 2: Impact of negative advertising on election outcomes

$\hat{\beta}$	KOW		Logistic		CBPS	
SE	$\mathcal{K}_1$	$\mathcal{K}_2$	IPTW	sIPTW	Full	Approx
Intercept	54.54* (2.15)	53.84* (2.38)	53.05* (2.88)	47.46* (2.98)	51.25* (2.70)	52.17* (2.39)
Negative ₁	2.43 (1.86)	3.27 (1.86)	4.41 (2.56)	7.62* (3.26)	5.95* (2.49)	4.81* (2.22)
Negative ₂	3.73 (2.18)	3.24 (2.22)	5.51* (2.38)	3.17 (3.19)	3.55 (2.73)	2.65 (2.42)
Negative ₃	-2.51 (2.34)	-2.39 (2.45)	-4.37 (2.54)	-8.32* (3.84)	-6.50* (3.20)	-6.31* (3.24)
Negative ₄	-7.16* (2.57)	-7.22* (2.75)	-8.77* (1.54)	2.34 (3.11)	-3.55 (3.71)	-3.12 (3.59)
Negative ₅	-2.75* (1.42)	-1.79 (1.59)	-3.19 (2.19)	-3.62 (2.59)	-1.92 (1.62)	-1.96 (1.54)
Intercept	51.40* (2.45)	50.56* (2.63)	58.29* (4.29)	42.63* (4.15)	49.38* (2.68)	50.28* (2.49)
Cumulative	-0.59 (0.58)	-0.37 (0.64)	-0.93 (1.57)	1.91 (1.15)	-0.28 (0.65)	-0.41 (0.77)

Note: $\hat{\beta}$ is the estimated effect of negative advertising. SE is the estimated robust standard error. Weights were obtained by using, KOW ($\mathcal{K}_1$): a product of two linear kernels, one for the history of negative advertising and one for the confounder history; KOW ($\mathcal{K}_2$): a product between a linear kernel for the history of negative advertising and a polynomial kernel of degree 2 for the confounder history; IPTW: a logistic model conditioned on the confounder history and two time-lags from the treatment history; sIPTW: stabilized IPTW; CBPS-full: CBPS with full covariance matrix; CBPS-approx: CBPS with low-rank approximation.

* indicates statistical significance at the 0.05 level.

Table 2 summarizes the results of our analysis, reporting robust standard errors [29]. The first six rows of Table 2 show the effect of the time-specific negative advertising. The last two rows present the effect of the cumulative effect of negative advertising. KOW ($\mathcal{K}_1$ and $\mathcal{K}_2$) and IPTW showed similar effects, with increased precision when using KOW except for time 4, in which both methods showed a significant negative effect but with greater precision when using IPTW. sIPTW, CBPS-full, and CBPS-approx showed a significant negative effect at time 3 with similar precision. No significant results were obtained when considering the cumulative effect of negative advertising. All except sIPTW showed a negative cumulative effect. KOW ($\mathcal{K}_1$) was the most precise. We conclude that the impact of negative advertising in the majority of the time periods and its cumulative effect on election outcomes are not statistically significant.

8 Conclusion

In this article, we presented KOW, which finds weights for fitting an MSM with the aim of balancing time-dependent confounders while controlling for precision. That KOW uses mathematical optimization to directly and fully balance covariates as well as optimize precision explains the better performance of KOW over IPTW, sIPTW, and CBPS observed in our simulation study. In addition, as shown in Sections 3.2, 5, and 6, the proposed methodology only needs to minimize a number of discrepancies that grow linearly in the number of time periods, mitigates the possible misspecification of the treatment assignment model, allows balancing non-additive covariate relationships, and can be extended to control for informative censoring, which is a common feature of longitudinal studies.

Alternative formulations of our imbalance-precision optimization problem, equation (10), may be investigated. For example, additional linear constraints can be added to the optimization problem, as shown in the empirical application of Section 7.1, and different penalties can be considered to control for extreme weights. For instance, in equation (10), at the cost of no longer being able to use convex-quadratic optimization, one may directly penalize the covariance matrix of the WLS estimator rather than use a convex-quadratic surrogate as we do.

One may also change the nature of precision control. Here, we suggested penalization in an attempt to target total error. Alternatively, similar to ref. [49], we may reformulate equation (10) as a constrained optimization problem where the precision of the resulting estimator is constrained by an upper bound ξ , thus seeking to minimize imbalances subject to having a bounded precision. In our convex formulation, the two are equivalent by Lagrangian duality in that for every precision penalization λ there is an equivalent precision bound ξ . However, it may make specifying the parameters easier depending on the application as it may be easier for a practitioner to conceive of a desirable bound on precision. There may also be other ways to choose the penalization parameter. Here we suggested using maximum marginal likelihood but cross validation based on predicting outcomes and their partial means may also be possible. In this article, we based the penalty on extreme weights on the L_2 norm of (additive) deviations from uniformity. While we found this to be one reasonable choice, there are many other options, e.g., negative entropy or negative empirical likelihood.

The flexibility of our approach is that any of these changes amount to simply modifying the optimization problem that is fed to an off-the-shelf solver. Indeed, we were able to extend KOW from the standard longitudinal setting to also handle both repeated observations of outcomes and informative censoring. In addition to offering flexibility, the optimization approach we took, which directly and fully minimized our error objective phrased in terms of covariate imbalances, was able to offer improvements on the state of the art.

Acknowledgements: This article is based upon work supported by the National Science Foundation under Grants Nos. 1656996 and 1740822.

Funding information: This article is based upon work supported by the National Science Foundation under Grants Nos. 1656996 and 1740822.

Conflict of interest: Authors state no conflict of interest.

References

- [1] Robins JM. Marginal structural models versus structural nested models as tools for causal inference. In: *Statistical models in epidemiology, the environment, and clinical trials*. Heidelberg, Germany: Springer; 2000. p. 95–133.
- [2] Robins JM, Hernan MA, Brumback B. Marginal structural models and causal inference in epidemiology. *Epidemiology*. 2000;11(5):550–60.
- [3] HIV-Causal Collaboration. When to initiate combined antiretroviral therapy to reduce mortality and AIDS-defining illness in HIV-infected persons in developed countries: an observational study. *Ann Inter Med*. 2011;154(8):509.
- [4] Hernán MA, Alonso A, Logan R, Grodstein F, Michels KB, Stampfer MJ, et al. Observational studies analyzed like randomized experiments: an application to postmenopausal hormone therapy and coronary heart disease. *Epidemiol (Cambridge, Mass)*. 2008;19(6):766.
- [5] Blackwell M. A framework for dynamic causal inference in political science. *Am J Politic Sci*. 2013;57(2):504–20.
- [6] Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika*. 1983;70(1):41–55.
- [7] Petersen ML, Porter KE, Gruber S, Wang Y, Van Der Laan MJ. Diagnosing and responding to violations in the positivity assumption. *Statist Meth Med Res*. 2012;21(1):31–54.
- [8] Kang JD, Schafer JL. Demystifying double robustness: a comparison of alternative strategies for estimating a population mean from incomplete data. *Statist Sci*. 2007;523–39.
- [9] Robins JM, Rotnitzky A, Zhao LP. Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *J Am Statist Assoc*. 1995;90(429):106–21.
- [10] Scharfstein DO, Rotnitzky A, Robins JM. Adjusting for nonignorable drop-out using semiparametric nonresponse models. *J Am Statist Assoc*. 1999;94(448):1096–120.
- [11] Lefebvre G, Delaney JA, Platt RW. Impact of mis-specification of the treatment model on estimates from a marginal structural model. *Statistics Med*. 2008;27(18):3629–42.
- [12] Cole SR, Hernán MA. Constructing inverse probability weights for marginal structural models. *Am J Epidemiol*. 2008;168(6):656–64.
- [13] Xiao Y, Moodie EE, Abrahamowicz M. Comparison of approaches to weight truncation for marginal structural Cox models. *Epidemiol Meth*. 2013;2(1):1–20.
- [14] Santacatterina M, García-Pareja C, Bellocco R, Sönnnerborg A, Ekström AM, Bottai M. Optimal probability weights for estimating causal effects of time-varying treatments with marginal structural Cox models. *Statist Med*. 2019;38(10):1891–902.
- [15] Imai K, Ratkovic M. Robust estimation of inverse probability weights for marginal structural models. *J Am Statist Assoc*. 2015;110(511):1013–23.
- [16] Imai K, Ratkovic M. Covariate balancing propensity score. *J R Statist Soc B (Statist Methodol)*. 2014;76(1):243–63.
- [17] Yiu S, Su L. Joint calibrated estimation of inverse probability of treatment and censoring weights for marginal structural models. *Biometrics*. 2020;1–13. doi: 10.1111/biom.13411.
- [18] Zubizarreta JR. Stable weights that balance covariates for estimation with incomplete outcome data. *J Am Statist Assoc*. 2015;110(511):910–22.
- [19] Chad H. Kernel Balancing: A flexible non-parametric weighting procedure for estimating causal effects. *Statistica Sinica*. 2020;30(3):1155–89.
- [20] Fong C, Hazlett C, Imai K. Covariate balancing propensity score for a continuous treatment: Application to the efficacy of political advertisements. *Ann Appl Statist*. 2018;12(1):156–77.
- [21] Wong RK, Chan KCG. Kernel-based covariate functional balancing for observational studies. *Biometrika*. 2018;105(1):199–213.
- [22] Zhao Q. Covariate balancing propensity score by tailored loss functions. *Ann Statist*. 2019;47(2):965–93.
- [23] Hirshberg DA, Wager S. Augmented minimax linear estimation. *arXiv: <http://arXiv.org/abs/arXiv:171200038>*; 2017.
- [24] Kallus N. Generalized optimal matching methods for causal inference. *J Mach Learn Res*. 2020;21(62):1–54.
- [25] Kallus N, Pennicooke B, Santacatterina M. More robust estimation of sample average treatment effects using Kernel Optimal Matching in an observational study of spine surgical interventions. *arXiv: <http://arXiv.org/abs/arXiv:181104274>*; 2018.

- [26] Imbens GW, Rubin DB. Causal inference in statistics, social, and biomedical sciences. Cambridge, England: Cambridge University Press; 2015.
- [27] Hernán MA, Robins JM. Causal inference. Boca Raton, FL: CRC; 2010.
- [28] Rubin DB. Randomization analysis of experimental data: The Fisher randomization test comment. *J Am Statist Assoc.* 1980;75(371):591–3.
- [29] Freedman DA. On the so-called “Huber sandwich estimator” and “robust standard errors”. *Am Statist.* 2006;60(4):299–302.
- [30] Hernán MA, Brumback B, Robins JM. Marginal structural models to estimate the joint causal effect of nonrandomized treatments. *J Am Statist Assoc.* 2001;96(454):440–8.
- [31] Karim ME, Petkau J, Gustafson P, Tremlett H, Group TBS. On the application of statistical learning approaches to construct inverse probability weights in marginal structural cox models: Hedging against weight-model misspecification. *Commun Statist-Simulat Comput.* 2017;46(10):7668–97.
- [32] Gruber S, Logan RW, Jarrín I, Monge S, Hernán MA. Ensemble learning of inverse probability weights for marginal structural modeling in large observational datasets. *Statist Med.* 2015;34(1):106–17.
- [33] Karim ME, Platt RW. Estimating inverse probability weights using super learner when weight-model specification is unknown in a marginal structural Cox model context. *Statist Med.* 2017;36(13):2032–47.
- [34] Schölkopf B, Smola AJ. Learning with kernels: support vector machines, regularization, optimization, and beyond. Cambridge, United States: MIT Press; 2002.
- [35] Berlinet A, Thomas-Agnan C. Reproducing kernel Hilbert spaces in probability and statistics. New York, United States: Springer Science & Business Media; 2011.
- [36] Kallus N. Optimal a priori balance in the design of controlled experiments. *J R Statist Soc B (Statist Methodol).* 2018;80(1):85–112.
- [37] Rasmussen CE, Williams CK. Gaussian processes for machine learning. vol. 1. Cambridge: MIT Press; 2006.
- [38] Rasmussen CE, Nickisch H. Gaussian processes for machine learning (GPML) toolbox. *J Mach Learn Res.* 2010;11(Nov):3011–5.
- [39] D’Agostino RB, Lee ML, Belanger AJ, Cupples LA, Anderson K, Kannel WB. Relation of pooled logistic regression to time dependent Cox regression analysis: the Framingham Heart Study. *Statist Med.* 1990;9(12):1501–15.
- [40] Hernán MA, Brumback BA, Robins JM. Estimating the causal effect of zidovudine on CD4 count with a marginal structural model for repeated measures. *Statist Med.* 2002;21(12):1689–709.
- [41] Hernán MÁ, Brumback B, Robins JM. Marginal structural models to estimate the causal effect of Zidovudine on the survival of HIV-Positive men. *Epidemiology.* 2000;11(5):561–70. Available from: <http://www.jstor.org/stable/3703998>.
- [42] Gurobi Optimization. Gurobi optimizer reference manual; 2018. Available from: <http://www.gurobi.com>.
- [43] Robins JM, Greenland S, Hu FC. Estimation of the causal effect of a time-varying exposure on the marginal mean of a repeated binary outcome. *J Am Statist Assoc.* 1999;94(447):687–700.
- [44] HIV-Causal Collaboration. The effect of combined antiretroviral therapy on the overall mortality of HIV-infected individuals. *AIDS (London, England).* 2010;24(1):123.
- [45] Lodi S, Costagliola D, Sabin C, delAmo J, Logan R, Abgrall S, et al. Effect of immediate initiation of antiretroviral treatment in HIV-positive individuals aged 50 years or older. *J Acquired Immune Deficiency Syndromes.* 2017;76(3):311–8.
- [46] van der Wal WM, Geskus RB. Ipw: an R package for inverse probability weighting. *J Stat Softw.* 2011;43(13):1–23.
- [47] Cameron DW, Heath-Chiozzi M, Danner S, Cohen C, Kravcik S, Maurath C, et al. Randomised placebo-controlled trial of ritonavir in advanced HIV-1 disease. *Lancet.* 1998;351(9102):543–9.
- [48] Hammer SM, Squires KE, Hughes MD, Grimes JM, Demeter LM, Currier JS, et al. A controlled trial of two nucleoside analogues plus indinavir in persons with human immunodeficiency virus infection and CD4 cell counts of 200 per cubic millimeter or less. *New England J Med.* 1997;337(11):725–33.
- [49] Santacatterina M, Bottai M. Optimal probability weights for inference with constrained precision. *J Am Statist Assoc.* 2018;113(523):983–91.

Appendix

Proof of Theorem 1. For clarity, we prove this for $T = 2$. The extension to $T > 2$ is by induction. Under consistency and Assumptions (1)–(2), we have

$$\begin{aligned}
 \mathbb{E}[W\mathbb{I}[\bar{A} = \bar{a}]Y] &= \mathbb{E}[W\mathbb{I}[A_1 = a_1]\mathbb{I}[A_2 = a_2]Y(\bar{a})] && \text{(consistency)} \\
 &= \mathbb{E}[W\mathbb{I}[A_1 = a_1]\mathbb{I}[A_2 = a_2]\mathbb{E}[Y(\bar{a})|A_1, A_2, X_1, X_2]] && \text{(iterated expectations)} \\
 &= \mathbb{E}[W\mathbb{I}[A_1 = a_1]\mathbb{I}[A_2 = a_2]\mathbb{E}[Y(\bar{a})|A_1, X_1, X_2]] && \text{(sequential ignorability)} \\
 &= \mathbb{E}[W\mathbb{I}[A_1 = a_1]\mathbb{E}[Y(\bar{a})|A_1, X_1, X_2]] + \delta_{\bar{a}}^{(2)}(W, g_{\bar{a}}^{(2)}) && \text{(definition of } \delta_{\bar{a}}^{(2)}, g_{\bar{a}}^{(2)}) \\
 &= \mathbb{E}[W\mathbb{I}[A_1 = a_1]\mathbb{E}[Y(\bar{a})|A_1, X_1]] + \delta_{\bar{a}}^{(2)}(W, g_{\bar{a}}^{(2)}) && \text{(iterated expectations)} \\
 &= \mathbb{E}[W\mathbb{I}[A_1 = a_1]\mathbb{E}[Y(\bar{a})|X_1]] + \delta_{\bar{a}}^{(2)}(W, g_{\bar{a}}^{(2)}) && \text{(sequential ignorability)} \\
 &= \mathbb{E}[\mathbb{E}[Y(\bar{a})|X_1]] + \delta_{\bar{a}}^{(1)}(W, g_{\bar{a}}^{(1)}) + \delta_{\bar{a}}^{(2)}(W, g_{\bar{a}}^{(2)}) && \text{(definition of } \delta_{\bar{a}}^{(1)}, g_{\bar{a}}^{(1)}) \\
 &= \mathbb{E}[Y(\bar{a})] + \delta_{\bar{a}}^{(1)}(W, g_{\bar{a}}^{(1)}) + \delta_{\bar{a}}^{(2)}(W, g_{\bar{a}}^{(2)}) && \text{(iterated expectations)}
 \end{aligned}$$

Proof of Theorem 2. Define $K_{ij} = \mathcal{K}_t((\bar{A}_{i,t-1}, \bar{X}_{it}), (\bar{A}_{j,t-1}, \bar{X}_{jt}))$. Then, by the representer property of the kernels and by self-duality of Hilbert spaces,

$$\begin{aligned}
 \Delta_{a_1}^{(1)}(W_{1:n})^2 &= \sup_{\|h^{(1)}\|_{(1)}^2 \leq 1} (\hat{\delta}_{a_1}^{(1)}(W_{1:n}, h^{(1)}))^2 \\
 &= \sup_{\|h^{(1)}\|_{(1)}^2 \leq 1} \left(\frac{1}{n} \sum_{i=1}^n \underbrace{(W_i\mathbb{I}[A_t = a_t] - 1)}_{z_i} h^{(1)}(X_{it}) \right)^2 \\
 &= \sup_{\|h^{(1)}\|_{(1)}^2 \leq 1} \left(\frac{1}{n} \sum_{i=1}^n z_i \langle \mathcal{K}_t((X_{i1}), \cdot), h^{(1)}(X_{i1}) \rangle \right)^2 \\
 &= \left\| \frac{1}{n} \sum_{i=1}^n z_i \mathcal{K}_t((X_{i1}), \cdot) \right\|_{(1)}^2 \\
 &= \left\langle \frac{1}{n} \sum_{i=1}^n z_i \mathcal{K}_t((X_{i1}), \cdot), \frac{1}{n} \sum_{i=1}^n z_i \mathcal{K}_t((X_{i1}), \cdot) \right\rangle \\
 &= \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n z_i z_j \mathcal{K}_t((X_{i1}), (X_{j1})) \\
 &= \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n (W_i\mathbb{I}[A_{i1} = a_1] - 1)(W_j\mathbb{I}[A_{j1} = a_1] - 1) K_{ij} \\
 &= \frac{1}{n^2} (I_{a_1}^{(1)} W_{1:n} - e)^T K_t (I_{a_1}^{(1)} W_{1:n} - e) \\
 &= \frac{1}{n^2} W_{1:n}^T I_{a_1}^{(1)} K_t I_{a_1}^{(1)} W_{1:n} - 2e^T K_t I_{a_1}^{(1)} W_{1:n} + e^T K_t e,
 \end{aligned}$$

$$\begin{aligned}
\Delta_{a_t}^{(t)}(W_{1:n})^2 &= \sup_{\|h^{(t)}\|_{(t)} \leq 1} (\hat{\delta}_{a_t}^{(t)}(W_{1:n}, h^{(t)}))^2 \\
&= \sup_{\|h^{(t)}\|_{(t)} \leq 1} \left(\frac{1}{n} \sum_{i=1}^n \underbrace{(\mathbb{I}[A_t = a_t] - 1) W_i}_{z_i} h^{(t)}(\bar{A}_{i,t-1}, \bar{X}_{it}) \right)^2 \\
&= \sup_{\|h^{(t)}\|_{(t)} \leq 1} \left(\frac{1}{n} \sum_{i=1}^n z_i \langle \mathcal{K}_t((\bar{A}_{i,t-1}, \bar{X}_{it}), \cdot), h^{(t)}(\bar{A}_{i,t-1}, \bar{X}_{it}) \rangle \right)^2 \\
&= \left\| \frac{1}{n} \sum_{i=1}^n z_i \mathcal{K}_t((\bar{A}_{i,t-1}, \bar{X}_{it}), \cdot) \right\|_{(t)}^2 \\
&= \left\langle \frac{1}{n} \sum_{i=1}^n z_i \mathcal{K}_t((\bar{A}_{i,t-1}, \bar{X}_{it}), \cdot), \frac{1}{n} \sum_{i=1}^n z_i \mathcal{K}_t((\bar{A}_{i,t-1}, \bar{X}_{it}), \cdot) \right\rangle \\
&= \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n z_i z_j \mathcal{K}_t((\bar{A}_{i,t-1}, \bar{X}_{it}), (\bar{A}_{j,t-1}, \bar{X}_{jt})) \\
&= \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n (W_i \mathbb{I}[A_{it} = a_t] - W_i)(W_j \mathbb{I}[A_{jt} = a_t] - W_j) K_{tij} \\
&= \frac{1}{n^2} (I_{a_t}^{(t)} W_{1:n} - W_{1:n})^T K_t (I_{a_t}^{(t)} W_{1:n} - W_{1:n}) \\
&= \frac{1}{n^2} W_{1:n}^T (I - I_{a_t}^{(t)}) K_t (I - I_{a_t}^{(t)}) W_{1:n}.
\end{aligned}$$

□

Proof of Theorem 3. For clarity, we prove this for $T = 2$. The extension to $T > 2$ is by induction. Under consistency, Assumptions (1)–(2) and (13),

$$\begin{aligned}
&\mathbb{E}[W \mathbb{I}[\bar{A} = \bar{a}] \mathbb{I}[C_2 = 0] Y] \\
&= \mathbb{E}[W \mathbb{I}[A_1 = a_1] \mathbb{I}[A_2 = a_2] \mathbb{I}[C_2 = 0] Y(\bar{a})] && \text{(consistency)} \\
&= \mathbb{E}[W \mathbb{I}[A_1 = a_1] \mathbb{I}[A_2 = a_2] \mathbb{I}[C_2 = 0] \mathbb{E}[Y(\bar{a}) | A_1, A_2, X_1, X_2, C_2, C_1]] && \text{(iterated expectations)} \\
&= \mathbb{E}[W \mathbb{I}[A_1 = a_1] \mathbb{I}[A_2 = a_2] \mathbb{I}[C_2 = 0] \mathbb{E}[Y(\bar{a}) | A_1, A_2, X_1, X_2]] && \text{(equation (13))} \\
&= \mathbb{E}[W \mathbb{I}[A_1 = a_1] \mathbb{I}[A_2 = a_2] \mathbb{I}[C_2 = 0] \mathbb{E}[Y(\bar{a}) | A_1, X_1, X_2]] && \text{(equation (2))} \\
&= \mathbb{E}[W \mathbb{I}[A_1 = a_1] \mathbb{I}[C_1 = 0] \mathbb{E}[Y(\bar{a}) | A_1, X_1, X_2]] + \delta_{\bar{a}}^{(2)}(W, g_{\bar{a}}^{(2)}) && \text{(definition of } \delta_{\bar{a}}^{(2)}, g_{\bar{a}}^{(2)}) \\
&= \mathbb{E}[W \mathbb{I}[A_1 = a_1] \mathbb{I}[C_1 = 0] \mathbb{E}[Y(\bar{a}) | A_1, A_2, X_1, X_2]] + \delta_{\bar{a}}^{(2)}(W, g_{\bar{a}}^{(2)}) && \text{(equation (2))} \\
&= \mathbb{E}[W \mathbb{I}[A_1 = a_1] \mathbb{I}[C_1 = 0] \mathbb{E}[Y(\bar{a}) | A_1, A_2, X_1, X_2, C_1, C_2]] + \delta_{\bar{a}}^{(2)}(W, g_{\bar{a}}^{(2)}) && \text{(equation (13))} \\
&= \mathbb{E}[W \mathbb{I}[A_1 = a_1] \mathbb{I}[C_1 = 0] \mathbb{E}[Y(\bar{a}) | A_1, X_1, C_1]] + \delta_{\bar{a}}^{(2)}(W, g_{\bar{a}}^{(2)}) && \text{(iterated expectations)} \\
&= \mathbb{E}[W \mathbb{I}[A_1 = a_1] \mathbb{I}[C_1 = 0] \mathbb{E}[Y(\bar{a}) | A_1, X_1]] + \delta_{\bar{a}}^{(2)}(W, g_{\bar{a}}^{(2)}) && \text{(equation (13))} \\
&= \mathbb{E}[W \mathbb{I}[A_1 = a_1] \mathbb{I}[C_1 = 0] \mathbb{E}[Y(\bar{a}) | X_1]] + \delta_{\bar{a}}^{(2)}(W, g_{\bar{a}}^{(2)}) && \text{(equation (2))} \\
&= \mathbb{E}[\mathbb{I}[C_0 = 0] \mathbb{E}[Y(\bar{a}) | X_1]] + \delta_{\bar{a}}^{(1)}(W, g_{\bar{a}}^{(1)}) + \delta_{\bar{a}}^{(2)}(W, g_{\bar{a}}^{(2)}) && \text{(definition of } \delta_{\bar{a}}^{(1)}, g_{\bar{a}}^{(1)}) \\
&= \mathbb{E}[Y(\bar{a})] + \delta_{\bar{a}}^{(1)}(W, g_{\bar{a}}^{(1)}) + \delta_{\bar{a}}^{(2)}(W, g_{\bar{a}}^{(2)}) && \text{(iterated expectations)}
\end{aligned}$$