Title
Robust Estimation Methods and Causal Inference for Time-Series

Author
MALENICA, IVANA

Publication Date
2022

Peer reviewed|Thesis/dissertation

Robust Estimation Methods and Causal Inference for Time-Series

by

Ivana Malenica

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Biostatistics

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Mark van der Laan, Chair
Professor Alan Hubbard
Professor Maya Petersen

Summer 2022

Robust Estimation Methods and Causal Inference for Time-Series

Abstract

Robust Estimation Methods and Causal Inference for Time-Series

by

Ivana Malenica

Doctor of Philosophy in Biostatistics

University of California, Berkeley

Professor Mark van der Laan, Chair

Intensive longitudinal data, defined as time-varying data collected frequently over time, holds immense promise to advance many healthcare and public health concerns. High quality use of time-series primarily depends on the efficient use of *causal inference methodology*, *online machine learning* and *sequential decision-making*. Today, causal inference is central to the study of the most impactful scientific questions. For long data streams and elaborate dependence, online estimation has become a paramount technique for learning in real-time, enabling estimation despite high computational cost. Finally, sequential decision-making is essential for practitioners and policy makers to learn *when*, in *what context*, and *what* exposures to assign to each person with the objective of optimizing desired outcome. For example, one might need to decide, with some confidence, *when to stop* administrating treatment if sufficient benefit is not observed, and what is the *best* alternative based on patient's current characteristics and adherence. Investing in the development of methodological approaches for intensive longitudinal data is paramount for advancement of fields such as precision health, where we use data to learn which components of successful strategies are essential to their success, and how best to tailor personalized exposures to meet the specific needs and contexts of individuals, clinics, and communities. Careful considerations and new, modern statistical methods are necessary in order to establish causality, deal with dependence (across time and samples), and estimate relevant parts of the process without imposing unnecessary assumptions.

This dissertation focuses on development of robust non/semi-parametric methods for complex parameters involving time-series data. Divided into five chapters, content discussed entails new methodological approaches for (1) online ensemble machine learning and (2) (causal) sequential decision-making in time-dependent settings. Common themes throughout the chapters deal with (1) different dependence structures (time and/or network) in realistic statistical models and (2) leveraging fully personalized target parameters (single time-series, "N-of-1" approaches) vs. relying on multiple samples. Ideas of studying asymp-

totics in time, samples or both are also explored.

We start with the idea of developing a "N-of-1" online ensemble machine learning algorithm in Chapter 1, denoted Personalized Online Super Learner. In particular, Chapter 1 studies an Online Super Learner which learns relevant parts of the likelihood while taking into account the amount of data collected (including dynamic enrollment), stationarity of the time-series, and the mutual characteristics/network of a group of trajectories. Further exploring the "N-of-1" paradigm, we propose a causal approach which assigns treatment conditional on the current context of the patient in Chapters 2 and 3, defined as the conditional (or context-specific) causal effects. Let $Y(t)$ denote the outcome, and $C_o(t)$ a fixed dimensional context at time t. A "N-of-1" statistical approach answers the following question: *"Averaged over times t, given $C_o(t)$, what is the distribution of $Y(t+s)$ had we intervened on treatment nodes between t and $t+s$, $s>0$, on sample i?"*. In Chapter 2, we propose a time-varying effect of interventions on multiple repeated nodes via context-specific average treatment effect in observation settings, and study the theoretical properties of the proposed estimator. In Chapter 3, we propose a method that learns an optimal treatment allocation for a single individual, adapting the randomization mechanism for future time-point experiments. We demonstrate that one can learn the optimal context defined rule based on a single sample, and thereby adjust the design at any point t with valid inference for the mean target parameter.

The high intensity exposure adaptation available in time-series data proves tremendous potential for infectious disease surveillance and control. For instance, due to the highly dynamic nature of most epidemic diseases, surveillance methods must adapt quickly in order to target individuals at the highest risk of infection. Instead of only considering dependence through time, sequential decision-making for infectious disease must account for the overall status of the epidemic in the population, including multiple trajectories with possible network dependence. In Chapter 4, we describe an adaptive surveillance design which optimizes testing allocation among a class of testing schemes based on the current status of the epidemic. While Chapter 4 focuses on adaptive monitoring for a closed community, the statistical problem is addressed within a model for the data-generating distribution that is completely nonparametric. As such, it represents a first step towards an important goal of developing adaptive sequential design for infectious disease surveillance in the general population under no assumptions on the dependence structure.

Finally, in order to develop data-driven, effective sequential interventions, it is crucial to learn new policies (series of treatment decisions) using existing data, and understand their long-term efficacy. In Chapter 5, we propose and analyze a novel double robust estimator for the "off-policy" evaluation problem in reinforcement learning. We show empirically that our estimator uniformly wins over existing off-policy evaluation methods, and characterize the asymptotic distribution and rate of convergence for the proposed estimator.

Ova disertacija se fokusira na razvoj robusnih neparametarskih i poluparametarskih metoda za kompleksne statističke parametre vremenskih serija. Podeljen u pet poglavlja, sadržaj disertacije čine novi metodološki pristupi za (1) mašinsko učenje onlajn ansambla i (2) (uzročno) sekvencijalno donošenje odluka u vremenski zavisnim okruženjima. Teme u poglavljima se bave (1) različitim strukturama zavisnosti (vreme i/ili mreža) u realističnim statističkim modelima i (2) korišcenjem potpuno personalizovanih parametara (jedna vremenska serija, N-od-1 pristup) naspram oslanjanja na više uzoraka. Ova disertacija takodje prezentuje teoretsku analizu statistickih parametra gde je asimptotika u vremenu, broju uzoraka ili oboje istovremeno.

Disertacija počinje sa idejom o N-od-1 onlajn algoritmu za mašinsko učenje ansambla. Konkretno, Poglavlje 1 proučava onlajn Super Learner (ansambl mašinskog učenja) koji uči relevantne delove verovatnoće uzimajuci u obzir količinu prikupljenih podataka — uključujuci dinamički upis, stacionarnost vremenske serije i medjusobne karakteristike/mrežu grupe vremenskih serija. Poglavlje 2 i 3 prezentuju kauzalni pristup koji dodeljuje intervenciju uslovljenu trenutnim kontekstom, definisanim kao uslovni (ili kontekstualno specifični) uzročni efekti. Neka $Y(t)$ bude ishod, a $C_o(t)$ fiksno-dimenzionalni kontekst u trenutku t. Statistički pristup N-od-1 odgovara na sledeće pitanje: *tokom vremena t, uzimajuci u obzir prošlost $C_o(t)$, sta je ishod $Y(t+s)$ ako utičemo na tretman uzorka i izmedju t i $t+s$, $s>0$?* U Poglavlju 2 analiziramo kontekstno-specifični prosečni efekt tretmana uzimajući u obzir prošlost, i proučavamo teoriju i asumptotiku predložene metode. U Poglavlju 3 analiziramo algoritam koja uči optimalnu alokaciju tretmana za jednu osobu, prilagodjavajući mehanizam randomizacije za buduće eksperimente.

Adaptacija vremenskih serija ima ogroman potencijal za nadzor i kontrolu zaraznih bolesti. Na primer, zbog veoma dinamične prirode većine epidemija, statističke metode se moraju brzo prilagoditi kako bi ciljale pojedince sa najvecim rizikom od infekcije. Sekvencijalno donošenje odluka mora da uzme u obzir sveukupni status epidemije u populaciji, uključujuci moguću mrežnu zavisnost vremenskih serija. U četvrtom poglavlju izučavamo adaptivni dizajn koji optimizuje raspodelu testiranja na osnovu trenutnog statusa epidemije. Teoretska analiza predloženog metoda uzima u obzir vremensku i mrežnu zavisnost u neparametarskom modelu procesa.

Da bi se razvile efikasne sekvencijalne intervencije, ključno je koristiti postojeće podatke i razumeti njihovu dugoročnu efikasnost. U 5. poglavlju predlažemo i analiziramo novi robusni metod za problem evaluacije znan kao "off-line" u reinforcement learning (pojačano učenje). Kao deo analize, mi teoretski (i u praksi) pokazujemo da predložen metod pobedjuje sve već predložene metode, i karakterišemo asimptotičku distribuciju i stopu konvergencije algoritma.

Za baba Stevanku. Uvek ces biti sa mnom, gde god da si.

Contents

List of Figures

List of Tables

Acknowledgments

This thesis is indebted to my advisor Mark van der Laan. First and foremost, I would like to thank Mark for the honor of learning from him. His clarity of thought, sharp insights, structured approach of thinking and solving problems is truly unmatched. I found myself in awe many times over the years by his ability to promptly distill the essence of a problem, dive into a new topic, or resume discussions despite having numerous students and collaborators. His dedication to sharing his insights, ideas, and time with others is truly inspirational, and continually restore my faith in good science. I have learned a lot from his statistical rigor, pushing me always to go back to the basics (define the problem from scratch, follow the Roadmap), see the big picture and learn to recognize/do good work. I will forever fondly recall our discussions on the balcony of BWW. In addition to fostering my intellectual growth, his immaculate work ethics, care for his students and passion for science have been an example and an inspiration. Lastly, I am extremely grateful to Mark for allowing me to have complete intellectual freedom over the years — letting me find, and pursue, various projects and directions in my own time. Thank you for the trust and encouragement.

I am equally grateful to Alan Hubbard, to whom I owe my first experiences with Targeted Learning and applied biostatistics work. Alan's perceptive insights, tireless support, guidance, patience, as well as endless optimism and humor have been an absolute pleasure to work with. My deepest gratitude also goes to Maya Petersen, for her encouragement over the years, and a push to pursue an academic path. I was lucky enough to get a chance to teach Causal II with Maya during my graduate studies, and share the love for causal inference. Maya's work ethic is truly something to inspire to, and her enthusiasm about her work and teaching is absolutely contagious. I am also thankful for many exciting and inspiring collaborators and mentors I have had the honor to work with. Romain Pirracchio, who has served as a dedicated and patient mentor to whom I owe much of the knowledge on how to work at the intersection of statistics and medicine. Antoine Chambaz, for his statistical rigor, enthusiasm, and clarity of writing. Michael Jordan, Sam Pimentel, Sergey Levine and John Colford for providing guidance and quality insight during my graduate experience and qualifying exam. I would also like to thank Kendall van Keuren Jensen and David Craig for their support and encouragement to pursue a graduate degree in biostatistics; finally Ilona Jeličić and Valerija Šetalo for support to study in America.

I thank my incredible fellow students and friends that surrounded me throughout my time at U.C. Berkeley. Suzanne, for being the only one to stay with me in STAT 201A during our first semester, and all the subsequent adventures that followed. Aurelien, for pushing me to implement and write a paper in a month during winter break. Rachael, Nima and Jeremy, for all the wonderful time we spent in and out of conferences and teaching workshops. I have truly enjoyed all the debates and hard work done together, and learned a lot about developing open source software from you. I was lucky to be surrounded by brilliant students in the biostatistics and statistics department over the years, and I will always fondly remember the time spent with Ale Benitez, Joe Borja, Wilson Cai, Mary Combs, Shalika Gupta, Steve Howard, Partow Imani, Chris Kennedy, Soren Kunzel, Andre

Kurepa-Waschka, Jonathan Levy, Caleb Miles, Henry Pinkard, Lucia Petito, George Shan, Simon Walter, Yuting Ye, Yue You and Chi Zhang, to name a few.

Lastly, I would like to thank my parents, sister, family and partner for always believing in me — even while 10,000 km away, and over rough and calm waters.

Chapter 1

Personalized Online Super Learner

In this chapter, we introduce the Personalized Online Super Learner (POSL): an online personalizable ensemble machine learning algorithm for streaming data and time-series. POSL optimizes predictions with respect to baseline covariates, so personalization can vary from completely individualized (i.e., optimization with respect to subject ID) to many individuals (i.e., optimization with respect to common baseline covariates). As an online algorithm, POSL learns in real-time. As a Super Learner, POSL can leverage a diversity of candidate algorithms: including online algorithms with different training and update times, fixed/offline algorithms that are never updated during the procedure, pooled algorithms that learn from many individuals' time series, and individualized algorithms that learn from within a single time series. POSL's ensembling of this hybrid of candidates can depend on the amount of data collected, the stationarity of the time series, and the mutual characteristics of a group of time series. In essence, POSL is able to adapt to learning across samples or through time, or both, depending on the underlying data-generating process and the information available in the data. For a range of simulations that reflect realistic forecasting scenarios and in a medical data application, we examine the performance of POSL relative to other current ensembling and online learning methods. We show that POSL is able to provide reliable predictions for both short and long time series, and it's able to adjust to changing data-generating environments. We further cultivate POSL's practicality by extending it to settings where time series enter and exit dynamically over time.

1.1 Introduction

Predictive analytics with large data streams is a common task across many fields, including physics, medicine, engineering and finance. The insights drawn from these data typically come in the form of forecasts (predictions about the future), and inform subsequent action by the machine or user. The predictions' usefulness often depends on their timeliness, accuracy, and uncertainty. For example, in a hospital's intensive care unit (ICU), it is imperative that any predictions derived from patient data streams are generated quickly enough for the

clinician to respond to them appropriately [20]. Drawing from the COVID-19 pandemic, obtaining accurate forecasts in a timely manner is crucial for making evolving policy decisions [2]. In these examples, and for real-world data streams in general, the observations are derived from dynamic environments, where time series are ever-growing and evolve in possibly unforeseeable ways.

In order for a machine to quickly adapt with the dynamic patterns in data streams, algorithmic strategies that regularly reassess the information learned from incoming data relative to historical data are essential. The traditional machine learning paradigm has been in the form of offline estimation, where a prediction algorithm is updated with new batches of data by first adding the new data to all, or part, of the existing data and then retraining the learner on the new training dataset. Because an offline algorithm's training dataset grows with each update, these strategies are generally not scalable when updates are frequent. Tools that assess the reliability of an algorithm, such as calibration diagnostics, can inform reactive updates for an offline algorithm. However, when new patterns are expected to emerge quickly and often in the time series, real-time (as opposed to reactionary) learning is important to maintain the reliability of the system. Online estimation has become a promising technique for learning from data streams in real-time, since it involves update procedures that do not require the revisiting of past training data. There is a growing body of literature on online algorithms and software, including online implementations of canonical time series algorithms [3, 58]. Some online implementations are ensemble-based, combining forecasts from multiple algorithms as part of their procedure, and such strategies have been shown to increase forecast accuracy. For instance, the most successful entries in the 2018 M4 Forecasting Competition were ensembling methods[115, 46, 114, 92]. Also, Hibon and Evgeniou [56] showed empirically that the best combination of forecasts performed as well as the best individual forecast.

Ensembling methods can help mitigate several longstanding challenges in applying online learning strategies. These strategies, just like all other types of machine learning, require data in order to perform well but in some settings data accumulation is a luxury that is not guaranteed. For instance, in order to forecast a hospital patient's trajectory, a purely online learning strategy would require following the patient for a long period of time before making predictions, which is not practical for in-hospital forecasting applications. Another limitation of online learning is when new information interferes with what the has already been learned: a phenomenon known as catastrophic forgetting/interference. This can result in sudden drops in performance and in overwriting prior knowledge that could be informative again in the future [72]. Constrained online learning/ensembling strategies offer the potential to reduce catastrophic forgetting events by restricting the degree in which an algorithm is allowed to adjust its parameters at each update. Online ensembling of a hybrid of offline and online algorithms also provides a means to address these limitations.

A principled methodology for algorithm selection and ensembling is warranted [114, 56]. However, despite the emerging popularity of online learning implementations and ensembling algorithms, literature at the intersection of these two fields is relatively scarce. Furthermore, a *personalized* online ensembling paradigm, to fit and evaluate the performance

of algorithms under an individualized optimization strategy, has only been described in the commercial/proprietary realm and has not yet been formally defined in the literature to the best of our knowledge. This chapter proposes such a paradigm and grounds it in statistical optimality theory. In this work, we introduce a novel online ensembling algorithm — Personalized Online Super Learner (POSL) — that utilizes a diversity of time series and ensembling methods, with the goal of optimizing baseline covariate-level (including individual-level) predictions. POSL leverages multiple candidate algorithms, including pooled (population-based), individualized, online and offline learners, and allows for the ensembling to depend on the amount of data collected, status of stationarity, and residual noise. As such, POSL is not hindered by the limitations of purely online or offline learning.

The data structure we consider for POSL consists of observing n units/subjects, which are possibly drawn from different data-generating distributions, for a finite amount of time. Each observation is comprised of baseline covariates, time-varying covariates, and a response. We consider this setup in fixed time series settings (i.e., all n units' time series enter and exit at the same time) and in dynamic settings (i.e., the n time series enter/exit at different chronological times and are observed for different lengths of time). We introduce new formulations for online cross-validation (CV) for single time series ($n = 1$) and multiple time series ($n > 1$) with varying dependence, and all of them can be used by POSL. Building on theoretical foundations proposed by Benkeser et al.[7], we apply the online oracle inequalities to multiple time series to show that the POSL candidate algorithm with the best CV performance is asymptotically equivalent with the performance of the oracle benchmark selector. As opposed to the original work by Benkeser et al.[7], we consider a different target parameter, which conditions on the shared baseline covariates. We also formulate the problem to include multiple time series with possible baseline dependence among samples, and extend the methodology to include CV schemes which handle dependence both across time and subjects. Formulating the problem in this way allows us to study its asymptotics properties across time, samples or both, depending on the dependence structure in the data. Most importantly, we extend the work by Benkeser et al.[7] to encompass dynamic enrollment, where time series start and exit at random times. This extension is particularly important in healthcare applications, where enrollment is not synchronized as in a trial. Lastly, we propose an adaptive ensembling step for POSL that allows for a continuum of personalization based on mutual characteristics of a group of time series, and aims to increase predictive power for shorter time series.

We formulate the methodology in subsection 1.2. In particular, in subsection 1.2 we specify the statistical estimation problem, which includes defining the likelihood, statistical model, statistical target parameter, and the loss-based paradigm for estimation. In subsection 1.3, we present several online CV schemes and propose extensions to the multiple time series. In subsection 1.4 we introduce the POSL, show how online CV is used to identify the best performing individual online algorithm and the best performing online ensemble of individual algorithms, and define the oracle selector for POSL. In subsection 1.5, we extend the current formulation of POSL to dynamic enrollment settings, where possibly different lengths of time series and numbers of subjects are observed at each chronological time point.

Our formulation is purposely general, in order to encompass varying CV schemes, loss functions, number of samples, and enrollment time. In section 1.4, we provide an example of one version of the POSL algorithm. In section 1.6, we conduct multiple simulation studies to compare POSL to various ensembling and online methods currently available in the literature. In section 1.7, we provide a data analysis example for blood pressure forecasting. We conclude with a short discussion in section 1.8.

1.2 Statistical Formulation of the Problem

In the following, we formalize the prediction task as an estimation problem, identifying the statistical target parameter as the minimizer of the risk induced by a valid, problem-specific loss function.

Data, Likelihood and the Statistical Model

We model a data structure under the shape of a random variable defined as $O_i = (O_i(t) : t = 1, \ldots, \tau)$, where $O_i(t)$ is a finite-dimensional variable for sample i that is indexed by time t. We refer to this sequence of variables across time, $O_i = (O_i(1), \ldots, O_i(\tau))$, as subject i's time series. We focus on situations where $O_i(t)$ decomposes as $O_i(t) = (Y_i(t), W_i(t))$, with $Y_i(t)$ defining a response variable for sample i occurring at time t and $W_i(t)$ defining a vector of time-varying covariates for sample i occurring after time t. We denote X_i as a vector of baseline covariates which, by definition, are initiated at $t = 0$ and not dependent on t. We view each X_i and $O_i = (O_i(t) : t = 1, \ldots, \tau)$ as the sample i-specific baseline covariates and time series, respectively. We observe n independent realizations of random variables denoted as $(X_1, O_1), \ldots, (X_n, O_n)$. For convenience, we also introduce $X^n = (X_i : i = 1, \ldots, n)$ and $O^n = (O_i : i = 1, \ldots, n)$ as the collection of the n subjects' baseline covariates and time series, respectively.

To motivate our statistical formulation of the problem, we consider an example from healthcare. When patients are admitted to the hospital, their medical history and all other administrative clinical data relative to their care is recorded. This electronic health record (EHR) is instantiated with baseline, time-invariant information (X_i), such as the patient's age, sex, ethnicity, insurance status, medical history, and other demographic factors. The EHR is maintained by the provider so relevant information occurring during the patient's stay, such as treatment plans, laboratory/test results, clinician notes, radiology images and vital signs, are added to the EHR over time (O_i). The EHR for patient i therefore contains their time series O_i and the baseline covariates X_i. For simplicity, consider a simplified EHR with O_i including blood pressure and heart rate (i.e., time-varying variables), and X_i including age, sex, ethnicity and race (i.e., time-invariant variables). Suppose the response variable of interest is blood pressure (Y_i), and the time-varying covariates are all other variables in the time series, which for this simple example was heart rate (W_i). Note that patient i's blood pressure at time t ($Y_i(t)$) is not influenced by their heart rate at time t

$(W_i(t))$, since $W_i(t)$ does not occur before $Y_i(t)$, but $Y_i(t)$ might be influenced by patient i's heart rate and blood pressure from previous time points.

Let $\mathcal{M}$ denote the statistical model: the set of laws from which (X^n, O^n) can be drawn. The more we know, or are willing to assume, about the experiment that produces the data, the smaller $\mathcal{M}$ will be. Let $P_0^n \in \mathcal{M}$ be the true probability distribution of (X^n, O^n). Moreover, let $P_{0,O_i|X_i}$ be the conditional distribution of O_i given X_i for each $i = 1, \ldots, n$. When conditioning on X_i, we use the short notation P_{0,X_i} instead of $P_{0,O_i|X_i}$ (not to be confused with $P_{0,X}$, the marginal distribution over the baseline covariates). We emphasize that P_{0,X_i} could be just unit i specific, as is the case when X_i is simply a function of i itself; alternatively, P_{0,X_i} could be a smooth function of X_i, allowing one to smooth across the subjects. We let p_0^n denote the density of P_0^n with respect to (w.r.t) a measure μ^n that dominates all elements of $\mathcal{M}$. The joint likelihood of (x^n, o^n) can be factorized according to the time-ordering as follows:

$$p_0^n(x^n, o^n) = \prod_{i=1}^{n} p_{0,x}(x_i) \prod_{t=1}^{\tau} p_{0,o_i(t)}(o_i(t) \mid x_i, \overline{o}_i(t-1)), \tag{1.1}$$

where $p_{0,x}$ marks the probability density for the baseline covariates, and $p_{0,o_i(t)}$ is the conditional density of $O_i(t)$ given X_i and all the observed past until time t for sample i. In particular, we define $\overline{O}_i(t-1)$ as the t-specific history of the time series for sample i, with $\overline{O}_i(t-1) = (O_i(1), \ldots, O_i(t-1))$ (note the convention $\overline{O}_i(0) = \emptyset$).

In order to allow learning from a dependent process, we must make a few assumptions on the law of the data, P_0^n, through restrictions made on the statistical model $\mathcal{M}$. In particular, we assume that each factor $p_{0,o_i(t)}(O_i(t)|X_i, \overline{O}_i(t-1))$ depends on the past through a fixed-dimensional summary measure $Z_i(t-1)$. For some applications, the fixed dimensional summary measure $Z_i(t)$ covers a limited history, such that the dependent process has a finite memory allowing us to learn through time. Similarly, we could have defined $Z_i(t)$ to be a function of finite memory of a finite number of other time series in unit i's network. Another example of summary measures is $Z_i(t) = t^{-1} \sum_{t'=1}^{t} O_i(t')$, with the means computed component-wise. The formulation of Z_i is general enough to allow for different trends (e.g., seasonality), because the definition of Z_i can involve i itself. Secondly, we define P_{0,O_i} as the *common* conditional probability distribution of $O_i(t)$ given $Z_i(t-1)$ and X_i under P^n. As we will describe later, this assumption is not crucial for the algorithm itself — if not enough time points are collected, we rely on performance based on the number of trajectories. However, if online learning is to be useful, some structure across time is necessary. We also stress that our formulation allows for P_{0,O_i} to be a function of time t, making it possible for the proposed procedure to learn how much to rely on conditional stationarity over time. Thus, in light of (1.1) and the two above mentioned assumptions, the joint likelihood of (x^n, o^n) under any element P^n of the (constrained) statistical model $\mathcal{M}$ decomposes as

$$p^n(x^n, o^n) = \prod_{i=1}^{n} p_x(x_i) \prod_{t=1}^{\tau} p_{o_i}(o_i(t) \mid x_i, z_i(t-1)), \tag{1.2}$$

where we extend the previously described notation with the substitution of P^n (p^n) for P_0^n (p_0^n). Note that P_{0,O_i} is subject specific, and we don't assume a common across i distribution; if all the time series are drawn from the same distribution, we let the algorithm learn that. In the rest of the manuscript, we will deal explicitly with P_{0,O_i} and P_{0,X_i}.

The above derivation of (1.2) hinges on independence across subjects. While we write the likelihood as a product of both the number of samples (n) and time points (t), we emphasize that for deriving our main results, dependent on the asymptotics in time, we do not need to assume anything about dependence among subjects. Our statistical model $\mathcal{M}$ is, in essence, a model for a single time series. Independence across subjects however, allows us to have asymptotics in the total number of time points observed across the n subjects (effectively having $n \times \tau$ samples). Network dependence could be allowed simply by letting each $Z_i(t-1)$ to summarize the whole past $\overline{O}^n(t-1) = \{\overline{O}_1(t-1), \ldots, \overline{O}_n(t-1)\}$ of O^n at time $(t-1)$, or a subgroup specific past of its network, as opposed to the i-specific past $\overline{O}_i(t-1)$.

Statistical Target Parameter

Most prediction-based literature focuses on parameters of the population distribution P_0^n or, as is the case for the time series literature, on unit-specific forecasts. Our goal is not to understand the population distribution P_0^n. Instead, we focus on the parameters of the unit-specific conditional distribution P_{0,X_i}. We define the relevant feature of the true data distribution we are interested in as the *statistical target parameter*. As in above, we assume that P_0^n belongs to a statistical model $\mathcal{M}$, defined as a collection of possible common conditional distributions P_{0,O_i} and marginal $P_{0,X}$ that could have given rise to the observed data. We define a parameter mapping, $\Psi : \mathcal{M} \to \mathcal{D}$, from the model $\mathcal{M}$ into a space $\mathcal{D}$; and a parameter value, $\psi := \Psi(P)$ of Ψ for a given $P \in \mathcal{M}$. The parameter space, corresponding to parameter mapping Ψ, is defined as $\boldsymbol{\Psi} := \{\Psi(P) : P \in \mathcal{M}\} \subseteq \mathcal{D}$.

In some cases, we might be interested in learning the entire conditional distribution P_{0,X_i}; however, frequently the actual goal is to learn a particular feature of the true distribution that satisfies a scientific question of interest. In particular, we are interested in forecasting – hence, we define our estimand for the i^{th} subject as:

$$\Psi(P_0^n)(X_i, Z_i(t-1), t) = E_{P_{0,X_i}}[Y_i(t)|Z_i(t-1)], \tag{1.3}$$

where the expectation on the right hand side is taken w.r.t the conditional distribution P_{0,X_i}, and $\psi_0(X_i, Z_i(t-1), t) := \Psi(P_0^n)(X_i, Z_i(t-1), t)$ is the prediction function evaluated at the truth for the i^{th} subject at time t. In particular, we want to learn $(X, Z(t-1), t) \mapsto \Psi(P_0^n)(X, Z(t-1), t)$, where $Z(t-1)$ is fixed dimensional, and thereby obtain a prediction function for each unit i that predicts $Y_i(t)$ with $\Psi(P)(X_i, Z_i(t-1), t)$ for $P \in \mathcal{M}$.

Recall the EHR example from above, where for patient i, we have a simplified EHR that contains their time series O_i, with blood pressure as Y_i and heart rate as W_i, and X_i includes age, sex, ethnicity and race. Suppose we are interested in predicting this patient's expected blood pressure at time t, given the information available up until time t, including data from

other samples. This setup represents a forecasting problem; we aim to use collected history (from all patients) in order to predict sample i's future blood pressure. For each time t, patient i's response is their blood pressure $Y_i(t)$ at time t and observed history includes all of the past blood pressure and heart rate information until time t. The estimand can then be defined identically to Equation 1.3.

Loss-based Parameter Definition and Estimation

We define L as a loss function; we emphasize that the chosen loss should be picked in accordance with the target parameter. Specifically, a valid loss function for a given parameter is defined as a function whose true conditional mean is minimized by the true value of the parameter. As such, let L be a loss function adapted to the problem, i.e. a function that maps every $\Psi(P)$ to $L(\Psi(P)) : (x_i, y_i(t), z_i(t-1)) \mapsto L(\Psi(P))(x_i, y_i(t), z_i(t-1))$. With that, we define $L(\Psi(P))(X_i, Y_i(t), Z_i(t-1))$ as a time t and subject i loss for $\Psi(P)$. Note that we could equivalently define L a function that maps every ψ to $L(\psi) : (x_i, y_i(t), z_i(t-1)) \mapsto L(\psi)(x_i, y_i(t), z_i(t-1))$ since $\psi := \Psi(P)$. As our parameter of interest is a conditional mean, we could use the square error to define the loss; then we have that $L(\psi)(X_i, Y_i(t), Z_i(t-1)) = c(i,t)(Y_i(t) - \psi(X_i, Z_i(t-1), t))^2$, where $c(i,t)$ is a subject- and time-specific weight (e.g., we might down-weight losses that are further away from t and up-weight losses that are closer to time t; alternatively, we might give weight only to a specific sample). Our emphasis on appropriate loss functions strives from their multiple uses within our framework — as a theoretical criterion for comparing an estimator and the truth, as well as a way to compare multiple estimators of the target parameter.

We define the true risk as the expected value of $L(\psi)(X_i, Y_i(t), Z_i(t-1))$ w.r.t the conditional distribution $P_{0,O}$ across all individuals and times:

$$\begin{aligned} R(P_0^n, \psi) &= \frac{1}{nt} \sum_{i=1}^{n} \sum_{t=1}^{\tau} E_{P_{0,O_i}}[L(\psi)(X_i, Y_i(t), Z_i(t-1))|X_i, Z_i(t-1)] \\ &= \frac{1}{nt} \sum_{i=1}^{n} \sum_{t=1}^{\tau} E_{P_{0,O_i}}[c(i,t)(Y_i(t) - \psi(X_i, Z_i(t-1), t))^2|X_i, Z_i(t-1)], \end{aligned} \tag{1.4}$$

where the second equality holds only when the loss function is valid for the target parameter; we simply illustrate what $R(P_0^n, \psi)$ would be with squared error as a loss in (1.4). The notation for true risk, $R(P_0^n, \psi)$, emphasizes that ψ is evaluated w.r.t. the true data-generating distribution. Finally, we define ψ_0 as the minimizer over the true risk of all evaluated ψ in the parameter space

$$\psi_0 = \underset{\psi \in \mathbf{\Psi}}{\operatorname{argmin}} R(P_0^n, \psi). \tag{1.5}$$

The corresponding true risk is denoted as $\theta_0 = R(P_0^n, \psi_0)$. In particular, the true risk establishes a true measure of performance for ψ, optimizing over all times. We note, however, that we could also define a true i-specific risk — where the i-specific risk would measure the performance of ψ for individual i across all time points. Note that ψ_0 implies $\psi_{0,i}$

by evaluating at X_i, as $\psi_{0,i}$ is a prediction function given X_i. The i-specific expected loss measures the performance of the prediction function Ψ for individual i across all time points, optimizing the following equation:

$$\psi_{0,i} = \underset{\psi \in \mathbf{\Psi}}{\operatorname{argmin}} \sum_{t=1}^{\tau} E_{P_{0,O_i}}[L(\psi)(X_i, Y_i(t), Z_i(t-1))|X_i, Z_i(t-1)], \tag{1.6}$$

with optimal risk, corresponding to $\psi_{0,i}$, defined as $\theta_{0,i} = R(P_0^n, \psi_{0,i})$.

The estimator mapping, $\hat{\Psi}$, is a function from the empirical distribution to the parameter space $\mathbf{\Psi}$. Let $P_{n,t}$ denote the empirical distribution of n time series collected until time t. In particular, $P_{n,t} \mapsto \hat{\Psi}(P_{n,t})$ represents a mapping from $P_{n,t}$, with n time series collected until time t, into a predictive function $\hat{\Psi}(P_{n,t})$. Further, the predictive function $\hat{\Psi}(P_{n,t})$ maps $(X_i, Z_i(t-1), t)$ into a time- and subject-specific outcome, $Y_i(t)$. We emphasize that $\hat{\Psi}(P_{n,t})$ can map any $(X_i, Z_i(s-1), s)$ into a time s prediction, even for $s > t$ under stationarity conditions; as such, we can forecast at any future time point using the collected data until time t. We can write $\psi_{n,t}(X_i, Z_i(t-1), t) \coloneqq \hat{\Psi}(P_{n,t})(X_i, Z_i(t-1), t)$ as the predicted outcome for unit i of the estimator $\hat{\Psi}(P_{n,t})$ at time t, based on $(X_i, Z_i(t-1), t)$. We define the conditional risk as the risk for $\psi_{n,t}$ with respect to the true, unknown data-generating distribution P_0^n, denoted as $\tilde{\theta}_n = R(P_0^n, \psi_{n,t})$. The naive risk is defined as $\hat{\theta}_n = R(P_{n,t}, \psi_{n,t})$. In order to obtain an unbiased estimate of the true conditional risk, we resort to appropriate CV for dependent data, as described in the next subsection.

1.3 Cross-validation for Dependent Data

Let $C(i, s, \cdot)$ denote, at minimum, the time s- and unit i-specific record $C(i, s, \cdot) = (X_i, Z_i(s-1), Y_i(s), \cdot)$. The general formulation of $C(i, s, \cdot)$ allows us to add identifying information (in addition to time and sample ID) needed to construct a valid CV scheme; for instance, for dynamic enrollment/exit dates, $C(i, s, \cdot)$ might include enrollment and exit time for a time series as well. If no additional information is included, we write $C(i, s, \cdot) = C(i, s)$. To derive a general representation for CV, we also define a time t specific split vector B_t, where t indicates the final time-point of the currently available data. Then for all $1 \leq i \leq n$, $B_t(i, \cdot) \in \{-1, 0, 1\}^t$. Let v be a particular CV fold, where v ranges from 1 to V. A realization of B_t defines a particular split of the learning set into corresponding three disjoint subsets,

$$B_t^v(i, s, \cdot) = \begin{cases} -1, & C(i, s, \cdot) \text{ not used} \\ 0, & C(i, s, \cdot) \text{ in the training set} \\ 1, & C(i, s, \cdot) \text{ in the validation set,} \end{cases}$$

where $B_t^v(i, s, \cdot)$ reflects, at minimum, unit i at time point s for fold v's split, B_t^v. Realizations of B_t therefore admit the CV folds. For each t, let $P_{n,t}^0$ denote the empirical distribution of the training set until time t. Similarly, we define $P_{n,t}^1$ as the empirical distribution of the validation set. Let $n_t^0 = \sum_{v=1}^{V} \sum_{i=1}^{n} \sum_{s=1}^{t} \mathbb{I}(B_t^v(s, i, \cdot) = 0)$ and

$n_t^1 = \sum_{v=1}^{V} \sum_{i=1}^{n} \sum_{s=1}^{t} \mathbb{I}(B_t^v(s,i,\cdot) = 1)$ denote the number of observations in the training and validation sets respectively, over all folds v until time t. For fold v admitted by realizing B_t, let $\mathcal{B}_{t,v}^0$ denote all the $(i,s,\cdot)$ indexes in the training set, and $\mathcal{B}_{t,v}^1$ all indexes in the validation set. In general, we use different time series CV schemes to evaluate how well an estimator trained on specific samples' past data is able to predict an outcome for specific samples in the future. We now give relevant CV schemes that are supported by the theoretical results for our proposed algorithm.

Rolling Origin Cross-validation

A rolling origin CV (ROCV) scheme lends itself to online CV-based ensemble learning [7]. In general, the ROCV scheme defines an initial training set and, with each iteration, the size of the training set grows by m observations until we reach time t for split B_t [103]. Whether or not the samples in the training set are also present in the validation set is optional, but classically, ROCV represents the scenario where the training and validation points are evaluated on the same time series. Regardless of which samples are included in the training and validation sets, time points included in the training set always occur before the validation set time points. Additionally, there might be a gap between the last training time and first validation times of size h. We define ROCV folds $v = 1, \ldots, V$ with B_t^v for a single unit as follows:

$$B_t^v(i,s,\cdot) = \begin{cases} -1, & C(i,s,\cdot) \text{ not used,} \\ & \quad s \in \{n_{t,v_1}^0 + m \times (v-1) + 1 : n_{t,v_1}^0 + m \times (v-1) + h\} \\ 0, & C(i,s,\cdot) \text{ in the training set,} \\ & \quad s \in \{1 : n_{t,v_1}^0 + m \times (v-1)\} \\ 1, & C(i,s,\cdot) \text{ in the validation set,} \\ & \quad s \in \ \{n_{t,v_1}^0 + m \times (v-1) + h + 1 \ : \ n_{t,v}^0 + m \times (v-1) + h + n_{t,v}^1\} \end{cases}$$

where n_{t,v_1}^0 is the size of the training set for the first fold ($v = 1$); $n_{t,v}^1$ is the size of the validation set for all folds v, $v = 1, \ldots, V$; m is the batch size, indicating the number of time points training set moves forward from one fold v to the next; h is the gap between the training and validation sets. An example of ROCV is illustrated in Figure 1.1. A variant of ROCV which accounts for sample dependence is the rolling-origin-V-fold CV (ROVFCV) scheme. In contrast to ROCV, samples in the training and validation set differ for a ROVFCV scheme, as it encompasses both V-fold CV for splitting across samples and ROCV for splitting across time (Appendix B Figure 1.9).

Rolling Window Cross-validation

Instead of adding more time points to the training set per each iteration, as in ROCV, the rolling window CV (RWCV) scheme "rolls" the training sample forward by m time units,

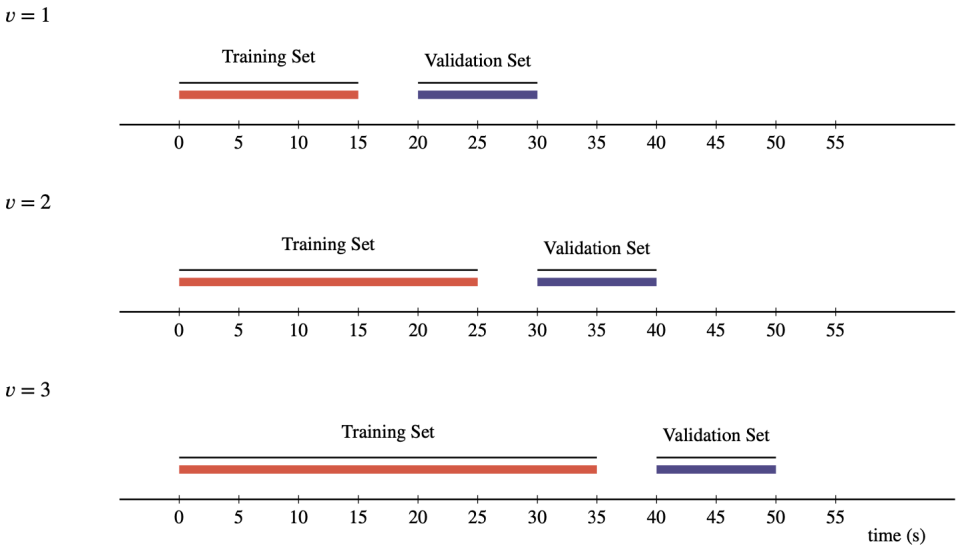


Figure 1.1: Rolling origin cross-validation (ROCV) scheme invoked for a sample i whose final time-point of their currently available data is $t = 50$, and with the following specification: first training set size $n^0_{t,v_1} = 15$, validation size $n^1_{t,v} = 10$, batch size $m = 10$, gap $h = 5$. Given the ROCV specification and the data provided, this scheme thus admits $V = 3$ ROCV folds.

such that the training set is the same size across all RWCV folds v, $v = 1, \ldots, V$ (i.e. the training sample size for each iteration of the RWCV scheme is always $n^0_{t,v}$ for all v). We define RWCV folds $v = 1, \ldots, V$ with B^v_t, as realization of B_t, and gap of size h for a single time series as follows:

$$B^v_t(i, s, \cdot) = \begin{cases} -1, & C(i, s, \cdot) \text{ not used,} \\ & \quad s \in \{n^0_{t,v} + m \times (v-1) + 1 : n^0_{t,v} + m \times (v-1) + h\} \\ 0, & C(i, s, \cdot) \text{ in the training set,} \\ & \quad s \in \{n^0_{t,v} + m \times (v-1) - n^0_{t,v} : n^0_{t,v} + m \times (v-1)\} \\ 1, & C(i, s, \cdot) \text{ in the validation set,} \\ & \quad s \in \{n^0_{t,v} + m \times (v-1) + h + 1 \ : \ n^0_{t,v} + m \times (v-1) + h + n^1_{t,v}\}. \end{cases}$$

where, for all folds v, $v = 1, \ldots, V$, $n^0_{t,v}$ is the size of the training set and $n^1_{t,v}$ is the size of the validation set; m is the batch size, indicating the number of time points training set moves forward from one fold v to the next; and h is the gap between the last training time and first validation time. We illustrate the canonical RWCV in Figure 1.2. The rolling-window-

V-fold CV (RWVFCV) scheme is a variant of RWCV which accounts for sample dependence (Appendix B Figure 1.9).

The RWCV scheme might be considered in parametric settings when one wishes to guard against moment or parameter drift that is difficult to model explicitly. It is also more efficient for computationally demanding settings (such as high-frequency streaming data), in which large amounts of training data cannot be stored. We emphasize that the RWCV could also be viewed as a subset of the ROCV, where only recent data are used for training. In fact, we could incorporate both ROCV and RWCV schemes within POSL, by considering in the library candidates that only learn from the recent past via RWCV and candidates that learn from the entire past via ROCV. In such a scenario, a ROCV scheme could be used to evaluate the final loss, but we might incorporate RWCV-based learners that train on fewer training time-points, as they can more quickly adjust to changes over time. In all the further sections and theoretical results, we consider RWCV as a subset of ROCV.

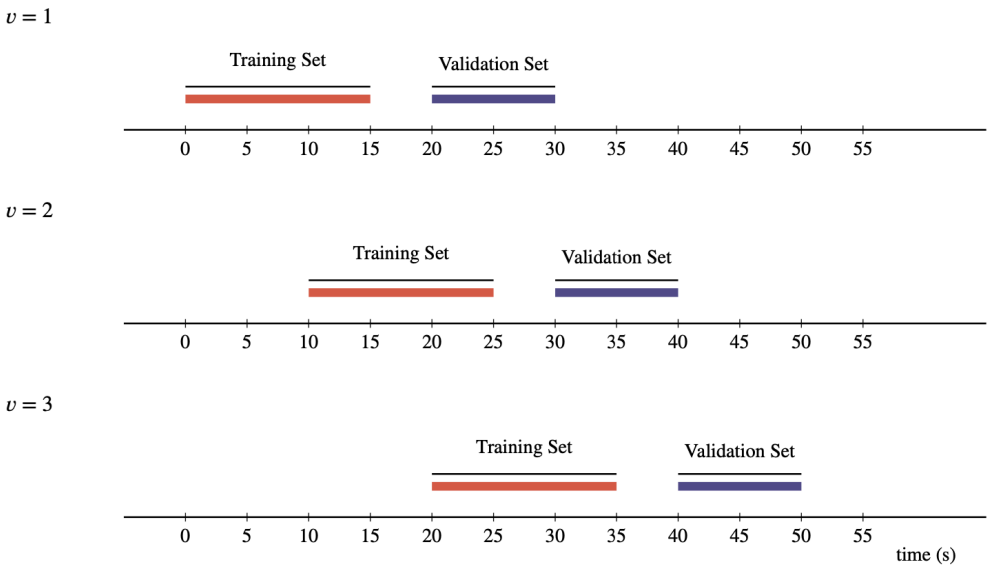


Figure 1.2: Rolling window cross-validation scheme invoked for a sample i whose final time-point of their currently available data is $t = 50$, and with the following specification: window size $n^0_{t,v_1} = 15$, validation size $n^1_{t,v} = 10$, batch size $m = 10$, gap $h = 5$. Given the RWCV specification and the data provided, this scheme thus admits $V = 3$ RWCV folds.

1.4 Personalized Online Super Learner

Online Cross-validation Selector

Suppose we have K candidate estimators, $\hat{\Psi}_k$, and recall the definition of an estimator from subsection 1.2. In order to evaluate performance of each $\hat{\Psi}_k$, we use CV for dependent data to estimate the average loss for each candidate. In particular, each $\hat{\Psi}_k$ is trained on the training set until time t, using $P^0_{n,t}$ and resulting in a predictive function $\psi^0_{n,t,k} \coloneqq \hat{\Psi}_k(P^0_{n,t})$ for $k = 1, \ldots, K$. We define the online CV risk for each candidate estimator as:

$$\begin{aligned} R_{CV}(P^1_{n,t}, \hat{\Psi}_k(\cdot)) &= \sum_{j=1}^{t} \sum_{v=1}^{V} \sum_{(i,s) \in \mathcal{B}^1_{j,v}} L(\hat{\Psi}_k(P^0_{n,j}))(X_i, Y_i(s), Z_i(s-1)) \\ &= \sum_{j=1}^{t} \sum_{v=1}^{V} \sum_{(i,s) \in \mathcal{B}^1_{j,v}} L(\psi^0_{n,j,k})(C(i,s)), \end{aligned} \tag{1.7}$$

where $R_{CV}(P^1_{n,t}, \hat{\Psi}_k(\cdot))$ is the cumulative performance of $\hat{\Psi}_k$ trained on training sets and evaluated on corresponding validation samples across all time points until time t. For instance, while $\hat{\Psi}_k(P^0_{n,t})$ is trained on the training set $P^0_{n,t}$, its performance will be over the validation set $P^1_{n,t}$. Additionally, if $\hat{\Psi}_k$ is an online estimator, then the online CV risk is also an online estimator. For the squared error loss mentioned in subsection 1.2, where $c(i,j) = 1$ (and is thus omitted), we can rewrite the above online CV risk as:

$$R_{CV}(P^1_{n,t}, \hat{\Psi}_k(\cdot)) = \sum_{j=1}^{t} \sum_{v=1}^{V} \sum_{(i,s) \in \mathcal{B}^1_{j,v}} (Y_i(s) - \hat{\Psi}_k(P^0_{n,j})(X_i, Z_i(s-1), s))^2. \tag{1.8}$$

The online CV risk estimates the following true online CV risk, denoted as $R_{CV}(P^n_0, \hat{\Psi}_k(\cdot))$ and expressed as

$$R_{CV}(P^n_0, \hat{\Psi}_k(\cdot)) = \sum_{j=1}^{t} \sum_{v=1}^{V} \sum_{(i,s) \in \mathcal{B}^1_{j,v}} E_{P_{0,O}}[L(\psi^0_{n,j,k})(C(i,s)) | X_i, Z_i(s-1)]. \tag{1.9}$$

Note that $R_{CV}(P^n_0, \hat{\Psi}_k(\cdot))$ reflects the true average loss for the candidate estimator with respect to the true conditional distribution P_{0,O_i}. As opposed to the true online CV risk, $R_{CV}(P^1_{n,t}, \hat{\Psi}_k(\cdot))$ gives an empirical measure of performance for each candidate estimator k trained on training data until time t. In light of that, we define the discrete online CV selector as the estimator that minimizes the online CV risk:

$$k_{n,t} = \arg\min_{k=1,\ldots,K} R_{CV}(P^1_{n,t}, \hat{\Psi}_k(\cdot)). \tag{1.10}$$

The discrete online Super Learner (SL) is the estimator that at each time point t uses the estimates from the discrete online CV selector — for time t, we have $\psi^0_{n,t,k_{n,t}} := \hat{\Psi}_{k_{n,t}}(P^0_{n,t})$. We emphasize that the discrete online SL can switch from one learner to another as t progresses, in response to accumulating more data and detecting changes in the time series.

Defining the Gold-Standard Oracle Selector

In order to study performance of an estimator of ψ_0, we construct loss-based dissimilarity measures. First, we define i-specific loss-based dissimilarities for the k^{th} estimator, $\hat{\Psi}_k$, trained until time t as

$$d_{0,t}(\psi_{n,t,k,i}, \psi_{0,i}) = \sum_{j=1}^{t} \sum_{v=1}^{V} \sum_{(s)\in\mathcal{B}^1_{j,v}} E_{P_{0,O_i}}\left[\Big(L(\psi_{n,j,k,i}) - L(\psi_{0,i})\Big)(C(i,s))\middle| X_i, Z_i(s-1)\right], \quad (1.11)$$

which compares performance of the CV estimator to the true parameter. Note that the training and validation sample is just sample i. We further define the measure $d_{0,t}(\psi_{n,t,k}, \psi_0) = \frac{1}{n}\sum_{i=1}^{n} d_{0,t}(\psi_{n,t,k,i}, \psi_{0,i})$ as an average of i-specific loss-based dissimilarities over all the samples until time t for the k^{th} estimator; $d_{0,t}(\psi_{n,t,k}, \psi_0)$ reflects how far $\psi_{n,t,k}$ is from ψ_0 over all available times and samples in terms of the chosen loss. We define the time t *oracle* selector as the unknown estimator that uses the candidate closest to the truth in terms of the defined dissimilarity measure:

$$\overline{k}_{n,t} = \arg\min_{k=1,\ldots,K} d_{0,t}(\psi_{n,t,k}, \psi_0). \quad (1.12)$$

Due to it being a function of the true conditional mean, the oracle selector cannot be computed in practice. However, we can utilize it as benchmark in order to describe performance of the online CV-based estimator. In Appendix Theorem 1, assuming conditional stationarity as proposed in subsection 1.2 (an assumption not necessary for the algorithm function), we extend work from Benkeser et al.[7] to multiple time series using the CV schemes described in subsection 1.3. In particular, Appendix Theorem 1 shows that the performance of the discrete POSL is asymptotically equivalent to that of the oracle selector. The result relies on the martingale finite-sample inequality by van Handel [54] to show that, as $t \to \infty$,

$$\frac{d_{0,t}(\psi_{n,t,k_{n,t}}, \psi_0)}{d_{0,t}(\psi_{n,t,\overline{k}_{n,t}}, \psi_0)} \to_p 1, \quad (1.13)$$

under conditional stationarity and additional conditions specified in the Appendix.

Ensemble of Candidate Estimators

In this section, we consider a more flexible online learner that generates a weighted combination of candidate estimators. Let $\hat{\Psi}_\alpha$ be a function of empirical distribution ($P_{n,t}$, at any

t) generating an ensemble of K estimators $(\hat{\Psi}_1, \ldots, \hat{\Psi}_K)$ indexed by a vector of coefficients α, where $\alpha = (\alpha_1, \ldots, \alpha_K)$. For example, $\hat{\Psi}_\alpha$ could represent a convex linear combination:

$$\hat{\Psi}_\alpha = \sum_{k=1}^{K} \alpha_k \hat{\Psi}_k,$$

such that $\sum_{k=1}^{K} \alpha_k = 1$ and for all α_k, $\alpha_k \geq 0$. We define conditional meta-learning by allowing the weight vector to depend on the baseline covariates X, where $\alpha(X) = \{\alpha_1(X), \ldots, \alpha_K(X)\}$ with $\sum_{k=1}^{K} \alpha_k(X) = 1$ and for all $\alpha_k(X)$, $\alpha_k(X) \geq 0$. For example, we can define $\alpha(X)$ by considering a parametric family $\mathcal{H} = \{\alpha_\beta : \beta \in \mathbb{B}\}$ where

$$\alpha_\beta(X) = \frac{\exp\left(\beta_{k,1} + \beta_{k,2} X\right)}{\sum_{k=1}^{K} \exp\left(\beta_{k,1} + \beta_{k,2} X\right)}.$$

To alleviate notation, we define α as an universal vector of coefficients (including conditional meta-learning) in further sections. Let $\hat{\Psi}_\alpha = \sum_{k=1}^{K} \alpha_k \hat{\Psi}_k$, so that the predictive function based on the training set $P_{n,t}^0$ is given by $\psi_{n,t,\alpha}^0 := \sum_{k=1}^{K} \alpha_k \hat{\Psi}_k(P_{n,t}^0)$ with $\alpha \in \mathcal{H}$. We define a $\mathcal{H}$-specific online CV selector for the ensemble as:

$$\begin{aligned} \alpha_{n,t} &= \underset{\alpha \in \mathcal{H}}{\operatorname{argmin}} R_{CV}(P_{n,t}^1, \hat{\Psi}_\alpha(\cdot)) \\ &= \underset{\alpha \in \mathcal{H}}{\operatorname{argmin}} \sum_{j=1}^{t} \sum_{v=1}^{V} \sum_{(i,s) \in \mathcal{B}_{j,v}^1} (Y_i(s) - \hat{\Psi}_\alpha(P_{n,j}^0)(X_i, Z_i(s-1), s))^2, \end{aligned} \tag{1.14}$$

where the loss is defined as the mean squared error. We can define an oracle selector for this class of estimators as the choice of weights that minimizes the true average of the loss-based dissimilarity:

$$\begin{aligned} \overline{\alpha}_{n,t} &= \underset{\alpha \in \mathcal{H}}{\operatorname{argmin}}\, d_{0,t}(\psi_{n,t,\alpha}^0, \psi_0) \\ &= \underset{\alpha \in \mathcal{H}}{\operatorname{argmin}} \sum_{j=1}^{t} \sum_{v=1}^{V} \sum_{(i,s) \in \mathcal{B}_{j,v}^1} E_{P_{0,O_i}} \left[\Big(L(\psi_{n,j,\alpha}^0) - L(\psi_0) \Big)(C(i,s)) \,\middle|\, X_i, Z_i(s-1) \right]. \end{aligned} \tag{1.15}$$

The results from Theorem 1 extend to all meta-learning, as the performance of the online CV ensemble is asymptotically equivalent to the oracle ensemble of candidate estimators as t goes to infinity

$$\frac{d_{0,t}(\psi_{n,t,\alpha_{n,t}}, \psi_0)}{d_{0,t}(\psi_{n,t,\overline{\alpha}_{n,t}}, \psi_0)} \to_p 1. \tag{1.16}$$

We note that one could also define a sequence of $\mathcal{H}_m$-specific online SLs, ranging from highly parametric to nonparametric for $m = 1, \ldots, M$, possibly stratified by the subject itself.

Then, the online ensemble would have candidate algorithms $\hat{\Psi}_k$ for $k = 1, \ldots, K$ augmented with a collection of online SLs indexed by weight classes $\mathcal{H}_m$, $m = 1, \ldots, M$. In this matter, the discrete online SL would adaptively determine the optimal level of data adaptivity of the meta-learner, based on many candidates for the online discrete SL considered. For example, depending on the number of subjects n and time t, the choice $k_{n,t}$ of the online CV selector might switch from discrete online SL based on K algorithms to more aggressive online SL indexed by a more flexible weight class over time.

Algorithm

Due to the continuously updating procedure that allows the algorithm to evolve over time, POSL generalizes to a diversity of data streams. POSL accommodates varying degrees of personalization, such as within-covariate or within-subject. It can handle multiple (potentially time-varying) dependence structures, from individual time series to networks of connected individuals. We delineate *one version* of POSL in Algorithm 1, which benefits from learning both from other subjects and from the history of the target individual's trajectory.

We define Historical learners as K_H learners generating a pooled (across individuals and time before t) estimator $\hat{\Psi}_k(P^0_{n,t})$ for algorithm $k \in K_H$, trained on samples $j = 1, \ldots n$ in the training sample. The motivation behind Historical learners is to provide an initial estimate based on previously collected trajectories (or multiple concurrently collected time series), convenient for forecasting early in the trajectory for individual i. Historical learners can be trained on time series data collected even before the trajectory of interest is sampled, and can be updated at specific time points t_s based on the computational efficiency. We note that Historical learners can generate a historical online SL as well, which thus provides another candidate online estimator. On the other hand, we define Individual learners as K_I learners applied to $P_{n,t}$, which stratify per subject when training. With that, we generate an individual estimator $\hat{\Psi}_{S,k}(P^0_{i,t})$ for algorithm $k \in K_I$, individualized to sample target i. Individual learners train on the training data by stratify by ID, and predict the outcome in the future according to the forecast horizon. The K_I candidate learners are possibly time series learners, and are frequently updated in order to accommodate the continuously incoming data. Like historical learners, individual learners can also form an individual online SL, which becomes another candidate in the POSL library.

With this formulation, we allow the POSL to leverage CV in order to choose between pooled and individual fits at each time point t — in essence, allowing the algorithm to choose an appropriate structure from the data (learning from samples if no conditional stationarity is present or learning through time). This results in a natural adaptation to the amount of available data and the stationarity of the individual target time series. The final discrete and full online SL for the POSL is generated based on all the samples until specified time t. The CV selector $k_{n,t}$ reflects an optimized ensemble among all the available learners; the candidate learners reflect a collection of online algorithms which range in how much they use the current time series (stratify by ID) and use the historical data (pool across all

available time series). All simulations in Section 1.6 test the version of the POSL described in Algorithm 1.

Algorithm 1 Personalized Online Super Learner

t_s: time steps at which Historical learner fit is updated.
K_H: Historical candidate learners.
K_I: Individual candidate learners.
K: All candidate learners, $K_H \cup K_I$.
k: Any learner among candidate learners.

Procedure HISTORICAL LEARNER(n,t)
Return $\hat{\Psi}(P^0_{n,t})$, trained using using all available units ($i = 1, \ldots, n$) for any t.

Procedure INDIVIDUAL LEARNER(i,t)
Return $\hat{\Psi}_S(P^0_{n,t})$, for any t where we stratify by sample i.

while $t < \tau$ **do**
  for $k \in K_H$ **do**
    if $t \in t_s$ **then**
      Run HISTORICAL LEARNER(n,t), return $\psi^H_{n,t,k} = \hat{\Psi}_k(P^0_{n,t})$.
    else
      $\psi^H_{n,t,k} = \hat{\Psi}_k(P^0_{n,t_{s-1}})$.
    end
  end
  for $k \in K_I$ **do**
    Run INDIVIDUAL LEARNER(i,t), return $\psi^I_{i,t,k} = \hat{\Psi}_{S,k}(P^0_{n,t})$.
  end
  if DISCRETE ONLINE SUPER LEARNER **then**
    Return $\psi_{n,t,k_{n,t}}$, where $k_{n,t} = \text{argmin}_k \, R_{CV}(P_{n,t}, \{\psi^H_{n,t,k}, \psi^I_{i,t,k}\})$.
  end
  if ENSEMBLE ONLINE SUPER LEARNER **then**
    Return $\psi_{n,t,\alpha_{n,t}}$, where $\alpha_{n,t} = \text{argmin}_\alpha \, d_{0,t}(\{\psi^H_{n,t,k}, \psi^I_{i,t,k}\}, \psi_0)$.
  end
end

1.5 Personalized Online Super Learner with Dynamic Streams

In most practical settings, time series data exhibit a heterogeneous streaming profile comprised of varied length of the series and diverse start and exit times. In order to accommodate a manifold of different applications, including dynamic enrollment and exit (collectively referred to as dynamic streams), we extend the formulation of the estimation problem described in the subsection 1.4 and the POSL algorithm, in this subsection. In particular, we redefine the observed data, statistical model, and the target parameter below, taking into account the possibly dynamic and disparate tracking of each collected sample. Afterwards, we describe the appropriate CV for dynamic streams, redefine the loss, online CV selector, ensemble of candidate estimators, and define a new prediction function for dynamic enrollment streaming settings. In Figure 1.3 we provide examples of dynamic streams, introducing subject-specific time and its relationship to chronological time, and illustrate various ways in which this heterogeneous streaming profile can evolve.

Formulation of the Estimation Problem with Dynamic Streams

Let E_i be an entry time for each new time series, corresponding to the chronological time domain $t = 1, \ldots, \tau$, for $i = 1, \ldots, n$. We assume a natural ordering for all E_i, with $0 \leq E_1 \leq \ldots \leq E_n$ even if multiple samples enroll around the same time. Our assumption on the strictly monotone increasing entry times follows from the fluid definition of t; as we put no restrictions on the time definition, we note that for sufficiently small t, no subjects exhibit $E_i = E_{i+1}$ even if enrolling a group of units.

Suppose each unit i is tracked over M_i time points starting at E_i, where the final chronological time T_i and duration M_i are within the $(0, \tau)$ range. The one-to-one mapping from t-chronological time to m-individual time is given by $h_i(t) = t - E_i$, resulting in $m \in \{0, \ldots, M_i\}$; with that, let $h_i(E_i) = 0$ and $h_i(T_i) = M_i$. The function h_i is subject specific as it depends on individual i's start time E_i; writing just h denotes a general function which can take a vector of start times. We define the process on each unit i as $m \mapsto O_i^{'}(m)$ with $O_i^{'} = (O_i(m) : m = 0, \ldots, M_i)$ being the full observed time series on subject i. Let $O_i^{'}(m) = O_i(h_i(t))$ be the observed time series on subject i at chronological time t; note that, in order to define $O_i(h_i(t))$, we need the current time t and the subject's entry time, E_i. As in subsection 1.2, the time series decomposes as $O_i^{'}(m) = (Y_i^{'}(m), W_i^{'}(m))$, equivalently written as $O_i(h_i(t)) = (Y_i(h_i(t)), W_i(h_i(t)))$ in chronological time. Here, $Y_i(h_i(t))$ is a response variable and $W_i(h_i(t))$ is a vector of time-varying covariates for subject i at collected point $h_i(t)$ in chronological time t. We similarly define X_i as the vector of baseline covariates collected at entry time E_i for subject i. We define the total number of subjects in the study at chronological time t as $n(t) = \sum_{i=1}^{n} \mathbb{I}(E_i \leq t)$, reflecting all trajectories started before (or at) time t. Equivalently, we also let $n_m(t)$ denote the number of samples with m points up to t, where $n_m(t) = \sum_{i=1}^{n(t)} \sum_{s=1}^{t} \mathbb{I}(h(s) = m)$. We can represent the observed data

coming from dynamic streams as a single time series through chronological time t by defining a process F such that $F(t) = F^{n(t)}(t) \equiv \{O_i(h_i(t)) : \text{all } i \text{ where } E_i \leq t\}$. Then, we have that $(F^{n(0)}(0), \ldots, F^{n(\tau)}(\tau))$ reflects a single time series we can learn from. We emphasize that, for dynamic streaming settings, $F(t)$ describes a collection of all observed time series enrolled at or before time point t.

In the previous sections we defined time t-specific and sample i-specific history as $\overline{O}_i(t-1) = (O_i(1), \ldots, O_i(t-1))$. For dynamic streams, we let the history of the i-th time series until time t be defined as $\overline{O}_i(h_i(t-1)) = (O_i(h_i(0)), \ldots, O_i(h_i(t-1)))$. We define the complete history for all samples until chronological time t as $\overline{O}(h(t-1))$, which includes all trajectories observed by time t.

$$\overline{O}(h(t-1)) = \bar{F}(t-1) = \{\overline{O}_i(h_i(t-1)) : \text{ all } i \text{ where } E_i \leq t\}.$$

Analogue to subsection 1.2, let $Z_i^{'}(m-1) = Z_i(h_i(t-1))$ denote the fixed dimensional summary measure of the form $Z_i^{'}(m-1) = Z_i(h_i(t-1)) = f_i(\overline{O}(h(t-1)) \in \mathbb{R}^k$; note that with this formulation, $Z_i(h_i(t-1))$ could support both time and sample dependence, as discussed in previous sections. We define the estimand as a time m prediction problem for the i^{th} subject:

$$\Psi(P_0)(X_i, Z_i^{'}(m-1), m) = \Psi(P_0)(X_i, Z_i(h_i(t-1)), h_i(t-1)) = E_{P_{0,X_i}}[Y_i^{'}(m)|Z_i^{'}(m-1)], \tag{1.17}$$

where the expectation on the right hand side is taken w.r.t the conditional distribution P_{0,X_i}, and $\Psi(P_0)(X_i, Z_i^{'}(m-1), m)$ is the prediction function for the i^{th} subject at time series time m (equivalent to chronological time $h_i(t)$). In particular, we want to learn $(X, Z^{'}(m-1), m) \mapsto \Psi(P_0)(X, Z^{'}(m-1), m)$, and thereby obtain a prediction function for each unit i that predicts $Y_i^{'}(m)$ with $\Psi(P_0)(X_i, Z_i^{'}(m-1), m)$. Further, let $P_{n,t} \mapsto \hat{\Psi}(P_{n,t})$ represent a mapping from $P_{n,t}$ into a predictive function $\hat{\Psi}(P_{n,t})$. We define $\hat{\Psi}(P_{n,t})(X_i, Z_i^{'}(m-1), m)$ as an estimator of $\Psi(P_0)(X_i, Z_i^{'}(m-1), m)$, and write $\hat{\Psi}(P_{n,t})(X_i, Z_i^{'}(m-1), m) = \psi_{n,t}(X_i, Z_i^{'}(m-1), m)$ as the predicted outcome for unit i of the estimator $\hat{\Psi}(P_{n,t})$ based on $(X_i, Z_i^{'}(m-1), m)$.

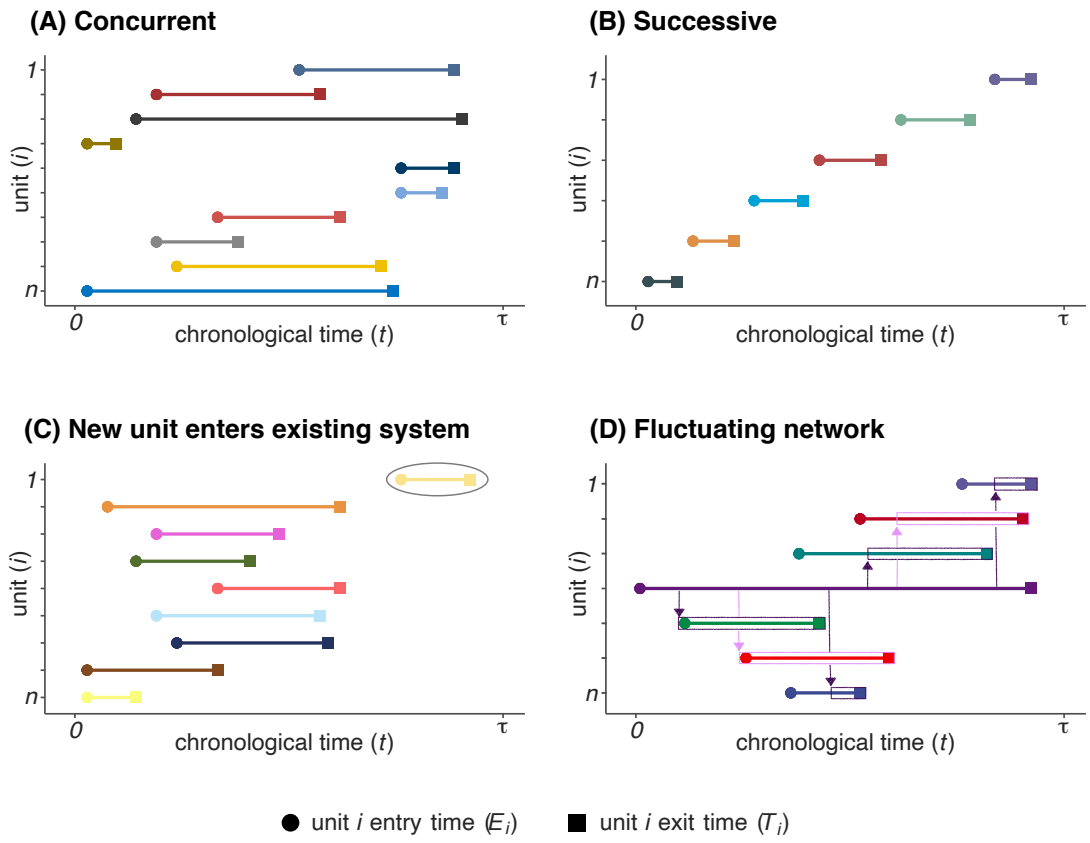


Figure 1.3: Examples of dynamically streaming time series data structures. The depicted scenarios display a window of chronological time $[0, \tau]$ where units i have varying entry times E_i, exit times T_i, and observation periods M_i. A classic example showing units with various start and exit times over a window of chronological time is shown in (A), and an example of this setting includes website or sensor monitoring. In (B) streams of non-overlapping time series, such as patients seen by a doctor, are displayed. In (C) a dynamic setting with large amounts of historical information relative to recently entered time series (unit 1, circled) is shown; in this scenario training algorithms on historical information might be particularly useful for making forecasts in the beginning of subject 1's trajectory. An example of a network of dynamically streaming time series is provided in (D); in particular, the network for a single subject (unit 4) is shown. This unit's network is constrained in that they can only interact with at most two other units at any given time, and this restriction is illustrated by the two colors for arrows and boxes that branch from the middle subject's time series.

Personalized Online Super Learner for Dynamic Streams

Let $C_h(i,t,\cdot) = C_h(i,s,E_i,T_i) = (X_i, Y_i(h_i(t)), Z_i(h_i(t-1)), E_i, T_i)$ denote the subject i and chronological time t observed data. As previously defined, B_t defines a time-specific split vector such that, for all $1 \leq i \leq n$, $B_t(i,\cdot,E_i,T_i) \in \{-1,0,1\}^t$. We define the following split of the learning set into corresponding three disjoint subsets,

$$B_t^v(i,s,E_i,T_i) = \begin{cases} -1, & C_h(i,s,E_i,T_i) \text{ not used} \\ 0, & C_h(i,s,E_i,T_i) \text{ in the training set} \\ 1, & C_h(i,s,E_i,T_i) \text{ in the validation set,} \end{cases}$$

where our CV scheme now takes into account if we have yet to observe sample i (by chronological time t) and how long is its trajectory. Knowing E_i and T_i proves important shortly, as we define how the loss, and the corresponding (online) CV risk are defined and evaluated.

Let $L(\psi)(C_h(i,t,E_i,T_i))$ denote the loss function for the data record $C_h(i,t,E_i,T_i)$ for subject i, where

$$(X_i, Y_i(h_i(t)), Z_i(h_i(t-1)), E_i, T_i) \mapsto L(\psi)(X_i, Y_i(h_i(t)), Z_i(h_i(t-1)), E_i, T_i).$$

For example, given the prediction function ψ, we might want to evaluate its performance using the squared error loss

$$L(\psi)(C_h(i,t,E_i,T_i)) = c(i,h_i(t),E_i,T_i)(Y_i(h_i(t)) - \psi(C_h(i,t,E_i,T_i)))^2, \tag{1.18}$$

where $c(i,h_i(t),E_i,T_i)$ represents a weight function dependent on sample i, time $h_i(t)$ and the unit's entry and exit time. In particular, if $t \leq E_i$ or $t \geq T_i$, we might define $c(i,h_i(t),E_i,T_i) = 0$ resulting in $L(\psi)(C_h(i,t,E_i,T_i)) = 0$. The $c(i,h_i(t),E_i,T_i)$ weight might additionally represent a weight function that down-weights the losses for points $h_i(t)$ for which $n_l(t)$ is small; with that, our prediction function would not be penalized for not having enough data collected up until certain times l.

Let $\mathcal{I} = \{i : E_i \leq t, t \leq T_i\}$ denote a set of all samples with start date before current chronological time t with data still being collected ($t \leq T_i$). We define the true risk as the expected value of $L(\psi)(C_h(i,t,E_i,T_i))$ w.r.t the true conditional distribution P_{0,O_i} for each sample i:

$$\begin{aligned} R(P_0,\psi) &= \sum_{t=1}^{\tau} \sum_{i \in \mathcal{I}} E_{P_{0,O_i}}[L(\psi)(C_h(i,t,E_i,T_i))|X_i, Z_i(h_i(t-1))] \\ &= \sum_{t=1}^{\tau} \sum_{i \in \mathcal{I}} E_{P_{0,O_i}}[(Y_i(h_i(t)) - \psi(C_h(i,t,E_i,T_i)))^2|X_i, Z_i(h_i(t-1))], \end{aligned} \tag{1.19}$$

defined for all subjects that had their start E_i before chronological time t, and end date after t. Note that, if sample i with start date $E_i \leq t$ also had their end date before time t - then

their loss would be undefined, unless an appropriate weighting is part of the loss definition (as discussed above). We note that $R(P_0, \psi)$ is an average of all appropriate i-specific losses measuring the performance of ψ across all available time points. One might instead be interested in defining a m-specific prediction function up until time τ. In particular, the true risk of the m-specific prediction function can be written as:

$$R_m(P_0, \psi) = \sum_{t=1}^{\tau} \sum_{i \in \mathcal{I}} \mathbb{I}(h_i(t) = m) E_{P_{0,O_i}}[L(\psi)(C_h(i, t, E_i, T_i))|X_i, Z_i(h_i(t-1))] \tag{1.20}$$

reflecting an average over all available active samples and times with m time points.

Suppose we have K candidate estimators $\hat{\Psi}_k$, where we denote $\psi^0_{n,t,k}$ as $\psi^0_{n,t,k} \coloneqq \hat{\Psi}_k(P^0_{n,t})$ for $k = 1, \ldots, K$. In order to evaluate the time specific performance of each $\hat{\Psi}_k$, we use CV for dependent dynamic steams (which takes into account E_i and T_i) in order to estimate the average loss for each candidate k over time. The online CV risk of an online estimator is computed at each time point in chronological time and defined as follows

$$\begin{aligned} R_{CV}(P^1_{n,t}, \hat{\Psi}_k(\cdot)) &= \sum_{j=1}^{t} \sum_{v=1}^{V} \sum_{(i,s,E_i,T_i) \in \mathcal{B}^1_{j,v}} L(\hat{\Psi}_k(P^0_{n,j}))(C_h(i, s, E_i, T_i)) \\ &= \sum_{j=1}^{t} \sum_{v=1}^{V} \sum_{(i,s,E_i,T_i) \in \mathcal{B}^1_{j,v}} c(i, h_i(s), E_i, T_i)(Y_i(m_i(s)) \\ &\quad - \hat{\Psi}_k(P^0_{n,j})(C_h(i, s, E_i, T_i)))^2, \end{aligned} \tag{1.21}$$

with mean squared error as the loss. One could define the online CV risk $R_{CV,m}(P^1_{n,t}, \hat{\Psi}_k(\cdot))$ of $\hat{\Psi}_k$ for evaluating the CV performance of the prediction function $\psi^0_{n,t,k}$ at time m as:

$$R_{CV,m}(P^1_{n,t}, \hat{\Psi}_k(\cdot)) = \sum_{j=1}^{t} \sum_{v=1}^{V} \sum_{(i,s,E_i,T_i) \in \mathcal{B}^1_{j,v}} \mathbb{I}(h_i(s) = m) L(\hat{\Psi}_k(P^0_{n,j}))(C_h(i, s, E_i, T_i)) \tag{1.22}$$

We define the total online CV risk of m-specific prediction functions as m-specific risks, with:

$$R_{CV}(P^1_{n,t}, \hat{\Psi}_k(\cdot)) = \sum_{m} R_{CV,m}(P^1_{n,t}, \hat{\Psi}_k(\cdot)). \tag{1.23}$$

The online CV risk $R_{CV}(P^1_{n,t}, \hat{\Psi}_k(\cdot))$ gives an empirical measure of performance for candidate estimator k trained on training data until chronological time t. We define the time t discrete online CV selector as:

$$k_{n,t} = \arg\min_{k=1,\ldots,K} R_{CV}(P^1_{n,t}, \hat{\Psi}_k(\cdot)), \tag{1.24}$$

reflecting the discrete online SL for all m. Instead, we could define a separate selector for the different time points m, with the discrete online SL stratifying the selector by m,

$$k_{n,t,m} = \arg\min_{k=1,\ldots,K} R_{CV,m}(P^1_{n,t}, \hat{\Psi}_k(\cdot)). \tag{1.25}$$

Finally, we consider a more flexible online learner that generates a weighted combination of candidate estimators at each time point. Let $\hat{\Psi}_\alpha$ be a function of empirical distribution generating an ensemble of K estimators $\{\hat{\Psi}_1, \ldots, \hat{\Psi}_K\}$ indexed by a vector of coefficients α. Let $\mathcal{H}$ define a class of weight functions, where $\alpha = \alpha(X) = (\alpha_1(X), \ldots, \alpha_K(X))$ is a collection of K weights that might depend on the baseline covariates X, with $\sum_{k=1}^K \alpha_k = 1$ and $\forall \alpha_k,\ \alpha_k \geq 0$. Let $\hat{\Psi}_\alpha = \sum_{k=1}^K \alpha_k \hat{\Psi}_k$, so that the predictive function based on the training set $P^0_{n,t}$ is given by $\psi^0_{n,t,\alpha} := \sum_{k=1}^K \alpha_k \hat{\Psi}_k(P^0_{n,t})(C_h(i,t,E_i,T_i))$ with $\alpha \in \mathcal{H}$. We define a $\mathcal{H}$-specific online CV selector for the ensemble as:

$$\begin{aligned} \alpha_{n,t} &= \operatorname*{argmin}_{\alpha\in\mathcal{H}} R_{CV}(P^1_{n,t}, \hat{\Psi}_\alpha(\cdot)) \\ &= \operatorname*{argmin}_{\alpha\in\mathcal{H}} \sum_{j=1}^{t} \sum_{v=1}^{V} \sum_{(i,s,E_i,T_i)\in\mathcal{B}^1_{j,v}} (Y_i(h_i(t)) - \hat{\Psi}_\alpha(P^0_{n,t})(C_h(i,t,E_i,T_i)))^2, \end{aligned} \tag{1.26}$$

where the loss is defined as the mean squared error. Alternatively, we could compute an online CV selector $\alpha_{n,t,m}$ for each m, where

$$\alpha_{n,t,m} = \operatorname*{argmin}_{\alpha\in\mathcal{H}} R_{CV,m}(P^1_{n,t}, \hat{\Psi}_\alpha(\cdot)). \tag{1.27}$$

1.6 Simulations

We used simulations to evaluate the POSL implementation described in Algorithm 1, testing its performance and adaptivity over time for several common time series settings. For all scenarios, we simulated a total of 31 time series with $\tau = 540$ time-points, and repeated the entire procedure 50 times for a total of 1550 trajectories. We used a random sample of 30 time series to train a Historical SL, and the remaining random sample for the Individual learners. For all simulations described below, we used the same library consisting of a grid of XGBOOST, GLM and ARIMA learners for the Historical and Individual learners implemented in SL3 R package[23, 27, 60, 99]. The Historical SL was fit once on the pooled data across individuals. The Individual learners and POSL were updated every 20 time points, resulting in possibly different fits and weights at different times. In particular, we sequentially trained over incoming batches of 20 time points after start time 10, $\{10, 30, 50, \ldots, 470, 490, 510\}$, and we evaluated the loss over the last five time points of the time series i for which we forecast. The POSL was evaluated using four different simulation scenarios, which were based on autoregressive integrated moving average (ARIMA) models and Gaussian mixture autoregressive (AR) models in Simulations A–C and Simulation D, respectively. Each of

the simulations reflect different degrees/forms of similarity between the data-generating processes (DGP) used to generate the historical subjects and the individual subject. For each simulation study, the time series generated by these different DGPs is shown in Figure 1.4; the corresponding density plots can be found in the Appendix Figure 1.9.

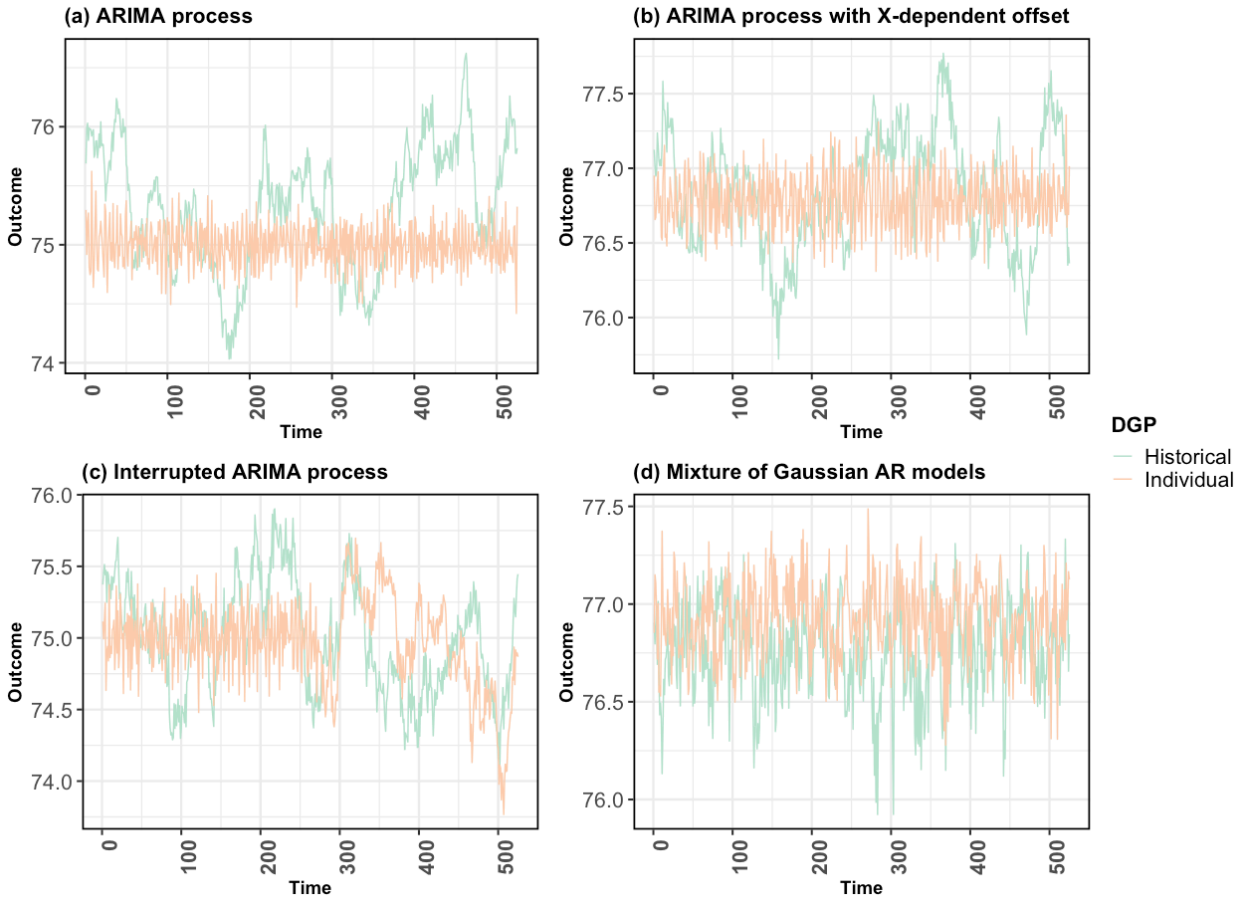


Figure 1.4: Mean outcome for the Historical and Individual data-generating processes (DGP) over time. Each panel corresponds to a simulation described in section 4, depicting both the Historical DGP and Individual DGP for simulation studies A–D. For each simulation, the mean value over time was obtained based on 100 simulated time series from both DGPs, with a final time point of $\tau = 540$.

In Figure 1.5, we report an average over 50 simulations of POSL's ensembling weights, which are assigned at each update to K_H Historical learners and K_I Individual candidate learners. Here we emphasize that POSL used a convex non-negative least squares (NNLS) regression as its meta-learner. Additionally, we compared the performance of the POSL to

purely online and purely offline ensemble SL methods using the same library of candidate learners, meta-learner, data, and test set. Specifically, we compared POSL to the canonical online SL algorithm [7] and the V-fold CV-based SL for independent and identically distributed data [136]. The online SL was trained using all the samples in a sequential manner: the loss was evaluated over a future five time-point window not seen by the learners, and then the fit was updated with new data. The evolution of the mean squared error (MSE) is shown in Figure 1.6.

Simulation A: Different ARIMA Processes

A simple scenario was considered as a first step to make sure that the POSL could learn the conditional mean under the correct DGP over time for sample i of interest. In particular, in this simulation historical subjects and the individual subject were sampled from different DGPs. We sampled 30 time series from a fifth-order ARIMA model, ARIMA(5,0,0), and this represented the historical subjects used to train the Historical SL. We sampled a single time series i, who is of interest for optimizing predictions, from an ARIMA(1,0,5) model.

Simulation B: ARIMA Processes with X-dependent Common Offset

In this simulation we built on Simulation A by adding a common component to the Historical and Individual DGPs. This permitted an investigation of the performance and behavior of the POSL algorithm in situations where there is considerable similarity in the Historical and Individual DGPs, but the Individual DGP is different enough that the POSL should be able to pick up on this asymptotically. We simulated baseline covariates $X = \{X_1, X_2, X_3\}$ with

$$\begin{aligned} X_1 &\sim \text{Binomial}(0.5), \\ X_2 &\sim \text{Uniform}(19, 90), \\ X_3 &\sim \text{Uniform}(0, 2). \end{aligned}$$

The distribution of the baseline covariates here was motivated by the data on sex, age, and care unit considered in the Clinical Data Application Section. We defined the X-dependent offset as a function of X_1, X_2 and X_3 with $f(X_i) = 0.5X_{1,i} + 0.02X_{2,i} + 0.5X_{3,i}$ being the offset for sample i. We sampled 30 time series from $f(X)$ + ARIMA(5,0,0) process, reflecting the Historical DGP. We generated sample i from $f(X_i)$ + ARIMA(0,0,5), so the trajectory evolves as a MA process with offset $f(X_i)$.

Simulation C: Interrupted ARIMA Processes

We continued to build on Simulation B by generating an interrupted time series as the DGP for sample i we want to create forecasts for. The intention for this simulation was to test if POSL can detect changes in the underlying stream of data and adjust accordingly. We sampled 30 time series from $f(X)$ + ARIMA(5,0,0) process, representing the Historical DGP. As in Simulation B, we defined $f(X)$ as a X-dependent function with $f(X_i) = 0.5W_{1,i} +$

$0.02W_{2,i} + 0.5W_{3,i}$, characterizing the sample i offset. In contrast, we sampled trajectory i as an interrupted time series, with the first half drawn from $f(X)$ + ARIMA(0,0,5) process and the second half drawn from the same process as the Historical DGP, $f(X)$ + ARIMA(5,0,0).

Simulation D: Finite Mixture of Gaussian Autoregressive Processes

In Simulation D, we simulated sets of time series using the GRATIS R package, which was developed to expedite simulation of dependent data with controllable features and to provide a basis for time series benchmarking [66]. The general approach employed is based on Gaussian mixture AR models to generate a wide range of non-Gaussian and nonlinear time series. First developed by Nhu et al.[71], mixture transition distribution models were used to capture many non-Gaussian and nonlinear features, and were later generalized to Gaussian mixture AR models by Wong and Li[146]. In addition to supporting generation of heterogeneous sets of time series, GRATIS also provides options for simulating from a random population of mixture AR models with specified features[66]. We used this software to specify common features of the Historical and Individual DGPs, including entropy and the smoothed trend component for the Seasonal and Trend decomposition using Loess (STL decomposition). We differentiated the series based on their stability, defined as the variance of non-overlapping window means and the largest mean shift between two consecutive windows. With that, the Historical and Individual DGPs exhibited the same trend and amount of information, but different variance.

Simulation Results

From Figure 1.5A, we can see how POSL assigned its convex NNLS-based ensembling weights over time for Simulation A. The POSL gave more weight to the Historical learners in the beginning as there were not enough time points to learn solely from sample i. As more data on sample i is collected and the time series progresses, POSL progressively and consistently gave more weight to the Individual learners. As shown in Figure 1.6A, the POSL demonstrated good forecasting performance in terms of the MSE at all training times for this simulation, with V-fold SL as a close second.

The evolution of meta-learning weights for the Historical and Individual learners over time is shown for Simulation B in Figure 1.5B. As seen in Simulation A results, the POSL gives more weight to the Historical fit in beginning, due to the scarce number of time points collected for time series i. As data becomes more abundant, individualized learners are better able to characterize the conditional mean for sample i, and POSL therefore gives more weight to the Individual learners as time progresses. However, due to the common offset in Simulation B, we can see that POSL does not start giving more weight to the Individual learners until about 100 time points are included in the training set — much further than seen in Simulation A. This shows that POSL is able to pick up on the resemblance between the Historical and Individual DGPs, but also is able to distinguish the Individual DGP

over time. As shown in Figure 1.6B, POSL demonstrates uniformly the best forecasting performance relative to comparators for Simulation B.

From Figure 1.5C we can see that the POSL is able to detect changes in the time series data for sample i as time progresses and more data is collected. As in Simulation B, the POSL starts with giving more weight to the Historical learners (until about $t = 100$), but quickly learns to start giving more weight to the Individual learners. At time $t = 270$, at roughly about half of the training time, we can see that POSL responds to the simulated interruption in the DGP as it reverts back to giving more weight to the Historical learners. The distribution of weights assigned to the Individual learners continues to decrease until the end of training, as demonstrated in Figure 1.5C, showing that POSL is able to quickly adapt to changes in time series as time progresses. In Figure 1.6C, we can see that POSL outperforms all other tested algorithms for Simulation C, except at rare points in the later part of the time series i when V-fold SL slightly outperforms (or performs as well) as POSL. This can be explained by the fact that V-fold SL fit is trained only on the samples sampled from $f(X)$ + ARIMA(5,0,0) process, and POSL has to learn the correct current form with a slight delay due to the small batch sizes used for training.

For Simulation D (Figure 1.5D), we can see that POSL once again starts with giving more weight to the Historical learners, but eventually switches completely to the Individual learners as more data on the sample in question is collected. In terms of the MSE for this simulation, POSL shows uniformly the best forecast performance across all tested times compared to the other SL algorithms considered (Figure 1.6D).

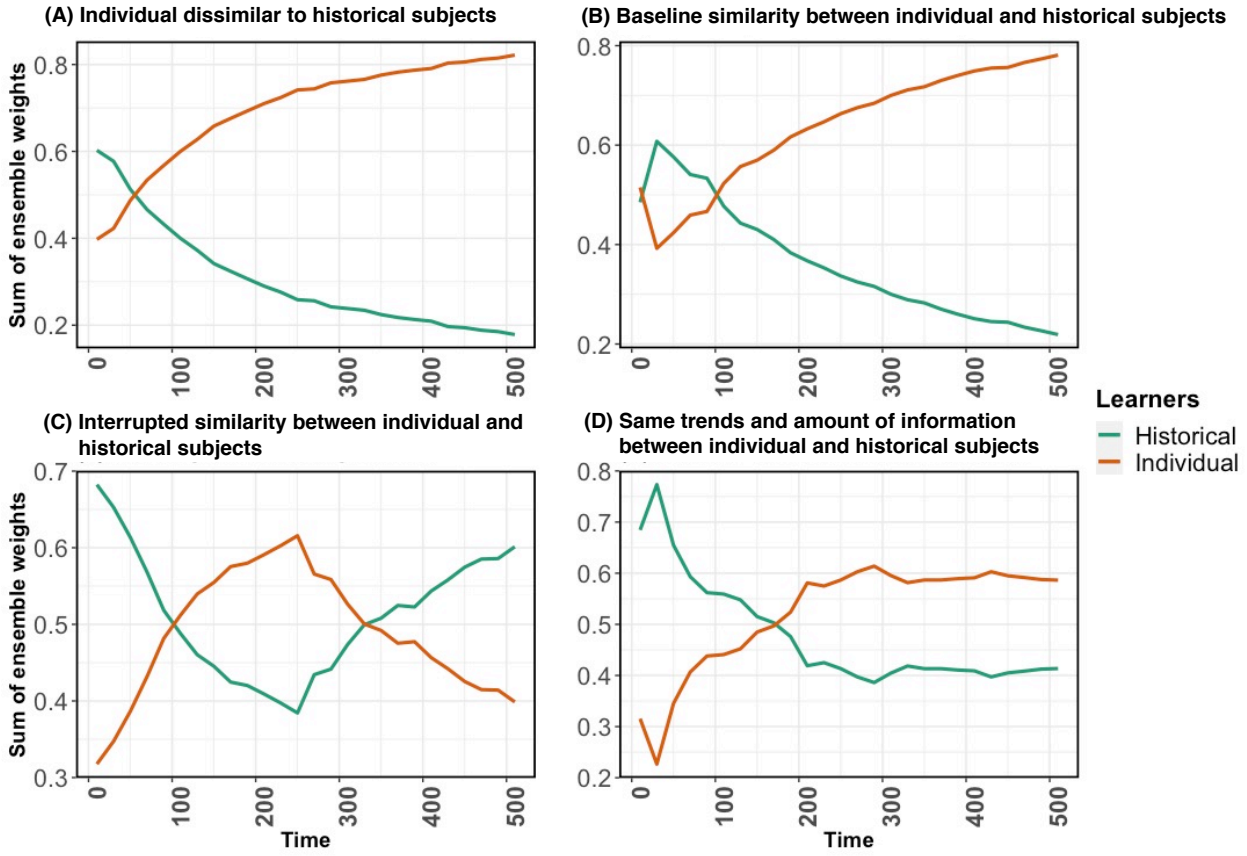


Figure 1.5: Sum of the Personalized Online Super Learner's (POSL) ensembling weights over time, stratified by Historical and Individual learners, for four simulation studies. Evolution over time of POSL's ensembling weights that were assigned to candidate learners at each training time by its meta-learner, a convex non-negative least squares regression. The weights assigned to each candidate learner were grouped by the learner type (either Historical or Individual for learners trained on the historical subjects or individual subject, respectively). The four simulation studies described in section 4 were considered. The results for Simulations A–D are summarized in parts (A)–(D) across 50 runs of each simulation.

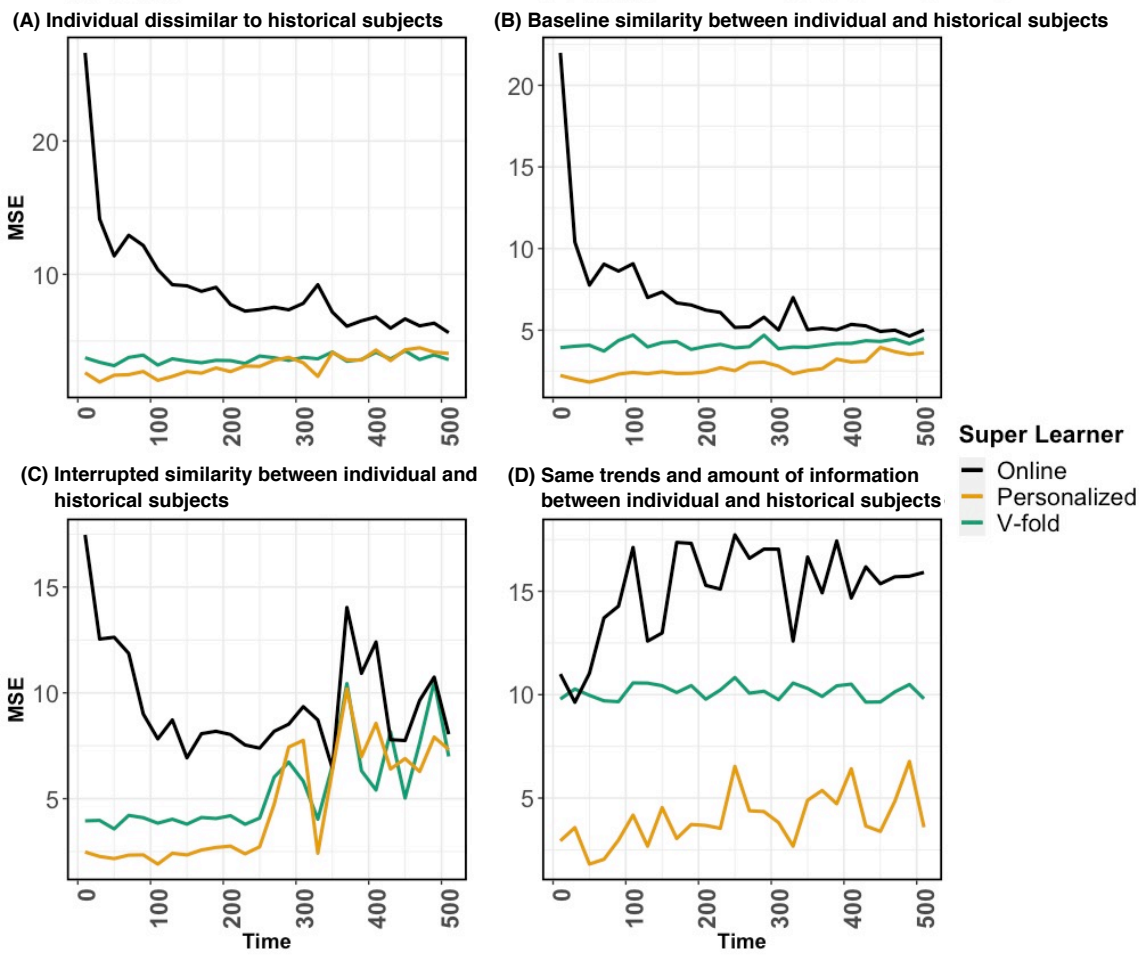


Figure 1.6: Predictive performance over time for three different Super Learners (SL) in four simulation studies. Evolution over time of the mean squared error (MSE) of the Personalized Online SL ("Personalized"), online SL ("Online"), and offline SL trained under a V-fold cross-validation scheme ("V-fold") for four simulation studies. The "V-fold" SL was trained once on the simulated data, whereas "Personalized" and "Online" were trained in an online fashion; the loss was evaluated over a five time-point window that was not seen by the any of them. The data for seen by the SLs was comprised of 31 subjects, and each subject's time series consisted of 540 time points. One individual subject was considered for making predictions and this subject's time series was sampled from a different data-generating process (DGP) than the other 30 subjects, which represented a set of historical / previously observed subjects that could be used to assist in making predictions for the individual of interest. The four simulation studies described in section 4 were considered, and each simulation was ran 50 times. The results for Simulations A–D are summarized in parts (A)–(D).

1.7 Clinical Data Application

We illustrate the POSL algorithm in an application for five-minute ahead forecasting of an individual's mean arterial pressure (MAP), which is one of the most important vital signs in the intensive care unit (ICU). Data obtained from the MIMIC II database (Multiparameter Intelligent Monitoring in Intensive Care) included 370 subjects' baseline covariates (age, sex, body mass index, ICU subunit, SAPS II and SOFA mortality scores, ICU admission type), time-varying binary exposures (vasopressors, ventilation, sedation), and time-varying continuous vitals (pulse, heart rate, systolic and diastolic blood pressures, and MAP outcome) [47, 111]. Additional covariates were derived from this set of variables, most of them from the time-varying variables, including lagged values of the time-series and summary measures over at most one hour of history.

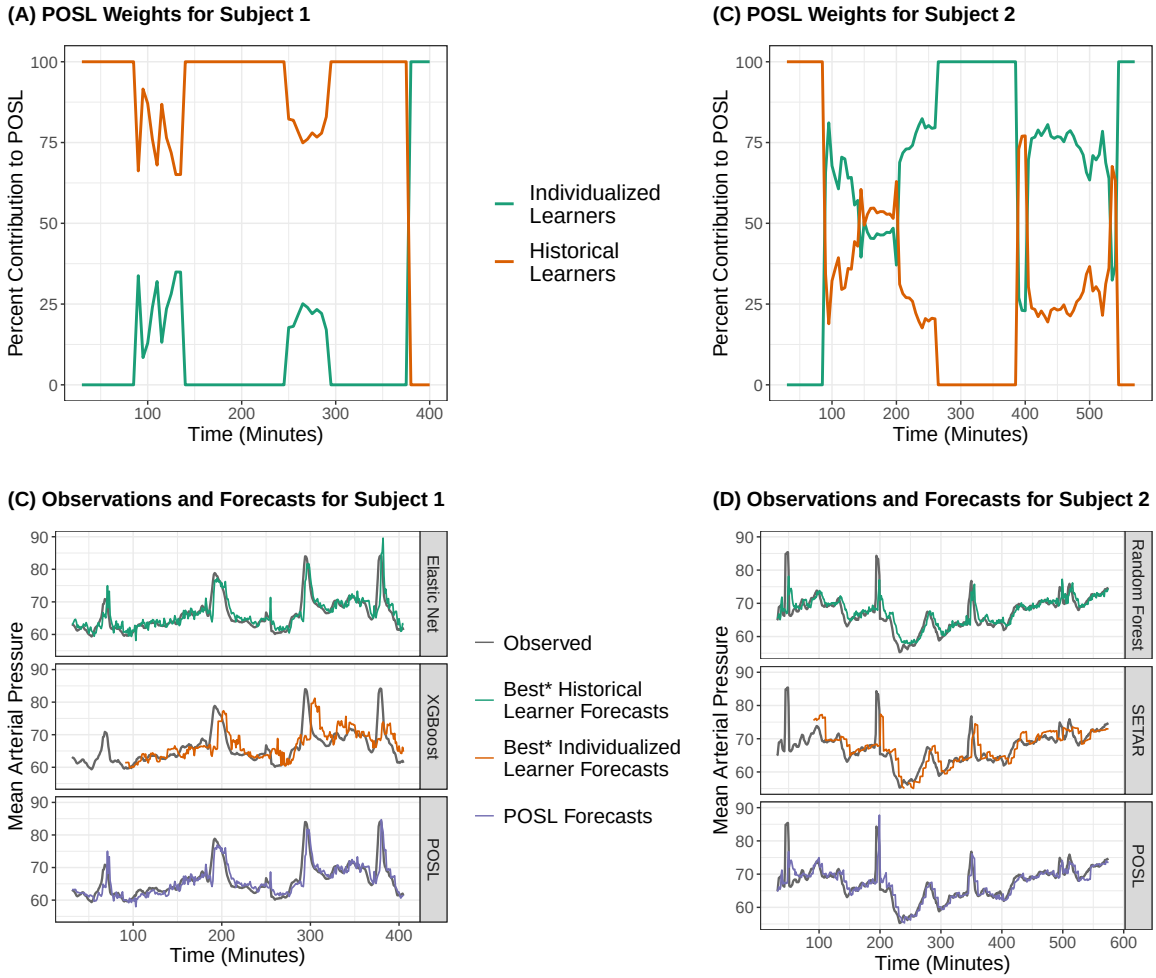
A total of 368 subjects were used for training the Historical learners using the SL3 R package [27, 99]. The library of Historical learners included the following: multiple variations of gradient boosted decision trees (XGBOOST), random forests (RANGER), and elastic net generalized linear models (GLMNET); a discrete Bayesian additive regression trees model (DBARTS), a Bayesian generalized linear model (BAYESGLM), and a linear regression (GLM) [23, 147, 41, 24, 43]. This library was fit after reducing the number of time-varying covariates with a pre-screening step that selected the 200 "most important" time-varying covariates according to a RANGER random forest variable importance metric, and then those 200 time-varying covariates and the baseline covariates were passed on to the library of Historical learners. The two patients that were not selected for training the Historical learners were used to train, separately, a library of Individual online learners; the selection of the two patients was random. Individual learners were updated with each batch of five observations (i.e., updated every five minutes), following accumulation of an initial training size consisting of 60 observations. The Individual learners included the following: multiple variations of nonlinear time-series models (TSDYN) and elastic net generalized linear models (GLMNET); a linear regression (GLM), a gradient-boosted decision tree model (XGBOOST), a random forest model (RANGER), and an ARIMA model with automated tuning (AUTO.ARIMA) [38, 41, 23, 147, 60]. The linear regression and ARIMA Individual learners were fit following a pre-screening step involving lasso regression, in which the variables with non-zero lasso regression coefficients were selected and then passed to these Individual learners.

At each subject-specific 5-minute update, POSL selects the candidate with the lowest online cross-validated risk, where the set of candidates included the Individual and Historical learners, as well as ensembles of them. For both subjects, POSL's risk function was the weighted mean squared error (expectation of the weighted squared error loss), where the weights decreased as a function of time. For losses obtained 180 minutes or more from the subject's current time m, the weights were set to 0 when calculating the weighted mean loss. For losses obtained 30 minutes or less from the subject's current time m, the weights assigned to those losses were set to 1. The weights assigned to losses that were obtained more than 30 minutes but less than 180 minutes from the subject's current time m decayed as $(1 - 0.001)^{m-m_L}$, where m_L is the time when the loss was measured, $m_L = 0, \ldots, m$, so

the difference $m - m_L$ is the lag in time fromloss's time andcurrent time. Let $w(m_L)$ denote the weight assigned to a loss measured at time m_L, then this strategy to weight losses based on their lag from the current time can also be expressed as

$$w(m_L) = \begin{cases} 0 & \text{if } m_L \leq m - 180 \\ (1 - 0.001)^{m - m_L} & \text{if } m - 180 < m_L < m - 30 \\ 1 & \text{if } m_L \geq m - 30. \end{cases}$$

In Figure 1.7, we illustrate the application of POSL to the ICU data problem to obtain five-minute ahead forecasts of an individual's MAP, summarizing POSL's performance for the two subjects that were not used in training the Historical learners. In Figure 1.7A and Figure 1.7B we show how POSL assigned weight to the Historical and Individual candidate learners over time. For each subject, we identified the Individual learner and the Historical learner that had the lowest MSE when averaged across the individual's time series, and these "best" Individual and Historical learners varied across the subjects. We present POSL's forecasts and the "best" Historical and Individual learners' forecasts alongside the observed mean arterial pressure in Figure 1.7C and Figure 1.7D. In Figure 1.7E we present the MSE of learner forecasts plotted in Figure 1.7C and Figure 1.7D, which displays for each subject, the performance of the learners that performed best for both subjects. This table highlights the variability of the candidate learners' performance across subjects and the stability of POSL's performance across subjects, and demonstrates POSL's ability to adapt to an individual's time series and to perform better than any of its candidates.



(E) Mean Squared Error of POSL and Best* Learners for Subjects 1 and 2

Subject	POSL	Historical Elastic Net	Historical Ranger	Individualized XGBoost	Individualized SETAR
1	7.596	7.69	9.513	19.742	49.067
2	7.397	10.655	8.696	21.094	12.495

**Best denotes lowest mean squared error*

Figure 1.7: Five-minute ahead forecasting of mean arterial pressure for two intensive care unit patients with the Personalized Online Super Learner (POSL). In (A) and (B), the POSL ensembling weights assigned at each five-minute update time, and grouped by the Historical and Individual candidate learners, are plotted for subjects 1 and 2, respectively. In (C), subject 1's observed mean arterial pressure (MAP) is plotted alongside the forecasts from the following: subject 1's best-performing Historical learner (elastic net regression), subject 1's best-performing Individual online learner (gradient boosted regression tree, XGBoost), and POSL. In (D), subject 2's observed MAP is plotted alongside the forecasts from the following: subject 2's best-performing Historical learner (random forest), subject 2's best-performing Individual online learner (self-exciting threshold autoregressive model, SETAR), and POSL. In (E), the mean squared error of all the learners plotted in (C) and (D) is displayed for both subjects.

1.8 Discussion

In this work, we consider the problem of generating personalized forecasts in the data streaming setting with multiple time series of unknown underlying structure. The Personalized Online Super Learner (POSL) is an online ensembling machine learning algorithm which utilizes multiple time series and ensembling combination methods with the goal of optimizing personalized forecasts. The POSL is regularly updated over time using batches of streaming data, and leverages both online pooled (learning across individuals) and individual (learning through time) learners at each time step, allowing for the ensemble weights to depend on the amount of data collected, stationarity, and degree of noise. The scenario studied considers observing n units'/subjects' time series over a finite number of time points. Each observation is comprised of baseline and time-varying covariates, and a response. The n time series observed might be subject to intra- and inter-dependence sampled from different data-generating processes. Also, in dynamic settings, the n time series observed might be of varying length at chronological time t, while across t comprised of different numbers and types of units.

We present multiple CV schemes relevant for different streaming settings, and advocate for an adaptive meta-learning step, where the final weights of the ensemble learner are based on mutual characteristics of a group of time series, or completely individualized. Finally, under stronger conditions then necessary for the setup we describe, we apply the results established by Benkeser et al.[7] in a more general time series setting. Opposed to the work by Benkeser et al.[7], we consider a different target parameter, and formulate the problem to include multiple time series with possible baseline dependence among samples. In addition, we extend the results to different CV schemes supported by the problem setup (dictated by the different dependence structure), and asynchronous enrollment of subjects across time. The established result shows that the performance of the CV-based best algorithm is asymptotically equivalent with the performance of the best unknown candidate learner — providing a powerful way to optimally, and in a personalized way, combine multiple estimators in an online, dependent setting [36].

We note that the POSL can be used for estimation of any parameter of the conditional distribution of O_i, given its past and past of other (possibly different) time series, that can be defined as a minimizer of an empirical risk. Thinking of all the trajectories as a single ordered time series provides an interesting opportunity to study the asymptotic behavior of the proposed online SL in a variety of settings, including dynamic streams. Depending on the number of time points, type of enrollment and dependence across subjects, it is possible to consider asymptotics in time t, number of subjects n, or a combination. For example, one can study asymptotics in time t only for a fixed number of dependent subjects, or asymptotics in time t for time series sampled from different data-generating distributions. Alternatively, we could rely on the number of samples only, which might be practically useful when subjects are followed up for a limited time frame, when there is no common structure through time, or when the entry times are all concentrated in a finite chronological time interval. For low dependence settings where samples are followed for a long period of time, it is possible

to exploit asymptotics in both the total number of time points observed as well as across the n subjects. We emphasize that POSL is able to adapt to the underlying structure in data for all the mentioned settings — this allows the proposed methodology to pick between relying on structure through time, samples, or both, at each time point. As such, while we impose assumptions on our statistical model for the sake of obtaining oracle results, our true statistical model does not rely on conditional stationarity in order for POSL to perform well, which is in contrast to the canonical online SL[7]. Our proposed method is also constructed to provide optimal forecasts for unit i sampled from P_{0,O_i}, instead of a collection of time series.

Finally, we emphasize that the POSL represents theoretically proven, flexible, open-source algorithm for many canonical and custom made time series prediction problems. While motivated by precision medicine, POSL has a wide range of applications that could be considered, including infectious disease forecasting and stock market forecasting. The general algorithm described encompasses various forecasting horizons, CVs, dependencies across time, enrollment/exit times, ensembling methods, and combinations of individual time series and pooled algorithms. Our simulation results show superior performance over current state-of-the-art online and ensembling algorithms in terms of MSE across a wide range of forecasting scenarios. In future work, we explore formulations of the POSL for adaptive peak detection and safe update procedures under data drift.

1.9 Appendix

Lemma 1. *The difference between the online cross-validated (CV) risk (minimized by $k_{n,t}$) and the online CV true risk (minimized by $\overline{k}_{n,t}$) is a discrete martingale.*

Proof. Let $M_n(f) = (R_{CV}(P^1_{n,t}, \hat{\Psi}_k(\cdot)) - R_{CV}(P_0, \hat{\Psi}_k(\cdot)))$. The difference between centered CV risk $R_{CV}(P^1_{n,t}, \hat{\Psi}_k(\cdot))$ and the true CV risk $R_{CV}(P_0, \hat{\Psi}_k(\cdot))$ conditional on the filtration defined by the training set is a discrete martingale:

$$\begin{aligned}
M_n(f) &= \sum_{j=1}^{t} \sum_{(i,s)\in\mathcal{B}^1_j} [L(\psi^0_{n,j,k}) - L(\psi_0)(C(i,s))] \\
&\quad - \sum_{j=1}^{t} \sum_{(i,s)\in\mathcal{B}^1_j} E_{P_{0,O_i}}[L(\psi^0_{n,j,k}) - L(\psi_0)(C(i,s))|X_i, Z_i(s-1)] \\
&= \sum_{j=1}^{t} \sum_{(i,s)\in\mathcal{B}^1_j} L(\psi^0_{n,j,k})(C(i,s)) - E_{P_{0,O_i}}[L(\psi^0_{n,j,k})(C(i,s))|X_i, Z_i(s-1)] \\
&= \sum_{j=1}^{t} \sum_{(i,s)\in\mathcal{B}^1_j} f(C(i,s)) - E_{P_{0,O_i}}[f(C(i,s))|X_i, Z_i(s-1)].
\end{aligned}$$

□

A1. *There exists a $M_1 < \infty$ for any valid loss function L and $\psi \in \mathbf{\Psi}$ such that*

$$\sup_{\psi\in\mathbf{\Psi}} \sup_{C(i,s)} |L(\psi)(C(i,s)) - L(\psi_0)(C(i,s))| \leq M_1.$$

A2. *There exists a $M_2 < \infty$ for $\psi \in \mathbf{\Psi}$ so that with probability 1,*

$$\sup_{\psi\in\mathbf{\Psi}} \frac{P_{0,O_i}[L(\psi) - L(\psi_0)]^2}{P_{0,O_i}[L(\psi) - L(\psi_0)]} \leq M_2 < \infty$$

A3. *There exists a slowly increasing sequence $M_3 < \infty$ such that with probability tending to 1, we have*

$$\frac{1}{M_3} < \frac{d_{0,t}(\psi_{n,t,k_{n,t}}, \psi_0)}{E_{P_{0,O_i}}[d_{0,t}(\psi_{n,t,k_{n,t}}, \psi_0)]} < M_3$$

and

$$\frac{1}{M_3} < \frac{d_{0,t}(\psi_{n,t,\bar{k}_{n,t}}, \psi_0)}{E_{P_{0,O_i}}[d_{0,t}(\psi_{n,t,\bar{k}_{n,t}}, \psi_0)]} < M_3.$$

A4. *Given that M_3 is a sequence that grows arbitrarily slow to infinity,*

$$tM_3{}^{-3} \min_k E_{P_{0,O_i}}[d_{0,t}(\psi_{n,t,k}), \psi_0)] \to \infty$$

as $t \to \infty$.

Theorem 1. *Let P_0^n describe the true data-generating distribution $P_0^n \in \mathcal{M}$, with the target parameter defined as $\Psi : \mathcal{M} \to \mathbf{\Psi}$ evaluated at a particular $P \in \mathcal{M}$. We establish the CV selector $k_{n,t}$ as the minimizer of the CV risk, and the oracle selector $\overline{k}_{n,t}$ as the minimizer of the true CV risk. Under assumptions A1–A4, there exists a constant $C(M_1) < \infty$ such that:*

$$E_{P_{0,O_i}}[(d_{0,t}(\psi_{n,t,k_{n,t}}, \psi_0))] \leq E_{P_{0,O_i}}[(d_{0,t}(\psi_{n,t,\overline{k}_{n,t}}, \psi_0))] + C(M_1)[\frac{\log(1 + K(t * n))}{t * n}]^{1/2}$$

Proof. Under Lemma 1, the proof is a direct generalization of the oracle inequality for a single time series proved in Benkeser et al.[7] to multiple time series under CV schemes described in subsection 1.3, assuming conditional stationarity. □

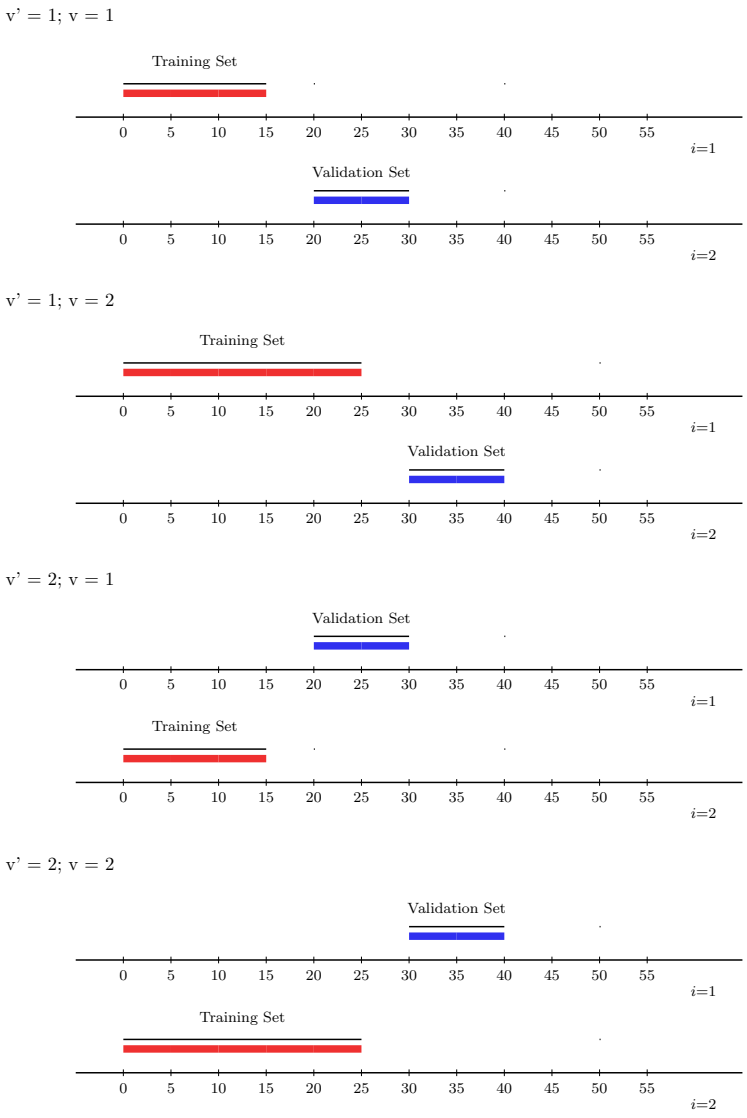


Figure 1.8: Rolling origin V-fold cross-validation (ROVFCV) scheme invoked for two unique subjects' time series, both whose final time-point of their currently available data is $t = 40$. The ROVFCV scheme is invoked with the following specification: subject identifier ($i = 1$ and $i = 2$), the initial training set size $n^0_{t,v} = 15$, validation set size $n^1_{t,v} = 10$, batch size $m = 10$, and gap $h = 5$. Given the specification and the data provided, this CV scheme thus admits $V = 4$ ROVFCV folds, where two of the folds are admitted due to splitting across samples (i.e. via V-fold cross-validation, where each unique V-fold cross-validation fold is denoted by v') and the other two folds are admitted due to splitting across time (i.e. via rolling origin cross-validation, where each unique rolling origin cross-validation fold is denoted by v). For a ROVFCV scheme, the predictions are evaluated on the future times of subjects' whose time series were not seen during training, allowing for dependence across time and samples.

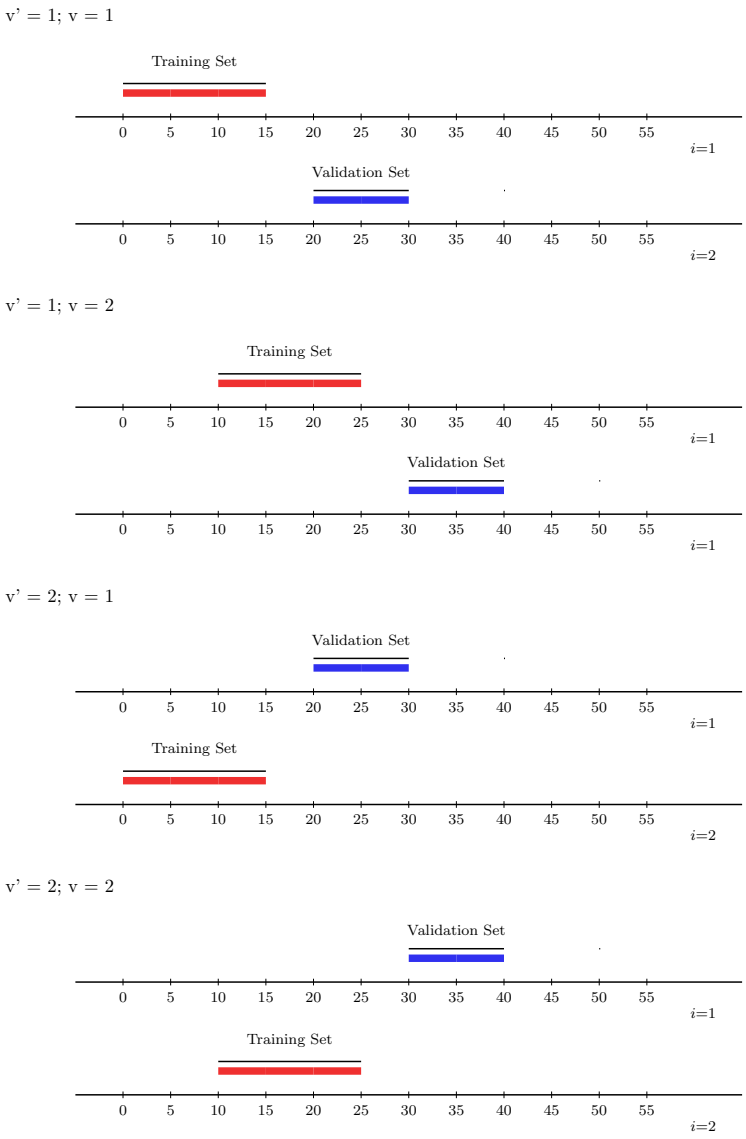


Figure 1.9: Rolling window V-fold cross-validation (RWVFCV) scheme invoked for two unique subjects' time series, both whose final time-point of their currently available data is $t = 40$. The RWVFCV scheme is invoked with the following specification: subject identifier ($i = 1$ and $i = 2$), the training set size $n^0_{t,v} = 15$, validation set size $n^1_{t,v} = 10$, batch size $m = 10$, and gap $h = 5$. Given the specification and the data provided, this CV scheme thus admits $V = 4$ RWVFCV folds, where two of the folds are admitted due to splitting across samples (i.e. via V-fold cross-validation, where each unique V-fold cross-validation fold is denoted by v') and the other two folds are admitted due to splitting across time (i.e. via rolling window cross-validation, where each unique rolling window cross-validation fold is denoted by v). For a RWVFCV scheme, the predictions are evaluated on the future times of subjects' whose time series were not seen during training, allowing for dependence across time and samples.

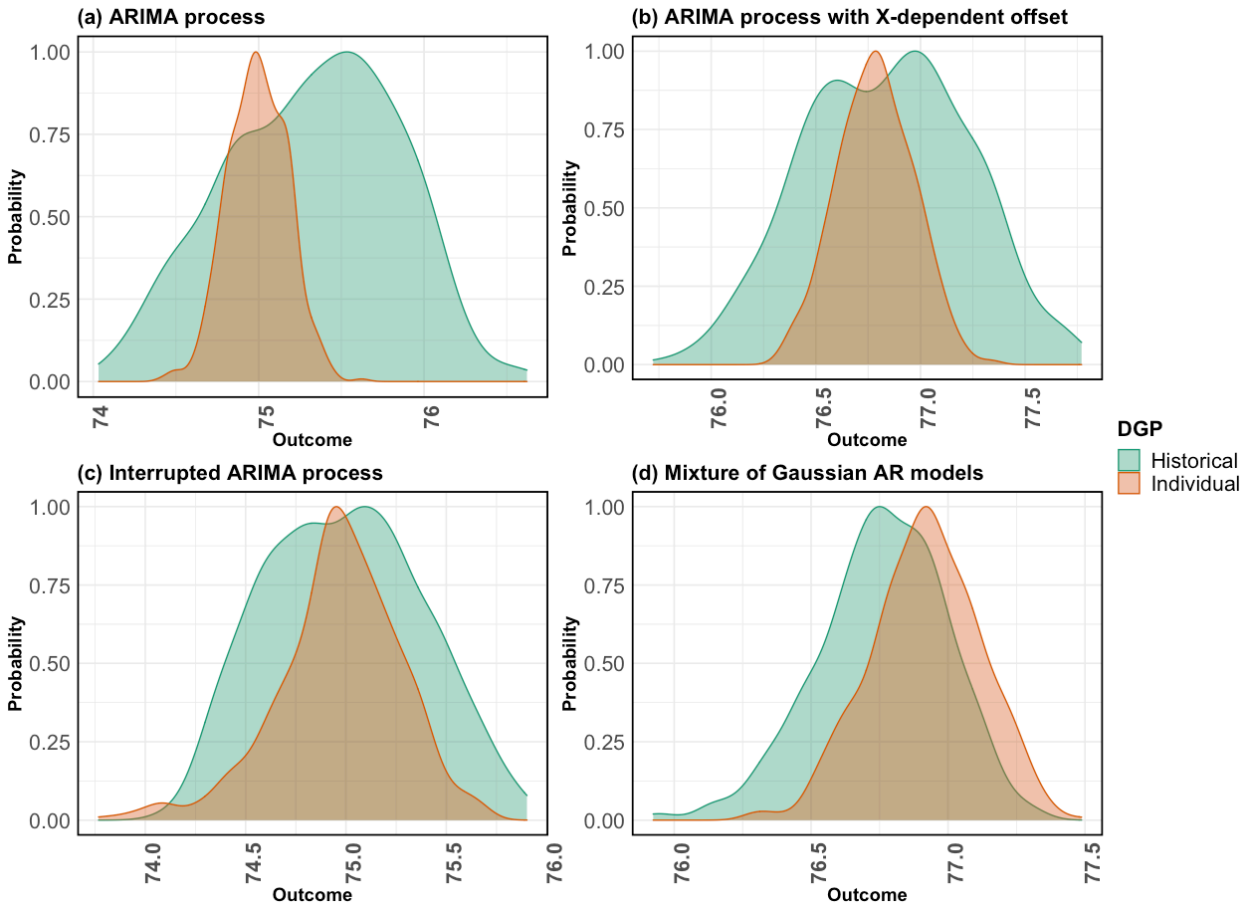


Figure 1.10: Density for both the Historical and Individual data-generating process (DGPs) considered in each simulation study, depicting the Historical DGP alongside Individual DGP, and where panels (A)–(D) correspond to simulation studies A–D described in section 4. The plotted density per panel was obtained based on a 100 simulated time series from both DGPs with a final time point of $\tau = 540$.

Chapter 2

Conditional Causal Effect for a Single time-series

Consider the case that one observes a single time-series, where at each time t one observes a data record $O(t)$ involving treatment nodes $A(t)$, possible covariates $L(t)$ and an outcome node $Y(t)$. We assume that the conditional distribution of $O(t)$, given the observed past, is described by a common function only depending on a fixed dimensional summary measure of the past ($C_o(t)$). The data record at time t carries information for an (potentially causal) effect of the treatment $A(t)$ on the outcome $Y(t)$, in the context defined by $C_o(t)$. The conditional distribution of $O(t)$ is characterized by a conditional distribution of the treatment nodes and the conditional distribution of possibly time-dependent covariates and outcome. We consider the case when the (possibly causal) effects can be estimated in a double robust manner, analogue to double robust estimation of effects in the i.i.d. causal inference literature. Previous work on the marginal distribution of counterfactual outcomes, such as the marginal distribution of the outcome at a particular time point under a certain intervention on one or more of the treatment nodes, cannot be estimated in a double robust manner [131, 137]. Instead, in this chapter, we propose a general class of averages of conditional (context-specific) causal parameters that can be estimated in a double robust manner. We propose a targeted maximum likelihood estimator (TMLE) of these causal parameters, and study the asymptotic theoretical properties of the TMLE. We demonstrate the favorable statistical properties of our estimator through various simulation studies. This work opens up robust statistical inference for causal questions based on observing a single time-series on a particular unit.

2.1 Introduction

Suppose one observes a single time series, where at each time t, one observes a data record $O(t)$ involving treatment nodes $A(t)$, an outcome node $Y(t)$, and possibly other covariates $L(t)$. In order to talk about causality, we assume that the time-ordering within $O(t)$ with

respect to the treatment nodes is known. In most of our examples, $A(t)$ is a single treatment node, but it could also be a vector of time-ordered treatment nodes, alternated with components of $L(t)$. We assume that the conditional distribution of $O(t)$ given the observed past is described by a common unknown function $(o(t), C_o(t)) \rightarrow \theta(o(t), C_o(t))$, that only depends on the past $O(1), \ldots, O(t-1)$ through a fixed dimensional summary measure $C_o(t)$. For example, one might assume that the conditional density of $O(t)$, given $O(1), \ldots, O(t-1)$, equals a conditional density $\theta(o(t) \mid C_o(t))$ for a common function θ, where this function is otherwise unspecified. More generally, we have that the conditional distribution $P_{\theta, C_o(t)}$ is determined by a common function θ.

The density of $O(t)$ is characterized by the conditional density of treatment nodes and conditional density of the outcome and covariate nodes. One might know, by design, that the conditional density of the treatment node is known (under control of the experimenter) while the other conditional densities are unknown. In that case, one would assume a common conditional density for the outcome and covariate nodes. This setup again describes a model for the distribution of the time-series, indexed by common (in time) conditional densities, analog to the standard conditional stationarity assumptions in time-series literature. For certain target parameters it is also necessary to assume a limited memory in the sense that $C_o(t)$ is only a function of a limited past $O(t-k), \ldots, O(t-1)$ for some fixed dimensional k. We are interested in models for the probability distribution of the time-series that refrain from making unrealistic parametric assumptions. In particular, we concentrate on models that only make a conditional stationarity assumption for relevant parts of the likelihood. Since the likelihood of the data is parameterized by a function $\theta \in \Theta$, one can consistently estimate this common function θ and thereby the probability distribution of the time-series. For example, one might use likelihood based estimation combined with online cross-validation, such as an online super learner [7]. We note that the standard maximum likelihood estimation would break down for infinite dimensional parameter spaces Θ, due to the curse of dimensionality.

While possible, our goal is not to estimate the whole mechanism θ, and thereby the whole density of the time-series. We are concerned with statistical inference about causal impacts of treatment on the outcome nodes, reflecting a certain part of the distribution. For example, one might want to know what the distribution of the outcome at time τ, $(Y(\tau))$, would have been had we intervened on some of the past treatment nodes in the time-series. These type of marginal parameters with the corresponding efficient influence curve and targeted maximum likelihood estimator were developed and proposed in the previous work [131, 137]. The asymptotic normality of these estimators relies on consistent estimation (e.g., at an appropriate rate faster than $\tau^{-1/4}$) of the part of θ the efficient influence curve depends upon. However, the efficient influence curve of the marginal time-series parameter relies on the whole mechanism θ in a non-double robust manner [131, 137]. Therefore, even for the situation where the treatment nodes were randomly assigned and known, the inference will still rely on consistent (at rate) estimation of the conditional distributions of the covariate and outcome nodes. This is in stark contrast to the independent and identically distributed case with nonparametric model for the common distribution, where the TMLE of such parameters would be completely robust against misspecification if the treatment mechanism

is known. The lack of robustness of the efficient influence function for the marginal time-series parameter is due to its dependence on the density of the marginal distribution of $C_o(t)$ across time t, a complex function of the common stationary mechanism $P_{O(t)|C_o(t),\theta}$. As such, estimation of the efficient influence curve of the marginal time-series parameter, and thereby the construction of a TMLE, is quite involved and computer intensive.

This raises the question if there are causal parameters of the time-series data distribution which are possibly easier to estimate efficiently, and which exhibit robust inference when the treatment mechanism is known. We provide a confirmatory answer in this work. Specifically, we propose a class of statistical target parameters $\Psi(\theta)$ defined as the average over time t of $C_o(t)$-specific pathwise differentiable target parameters $\Psi_{C_o(t)}(\theta)$ of the conditional distribution of $P_{\theta,C_o(t)}$. That is, for context $C_o(t)$, one defines a desired target parameter of the distribution of $O(t)$, given $C_o(t)$, as if we were able to observe many observations from this distribution. Interestingly and importantly, one could make the choice $\Psi_{C_o(t)}(\theta)$ of target parameter of the conditional distribution of $O(t)$ given $C_o(t)$ depend on the context $C_o(t)$, allowing one to adapt the choice of target parameter over time in response to $C_o(t)$.

We emphasize that statistical target parameters $\Psi(\theta)$ are data-dependent, since they are defined as an average over time of parameters of the conditional distribution of $O(t)$ given the observed realization of $C_o(t)$. As such, $\Psi(\theta)$ depends on the actual realization of the time-series, specifically $(C_o(1), \ldots, C_o(\tau))$. We also note that since the efficient influence function $D_{C_o(t)}(\theta)$ of each $C_o(t)$-specific target parameter is double robust, it follows that we can estimate the average of $C_o(t)$-specific causal effects in a double robust manner as well. In addition, the linear approximation of the TMLE is a martingale sum $\frac{1}{\tau}\sum_{t=1}^{\tau} D^{C_o(t)}(\theta)(O(t))$, allowing for the asymptotic normality of the TMLE to be established based on the martingale central limit theorem and general results for martingale processes.

2.2 Formulation of the Estimation Problem

Data and Likelihood

We model data under the shape of a random variable O, where the observed data represents a single copy of a longitudinal structure

$$(W(0), A(1), Y(1), W(1), \ldots, A(\tau), Y(\tau), W(\tau)), \tag{2.1}$$

which corresponds to observations from time $t = 0$ to the final $t = \tau$. Let $O(t)$ denote data at a specific time point t, such that we can write $O^\tau = \{O(t)\}_{t=0}^{t=\tau}$ for the full observed trajectory, emphasizing the final time point τ. Further, we define $O(0) = (W(0))$ with $A(0) = Y(0) = \emptyset$ as a vector of baseline covariates and $O(-1) = \emptyset$ by definition. The time-varying part of the single sample trajectory at t, $O(t)$ for $t > 0$, decomposes as $O(t) = (A(t), Y(t), W(t))$. The observed data collected at each time point is of a fixed dimension, and an element of an Euclidean set O with domain $\mathcal{O} := \mathcal{A} \times \mathcal{Y} \times \mathcal{W}$; there are no restrictions on the dimension and support of O. In the following, we elaborate on each component of $O(t)$. First, let $A(t)$

denote the time-varying binary exposure such that $A(t) \in \mathcal{A} := \{0,1\}$. The subsequent observed outcome corresponds to $Y(t) \in \mathcal{Y}$, which, without loss of generality, is assumed to be either a binary outcome in $\{0,1\}$ or a bounded continuous response taking values in $(0,1)$. We define $W(t)$ as as a vector of all post-outcome time-varying covariates lying in an Euclidean set $\mathcal{W}$. To put it in context: $A(t)$ might denote a suggestion to exercise at time t, with $Y(t)$ being an indicator of exercising in the next 30 minutes and $W(t)$ containing weather information at t. Finally, we denote $\bar{O}(t-1)$ as the t-specific history of the trajectory, such that $\bar{O}(t-1) = (O(0), O(1), \ldots, O(t-1))$. With that, $\bar{O}(t-1)$ contains all the observed history of the patient up until time t. Similarly, we define $\bar{A}(t-1) = (A(1), \ldots, A(t-1))$, $\bar{Y}(t-1) = (Y(1), \ldots, Y(t-1))$ and $\bar{W}(t-1) = (W(0), \ldots, W(t-1))$ as histories for A, Y and W processes until t.

Let P_0^τ denote the true probability distribution of O^τ such that $O^\tau \sim P_0^\tau$. Throughout the manuscript we use the naught subscript to indicate *true* probability distributions, or components thereof. Let p_0^τ denote the density of P_0^τ with respect to (w.r.t) a dominating measure μ over $\mathcal{O}$. In particular, we can write μ as a product measure $\mu = \times_{t=1}^{\tau}(\mu_A \times \mu_Y \times \mu_W)$ — where μ_A, μ_Y and μ_W are measures over $\mathcal{A}$, $\mathcal{Y}$ and $\mathcal{W}$, respectively. We denote realizations of a random variable O^τ as lower case letters, o^τ. We also define $P_{O(t)|\bar{O}(t-1)}$ as the conditional probability of $O(t)$ given the past until time t. In the rest of the manuscript, we will mostly be dealing with $P_{O(t)|\bar{O}(t-1)}$. The likelihood of realization o^τ of O^τ under the true data distribution P_0^τ can be factorized according to the time-ordering as follows:

$$\begin{aligned} p_0^\tau(o^\tau) = &\prod_{t=1}^{\tau} p_{0,a(t)}(a(t) \mid \overline{o}(t-1)) \times \prod_{t=1}^{\tau} p_{0,y(t)}(y(t) \mid a(t), \overline{o}(t-1)) \\ &\times \prod_{t=0}^{\tau} p_{0,w(t)}(w(t) \mid y(t), a(t), \overline{o}(t-1)) \\ = &\prod_{t=1}^{\tau} g_{0,a(t)}(a(t) \mid \overline{o}(t-1)) \times \prod_{t=1}^{\tau} q_{0,y(t)}(y(t) \mid a(t), \overline{o}(t-1)) \\ &\times \prod_{t=0}^{\tau} q_{0,w(t)}(w(t) \mid y(t), a(t), \overline{o}(t-1)), \end{aligned} \tag{2.2}$$

where $a(t) \mapsto p_{0,a(t)}(a(t) \mid \bar{o}(t-1)$, $y(t) \mapsto p_{0,y(t)}(y(t) \mid a(t), \bar{o}(t-1))$, and $w(t) \mapsto p_{0,w(t)}(w(t) \mid y(t), a(t), \bar{o}(t-1))$ are conditional densities w.r.t. the dominating measures μ_A, μ_Y, and μ_W. We use shorthand notation for conditional densities and distributions of the relevant nodes. In particular, we write $q_{0,y(t)}$ and $Q_{0,y(t)}$ as the true time t-specific conditional density and probability distribution of $Y(t)$ given the observed past until time t. In order to denote a conditional expectation of $Y(t)$ given the observed past, we write $\bar{Q}_{0,t}(A(t), \bar{O}(t-1)) = \mathbb{E}_0(Y(t) \mid A(t), \bar{O}(t-1))$. At time t, $g_{0,a(t)}$ reflects the true probability of $A(t)$ conditional on the past until time t, $\bar{O}(t-1)$. We write $Q_{0,w(t)}$ and $q_{0,w(t)}$ as the conditional distribution and density of $W(t)$ given the past $(Y(t), A(t), \bar{O}(t-1))$. To end the description of the data, here we emphasize that time-specific random variables $\{O(t)\}_{t=0}^{\tau}$ are not independent draws

from the same law: instead, they represent a dependent sequence, constituting a single draw from P_0^τ.

Statistical Model

We define a *statistical model* $\mathcal{M}$ for the probability distribution of the data P_0^τ, such that $P_0^\tau \sim \mathcal{M}$. The more we know (or are willing to assume) about the experiment that produces the data, the smaller the statistical model $\mathcal{M}$. For example, if the treatment is randomized with known randomization probabilities, then $\mathcal{M}$ should incorporate knowledge about the treatment mechanism. Referring back to the likelihood expression in (2.2), the decomposition presented places no restrictions on the type of time dependence possible. With the structure of dependence unknown, O^τ represents a time-indexed sequence of successive observations collected on a single patient — a *single time-series*. As such, the observed data reduces to a single draw from P_0^τ. In order to learn any relevant part of the data generating distribution, we have to put some restrictions on the statistical model $\mathcal{M}$.

We start by restricting the complexity of dependence allowed for the unknown time structure. In particular, we assume the conditional distribution of $O(t)$ given $\bar{O}(t-1)$, $P_{O(t)|\bar{O}(t-1)}$, depends on the observed past only through a fixed dimensional summary measure $C_o(t)$; we write $C_o(t) = C_o(\bar{O}(t-1)) \in C$ for some function C_o which takes $\bar{O}(t-1)$ as input. This is in contrast to assuming that $O(t)$ depends on the full observed history, $(O(0), O(1), \ldots, O(t-1))$. For some applications, the summary measure might cover a finite number of previous time points, analogue to a Markov order assumption. Then, $C_o(t)$ is a fixed dimensional extraction from the complete history, such that $C_o(t) = C_o(\bar{O}(t-1)) \in \mathbb{R}^k$ of the form $C_o(t) = \{O(s) : s = t-1, t-2, \cdots, t-k\}$. Alternatively, the fixed dimensional summary measure could encompass dependence structures described by summary measures of the time series pattern (e.g.: moving average, finite memory, STL decomposition, spectral entropy, Hurst coefficient), or it can depend on particular components on $\bar{O}(t-1)$ (e.g., only parts of the $\bar{W}(t-1)$ process). We impose no restrictions or assumptions on what $C_o(t)$ is, other than it must be fixed dimensional. As in the previous section, we write realizations of $C_o(t)$ as lower case letters.

We denote the conditional distribution $P_{O(t)|\bar{O}(t-1)}$ as $P_{C_o(t)}$ when making assumptions on the complexity of dependence allowed (short for $P_{O(t)|C_o(t)}$). Similarly, we define $p_{C_o(t)}$ as the conditional density $(o, C_o) \to p_{C_o(t)}(o \mid C_o)$ with respect to a dominating measure $\mu_{C_o(t)}$. As such, for each value of $C_o(t)$, we have that $\int p_{C_o(t)}(o \mid C_o(t)) d\mu_{C_o(t)}(o) = 1$. Therefore, for every $t \in [\tau]$, $O(t)$ is independent of its past conditional on $C_o(t)$. We can make similar assumptions on the complexity of dependence allowed for each component of the likelihood, instead of the whole $O(t)$. In particular, we let $q_{0,y(t)}$ denote the true conditional density of $Y(t)$ given a fixed dimensional summary $C_y(t) = C_y(A(t), \bar{O}(t-1))$, such that $q_{0,y(t)}(y(t) \mid a(t), \bar{o}(t-1)) = q_{0,y(t)}(y(t) \mid c_y(t))$. Similarly, we define $C_a(t)$ and $C_w(t)$ as fixed dimensional summaries derived as $C_a(t) = C_a(\bar{O}(t-1))$ and $C_w(t) = C_w(Y(t), A(t), \bar{O}(t-1))$ for the treatment and covariate processes, respectively. We define $C_y(t)$, $C_a(t)$ and $C_w(t)$ separately in order to emphasize that the conditioning set can be different for each node.

We write assumptions on fixed dimensional summary measures as Assumptions 1, 2 and 3, as stated below.

Assumption 1 (Decomposition of the fixed dimensional summary)**.** *For every $t \in [\tau]$, the fixed dimensional summary measure $C_y(t)$ can be written as*

$$C_y(t) = (A(t), C_o(t)),$$

where $C_o(t) = C_o(\bar{O}(t-1)) \in C$.

Assumption 2 (Conditional independence of $Y(t)$ given a summary measure)**.** *For every $t \in [\tau]$ and under Assumption 1,*

$$q_{0,y(t)}(y(t) \mid a(t), \bar{o}(t-1)) = q_{0,y(t)}(y(t) \mid c_y(t)) = q_{0,y(t)}(y(t) \mid a(t), c_o(t)).$$

Assumption 3 (Conditional independence of $A(t)$ given a summary measure)**.** *For every $t \in [\tau]$ we have that*

$$g_{0,a(t)}(a(t) \mid \bar{o}(t-1)) = g_{0,a(t)}(y(t) \mid c_a(t)) = g_{0,a(t)}(y(t) \mid c_o(t)).$$

Additionally, we assume that $p_{C_o(t)}$ is parameterized by a *common-in-time* function $\theta \in \Theta$, where $\theta : \mathcal{C} \times \mathcal{O} \to \mathbb{R}$. Therefore, we have that $(c, o) \to \theta(c, o)$, and $p_{C_o(t)}$ depends on θ only through $\theta(C_o(t), \cdot)$. In the following, we write interchangeably $p_{C_o(t)}$ and $p_{\theta, C_o(t)}$, depending on whether we want to emphasize dependence on θ in the subscript. As $p_{C_o(t)}$ factors into multiple conditional densities, we can parse this assumption by the components of the likelihood. Therefore, we assume the conditional density of $Y(t)$ given the observed fixed dimensional summary of the history is a constant function across time. As such, there exists a common conditional density $q_{0,y}$ such that $q_{0,y(t)} = q_{0,y}$. Similarly, we define $\bar{Q}_0$ as the common conditional expectation of $Y(t)$ given $C_y(t)$, such that $\bar{Q}_0(A(t), C_o(t)) = \int y q_{0,y}(y \mid A(t), C_o(t)) d\mu_y(o)$. In an observational study, where the treatment mechanism is not known, we also need to make conditional stationarity assumptions on $g_{0,a(t)}$. Therefore, unless in a randomized trial, we assume that $g_{0,a(t)} = g_{0,a} = g_0$ where $g_{o,a}$ is a common conditional probability of $A(t)$ given $C_o(t)$. We make no conditional stationarity assumptions on $q_{0,w(t)}$, allowing it to change over time; we elaborate on this point in the later sections, once the target parameter is defined. Other than assuming conditional stationarity given a fixed dimensional summary measure of the past, we assume no knowledge of the structural form of $q_{0,y}$ or g_0; both $q_{0,y}$ and g_0 could exhibit complex relationships between the outcome (treatment) and the fixed dimensional past. We write assumptions on $q_{0,y(t)}$ and $g_{0,a(t)}$ under the *common-in-time* model as Assumptions 4 and 5.

Assumption 4 (Common in t conditional density of outcome)**.** *There exists a common across time (t) conditional density of $Y(t)$ given the fixed dimensional summary measure $C_y(t)$, such that $q_{0,y(t)} = q_{0,y}$ for every $t \in [\tau]$.*

$$q_{0,y(t)}(y(t) \mid C_y(t)) = q_{0,y}(y(t) \mid C_y(t)).$$

Assumption 5 (Common in t conditional probability of treatment)**.** *There exists a common across time (t) conditional probability of $A(t)$ given the fixed dimensional summary measure $C_o(t)$, such that $g_{0,a(t)} = g_{0,a} = g_0$ for every $t \in [\tau]$.*

$$g_{0,a(t)}(a(t) \mid C_o(t)) = g_0(a(t) \mid C_o(t)).$$

With a slight abuse of notation, we write $\theta = (g, \bar{Q})$ and let $\Theta = \mathcal{G} \times \mathcal{Q}$ be the Cartesian product of two nonparametric spaces for g and $\bar{Q}$. As previously defined, we write $p_{\theta,C_o(t)}$ (short: $p_{C_o(t)}$) and p^τ_θ (short: p^τ) as the density for $O(t)$ given $C_o(t)$ and O^τ implied by θ. This defines a statistical model $\mathcal{M}^\tau = \{P^\tau_\theta : \theta\}$ for P^τ, the probability measure for the time-series. By construction, $\mathcal{M}^\tau$ is a statistical model for P^τ, and includes the true probability distribution P^τ_0. In addition, we define a statistical model $\mathcal{M}(C_o(t)) = \{P_{\theta,C_o(t)} : \theta\}$ as a model for the distribution of $O(t)$ given $C_o(t)$ at time t. The model $\mathcal{M}(C_o(t))$ includes $P_{0,C_o(t)}$ in its interior. We formally define statistical models $\mathcal{M}^\tau$ and $\mathcal{M}(C_o(t))$ in Definition 1 and 2. We can rewrite the likelihood presented in equation (2.2) under statistical model $\mathcal{M}^\tau$ as follows:

$$p^\tau_\theta(o^\tau) = \prod_{t=1}^{\tau} g_t(a(t) \mid C_a(t)) \prod_{t=1}^{\tau} q_y(y(t) \mid C_y(t)) \prod_{t=0}^{\tau} q_{w(t)}(w(t) \mid C_w(t)). \tag{2.3}$$

Definition 1 (Statistical Model $\mathcal{M}(C_o(t))$)**.** *We define a statistical model $\mathcal{M}(C_o(t))$ as the set of distributions $P_{\theta,C_o(t)}$ over the domain $\mathcal{O}$ that satisfy Assumptions 1, 2, 3, 4 and 5. In case of a randomized trial, we don't need Assumption 5 as the treatment mechanism is known.*

Definition 2 (Statistical Model $\mathcal{M}^\tau$)**.** *We define a statistical model $\mathcal{M}^\tau$ as the set of distributions P^τ_θ over the domain $\mathcal{O}$ that satisfy Assumptions 1, 2, 3, 4 and 5. In case of a randomized trial, we don't need Assumption 5 as the treatment mechanism is known.*

2.3 Target Parameter and Identification

Causal Target Parameter

Translation of the scientific question of interest into a causal parameter is facilitated by the use of a structural causal model (SCM; equivalently, structural equation model (SEM)) [93]. By specifying a SCM, we assume that each component of the data structure is a function of the observed time-specific data ("endogenous variables") and an unmeasured term ("exogenous variables") [93]. In the following, we denote all endogenous variables as O, and exogenous variables as U. By the SCM in Equation (2.4), we assume data structure at each time point t is a function of an observed, fixed-dimensional history and an unmeasured

exogenous term such that

$$\begin{aligned} W(0) &= f_{w(0)}(U_W(0)), \\ A(t) &= f_{a(t)}(C_o(\bar{O}(t-1)), U_a(t)), \\ Y(t) &= f_{y(t)}(C_y(A(t), \bar{O}(t-1)), U_y(t)), \\ W(t) &= f_{w(t)}(C_w(Y(t), A(t), \bar{O}(t-1)), U_w(t)), \end{aligned} \tag{2.4}$$

where $(f_a(t) : t = 1, \ldots, \tau)$, $(f_y(t) : t = 1, \ldots, \tau)$ and $(f_w(t) : t = 0, \ldots, \tau)$ are unspecified, deterministic functions. We denote a vector of exogenous variables as $U = (\{U(t)\}_{t=0}^{t=\tau}) = (\{(U_a(t), U_y(t), U_w(t))\}_{t=0}^{t=\tau})$ (note: $U(0) = U_w(0)$ with $U_a(0) = U_y(0) = \emptyset$), sampled from a probability distribution P_U. Given an input (O, U), structural equations $f_{a(t)}$, $f_{y(t)}$ and $f_{w(t)}$ for each time $t \in [\tau]$ deterministically assign a value to each of the nodes.

Let $\mathcal{M}^F$ define the *causal model*, which is a set of all probability distributions P^F over the domain of (O, U) that are compatible with the SCM defined in (2.4). We denote P_0^F as the true probability distribution of (O, U), where $P_0^F \in \mathcal{M}^F$. In order to refer to any distribution in $\mathcal{M}^F$, we write P^F. There is a clear connection between the full and observed data: any distribution P^F on the domain of the full data determines a corresponding distribution P on the domain of the observed data. The causal model $\mathcal{M}^F$ encodes all knowledge about the data-generating process, and implies a model for the distribution of the counterfactual random variables. As such, we conveniently define causal effects in terms of hypothetical interventions on the SCM. For instance, we can define a static intervention as $A(t) = 1$, which determnistically assigns treatment at time t. With that, $O^*(t)$ is the counterfactual full data generated from the SCM described in (2.4) by replacing the equation associated with the exposure node by the counterfactual intervention at time t,

$$\begin{aligned} A^*(t) &= 1, \\ Y^*(t) &= f_{y(t)}(C_y(1, \bar{O}(t-1)), U_y(t)), \\ W^*(t) &= f_{w(t)}(C_w(Y^*(t), 1, \bar{O}(t-1)), U_w(t)). \end{aligned}$$

We denote as $O^{*,\tau}$ the counterfactual random variable over all the times t where $O^{*,\tau} = (O^*(t) : t = 1, \ldots, \tau)$ and $(O^*(t) = (A^*(t), Y^*(t), W^*(t))$. With that, $O^{*,\tau}$ denotes a counterfactual time-series.

We also define $\mathcal{M}^F(C_o(t))$ as the *time- and context- specific causal model*. In contrast to $\mathcal{M}^F$, $\mathcal{M}^F(C_o(t))$ has the true conditional probability distribution $P^F_{0,C_o(t)}$ in its interior. In particular, $\mathcal{M}^F(C_o(t))$ contains all probability distributions compatible with the SCM in (2.4) over the domain of (O, U) where $C_a(t) = C_o(t) = c_o$:

$$\begin{aligned} W_{c_o}(0) &= f_{w(0)}(U_W(0)), \\ A_{c_o}(t) &= f_{a(t)}(c_o, U_a(t)), \\ Y_{c_o}(t) &= f_{y(t)}(C_y(A_{c_o}(t), c_o), U_y(t)), \\ W_{c_o}(t) &= f_{w(t)}(C_w(Y_{c_o}(t), A_{c_o}(t), c_o), U_w(t)). \end{aligned} \tag{2.5}$$

Let $O^*_{C_o(t)}(t)$ denote the counterfactual observation at time t, obtained by substituting input to $f_{a(t)}$ with a deterministic treatment for $A(t)$ for the SCM depicted in (2.5):

$$\begin{aligned} A^*_{c_o}(t) &= 1, \\ Y^*_{c_o}(t) &= f_{y(t)}(C_y(1, c_o), U_y(t)), \\ W^*_{c_o}(t) &= f_{w(t)}(C_w(Y^*_{c_o}(t), 1, c_o), U_w(t)). \end{aligned}$$

We write $(O^*_{C_o(t)}(t), U(t))$ as the full post-intervention data at time t, with the post-intervention distribution denoted as $P^{F*}_{C_o(t)}$. Consequently, $Y^*_{c_o}(t)$ then reflects the counterfactual outcome given $C_o(t) = c_o$ had the treatment been deterministically assigned at t, possibly contrary to the fact. We define our causal parameter of interest as

$$\Psi^F_{C_o(t)}(P^F_{C_o(t)}) := \mathbb{E}_{P^{F*}_{C_o(t)}}\left[Y^*_{c_o}(t)\right], \tag{2.6}$$

which is the expectation of the counterfactual random variable $Y^*_{c_o}(t)$ generated by the modified SCM as stated in equation (2.5). In words, our causal target parameter is the mean outcome we would have obtained after one time-step, if, starting at time t given the observed past, we had assigned treatment deterministically.

Identification and the Statistical Target Parameter

We lay the groundwork for addressing identifiability through P^τ_0 by providing a link between the causal model and the observed data. As a first step, we define the causal quantity of interest in terms of a deterministic intervention on the SCM, as outlined in (2.5). Further, we rely on the G-computation formula under the sequential randomization and positivity assumptions to identify the distribution of the time- and context- specific observation $O^*_{C_o(t)}(t)$, as well as the full counterfactual time-series $O^{*,\tau}$ [104, 106]. The two key Assumptions, sequential randomization and positivity, are stated below as Assumption 6 and 7. We note that, in case treatment mechanism is randomized at each t, Assumption 7 is satisfied by design.

Assumption 6 (Sequential Randomization)**.** *For any $t \in [\tau]$, we have that*

$$A(t) \perp\!\!\!\perp Y^*_{c_o}(t) \mid C_o(t) \quad \textit{and} \quad A^*_{c_o}(t) \perp\!\!\!\perp Y^*_{c_o}(t) \mid C_o(t).$$

Assumption 7 (Positivity)**.** *Under the treatment mechanism $g_{0,t}$, each treatment value $a \in \{0, 1\}$ has a positive probability of being assigned given the observed history. For every $t \in [\tau]$ with $P_0(C_o(t) = c_o) > 0$,*

$$g_{0,t}(A(t) \mid C_o(t) = c_o) > 0.$$

Under identification Assumptions 6 and 7, we can write the causal target parameter $\Psi^F_{C_o(t)}(P^F_{C_o(t)})$ defined in Equation (2.6) as a function of the true conditional data-generating distribution $P_{0,C_o(t)}$ where

$$\Psi^F_{C_o(t)}(P^F_{0,C_o(t)}) = \Psi_{C_o(t)}(P_{0,C_o(t)}) = \Psi_{C_o(t)}(\theta) := \mathbb{E}_{P_{0,C_o(t)}}[Y(t) \mid A(t) = 1, C_o(t)]. \tag{2.7}$$

In words, we define the target parameter as the causal effect (under Assumptions 6 and 7) of assigning treatment at t on the subsequent outcome $Y(t)$ in the context of the observed history until the current time point. Therefore, for a given observed summary $C_o(t)$, we define the target mapping $\Psi_{C_o(t)} : \mathcal{M}(C_o(t)) \to \mathbb{R}$, where $\Psi_{C_o(t)}$ is pathwise differentiable with a canonical gradient $D_{C_o(t)}(P_{C_o(t)})(o)$ at $P_{C_o(t)}$ in $\mathcal{M}(C_o(t))$. The variance of the canonical gradient (also known as the efficient influence function, EIC) gives the generalized Cramer-Rao lower bond for the variance of any regular and asymptotically linear estimator, providing a way to build efficient estimators [127, 138, 137]. As stated in Section 2.2, we can write the canonical gradient at $P_{C_o(t)}$ in $\mathcal{M}(C_o(t))$ interchangeably as $D_{C_o(t)}(P_{C_o(t)})(o)$ and $D_{C_o(t)}(\theta)(o)$, emphasizing dependence on $\theta = (g, \bar{Q})$. By the definition of a canonical gradient, $D_{C_o(t)}(\theta)$ is a function of the observed data o with conditional mean zero w.r.t. $P_{C_o(t)}$ [127]. In the following theorem we provide the exact form of the canonical gradient corresponding to the target parameter in Equation (2.7), along with its first order expansion and the double-robust second order term.

Theorem 2 (Time- and Context-Specific Target Parameter)**.** *We define the time- and context-specific parameter $\Psi_{C_0(t)}(P_{0,C_o(t)})$ under statistical model $\mathcal{M}(C_o(t))$ as*

$$\Psi_{C_o(t)}(P_{0,C_o(t)}) := \mathbb{E}_{P_{0,C_o(t)}}[Y(t) \mid A(t) = 1, C_o(t)],$$

or equivalently for $\theta = (g, \bar{Q})$ as

$$\Psi_{C_o(t)}(\theta_0) = \Psi_{C_o(t)}(g_0, \bar{Q}_0) := \mathbb{E}_{P_{0,C_o(t)}}[Y(t) \mid A(t) = 1, C_o(t)].$$

Under Assumption 7, the target parameter mapping $\Psi_{C_0(t)} : \mathcal{M}(C_o(t)) \to \mathbb{R}$ is pathwise differentiable w.r.t. $\mathcal{M}(C_o(t))$ and has a canonical gradient defined as

$$D_{C_o(t)}(\theta)(o) = \frac{1(A(t) = 1)}{g(A(t) = 1 \mid C_o(t))} \left(Y(t) - \bar{Q}(A(t) = 1, C_o(t))\right).$$

The time- and context-specific parameter admits the following first order expansion

$$\Psi_{C_o(t)}(\theta) - \Psi_{C_o(t)}(\theta_0) = -\mathbb{E}_{P_{0,C_o(t)}}[D_{C_o(t)}(\theta)] + R_{C_o(t)}(\bar{Q}, \bar{Q}_0, g, g_0),$$

where $R_{C_o(t)}$ is a second order remainder that is doubly-robust, with $R_{C_o(t)}(\bar{Q}, \bar{Q}_0, g, g_0) = 0$ if either $\bar{Q} = \bar{Q}_0$ or $g = g_0$.

Instead of the treatment specific mean (TSM) described in Equation (2.7) adapted to the observed time and context, one could instead focus on the time- and context-specific average treatment effect (ATE) — another common parameter of interest in causal inference literature. By an easy extension, we define a parameter mapping $\Psi^{\text{ATE}}_{C_0(t)} : \mathcal{M}(C_o(t)) \to \mathbb{R}$ with a canonical gradient $D^{\text{ATE}}_{C_o(t)}(\theta)(o)$ at $P_{C_o(t)}$ in $\mathcal{M}(C_o(t))$. In particular, we define $\Psi^{\text{ATE}}_{C_0(t)}(P_{0,C_o(t)}) = \Psi^{\text{ATE}}_{C_0(t)}(\theta_0)$ as

$$\Psi^{\text{ATE}}_{C_0(t)}(P_{0,C_o(t)}) := \mathbb{E}_{P_{0,C_o(t)}}[Y(t) \mid A(t) = 1, C_o(t)] - \mathbb{E}_{P_{0,C_o(t)}}[Y(t) \mid A(t) = 0, C_o(t)],$$

with canonical gradient defined as

$$D^{\text{ATE}}_{C_o(t)}(\theta)(o) = \left[\frac{1(A(t)=1)}{g(A(t)=1 \mid C_o(t))} - \frac{1(A(t)=0)}{g(A(t)=0 \mid C_o(t))}\right]\left(Y(t) - \bar{Q}(A(t)=1, C_o(t))\right).$$

Then, the time- and context-specific ATE is interpreted as the causal difference at t between assigning treatment ($A(t)=1$) and control ($A(t)=0$) on subsequent outcome $Y(t)$ in the context of the observed history until the current time point. As the theoretical analysis of target parameters $\Psi_{C_o(t)}(P_{0,C_o(t)})$ and $\Psi^{\text{ATE}}_{C_0(t)}(P_{0,C_o(t)})$ is the same, in the following we focus on the parameter described in Equation (2.7).

The targets defined so far are parameters defined for a specific time point t intervention given the observed past. However, we can also formulate target parameters as summaries of single time-point interventions over time. In particular, we propose a class of statistical target parameters $\Psi(\theta)$ defined as the counterfactual average of single time-point and context-specific averages over time. As such, we define a target parameter mapping $\Psi : \mathcal{M}^\tau \to \mathbb{R}$ of the data distribution $P^\tau \in \mathcal{M}^\tau$ written as

$$\begin{aligned}\Psi(\theta_0) &:= \frac{1}{\tau}\sum_{t=1}^{\tau} \Psi_{C_o(t)}(P_{0,C_o(t)}) \\ &= \frac{1}{\tau}\sum_{t=1}^{\tau} \mathbb{E}_{P_{0,C_o(t)}}[Y(t) \mid A(t)=1, C_o(t)].\end{aligned} \tag{2.8}$$

The average over time of $C_o(t)$-specific targets is a *data-dependent parameter*, as its value depends on the realization of the fixed dimensional summary measure. As such, $\Psi(\theta_0)$ reflects an average over all possible contexts observed over time. In the running example introduced in Section 2.2, if $C_o(t)$ were low dimensional weather summaries over a short history ("sunny", "rain", "cloudy"), then $\Psi(\theta_0)$ represents the average exercise 30 minutes after an intervention over all observed weather conditions so far. As $\Psi(\theta_0)$ is an average over time of $C_o(t)$-specific target parameters, its canonical gradient is also an average of $C_o(t)$-specific EIFs. Therefore, we have that $D(\theta)(o) = \frac{1}{\tau}\sum_{t=1}^{\tau} D_{C_o(t)}(\theta)(o)$, and $\Psi(\theta_0)$ can be estimated in a double robust matter as well. We reiterate the canonical gradients and first order expansion corresponding to the target parameter $\Psi(\theta_0)$ in $\mathcal{M}^\tau$ in Theorem (3) below.

Theorem 3 (Average over time Target Parameter)**.** *We define the average over time parameter under statistical model $\mathcal{M}^\tau$ as*

$$\Psi(\theta_0) := \frac{1}{\tau}\sum_{t=1}^{\tau} \mathbb{E}_{P_{0,C_o(t)}}[Y(t) \mid A(t)=1, C_o(t)].$$

Under Assumption 7, the target parameter mapping $\Psi : \mathcal{M}^\tau \to \mathbb{R}$ is pathwise differentiable w.r.t. $\mathcal{M}^\tau$ and has a canonical gradient defined as

$$D(\theta)(o) = \frac{1}{\tau}\sum_{t=1}^{\tau}\left[\frac{1(A(t)=1)}{g(A(t)=1 \mid C_o(t))}\left(Y(t) - \bar{Q}(A(t)=1, C_o(t))\right)\right].$$

The average over time target parameter admits the following first order expansion

$$\Psi(\theta) - \Psi(\theta_0) = -\frac{1}{\tau}\sum_{t=1}^{\tau} \mathbb{E}_{P_{0,C_o(t)}}[D_{C_o(t)}(\theta)] + \frac{1}{\tau}\sum_{t=1}^{\tau} R_{C_o(t)}(\theta, \theta_0).$$

2.4 Estimation Procedure

Targeted Maximum Likelihood Estimator

We build efficient estimators of the proposed estimands using the Targeted Maximum Likelihood methodology [139, 138, 137]. As a plug-in estimator, the Targeted Maximum Likelihood Estimator (TMLE) doesn't generate estimates that fall outside of their natural range — as may be the case with one-step and estimating equation approaches [138, 137]. First, we define the parameter of interest as a smooth functional Ψ evaluated at a law P_0^τ; our estimand, with its corresponding canonical gradient, is defined in Theorem 3. Next, we generate a possibly highly data-adaptive initial estimator $\theta_\tau = (g_\tau, \bar{Q}_\tau)$ of $\theta_0 = (g_0, \bar{Q}_0)$, using ensemble machine learning for dependent data ("Super Learner" or "Online Super Learner") [136, 7, 79]. In order to generate an initial estimate, we define $L(\theta_0)$ as a valid loss function for θ_0. Specifically, a valid loss function for a given parameter is defined as a function whose true conditional mean is minimized by the true value of the parameter. Let L be a loss function adapted to the problem, i.e. a function that maps every θ to $L(\theta) : (O(t), C_o(t)) \mapsto L(\theta)(O(t), C_o(t))$. As $\theta_0 = (g_0, \bar{Q}_0)$, we let $L(\bar{Q})$ and $L(g)$ denote loss functions for $\bar{Q}_0$ and g_0, respectively. In the rest of the section we focus on loss-based estimation of $\bar{Q}_0$, as the TMLE of the target parameter in Equation (2.8) requires an update of the initial estimate of $\bar{Q}_0$. We define a parameter mapping $L(\bar{Q}) : \mathcal{O} \times \mathcal{C} \to \mathbb{R}$ such that for all t,

$$\mathbb{E}_{P_{0,C_o(t)}} L(\bar{Q}_0) = \underset{\bar{Q} \in \mathcal{Q}}{\operatorname{argmin}}\, \mathbb{E}_{P_{0,C_o(t)}} L(\bar{Q}).$$

Therefore, the true $\bar{Q}_0$ minimizes risk under the true conditional probability distribution $P_{0,C_o(t)}$. For instance, we could define $L(\bar{Q})$ as the negative log-likelihood loss, written as

$$L(\bar{Q})(O(t), C_o(t)) = -\log \bar{Q}(A(t), C_o(t)).$$

or the mean squared error

$$L(\bar{Q})(O(t), C_o(t)) = (Y(t) - \bar{Q}(A(t), C_o(t)))^2.$$

Next, we define a *least favorable submodel* through the initial estimate $\bar{Q}_\tau$ whose score spans the canonical gradient at $\bar{Q}_\tau$ of the derivative of Ψ. In particular, we define a parametric working model $\{\bar{Q}_{\tau,\epsilon} : \epsilon\}$ through $\bar{Q}_\tau$ with a finite dimensional parameter ϵ; note that at $\epsilon = 0$, $\bar{Q}_{\tau,\epsilon} = \bar{Q}_\tau$. Intuitively, we construct a path starting at the initial estimator going in the direction defined by the canonical gradient. As the canonical gradient is in the tangent

space and therefore a score, it represents a valid direction for a path. Moreover, we have that the linear combination of the components of the derivative of the loss at $\epsilon = 0$ span the canonical gradient at the initial estimator

$$\left\langle \frac{d}{d\epsilon} L(\bar{Q}_{\tau,\epsilon}) \Big|_{\epsilon=0} \right\rangle \supset D(\bar{Q}_\tau),$$

as a path in the direction of the score. Here, we use the notation $\langle S \rangle$ for the linear span of the components of the function S. Alternatively, we could define the *universally least favorable submodel*: here, we define the path at all points along it, not just at the beginning [134]. For the universally least-favorable submodel, the derivative of the loss evaluated at any ϵ equals the canonical gradient at the fluctuated initial estimator $\bar{Q}_{\tau,\epsilon}$, such that

$$\frac{d}{d\epsilon} L(\bar{Q}_{\tau,\epsilon}) = D(\bar{Q}_{\tau,\epsilon}).$$

As the fluctuation $\{\bar{Q}_{\tau,\epsilon} : \epsilon\}$ through $\bar{Q}_\tau$ reflects a parametric model space with a single finite dimensional parameter ϵ, we can compute the minimum loss estimator (MLE) of ϵ as

$$\epsilon_\tau = \underset{\epsilon}{\operatorname{argmin}} \sum_{t=1}^{\tau} L(\bar{Q}_{\tau,\epsilon})(O(t), C_o(t)).$$

Maximizing the log-likelihood over the fluctuation model (or minimizing any valid loss L) defines an updated estimator $\bar{Q}_{\tau,\epsilon_\tau}$. In particular, we update the initial estimate using the logistic fluctuation model

$$\operatorname{logit}(\bar{Q}_{\tau,\epsilon}) = \operatorname{logit}(\bar{Q}_\tau) + \epsilon \frac{1(A(t) = 1)}{g(A(t) = 1 \mid C_o(t))},$$

where ϵ is a coefficient in front of the "clever covariate", which is specific to the target parameter. Under the local least favorable submodel, the process might need to be iterated; at each iteration, we replace the previous plug-in estimate and its canonical gradient with the result of the previous iteration. The final update at which $\epsilon_\tau \approx 0$ is denoted as $\bar{Q}^*_\tau = \bar{Q}_{\tau,\epsilon_\tau}$. We define TMLE as the plug-in estimator obtained at the last update of the estimator of $\bar{Q}_0$, $\bar{Q}^*_\tau$. By construction, TMLE solves the canonical gradient estimating equation as follows

$$\sum_{t=1}^{\tau} D_{C_o(t)}(\bar{Q}^*_\tau)(O(t)) \approx 0.$$

Under the universally least favorable model, we have that the score equation of the MLE of ϵ where $\bar{Q}^*_\tau = \bar{Q}_{\tau,\epsilon_n}$ immediately yields

$$\sum_{t=1}^{\tau} D_{C_o(t)}(\bar{Q}^*_\tau)(O(t)) = 0.$$

Highly Adaptive Lasso as Initial Estimator

For ATE parameters in observational data, a sufficient condition for nonparametric efficiency of the TMLE estimator is that g_τ and $\bar{Q}_\tau$ (combined, θ_τ) converge to the truth (θ_0) in root-mean-square error at a rate of $o_p(\tau^{1/4})$ each [127] As such, statistical inference in $\mathcal{M}^\tau$ for estimating θ_0 relies on fast-converging algorithms in a large space; we elaborate on this point in the next section. The Highly Adaptive Lasso (HAL) is a nonparametric regression estimator that does not rely on local smoothness assumptions, in contrast to most machine learning algorithms [6, 130]. Instead, HAL assumes the true function is right-hand continuous with left-hand limits (cadlag) and a bounded variation norm. In essence, HAL restricts the behavior of the true function globally, instead of locally where (1) cadlag functions are very general, even allowing discontinuities; (2) the variation norm can be made arbitrarily large [112]. As stated in [6], functions with infinite variation norm tend to be pathological, with $\cos(1/x)$ as an example. To the best of our knowledge, HAL is the only algorithm with fast-enough convergence rates to allow for efficient inference in nonparametric statistical models [130, 112]. In the following, we generalize HAL to our problem setting, and show its convergence rate in an online setup.

Let $L(\theta)(O(t), C_o(t))$ denote a valid loss function for θ evaluated at $(O(t), C_o(t))$, where the true $\theta \in \Theta$, θ_0, is the minimizer of the risk w.r.t. $P_{0,C_o(t)}$ such that

$$\theta_0 = \underset{\theta \in \Theta}{\operatorname{argmin}} \frac{1}{\tau} \sum_{t=1}^{\tau} \mathbb{E}_{P_{0,C_o(t)}} L(\theta).$$

Suppose that Θ is contained in a class of multivariate (e.g., d-variate) real valued cadlag functions $D[0, \upsilon]$ on a cube $[0, \upsilon] \subset \mathbb{R}^d_{\geq 0}$; we denote such a class as the "HAL" class or $\mathcal{H}$ in all further sections. By setup, any arbitrary function ϕ in $D[0, \upsilon]$ is right-continuous with left-hand limits and a sectional variation norm bounded by a universal constant. Instead of the usual definition of the variation norm where $\|\phi\|_v \equiv \int_{[0,\upsilon]} |\phi(du)|$, we write it as a sum of the variation norm over sections of the truth, ϕ_0 [6]. As such, we define the uniform sectional variation norm of a multivariate real valued cadlag function ϕ as

$$\|\phi\|_v \equiv \phi(0) + \sum_{s \subset \{1,\ldots,d\}} \int_{0_s}^{\upsilon_s} |\phi_s(du_s)|,$$

where the sum is taken over all subsets s of $\{1, \ldots, d\}$. For any function $\phi \in D[0, \upsilon]$ and subset s, we define $u_s = (u_i : i \in s)$, $u_{-s} = (u_i : i \notin s)$, with $\phi_s(u_s) \equiv \phi(u_s, 0_{-s})$. Then, $\phi_s(u_s)$ varies along the variables u_s according to ϕ, but sets variables in u_{-s} to zero. Consequently, any function for which $\|\phi\|_v < \infty$ can be represented as $\phi(0) + \sum_{s \subset \{1,\ldots,d\}} \int_{0_s}^{\upsilon_s} \phi_s(du_s)$ [45]. Therefore, a class of functions to which ϕ belongs to is a convex hull of the indicator class; by [126], we know that a convex hull of a Donseker class is a Donsker class as well. This will prove very important as we derive our theoretical results.

Let $\{L(\theta) : \theta \in \Theta\}$ be contained in a class of d-variate real-valued cadlag functions on a cube $[0, \upsilon]$ bounded by a universal constant $M^u < \infty$, or the $\mathcal{H}$ class. Further, let

$\Theta(M) = \{\theta \in \Theta : \|\theta\|_v \leq M\}$ denote a constrained subset of Θ, with a sectional variation norm smaller or equal to the known upper bound M^u. We then define the M-specific MLE as

$$\theta_{M,\tau} = \underset{\theta \in \Theta(M)}{\operatorname{argmin}} \frac{1}{\tau} \sum_{t=1}^{\tau} L(\theta)(O(t), C_o(t)),$$

where the true M-specific parameter $\theta_{M,0}$ is written as

$$\theta_{M,0} = \underset{\theta \in \Theta(M)}{\operatorname{argmin}} \frac{1}{\tau} \sum_{t=1}^{\tau} \mathbb{E}_{P_{0,C_o(t)}} L(\theta).$$

Consequently, we also define a loss-based dissimililarity measure between θ and θ_0 implied by the loss function $L(\theta)$ as

$$d_0(\theta, \theta_0) = \frac{1}{\tau} \sum_{t=1}^{\tau} \mathbb{E}_{P_{0,C_o(t)}} (L(\theta) - L(\theta_0)).$$

For quadratic loss-based dissimilarity, we also write $d_0(\theta, \theta_0) = \|\theta - \theta_0\|^2_{P_{0,C_o(t)}}$. The following Theorem 4 establishes the rate of convergence of the MLE w.r.t. the loss-based dissimilarity $d_0(\theta, \theta_0)$ under Assumptions 8 and 9. In particular, it establishes that the MLE $\theta_{M,\tau}$ converges to its M-specific truth $\theta_{M,0}$ at a rate no slower than $\tau^{-1/2}$, regardless of the dimension of the time-series.

Assumption 8. *Let $L(\theta)$ denote a valid loss function for θ, where $L(\theta)$ is a multivariate cadlag function whose sectional variation norm can be bounded by the sectional variation norm of θ in a sense that*

$$\sup_{\theta \in \Theta} \|L(\theta)\|_v / \|\theta\|_v < \infty.$$

Assumption 9. *Let $d_0(\theta_{M,\tau}, \theta_{M,0})$ denote the dissimilarity measure corresponding to loss $L(\theta)$ for $\theta \in \Theta(M)$. If $d_0(\theta_{M,\tau}, \theta_{M,0}) \to_p 0$, then*

$$\frac{1}{\tau} \sum_{t=1}^{\tau} \mathbb{E}_{P_{0,C_o(t)}} [L(\theta_{M,\tau}) - L(\theta_{M,0})]^2 \to_p 0.$$

Theorem 4 (Minimum loss-based estimator in the HAL class)**.** *Let $L(\theta)(O(t), C_o(t))$ denote a valid loss function for θ evaluated at $(O(t), C_o(t))$, with $d_0(\theta_\tau, \theta_0)$ being the loss-based dissimilarity measure for $L(\theta_\tau)(O(t), C_o(t))$ defined as*

$$d_0(\theta_\tau, \theta_0) = \frac{1}{\tau} \sum_{t=1}^{\tau} \mathbb{E}_{P_{0,C_o(t)}} (L(\theta_\tau) - L(\theta_0)).$$

Let $\Theta(M)$ define a set of cadlag functions with variation norm smaller or equal to M, such that $\Theta(M) = \{\theta \in \Theta : \|\theta\|_v \leq M\}$. Under Assumptions (8) and (9), we have that

$$d_0(\theta_{M,\tau}, \theta_{M,0}) = o_p(\tau^{-1/2}).$$

2.5 Theoretical Analysis

By Theorem 2, we have that the time- and context- specific parameter admits the following von Mises expansion

$$\begin{aligned}\Psi_{C_o(t)}(P_{\theta,C_o(t)}) - \Psi_{C_o(t)}(P_{\theta_0,C_o(t)}) &= \Psi_{C_o(t)}(\theta) - \Psi_{C_o(t)}(\theta_0) \\ &= -\mathbb{E}_{P_{\theta_0,C_o(t)}}[D_{C_o(t)}(\theta)] + R_{C_o(t)}(\theta,\theta_0),\end{aligned} \tag{2.9}$$

which is a natural consequence of the pathwise diffentiability of $\Psi_{C_o(t)}$. As von Mises expansion is an approximation, we have that the second order difference between $\Psi_{C_o(t)}(P_{\theta,C_o(t)})$ and $\Psi_{C_o(t)}(P_{\theta_0,C_o(t)})$ for a given $C_o(t)$ corresponds to

$$R_{C_o(t)}(\theta,\theta_0) \equiv \Psi_{C_o(t)}(\theta) - \Psi_{C_o(t)}(\theta_0) + \mathbb{E}_{P_{\theta_0,C_o(t)}}[D_{C_o(t)}(\theta)],$$

which is discussed further in following Section 2.5. As noted in Theorem 3, the same expansion and second order remainder definition follows for the target parameter defined in Equation (2.8), but averaged over time.

In the following, at times, it proves useful to use notation from empirical process theory; specifically, we define Pf to be the empirical average of the function f w.r.t. the distribution P, that is, $Pf = \int f(o)dP(o)$. Further, we ease notation by defining a centered martingale process $(M_\tau(f) : f)$ for a function f in a class of multivariate real valued functions of $(O, C) \in \mathcal{O} \times \mathcal{C}$ as

$$M_\tau(f) = \frac{1}{\tau}\sum_{t=1}^{\tau}[f(C_o(t), O(t)) - P_{0,C_o(t)}f],$$

where for all f, $\tau M_\tau(f)$ is a discrete martingale. For a centered martingale difference process, we use the following notation

$$M_\tau(f,h) = \frac{1}{\tau}\sum_{t=1}^{\tau}(\delta_{C_o(t),O(t)} - P_{0,C_o(t)})[f(C_o(t), O(t)) - h(C_o(t), O(t))]$$

with function h in a class of multivariate real valued functions of $(O, C) \in \mathcal{O} \times \mathcal{C}$.

In general, weak convergence of $(\sqrt{\tau}M_\tau(f) : f)$ to a Gaussian process is equivalent with convergence of all finite dimensional distributions and an asymptotic equicontinuity condition. As such, our theoretical analysis relies on the fact that the difference between the TML estimator and the estimand can be decomposed as a sum of (1) an average of martingale difference sequence, (2) a martingale process for which we can show an equicontinuity result, (3) second order remainder. We present formally this decomposition in Theorem 5.

Theorem 5. *Let θ^*_τ be the one-step TMLE satisfying $\frac{1}{\tau}\sum_{t=1}^{\tau} D_{C_o(t)}(\theta^*_\tau)(O(t)) = 0$ for $\theta^*_\tau = (g_\tau, \bar{Q}^*_\tau)$. Further, we define $\theta_l = (g_0, \bar{Q}_l)$ as the limit of $\theta^*_\tau = (g_\tau, \bar{Q}^*_\tau)$ for $\theta_l \in \Theta$, and $\theta_0 = (g_0, \bar{Q}_0)$ as the truth. Then the difference between the TMLE and its estimand decomposes*

as

$$\Psi(\theta_\tau^*) - \Psi(\theta_0) = \underbrace{M_{1,\tau}(\theta_l)}_{\textit{Term 1}} + \underbrace{M_{2,\tau}(\theta_\tau^*, \theta_l)}_{\textit{Term 2}} + \underbrace{\frac{1}{\tau}\sum_{t=1}^{\tau} R_{C_o(t)}(\theta_\tau^*, \theta_0)}_{\textit{Term 3}},$$

with

$$\begin{aligned}
M_{1,\tau}(\theta_l) =& \frac{1}{\tau}\sum_{t=1}^{\tau}[D_{C_o(t)}(\theta_l)(O(t)) - \mathbb{E}_{P_{0,C_o(t)}} D_{C_o(t)}(\theta_l)],\\
M_{2,\tau}(\theta_\tau^*, \theta_l) =& \frac{1}{\tau}\sum_{t=1}^{\tau}[D_{C_o(t)}(\theta_\tau^*)(O(t)) - \mathbb{E}_{P_{0,C_o(t)}} D_{C_o(t)}(\theta_\tau^*)]\\
&- \frac{1}{\tau}\sum_{t=1}^{\tau}[D_{C_o(t)}(\theta_l)(O(t)) + \mathbb{E}_{P_{0,C_o(t)}} D_{C_o(t)}(\theta_l)]\\
&= \frac{1}{\tau}\sum_{t=1}^{\tau}(\delta_{C_o(t),O(t)} - P_{0,C_o(t)})(D_{C_o(t)}(\theta_\tau^*) - D_{C_o(t)}(\theta_l)).
\end{aligned}$$

We allocate the proof of Theorem 5 to the Appendix. In the following, we study each term from the decomposition presented in Theorem 5 separately.

Analysis of Term 1: $M_{1,\tau}$

The first term in the decomposition in Theorem 5, $M_{1,\tau}(\theta_l)$, is an average of a martingale difference sequence. Therefore, $M_{1,\tau}(\theta_l)$ can immediately be analyzed with the classical martingale central limit theorem. A set of sufficient conditions for the asymptotic normality of the first term in Theorem 5 includes (1) strong positivity, assuring $D_{C_o(t)}(\theta_l)(O(t))$ remains bounded; (2) stabilization of the conditional variance, assuring the variance converges in distribution as $t \to \infty$. By Theorem 6 stated below, we have that $\sqrt{\tau}M_{1,\tau}(\theta_l)$ converges to a normal distribution.

Assumption 10 (Strong positivity)**.** *There exists* $\delta > 0$ *such that*

$$g_0(A(t) \mid C_o(t)) \geq \delta, P_0\text{-a.s.}$$

Assumption 11 (Stabilization of the mean of conditional variances)**.** *There exists* $\sigma_l^2 \in (0, \infty)$ *such that*

$$\frac{1}{\tau}\sum_{t=1}^{\tau} \mathbb{E}_{P_{0,C_o(t)}} \left(D_{C_o(t)}(\theta_l)(O(t)) \mid C_o(t)\right)^2 \xrightarrow{d} \sigma_l^2.$$

Theorem 6. *Suppose that Assumption* (10) *and* (11) *hold. Then*

$$\sqrt{\tau}M_{1,\tau}(\theta_l) \xrightarrow{d} \mathcal{N}(0, \sigma_l^2). \tag{2.10}$$

Proof. The result follows directly from the martingale central limit theorem (e.g. Theorem 2 in [17]). □

Analysis of Term 2: $M_{2,\tau}$

The second term in the decomposition in Theorem 5 is a martingale process indexed by $\theta \in \Theta$, and evaluated at $\theta = \theta^*_\tau$. The analysis of $M_{2,\tau}$ entails showing asymptotic equicontinuity/tightness under a complexity condition for a process derived from a function class $\{D_{C_o(t)}(\theta) : \theta \in \Theta\}$, which implies that if $\frac{1}{\tau}\sum_{t=1}^{\tau} \mathbb{E}_{P_{0,C_o(t)}}[D_{C_o(t)}(\theta^*_\tau) - D_{C_o(t)}(\theta_l)]^2 \to_p 0$, then $M_{2,\tau}(\theta^*_\tau, \theta_l) = o_P(1/\sqrt{\tau})$. In particular, we analyze the martingale process $\{M_{2,\tau}(\theta, \theta_l) : \theta \in \Theta\}$ under a measure of complexity introduced by [128], and denoted as *sequential bracketing entropy* in [78]. In the following, we state the definition of the sequential bracketing entropy adopted to the setting studied, and provide main results with a brief discussion of the empirical process term analysis.

Definition 3 (Sequential bracketing entropy)**.** *Consider a stochastic process of the form $\Xi_\tau := \{(\xi_t(f))_{t=1}^{\tau} : f \in \mathcal{F}\}$ where $\mathcal{F}$ is an index set such that, for every $f \in \mathcal{F}$, $t \in [\tau]$, $\xi_t(f)$ is an $\bar{O}(t)$-measurable real valued random variable. We say that a collection of random variables of the form $\mathcal{B} := \{(\Lambda_t^j, \Upsilon_t^j)_{t=1}^{\tau} : j \in [J]\}$ is an $(\epsilon, b, \bar{O}(\tau))$ bracketing of Ξ_τ if*

1. *for every $t \in [\tau]$, and $j \in [J]$, $(\Lambda_t^j, \Upsilon_t^j)$ is $\bar{O}(t)$-measurable,*
2. *for every $f \in \mathcal{F}$ there exists $j \in [J]$ such that $\forall t \in [J]$, $\Lambda_t^j \leq \xi_t(f) \leq \Upsilon_t^j$,*
3. *for every $t \in [\tau]$, $j \in [J]$, $|\Lambda_t^j - \Upsilon_t^j| \leq b$ a.s.,*
4. *for every $j \in [J]$,*

$$\frac{1}{\tau}\sum_{t=1}^{\tau} \mathbb{E}\left[(\Upsilon_t^j - \Lambda_t^j)^2 \mid \bar{O}(t-1)\right] \leq \epsilon^2.$$

We denote $\mathcal{N}_{[]}(\epsilon, b, \Xi_\tau, \bar{O}(\tau))$ as the minimal cardinality of an $(\epsilon, b, \Xi_\tau, \bar{O}(\tau))$-bracketing.

Using Definition (3), we can see that $\{M_{2,\tau}(\theta, \theta_l) : \theta \in \Theta\}$ is derived from the process

$$\Xi_\tau := \left\{\left(D_{C_o(t)}(\theta)(O(t)) - D_{C_o(t)}(\theta_l)(O(t))\right)_{t=1}^{\tau} : \theta \in \Theta\right\}, \tag{2.11}$$

where $\mathcal{N}_{[]}(\epsilon, b, \Xi_\tau, \bar{O}(\tau))$ is the sequential bracketing number of the canonical gradient process Ξ_τ, corresponding to brackets of size ϵ. For results on how to formalize the connection between the sequential bracketing entropy of the process Ξ_τ to a traditional bracketing entropy measure, we refer the interested reader to the Appendix Section 8.3 of [78]. In particular, [78] show how to characterize the sequential bracketing entropy of Ξ_τ in terms of the bracketing entropy w.r.t. the norm $L_2(P_{\theta_0,h_\tau})$, where h_τ is the empirical measure defined as $h_\tau = \frac{1}{\tau}\sum_{t=1}^{\tau} \delta_{C_o(t)}$. One specific function class for which we know how to bound the latter, is the "HAL" class (denoted $\mathcal{H}$) discussed in subsection 2.4. The equicontinuity result presented in Lemma 2 is the sequential equivalent of its i.i.d. counterpart studied in [126], relying on a sequential Donsker-like condition.

Assumption 12 (Sequential Donsker condition)**.** *Define the sequential bracketing entropy integral as* $J_{[]}(\epsilon, b, \Xi_\tau, \bar{O}(\tau)) := \int_0^\epsilon \sqrt{\log(1 + \mathcal{N}_{[]}(u, b, \Xi_\tau, \bar{O}(\tau))}du$*. Suppose that there exists a function* $a : \mathbb{R}^+ \to \mathbb{R}^+$ *that converges to* 0 *as* $\delta \to 0$*, such that*

$$J_{[]}(\epsilon, b, \Xi_\tau, \bar{O}(\tau)) \leq a(\delta).$$

Assumption 13 (L_2 convergence)**.** *It holds that* $\|\theta^*_\tau - \theta_l\|_{2,h_\tau} = o_P(1)$*, where* h_τ *is the empirical measure* $h_\tau := \frac{1}{\tau}\sum_{t=1}^{\tau} \delta_{C_o(t)}$*.*

Lemma 2 (Equicontinuity of the martingale process)**.** *Consider the process* Ξ_τ *defined in equation* (2.11)*. Under Assumptions* (10)*,* (12) *and* (13)*, we have that*

$$M_{2,\tau}(\theta^*_\tau, \theta_l) = o_P(\tau^{-1/2}).$$

Proof. The proof is a direct application of Theorem 4 in [78] to the current problem setting. □

Analysis of Term 3: $R_{C_o(t)}$

Finally, we discuss term 3 of the decomposition presented in Theorem 5. The last term is an average over time of the second order remainder given $C_o(t)$, corresponding to

$$R_{C_o(t)}(\theta, \theta_0) \equiv \Psi_{C_o(t)}(\theta) - \Psi_{C_o(t)}(\theta_0) + \mathbb{E}_{P_{\theta_0, C_o(t)}}[D_{C_o(t)}(\theta)]$$

for any $\theta \in \Theta$. Unlike the empirical process term (term 2), the second order remainder is specific to the target parameter and the statistical model. Given the above expression for $R_{C_o(t)}(\theta, \theta_0)$ and the EIF, applying Cauchy-Schwarz we can show that

$$\begin{aligned} R_{C_o(t)}(\theta, \theta_0) &= \frac{g(1 \mid C_o(t)) - g_0(1 \mid C_o(t))}{g(1 \mid C_o(t))} \left[\bar{Q}(1, C_o(t)) - \bar{Q}_0(1, C_o(t))\right] \\ &\leq \|g - g_0\| \|\bar{Q} - \bar{Q}_0\|, \end{aligned}$$

elucidating the double robust nature of the second order remainder, akin to the well studied literature on TSM parameters in observational studies and i.i.d. settings [138, 137].

While studying theoretical properties of estimators, a typical condition on the second order remainder is that it converges to zero — therefore assuming it is negligible in large samples [138, 137]. In our problem setup, we assume that the remainder term can be represented by a sum of a martingale and a second order term that is $o_P(1/\sqrt{\tau})$. In the following Assumption (14), we state the exact form of asymptotic linearity of the second order remainder term necessary for further asymptotic normality of the TMLE.

Assumption 14 (Negligible Second Order Remainder)**.** *Let the second order remainder be written as* $R_{C_o(t)}(\theta, \theta_0)$*, where* $\theta = (g, \bar{Q})$*. We further define* $\theta_l = (g_0, \bar{Q}_l)$ *as the limit of* $\theta^*_\tau = (g_\tau, \bar{Q}^*_\tau)$ *for* $\theta_l \in \Theta$*, and* $\theta_0 = (g_0, \bar{Q}_0)$ *as the truth. We set the following*

$$\frac{1}{\tau}\sum_{t=1}^{\tau} \frac{g_\tau - g_0}{g_\tau}(\bar{Q}^*_\tau - \bar{Q}_l) = o_P(1/\sqrt{\tau}),$$
$$\frac{1}{\tau}\sum_{t=1}^{\tau} \frac{(g_\tau - g_0)^2}{g_\tau g_0}(\bar{Q}_l - \bar{Q}_0) = o_P(1/\sqrt{\tau}).$$

Additionally, we assume that following function of $(g_\tau, g_0, \bar{Q}_l, \bar{Q}_0)$ *can be represented by a martingale sum for some function* f *such that*

$$\frac{1}{\tau}\sum_{t=1}^{\tau} \frac{g_\tau - g_0}{g_0}(\bar{Q}_l - \bar{Q}_0) = \frac{1}{\tau}\sum_{t=1}^{\tau} f(C_o(t))(A(t) - g_0(A(t) \mid C_o(t)) + o_P(1/\sqrt{\tau}).$$

Note that, if g_τ is an MLE according to a parametric model, then the martingale approximation in Assumption (14) is true under weak regularity conditions. We also emphasize that, if $\bar{Q}_\tau$ is consistent for $\bar{Q}_0$, we only need to assume that

$$\frac{1}{\tau}\sum_{t=1}^{\tau} \frac{g_\tau - g_0}{g_\tau}(\bar{Q}^*_\tau - \bar{Q}_0) = o_P(1/\sqrt{\tau}).$$

Asymptotic Normality of the TMLE

Finally, as an immediate consequence of the analysis presented in Section 2.5 (Theorem 6 with assumptions), Section 2.5 (Lemma 2 with assumptions) and Section 2.5 (Assumption (14)), we have the following asymptotic normality result for the TMLE of the target parameter in Equation (2.8). We allocate the proof to the Appendix Section 2.9.

Theorem 7 (Asymptotic normality of the TMLE)**.** *Let* θ^*_τ *be the one-step TMLE satisfying* $\frac{1}{\tau}\sum_{t=1}^{\tau} D_{C_o(t)}(\theta^*_\tau)(O(t)) = 0$ *for* $\theta^*_\tau = (g_\tau, \bar{Q}^*_\tau)$*. Further, we define* $\theta_l = (g_0, \bar{Q}_l)$ *as the limit of* $\theta^*_\tau = (g_\tau, \bar{Q}^*_\tau)$ *for* $\theta_l \in \Theta$*, and* $\theta_0 = (g_0, \bar{Q}_0)$ *as the truth. Under Assumptions* (10)*,* (11)*,* (12)*,* (13) *and* (14) *we have that*

$$\sqrt{\tau}\left(\Psi(\theta^*_\tau) - \Psi(\theta_0)\right) \xrightarrow{d} \mathcal{N}(0, \sigma_0^2),$$

where $\bar{f}_t = D_{C_o(t)}(\theta_l)(O(t)) + f(C_o(t))(A(t) - g_0(1 \mid C_o(t))$ *and* $\frac{1}{\tau}\sum_\tau \mathbb{E}_{P_{0,C_o(t)}} \bar{f}_t \xrightarrow{d} \sigma_0^2$.

2.6 Simulations

In the following, we present simulation results demonstrating theoretical properties of the estimator proposed in a single time-series setting. In particular, we focus on the average

over time of $C_o(t)$-specific causal effects of a single time-point intervention on the subsequent outcome. In the following simulations we consider binary outcome and treatment, but note that the results will be comparable for continuous outcome. Unless specified otherwise, all results are generated based on 500 Monte Carlo draws used to evaluate the performance of the TMLE estimator of the average over time context-specific causal effect of a single time intervention.

Simulation 1a (simple dependence)

We explore a scenario with binary treatment ($A(t) \in \{0,1\}$) and outcome ($Y(t) \in \{0,1\}$) first, with simple dependence extending to Markov order 2. We observe covariates $W_1(t)$, $W_2(t)$ and $W_3(t)$ for each $t = 1, \ldots, \tau$, with $W_1(t)$ and $W_3(t)$ drawn from a bernoulli distribution and $W_2(t)$ from a discrete uniform distribution. We note that for this scenario, $W(t) = (W_1(t), W_2(t))$ are drawn independently with respect to the observed past $\bar{O}(t)$. Further, let the treatment variable $A(t)$ be a function of the past up until $t-2$ and depend on $W_1(t-1), W_2(t-1), Y(t-1), A(t-1)$ and $W_3(t-2)$. The outcome variable $Y(t)$ exhibits dependence of order 2, as a function of $W_1(t-1), W_2(t-1), W_3(t-1), A(t), W_1(t-2)$ and $W_3(t-2)$. For notational convenience, we define $O(1:t)$ as $(O(1), \ldots O(t))$. The exact data-generating distribution used is as follows:

$$
\begin{aligned}
A(0:4) &\sim Bern(0.5) \\
Y(0:4) &\sim Bern(0.5) \\
W_1(0:4) &\sim Bern(0.5) \\
W_2(0:4) &\sim Unif(1,3) \\
W_3(0:4) &\sim Bern(0.5) \\
A(4:\tau) &\sim Bern(expit(0.25 * W_1(t-1) - 0.2 * W_2(t-1) \\
&\qquad + 0.3 * Y(t-1) - 0.2 * A(t-1) \\
&\qquad + 0.2 * W_3(t-2))) \\
Y(4:t) &\sim Bern(expit(0.3 - 0.8 * W_1(t-1) \\
&\qquad + 0.1 * W_2(t-1) + 0.2 * W_3(t-1) \\
&\qquad + A(t) - 0.5 * W_1(t-2) \\
&\qquad + 0.2 * W_3(t-2) \\
W_1(4:\tau) &\sim Bern(0.5) \\
W_2(4:\tau) &\sim Unif(1,3) \\
W_3(4:\tau) &\sim Bern(0.5).
\end{aligned}
$$

The initial estimates g_τ and $\bar{Q}_\tau$ were obtained using the online version of the Super-Learner algorithm [7]. In particular, our initial ensemble consisted of multiple algorithms, including simple generalized linear models, penalized regressions and extreme gradient boosting [27]. For cross-validation, we relied on the online cross-validation scheme, also known as the recursive scheme in the time-series literature. We report Wald-type confidence intervals, with

asymptotic variance based on the EIF. We report the coverage of the resulting asymptotic 95% confidence intervals to evaluate the performance of the proposed method in Table 2.6.

Simulation 1b (more elaborate dependence)

Next, we explore the setting where the single time-series exhibits a more elaborate dependence. Effectively, we are decreasing the sample size and therefore testing the performance of our estimator for different finite sample settings, including the most extreme case of $\tau = 100$. In addition, we consider each part of $O(t)$ to exhibit different levels of dependence, including all the covariates in $W(t)$. For this particular simulation, we treat $A(t)$ as randomized, with a simulation mimicking an observational study considered in Simulation 1c.

$$
\begin{aligned}
A(0:7) &\sim Bern(0.5) \\
Y(0:7) &\sim Bern(0.5) \\
W_1(0:7) &\sim Bern(0.5) \\
W_2(0:7) &\sim Normal(0,1) \\
A(7:\tau) &\sim Bern(0.5) \\
Y(7:\tau) &\sim Bern(expit(1.5 * A(t) - A(t-1) \\
&\qquad + 0.5 * Y(t-1) - 1.1 * W_1(t-1) \\
&\qquad + 0.7 * Y(t-3) - A(t-5) + W_1(t-7)) \\
W_1(7:\tau) &\sim Bern(expit(0.5 * W_1(t-1) - 0.5 * Y(t-1) + 0.1 * W_2(t-1)) \\
W_2(7:\tau) &\sim Normal(0.6 * A(t-1) + Y(t-1) - W_1(t-1), sd = 1).
\end{aligned}
$$

Simulation 1c (Observational study, more elaborate functions and dependence)

Finally, we consider a typical observational study with varying level of dependence and variable interactions. In particular, Simulation 1c considers a setting where each part of the likelihood exhibits some level of dependence, including all of the covariates grouped in $W(t)$. As in Simulation 1a and 1b, we keep τ at constant levels $\tau = (100, 500, 1000)$, and report performance of our estimator for very low effective sample size ($\tau = 100$). We include the highly adaptive lasso (HAL) as part of our Super Learner library, in addition to several glms, penalized regressions and extreme gradient boosting. In addition, we test the double robustness property of our estimator for all sample sizes considered. The exact

data-generating distribution used is as follows:

$$
\begin{aligned}
A(0:6) &\sim Bern(0.5) \\
Y(0:6) &\sim Bern(0.5) \\
W_1(0:6) &\sim Bern(0.5) \\
W_2(0:6) &\sim Normal(0,1) \\
A(6:\tau) &\sim Bern(expit(0.7 * W_1(t-2) - 0.3 * A(t-1) \\
&\qquad + 0.2 * sin(W_2(t-2) * A(t-3)) \\
Y(6:\tau) &\sim Bern(expit(1.5 * A(t) - (W_1(t-1) * A(t-2))^2 \\
&\qquad + 0.9 * sin(W_2(t-4)) * A(t-3) * cos(W_2(t-6)) \\
&\qquad - abs(W_2(t-5) > 0)) \\
W_1(6:\tau) &\sim Bern(expit(0.5 * W_1(t-1) - 0.5 * Y(t-1) + 0.1 * W_2(t-1)) \\
W_2(6:\tau) &\sim Normal(0.6 * A(t-1) + Y(t-1) - W_1(t-1), sd = 1).
\end{aligned}
$$

	τ	**Bias**	**Variance**	**Coverage**
Single time-point intervention (1a)	1000	-2.37e-3	9.02e-4	94.8
	500	2.02e-3	1.71e-3	96.2
	100	5.02e-3	1.02e-2	92.0
Single time-point intervention (1b)	1000	-7.09e-4	7.58e-4	94.0
	500	1.16e-2	2.07e-3	89.6
	100	1.73e-2	1.30e-2	77.4
Single time-point intervention (1c)	1000	4.79e-3	9.45e-4	91.2
	500	7.52e-3	1.92e-3	93.8
	100	3.71e-3	1.25e-2	81.8

Table 2.1: Bias, variance and 95% coverage of the TMLE of the average over time context-specific causal effects with a single time-point intervention for Simulations 1a, 1b and 1c at sample sizes $\tau = 1000$, $\tau = 500$ and $\tau = 100$, over 500 Monte Carlo draws.

	τ	**Bias**	**Variance**	**Coverage**
Qmgc	1000	1.43e-2	1.26e-3	88.4
Qcgm	1000	1.42e-2	1.25e-3	88.4
Qmgc	500	1.29e-2	2.63e-3	89.2
Qcgm	500	1.30e-2	2.62e-3	89.4
Qmgc	100	3.68e-2	1.47e-2	84.4
Qcgm	100	-2.62e-2	9.78e-3	85.8

Table 2.2: Illustration of the double robustness property of our estimator for Simulation 1c with misspecified (m) and correctly specified (c) models for g_τ and $\bar{Q}_\tau$ at sample sizes $\tau = (1000, 500, 100)$ over 500 Monte Carlo draws.

2.7 Data Analysis

The Insulin Dependent Diabetes Mellitus (IDDM), also known as type 1 diabetes, is characterized by pancreatic beta cell dysfunction and insulin depletion [125]. The results of the Diabetes Control and Complications Trial, as well as the Epidemiology of Diabetes Interventions and Complications follow-up study, demonstrated that most people with IDDM need to be treated intensively to achieve hemoglobin A_{1c} and post-meal blood glucose (BG)

levels close to normal (hemoglobin: $< 7.0\%$; blood glucose: $80-140$ mg/d) [89]. Continuous glucose monitoring, multiple injections of long-acting and short-acting insulin, along with healthy diet and exercise are just few of the common simultaneous interventions prescribed to IDDM patients. For insulin treatments in particular, error in dosing can lead to potentially life-threatening hyperglycemia (blood glucose: > 200 mg/dl) and hypoglycemia events (blood glucose: $40-80$ mg/dl, with < 40 mg/d including neuroglycopenic symptoms) [125]. Long lasting hyperglycemia increases the risk of reinopathy, neuropathy and nephropathy, and is overall associated with poor long-term outcomes [89].

The blood glucose concentration is highly variable, and will vary even in individuals with normal pancreatic hormonal function. While the gold standard for pre-meal blood glucose is $80-120$ mg/dl with $80-140$ mg/dl for post-meal, these ranges are highly controversial and subject to individual variability in diabetes mellitus. Insulin injections are the primary treatment for patients with IDDM, and work by increasing the uptake of glucose in many of the tissues. Typically a patient with IDDM needs to administer insulin injections multiple times per day, usually at a regular schedule (e.g., before a meal, at bedtime). Each insulin formulation has its own characteristic time of onset of effect, time of peak action, and effective duration that needs to be taken into account while administrating a dose. In addition to the type of insulin formulation, other typically prescribed interventions for IDDM can provide important context for insulin dosing and blood glucose management. For example, regular exercise in the mid-afternoon can be associated with low blood glucose levels after dinner; on the other hand, more than usual strenuous exercise can lead to transient increase in BG levels. Similarly, diet and meal proportions can provide context for insulin dosing. In particular, a larger than usual meal can lead to longer and possibly higher BL levels. On the other hand, smaller than usual, or completely missed meal, can result in higher risk of low blood glucose in the hours that follow. All of it combined, relevant context (e.g., exercising, ingestion, age) as well as individual blood glucose variability are important components in deciding if an insulin dose is necessary, as well as if it might lead to a hyperglycemic or hypoglycemic episode.

In this Section, we analyze the diabetes dataset from the AAAI Spring Symposium on Interpreting Clinical Data available as part of the UCI Machine Learning Repository [32]. The dataset contains time-ordered records of patients with IDDM, where each record corresponds to a time interval around meal time and sleep (breakfast, lunch, dinner, irregular meal, before going to bed). A total of 70 patients were monitored on a regular basis, with the follow up in the $32-616$ time point range (shortest follow up: 11/01/89-11/06/89; longest follow: 4/29/90 to 12/16/90). Collected data at each time point includes BG level, insulin dose and type (regular insulin, Neutral Protamine Hagedorn (NPH), or UltraLente), meal ingestion (typical, more-than-usual or less-than-usual) and exercise activity (typical, more-than-usual or less-than-usual). In the analysis, we code insulin injection as a binary variable, instead of use the insulin dose information. We consider two outcomes, Y_{hyper} as a hyperglycemic and Y_{hypo} for a hypoglycemic episode. The primary intervention was insulin administration at time t. Context corresponding to $C_o(t)$ for insulin injection treatment at t consisted of chronological time, time of the day (morning, midday, afternoon, night),

pre-meal measurement, post-meal measurement, ingestion and exercise status. We analyzed the top 12 time-series with the most data available (number of time points > 350), with the basic characteristics of each presented in the Table below.

						Blood glucose level mg/d			
ID	**Days**	τ	$A_{\text{insulin}}\%$	$Y_{\text{hyper}}\%$	$Y_{\text{hypo}}\%$	**Min**	**Mean**	**Max**	**SD**
29	232	616	75	17	4	40	161	327	45
55	149	606	92	21	8	33	158	421	66
30	151	591	75	16	7	28	153	319	51
65	137	545	73	55	5	15	208	400	77
68	288	527	31	2	8	44	122	361	33
56	125	499	76	38	18	30	177	450	95
28	170	474	74	14	12	32	142	340	54
67	120	465	73	51	7	41	209	487	92
27	271	455	74	5	17	43	127	280	45
20	135	451	91	37	17	28	173	463	93
1	135	369	79	29	14	35	159	343	69
54	103	362	73	23	24	35	140	300	70

Table 2.3: Descriptive and summary statistics for each sample used in the data analysis. The descriptive information consists of the patient id, number of days monitored, number of time points collected, percent time administering insulin during observation ($A_{\text{insulin}}\%$), percent time having an hyperglycemic ($Y_{\text{hyper}}\%$) or hypoglycemic ($Y_{\text{hypo}}\%$) episode per sample. Summary statistics of blood glucose level in mg/d includes the minimum, maximum, average and standard deviation over all available time points.

For each sample, we estimate $\bar{Q}_0$ and g_0 across collected time points using the Online Super Learner available as part of the SL3 software package [27]. The Super Learner library for both $\bar{Q}_0$ and g_0 consisted of: highly adaptive lasso (HAL9001), gradient boosted decision trees (XGBOOST), random forests (RANGER), elastic net generalized linear models (GLMNET), Bayesian generalized linear models (BAYESGLM), and main terms linear regression (GLM) [28, 23, 147, 41, 43]. For cross-validation, we split the data into a minimum of 5 folds by implementing the Rolling Origin CV scheme as implemented in the ORIGAMI software package [26, 25, 79]. We set the first window to $t = 200$ points for all sample with more than 450 collected time points, and $t = 150$ for the remaining 3 samples. The validation size was $t = 50$ for all samples, with a batch of size $t = 50$ and no gap. For $C_o(t)$ we consider the last 5 time periods, corresponding to about a day of observed history; the considered context includes time, time of the day, pre-/post- meal indicator, more-/less- ingestion, more-/less-activity and previous insulin and hyperglycemic/hypoglycemic episodes.

The TMLE and its 95% confidence interval for each of the 12 analyzed samples with hyperglycemic episode (Y_{hyper}) and hypoglycemic event (Y_{hypo}) are shown in Figure 2.1A and Figure 2.1B, respectively. Based on Figure 2.1, we can roughly rank how insulin dependent each sample is — the bigger the difference between regular insulin and no insulin, the more dependent the patient is. Based on this premise, samples most reactive to the change in insulin regime are patients 67, 65, 54, 30 and 20; on the other hand, sample 56 is not, indicating perhaps a need for a higher dose or version of insulin treatment (due to its high rate of hyperglycemis and hypoglycemic events). It's interesting to point out that samples with the highest proportion of less-then-usual ingestion and more-than-usual activity result in the smallest ATE (e.g., Sample 55). Most samples (e.g., 68,67,65,56,30,29,28) would have dramatically more hyperglycemic events on a less insulin-intensive regime, in particular Sample 30 and 28. On the other hand, even more insulin intensive regime results in a (less extreme) increase in hypoglycemic events, highlighting the tricky nature of insulin dependent treatments. Finally, we note that Sample 20 would actually benefit from a different intensive insulin regime — the trajectory of Sample 20 indicates that it's hypoglycemic events are largely caused by an intensive insulin treatment.

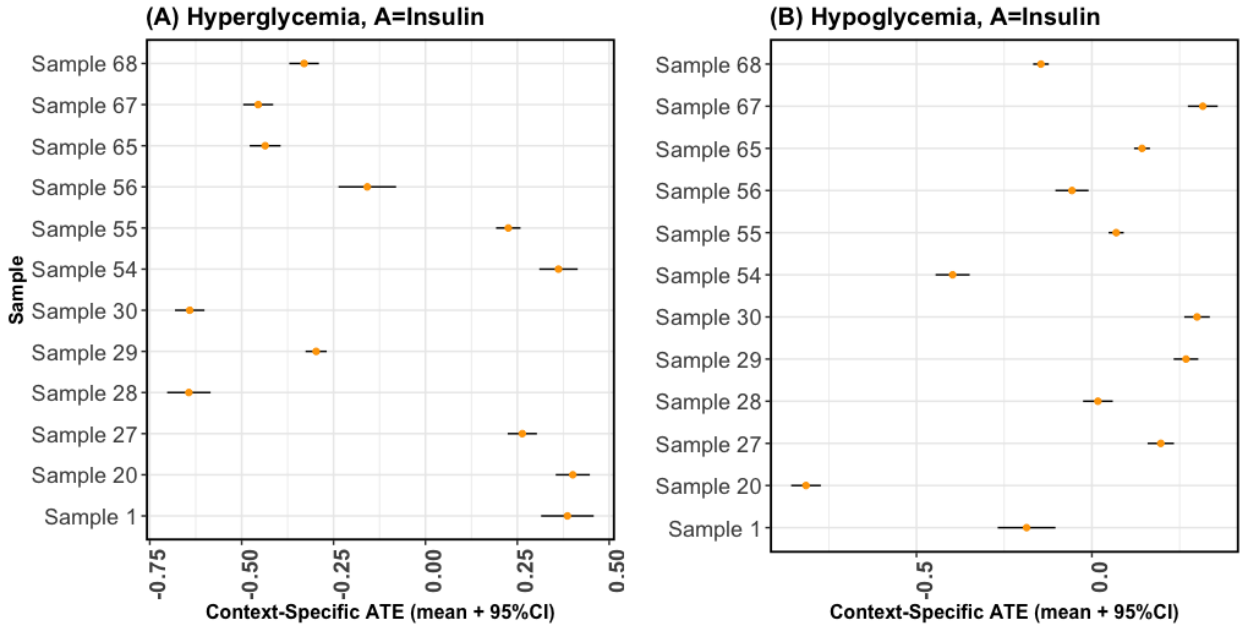


Figure 2.1: Context-Specific ATE with its 95% confidence interval for the 12 samples considered for the analysis. Panel (A) shows the mean and 95% confidence interval for Y_{hyper}, which is the outcome variable corresponding to a hyperglycemic episode. Panel (B) plots the mean and 95% confidence interval for each sample where the outcome is a hypoglycemic event, Y_{hypo}.

2.8 Discussion

In this chapter, we consider causal inference based on observing a single time series with asymptotic results derived over time t. The data setup constitutes a typical longitudinal data structure, where within each t-specific time-block one observes treatment and outcome nodes, and possibly time-dependent covariates in-between treatment nodes. Each t-specific data record $O(t)$ is viewed as its own experiment in the context of the observed history $C_o(t)$, carrying information about a causal effect of the treatment nodes on the outcome node. A key assumption necessary in order to obtain the presented results is that the relevant history for generating $O(t)$, given the past $\bar{O}(t-1)$, can be summarized by a fixed dimensional summary $C_o(t)$. We note that our conditions allow for $C_o(t)$ to be a function of the whole observed past, allowing us to avoid Markov-order type assumptions that limit dependence on recent past.

Due to the dimension reduction assumption, each t-specific experiment in the sequence of experiments corresponds with drawing from a conditional distribution of $O(t)$, given $C_o(t)$. We assume that this conditional distribution is either constant in time or is parametrized by a constant function. We concentrate on the first setting, as it covers all the applications presented in this chapter, but note the flexibility of our assumptions. Due to the conditional stationarity assumption, we can asymptotically learn the true mechanism that generates this time-series, even when the model for the mechanism is nonparametric. However, with the exception of parametric models allowing for maximum likelihood estimation, we emphasize that statistical inference for certain target parameters of the data generating mechanism is a challenging problem which requires targeted machine learning.

In previous work, [131] derive the TMLE for marginal causal parameters, which marginalize over the distribution of $C_o(t)$. For instance, one might be interested in the counterfactual mean of a future (e.g., long term) outcome under a stochastic intervention on a subset of the treatment nodes. This specific parameter addresses the important question regarding the distribution of the outcome at time t, had we intervened on some of the past treatment nodes in the time-series. While important, the TMLE of such target parameters are challenging to implement due to their reliance on the density estimation of the marginal density of $C_o(t)$ (averaged across t). Additionally, we remark that such marginal causal parameters cannot be robustly estimated if treatment is sequentially randomized, due to lack of double robustness of the second order remainder.

In this chapter, we instead focus on context-specific (or conditional) target parameter is order to explore robust statistical inference for causal questions based on observing a single time series on a particular unit. In particular, we note that for each given $C_o(t)$, any intervention-specific mean outcome $EY_{g^*}(t)$ with g^* being a stochastic intervention w.r.t. the conditional distribution of $P_{C_o(t)}$ represents a well studied statistical estimation problem based on observing τ i.i.d. copies. Due to this insight and formulation we are able to re-purpose known efficient influence curves and corresponding double robust second order expansions from the i.i.d. literature. Even though we do not have repeated observations from the $C_o(t)$-specific distribution at time t, due to the conditional stationarity assumption, the

collection $(C_o(t), O(t))$ across all time points represent the analogue of an i.i.d. data set $(C_o(t), O(t)) \sim_{iid} P_0$, where $C_o(t)$ can be viewed as a baseline covariate in this typical longitudinal causal inference data structure. Therefore, we estimate the sample-specific counterfactual mean (e.g., sample average treatment effect) $\frac{1}{\tau}\sum_{t=1}^{\tau} E(Y_{g^*}(t) \mid C_o(t))$ using the TMLE of EY_{g^*} developed for i.i.d. data. We note however that the initial estimation step of the TMLE should still respect the known dependence in construction of the initial estimator, by relying on appropriate estimation techniques developed for dependent data. In particular, we emphasize the importance of time-series based cross-validation schemes (rolling, recursive, fixed and hybrid, to name a few) instead of usual V-fold cross-validation commonly employed for i.i.d settings. Similarly, variance estimation can proceed as in the i.i.d case using the relevant i.i.d. efficient influence curve, while ignoring the component corresponding to the baseline covariate $C_o(t)$. This insight relies on the fact that the TMLE in this case allows for the same linear approximation as the TMLE for i.i.d. data, with the martingale central limit theorem applied to the linear approximation instead. Since the linear expansion of the time-series TMLE for context-specific parameter is an element of the tangent space of the statistical model, our derived TMLE is asymptotically efficient.

2.9 Appendix

Theorem 3: Minimum loss-based estimator in a class of cadlag functions with finite variation norm

Theorem 3 *Let $L(\theta)(O(t), C_o(t))$ denote a valid loss function for θ evaluated at $(O(t), C_o(t))$, with $d_0(\theta_\tau, \theta_0)$ being the loss-based dissimilarity measure for $L(\theta_\tau)(O(t), C_o(t))$ defined as*

$$d_0(\theta_\tau, \theta_0) = \frac{1}{\tau} \sum_{t=1}^{\tau} \mathbb{E}_{P_{0,C_o(t)}}(L(\theta_\tau) - L(\theta_0)).$$

Let $\Theta(M)$ define a set of cadlag functions with variation norm smaller or equal to M, such that $\Theta(M) = \{\theta \in \Theta : \|\theta\|_v \leq M\}$. Under Assumptions (8) and (9), we have that

$$d_0(\theta_{M,\tau}, \theta_{M,0}) = o_p(\tau^{-1/2}).$$

Proof. Let $M_\tau(\theta)$ define a martingale process indexed by a class of multivariate, real-valued cadlag functions with a uniform bound on the sectional variation norm. In particular, we define $M_\tau(\theta)$ as

$$M_\tau(\theta) = \frac{1}{\tau} \sum_{t=1}^{\tau} [L(\theta) - \mathbb{E}_{P_{0,C_o(t)}} L(\theta)],$$

which is a centered martingale of $L(\theta)$ process for any $\theta \in \Theta$. Then, we have that

$$\begin{aligned}
0 &\overset{(1)}{\leq} d_0(\theta_{M,\tau}, \theta_{M,0}) \\
&\overset{(2)}{=} \frac{1}{\tau} \sum_{t=1}^{\tau} \mathbb{E}_{P_{0,C_o(t)}}(L(\theta_{M,\tau}) - L(\theta_{M,0})) \\
&\overset{(3)}{=} \frac{1}{\tau} \sum_{t=1}^{\tau} \mathbb{E}_{P_{0,C_o(t)}}(L(\theta_{M,\tau}) - L(\theta_{M,0})) \\
&\quad - \frac{1}{\tau} \sum_{t=1}^{\tau} [L(\theta_{M,\tau}) - L(\theta_{M,0})] + \frac{1}{\tau} \sum_{t=1}^{\tau} [L(\theta_{M,\tau}) - L(\theta_{M,0})] \\
&\overset{(4)}{\leq} \frac{1}{\tau} \sum_{t=1}^{\tau} \mathbb{E}_{P_{0,C_o(t)}}(L(\theta_{M,\tau}) - L(\theta_{M,0})) - \frac{1}{\tau} \sum_{t=1}^{\tau} [L(\theta_{M,\tau}) - L(\theta_{M,0})] \\
&\overset{(5)}{=} -\frac{1}{\tau} \sum_{t=1}^{\tau} [L(\theta_{M,\tau}) - L(\theta_{M,0})] + \frac{1}{\tau} \sum_{t=1}^{\tau} \mathbb{E}_{P_{0,C_o(t)}}(L(\theta_{M,\tau}) - L(\theta_{M,0})) \\
&\overset{(6)}{=} -[M_\tau(\theta_{M,\tau}) - M_\tau(\theta_{M,0})].
\end{aligned}$$

The first two inequalities follow directly from the definition of $\theta_{M,\tau}$ which is a minimizer of the empirical risk $\frac{1}{\tau}\sum_{t=1}^{\tau} L(\theta)(O(t), C_o(t))$ over all $\theta \in \Theta(M)$. Line (3) equals (2) by adding and subtracting $\frac{1}{\tau}\sum_{t=1}^{\tau}[L(\theta_{M,\tau}) - L(\theta_{M,0})]$. Line (4) once again follows as the definition of $\theta_{M,\tau}$, with (5) and (6) due to rearranging and definition of $M_\tau(\theta)$. By Assumptions (8) and (9), we have that $d_0(\theta_{M,\tau}, \theta_{M,0}) = o_P(1)$, which implies $\frac{1}{\tau}\sum_{t=1}^{\tau} \mathbb{E}_{P_{0,C_o(t)}}[L(\theta_{M,\tau}) - L(\theta_{M,0})]^2 \to_p 0$. Consequently by Lemma 2 and asymptotic equicontinuity of the martingale process $M_\tau(\theta)$, we have that $M_\tau(\theta_{M,\tau}) - M_\tau(\theta_{M,0}) = o_P(\tau^{-1/2})$. It then follows that $d_0(\theta_{M,\tau}, \theta_{M,0}) = o_P(\tau^{-1/2})$ by the above set of inequalities, which proves our claim. □

Theorem 4: Decomposition of the difference between TMLE and the estimand

Theorem 4 *Let θ^*_τ be the one-step TMLE satisfying $\frac{1}{\tau}\sum_{t=1}^{\tau} D_{C_o(t)}(\theta^*_\tau)(O(t)) = 0$ for $\theta^*_\tau = (g_\tau, \bar{Q}^*_\tau)$. Further, we define $\theta_l = (g_0, \bar{Q}_l)$ as the limit of $\theta^*_\tau = (g_\tau, \bar{Q}^*_\tau)$ for $\theta_l \in \Theta$, and $\theta_0 = (g_0, \bar{Q}_0)$ as the truth. Then the difference between the TMLE and its estimand decomposes as*

$$\Psi(\theta^*_\tau) - \Psi(\theta_0) = M_{1,\tau}(\theta_l) + M_{2,\tau}(\theta^*_\tau, \theta_l) + \frac{1}{\tau}\sum_{t=1}^{\tau} R_{C_o(t)}(\theta^*_\tau, \theta_0),$$

with

$$\begin{aligned}
M_{1,\tau}(\theta_l) =& \frac{1}{\tau}\sum_{t=1}^{\tau}[D_{C_o(t)}(\theta_l)(O(t)) - \mathbb{E}_{P_{0,C_o(t)}} D_{C_o(t)}(\theta_l)], \\
M_{2,\tau}(\theta^*_\tau, \theta_l) =& \frac{1}{\tau}\sum_{t=1}^{\tau}[D_{C_o(t)}(\theta^*_\tau)(O(t)) - \mathbb{E}_{P_{0,C_o(t)}} D_{C_o(t)}(\theta^*_\tau)] \\
& - \frac{1}{\tau}\sum_{t=1}^{\tau}[D_{C_o(t)}(\theta_l)(O(t)) + \mathbb{E}_{P_{0,C_o(t)}} D_{C_o(t)}(\theta_l)] \\
& = \frac{1}{\tau}\sum_{t=1}^{\tau}(\delta_{C_o(t),O(t)} - P_{0,C_o(t)})(D_{C_o(t)}(\theta^*_\tau) - D_{C_o(t)}(\theta_l)).
\end{aligned}$$

Proof. Recall the von Mises expansion of $\Psi(\theta)$ given in Theorem 3, which gives a first order approximation of the difference between the estimator and the estimand:

$$\begin{aligned}
\Psi(P_\theta) - \Psi(P_{\theta_0}) &= \Psi(\theta) - \Psi(\theta_0) \\
&= -\frac{1}{\tau}\sum_{t=1}^{\tau} \mathbb{E}_{P_{0,C_o(t)}}[D_{C_o(t)}(\theta)] + \frac{1}{\tau}\sum_{t=1}^{\tau} R_{C_o(t)}(\theta, \theta_0).
\end{aligned}$$

As θ^*_τ is a TMLE, by definition we have that $\frac{1}{\tau}\sum_{t=1}^{\tau} D_{C_o(t)}(\theta^*_\tau)(O(t)) = 0$. With θ being a TMLE θ^*_τ, the first order expansion previously can be written as

$$\Psi(\theta^*_\tau) - \Psi(\theta_0) = \frac{1}{\tau}\sum_{t=1}^{\tau}[D_{C_o(t)}(\theta^*_\tau)(O(t) - \mathbb{E}_{P_{0,C_o(t)}} D_{C_o(t)}(\theta^*_\tau)] + \frac{1}{\tau}\sum_{t=1}^{\tau} R_{C_o(t)}(\theta^*_\tau, \theta_0).$$

By adding and subtracting $\frac{1}{\tau}\sum_{t=1}^{\tau}(D_{C_o(t)}(\theta_l)(O(t))-\mathbb{E}_{P_{0,C_o(t)}}D_{C_o(t)}(\theta_l))$, we achieve the stated decomposition. □

Theorem 5: Asymptotic Normality of the TMLE

Theorem 5 *Let θ_τ^* be the one-step TMLE satisfying $\frac{1}{\tau}\sum_{t=1}^{\tau}D_{C_o(t)}(\theta_\tau^*)(O(t))=0$ for $\theta_\tau^*=(g_\tau,\bar{Q}_\tau^*)$. Further, we define $\theta_l=(g_0,\bar{Q}_l)$ as the limit of $\theta_\tau^*=(g_\tau,\bar{Q}_\tau^*)$ for $\theta_l\in\Theta$, and $\theta_0=(g_0,\bar{Q}_0)$ as the truth. Under Assumptions* (10), (11), (12), (13) *and* (14) *we have that*

$$\sqrt{\tau}\left(\Psi(\theta_\tau^*)-\Psi(\theta_0)\right)\xrightarrow{d}\mathcal{N}(0,\sigma_0^2),$$

where $\bar{f}_t=D_{C_o(t)}(\theta_l)(O(t))+f(C_o(t))(A(t)-g_0(1\mid C_o(t))$ and $\frac{1}{\tau}\sum_\tau\mathbb{E}_{P_{0,C_o(t)}}\bar{f}_t\xrightarrow{d}\sigma_0^2$.

Proof. Recall the decomposition presented in Theorem 5, which states that the difference between the TMLE and its estimand decomposes as

$$\Psi(\theta_\tau^*)-\Psi(\theta_0)=M_{1,\tau}(\theta_l)+M_{2,\tau}(\theta_\tau^*,\theta_l)+\frac{1}{\tau}\sum_{t=1}^{\tau}R_{C_o(t)}(\theta_\tau^*,\theta_0).$$

Under Assumptions (10) and (11), we have by Theorem 6 that $\sqrt{\tau}M_{1,\tau}(\theta_l)$ converges to a normal distribution with mean zero and variance

$$\frac{1}{\tau}\sum_{t=1}^{\tau}\mathbb{E}_{P_{0,C_o(t)}}\left(D_{C_o(t)}(\theta_l)(O(t))\mid C_o(t)\right)^2\xrightarrow{d}\sigma_l^2.$$

Under Assumptions (12) and (13), Lemma 2 shows the asymptotic equicontinuity under a complexity condition for a process derived from a function class $\{D_{C_o(t)}(\theta):\theta\in\Theta\}$, which implies that if $\frac{1}{\tau}\sum_{t=1}^{\tau}\mathbb{E}_{P_{0,C_o(t)}}[D_{C_o(t)}(\theta_\tau^*)-D_{C_o(t)}(\theta_l)]^2\to_p 0$, then $M_{2,\tau}(\theta_\tau^*,\theta_l)=o_P(1/\sqrt{\tau})$. Therefore, we conclude that the second term in the decomposition, $M_{2,\tau}(\theta_\tau^*,\theta_l)$, is negligible under conditions outlined in Lemma 2. Finally, we consider the remainder term, which evaluated at $\theta=\theta_\tau^*$ can be written as

$$\begin{aligned}\frac{1}{\tau}\sum_{t=1}^{\tau}R_{C_o(t)}(\theta_\tau^*,\theta_0)&=\frac{1}{\tau}\sum_{t=1}^{\tau}\frac{g_\tau-g_0}{g_\tau}(1\mid C_o(t))\left[\bar{Q}_\tau^*-\bar{Q}_0\right](1,C_o(t))\\&=\underbrace{\frac{1}{\tau}\sum_{t=1}^{\tau}\frac{g_\tau-g_0}{g_\tau}(1\mid C_o(t))\left[\bar{Q}_\tau^*-\bar{Q}_l\right](1,C_o(t))}_{\text{Term A}}\\&\quad+\underbrace{\frac{1}{\tau}\sum_{t=1}^{\tau}\frac{g_\tau-g_0}{g_\tau}(1\mid C_o(t))\left[\bar{Q}_l-\bar{Q}_0\right](1,C_o(t))}_{\text{Term B}}.\end{aligned}$$

By Assumption (14), we have that Term A is $o_P(1/\sqrt{\tau})$; as such, g_τ and $\bar{Q}_\tau$ converge fast enough to their limits ($o_P(1/\sqrt{\tau})$). We can decompose Term B further, where we now have that

$$\frac{g_\tau - g_0}{g_\tau}(1 \mid C_o(t))\left[\bar{Q}_l - \bar{Q}_0\right](1, C_o(t)) = \underbrace{\frac{g_\tau - g_0}{g_0}(1 \mid C_o(t))\left[\bar{Q}_l - \bar{Q}_0\right](1, C_o(t))}_{\text{Term C}} + \underbrace{\frac{(g_\tau - g_0)^2}{g_\tau g_0}(1 \mid C_o(t))\left[\bar{Q}_l - \bar{Q}_0\right](1, C_o(t))}_{\text{Term D}}.$$

By Assumption (14), we have that Term D is $o_P(1/\sqrt{\tau})$. Assumption (14) also assumes that Term C, a smooth function of $g_\tau - g_0$, can be represented as a martingale sum for some function f where

$$\frac{1}{\tau}\sum_{t=1}^{\tau}\frac{g_\tau - g_0}{g_0}(\bar{Q}_l - \bar{Q}_0) = \frac{1}{\tau}\sum_{t=1}^{\tau} f(C_o(t))(A(t) - g_0(A(t) \mid C_o(t)) + o_P(1/\sqrt{\tau}).$$

The decomposition presented in Theorem 5 can then be written as

$$\begin{aligned}\Psi(\theta_\tau^*) - \Psi(\theta_0) &= M_{1,\tau}(\theta_l) + M_{2,\tau}(\theta_\tau^*, \theta_l) + \frac{1}{\tau}\sum_{t=1}^{\tau} R_{C_o(t)}(\theta_\tau^*, \theta_0) \\ &= \frac{1}{\tau}\sum_{t=1}^{\tau}[D_{C_o(t)}(\theta_l)(O(t)) - \mathbb{E}_{P_{0,C_o(t)}} D_{C_o(t)}(\theta_l)] + o_P(1/\sqrt{\tau}) \\ &+ \frac{1}{\tau}\sum_{t=1}^{\tau} f(C_o(t))(A(t) - g_0(A(t) \mid C_o(t)) + o_P(1/\sqrt{\tau}).\end{aligned}$$

Since both $M_{1,\tau}(\theta_l)$ and $\frac{1}{\tau}\sum_{t=1}^{\tau} f(C_o(t))(A(t) - g_0(A(t) \mid C_o(t))$ are discrete martingales whose standard version converges to a normal limit distribution, we conclude that

$$\sqrt{\tau}\left(\Psi(\theta_\tau^*) - \Psi(\theta_0)\right) \xrightarrow{d} \mathcal{N}(0, \sigma_0^2)$$

where $\bar{f}_t = D_{C_o(t)}(\theta_l)(O(t)) + f(C_o(t))(A(t) - g_0(1 \mid C_o(t))$ and $\frac{1}{\tau}\sum_\tau \mathbb{E}_{P_{0,C_o(t)}} \bar{f}_t \xrightarrow{d} \sigma_0^2$. □

Additional Data Analyses

In the following, we report results for additional data analyses where exposure is more-than-usual ingestion and more-than-usual activity. As before, a subset of the initial 70 samples is picked based on variability in exposure variable ($g_0(A(t) \mid \bar{O}(t-1)) > 0.05$), and length of observation ($\tau > 80$ time points). For each sample, we estimate $\bar{Q}_0$ and g_0 across collected time points using the Online Super Learner available as part of the SL3 software package [27]. The Super Learner library for both $\bar{Q}_0$ and g_0 consisted of: highly adaptive lasso (HAL9001), gradient boosted decision trees (XGBOOST), random forests (RANGER), elastic net generalized linear models (GLMNET), Bayesian generalized linear models (BAYESGLM), and main terms linear regression (GLM) [28, 23, 147, 41, 43]. For cross-validation, we split the data into folds by implementing the Rolling Origin CV scheme as implemented in the ORIGAMI software package [26, 25, 79]. For samples with more that 100 points, we set the first window to $t = 40$, with validation size of $t = 25$, no gap and batch of $t = 20$. For samples with less data ($\tau < 100$), we set the first window to $t = 32$, with validation size of $t = 25$, no gap and batch of $t = 20$. Here, the smaller initial fold sizes are motivated by shorter trajectories then seen in the original analysis where exposure is insulin injection. All analyses set $C_o(t)$ to the last 5 time periods of the relevant observed history: for more-than-usual ingestion that included time of the day, pre-/post- meal indicator, type of insulin, more-/less- activity and previous ingestion and hyperglycemic/hypoglycemic episodes. For more-than-usual activity as exposure, relevant fixed-dimensional summary measure was derived from the last 5 time periods of time of the day, pre-/post- meal indicator, type of insulin, more-/less- ingestion as well as previous activity and hyperglycemic/hypoglycemic episodes.

Ingestion as Exposure

The context-specific ATE over time with more-than-usual ingestion has a clear effect across all analyzed samples: more exposure (more times with higher than usual ingestion) results in higher proportion of hyperglycemic episodes over time. Similarly, less frequent more-than-usual ingestion leads to higher proportion of hyperglycemic episodes with the exception of Sample 58. While the effect has the same direction for most samples, the level of within-sample heterogeneity is significant, as exemplified in Figure 2.2. To highlight the importance of context in a single time-series context, Samples 12, 13 and 15 had the highest percent of hyperglycemic episodes on no more-than-usual ingestion, but had one of the highest more-than-usual activity, which reduces glucose blood levels for a significant period of time.

Activity as Exposure

The context-specific ATE over time with more-than-usual activity is not as clear as more-than-usual ingestion exposure, except for its effect on hypoglycemia. Across all samples but Sample 17, the effect of persistent more-than-usual activity leads to a higher rate of hypoglycemia episodes for individuals with IDDM. Sample 17 is interesting in that it has a higher rate of hypoglycemic episodes on no more-than-usual activity. With Y_{hyper} being the

						Blood glucose level mg/d			
ID	**Days**	τ	$A_{\text{Ingestion}}\%$	$Y_{\text{hyper}}\%$	$Y_{\text{hypo}}\%$	**Min**	**Mean**	**Max**	**SD**
12	44	89	24	39	35	31	163	436	106
13	36	97	20	33	31	31	161	461	104
16	56	93	19	41	17	23	179	422	102
15	53	93	18	34	30	23	159	384	97
58	30	115	12	54	11	33	218	501	104
59	31	119	12	56	13	31	208	442	99
33	29	118	10	27	15	45	154	349	68
36	33	122	10	39	14	39	173	501	86
35	29	120	10	22	23	40	141	302	66
3	38	141	7	17	22	22	136	303	64

Table 2.4: Descriptive and summary statistics for each sample used in the data analysis, with exposure being more-than-usual ingestion. The descriptive information consists of the patient id, number of days monitored, number of time points collected, percent time having more-than-usual ingestion during observation ($A_{\text{ingestion}}\%$), percent time having an hyperglycemic ($Y_{\text{hyper}}\%$) or hypoglycemic ($Y_{\text{hypo}}\%$) episode per sample. Summary statistics of blood glucose level in mg/d includes the minimum, maximum, average and standard deviation over all available time points.

outcome, most samples had more hyperglycemic episodes on the no more-than-usual activity regime, which physiologically corresponds to what we know about the relationship between diabetes mellitus and physical activity. Sample 50 and 49 had small to non-significant effect of more-than-usual activity on hyperglycemic episodes, despite being one of the most active sample during the observational period. Samples 18 and 19 were interesting in that they had more hyperglycemic episodes on more-than-usual activity. Sample 19 had persistently high blood glucose levels, indicating a problem that might not be possible to alleviate with physical activity (potentially in need of a more strict insulin treatment). Sample 18 on the other hand had the lowest percent of treatment, indicating that perhaps a longer trajectory was necessary to properly learn the treatment mechanism.

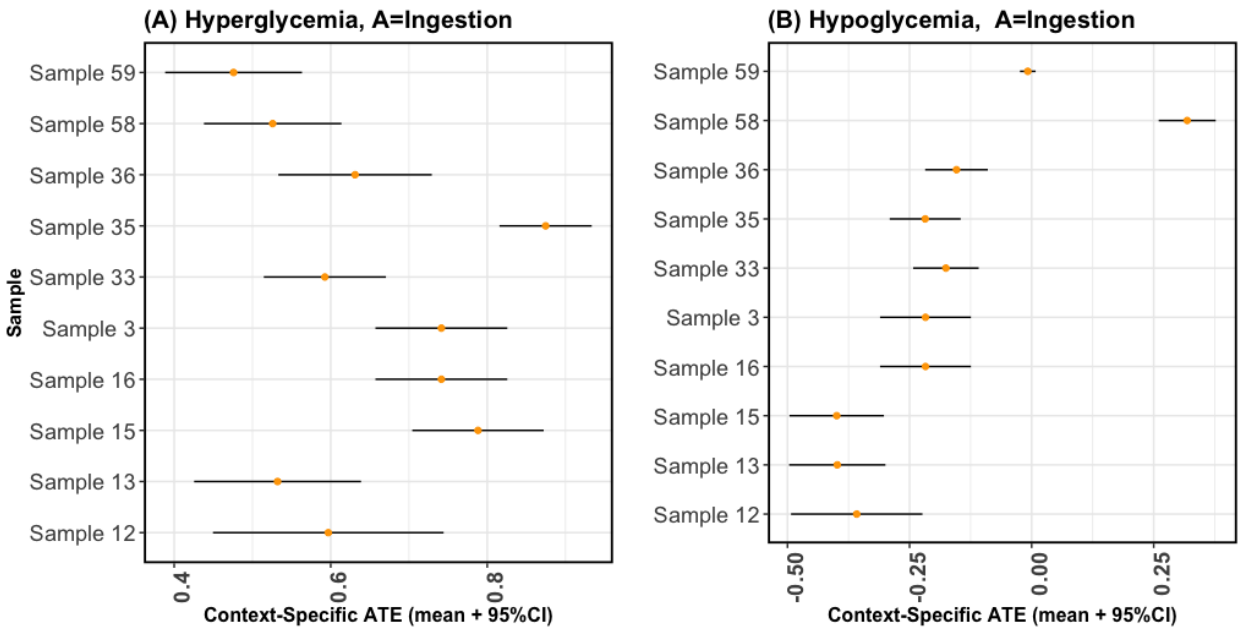


Figure 2.2: Context-Specific ATE and its 95% confidence interval for each sample with treatment corresponding to the more-than-usual Ingestion. Panel (A) shows the mean and 95% confidence interval for Y_{hyper}, which is the outcome variable corresponding to a hyperglycemic episode. Panel (B) plots the mean and 95% confidence interval for each sample where the outcome is a hypoglycemic event, Y_{hypo}.

						Blood glucose level mg/d			
ID	**Days**	τ	$A_{\text{Activity}}\%$	$Y_{\text{hyper}}\%$	$Y_{\text{hypo}}\%$	**Min**	**Mean**	**Max**	**SD**
50	60	249	7	1	44	31	86	207	33
19	31	126	8	29	21	41	161	501	89
18	32	134	4	22	23	43	141	370	75
17	31	103	5	23	30	43	142	380	76
49	45	189	6	2	40	31	92	231	38
51	39	151	6	13	36	35	114	278	60
33	29	118	7	27	15	45	154	349	68
52	47	172	5	16	37	35	128	501	83

Table 2.5: Descriptive and summary statistics for each sample used in the data analysis, with exposure being more-than-usual activity. The descriptive information consists of the patient id, number of days monitored, number of time points collected, percent time having more-than-usual activity during observation ($A_{\text{activity}}\%$), percent time having an hyperglycemic ($Y_{\text{hyper}}\%$) or hypoglycemic ($Y_{\text{hypo}}\%$) episode per sample. Summary statistics of blood glucose level in mg/d includes the minimum, maximum, average and standard deviation over all available time points.

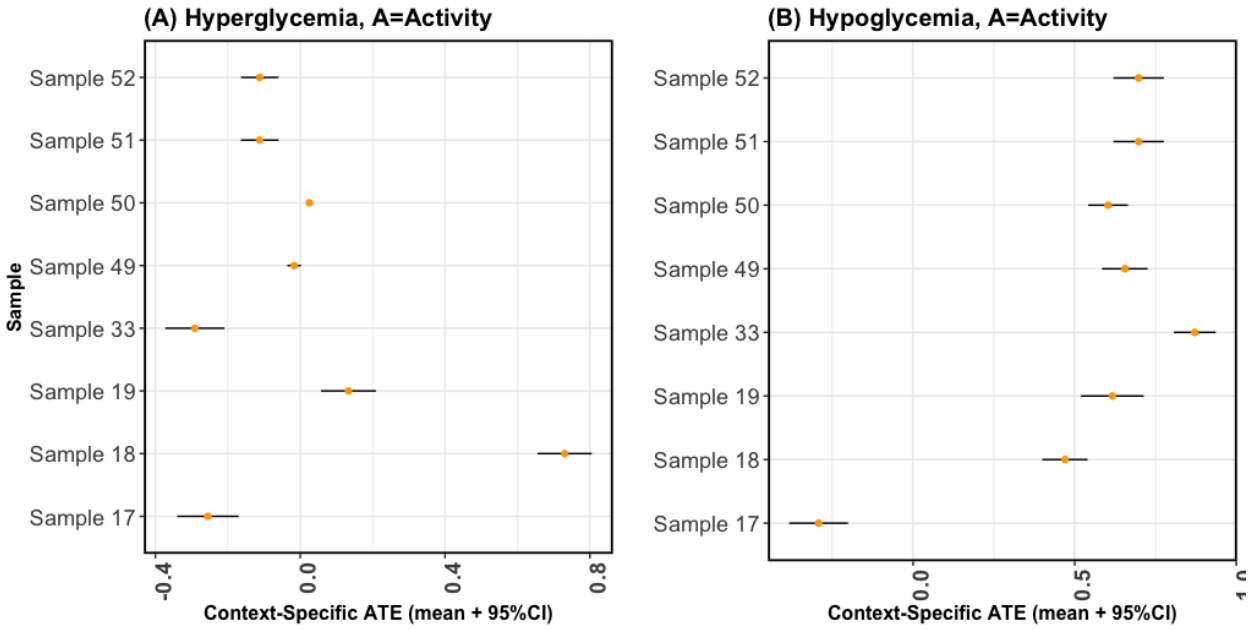


Figure 2.3: Context-Specific ATE and its 95% confidence interval for each sample with treatment corresponding to the more-than-usual Activity. Panel (A) shows the mean and 95% confidence interval for Y_{hyper}, which is the outcome variable corresponding to a hyperglycemic episode. Panel (B) plots the mean and 95% confidence interval for each sample where the outcome is a hypoglycemic event, Y_{hypo}.

Chapter 3

Adaptive Sequential Design for a Single time-series

The work described in this chapter is motivated by the need for robust statistical methods for precision medicine. In particular, it pioneers the concept of a sequential, adaptive design for a single individual. As such, we address the need for statistical methods that provide actionable inference for a single unit at any point in time. Consider the case that one observes a single time-series, where at each time t, we have a data record $O(t)$ involving treatment nodes $A(t)$, an outcome node $Y(t)$, and time-varying covariates $W(t)$. We aim to learn an optimal, unknown choice of the controlled components of the design in order to optimize the expected outcome; with that, we adapt the randomization mechanism for future time-point experiments based on the data collected on the individual over time. Our results demonstrate that one can learn the optimal rule based on a single sample, and thereby adjust the design at any point t with valid inference for the mean target parameter. We define a nonparametric model for the probability distribution of the time-series under few assumptions, and aim to fully utilize the sequential randomization in the estimation procedure via the double robust structure of the efficient influence curve. This work provides several contributions to the field of statistical precision medicine. First, we present multiple exploration-exploitation strategies for assigning treatment, and methods for estimating the optimal rule. Secondly, we present the study of the data-adaptive inference on the mean under the optimal treatment rule, where the target parameter adapts over time in response to the observed context of the individual. We characterize the limit distribution of our estimator under a Donsker condition expressed in terms of a notion of bracketing entropy adapted to martingale settings.

3.1 Introduction

There is growing scientific enthusiasm for the use, and development, of mobile health designs (mHealth) - broadly referring to the practice of health care mediated via mobile and wearable technologies [117, 80, 62, 61]. Numerous smartphones and Internet coupled devices,

connected to a plethora of mobile health applications, support continuous assembly of data-driven healthcare intervention and insight opportunities. Interest in mobile interventions spans myriad of applications, including behavioral maintenance or change [40, 85], disease management [55, 37, 84, 117], teaching and social support [69] and addiction management [35, 149]. In particular, Istepanian and Al-Anzi [61] refer to mHealth as one of the most transformative drivers for healthcare delivery in modern times. Recently, a new type of an experimental design termed micro-randomized trial (MRT) was developed in order to support just-in-time adaptive exposures - with an aim to deliver the intervention at the optimal time and location [29, 68]. To this date, multiple trials have been completed using MRT design, including encouraging regular physical activity [67] and engaging participation in substance use data gathering process in high-risk populations [100]. For both observational mHealth and MRT, the time-series nature of the collected data provides an unique opportunity to collect individual characteristics and context of each subject, while studying the effect of treatment on the outcome at a specified future time-point.

The generalized estimating equation (GEE) and random effects models are the most commonly employed approaches for the analysis of mobile health data [144, 119, 13, 51]. As pointed out in Boruvka et al. [14], these methods often do not yield consistent estimates of the causal effect of interest if time-varying treatment is present. As an alternative, Boruvka et al. [14] propose a centered and weighted least square estimation method for GEE that provides unbiased estimation, assuming linear model for the treatment effects. They tackle proximal and distal effects, with a focus on continuous outcome. On the other hand, Luckett et al. [74] propose a new reinforcement learning method applicable to perennial, frequently collected longitudinal data. While the literature on dynamic treatment regimes is vast and well-studied [87, 105, 18, 77, 75, 76], the unique challenges posed by mHealth obstruct their direct employment; for instance, mHealth objective typically has an infinite horizon. Luckett et al. [74] model the data-generating distribution as a Markov decision process, and estimate the optimal policy among a class of pre-specified policies in both offline and online setting.

While mHealth, MRT designs and the corresponding methods for their analysis aim to deliver treatment tailored to each patient, they are still not optimized with complete "N-of-1" applications in mind. The usual population based target estimands fail to ensnare the full, personalized nature of the time-series trajectory, often imposing strong assumptions on the dynamics model for the estimation purposes. To the best of our knowledge, Robins, Greenland, and Hu [106] provide the first step towards describing a causal framework for a single subject with time-varying exposure and binary outcome in a time-series setting. Focusing on full potential paths, Bojinov and Shephard [12] provide a causal framework for time-series experiments with randomization-based inference. Other methodologies focused on single unit applications rely on strong modeling assumptions, primarily linear predictive models and stationarity; see Bojinov and Shephard [12] for an excellent review of the few works on the topic. Alternatively, van der Laan, Chambaz, and Lendle [131] propose causal effects defined as marginal distributions of the outcome at a particular time point under a certain intervention on one or more of the treatment nodes. The efficient influence function of these estimators, however, relies on the whole mechanism in a non-double robust manner.

Therefore, even when the assignment function is known, the inference still relies on consistent (at rate) estimation of the conditional distributions of the covariate and outcome nodes.

The current work is motivated by the need for robust statistical methods for precision medicine, pioneering the concept of a sequential, adaptive design for a single individual. To the best of our knowledge, this is the first work on learning the optimal individualized treatment rule in response to the current context for a single subject. A treatment rule for a patient is an individualized treatment strategy based on the history accrued, and context learned, up to the most current time point. A reward is measured on the patient at repetitive units, and optimality is meant in terms of optimization of the mean reward at a particular time t. We aim to learn an optimal, unknown choice of the controlled components of the design based on the data collected on the individual over time; with that, we adapt the randomization mechanism for future time-point experiments. Our results demonstrate that one can learn the optimal, context defined rule based on a single sample, and thereby adjust the design at any point t with valid inference for the mean target parameter. We define models for the probability distribution of the time-series that refrains from making unrealistic parametric assumptions, and aims to fully utilize the sequential randomization in the estimation procedure. In particular, we present the study of the data-adaptive inference on the mean under the optimal treatment rule, where the target parameter adapts over time in response to the observed context of the individual. Our estimators are double robust and easier to estimate efficiently than previously proposed variations [131]. For inference, we rely on martingale Central Limit Theorem under a conditional variance stabilization condition and a maximal inequality for martingales with respect to an extension of the notion of bracketing entropy for martingale settings, initially proposed by [128], which we refer to as *sequential bracketing entropy*.

This structure of the chapter is as follows. In Section 3.2 we formally present the general formulation of the statistical estimation problem, consisting of specifying the statistical model and notation, the target parameter defined as the average of context-specific target parameters, causal assumptions and identification results, and the corresponding efficient influence curve for the target parameter. In Section 3.3 we discuss different strategies for estimating the optimal treatment rule and sampling strategies for assigning treatment at each time point. The following section, Section 3.4, introduces the Targeted Maximum Likelihood Estimator (TMLE), with Section 3.5 covering the theory behind the proposed estimator. In Section 3.6 we present simulation results for different dependence settings. We conclude with a short discussion in Section 3.7.

3.2 Statistical Formulation of the Problem

Data and Likelihood

Let $O(t)$ be the observed data at time t, where we assume to follow a patient along time steps $t = 1, \ldots, N$ such that $O^N \equiv (O(0), O(1), \ldots O(N)) = (O(t) : t = 0, \ldots, N)$. At each time

step t, the experimenter assigns to the patient a binary treatment $A(t) \in \mathcal{A} := \{0, 1\}$. We then observe, in this order, a post-treatment health outcome $Y(t) \in \mathcal{Y} \subset \mathbb{R}$, and then a post-outcome vector of time-varying covariates $W(t)$ lying in an Euclidean set $\mathcal{W}$. We suppose that larger values of $Y(t)$ reflect a better health outcome; without loss of generality, we also assume that $Y(t) \equiv (0, 1)$, with rewards being bounded away from 0 and 1. The ordering of the nodes matters, as $W(t)$ is an important part of post-exposure history to be considered for the next record, $O(t+1)$. Finally, we note that $O(0) = (W(0))$, where $O(-1) = A(0) = Y(0) = \emptyset$; as such, $O(0)$ plays the role of baseline covariates for the collected time-series, based on which exposure $A(1)$ might be allocated. We denote $O(t) := (A(t), Y(t), W(t))$ the observed data collected on the patient at time step t, with $\mathcal{O} := \mathcal{A} \times \mathcal{Y} \times \mathcal{W}$ as the domain of the observation $O(t)$. We note that $O(t)$ has a fixed dimension in time t, and is an element of an Euclidean set $\mathcal{O}$. Our data set is the time-indexed sequence $O^N \in \mathcal{O}^N$, or *time-series*, of the successive observations collected on a single patient. For any t, we let $\bar{O}(t) := (O(1), \ldots, O(t))$ denote the observed history of the patient up until time t. Unlike in more traditional statistical settings, the data points $O(1), \ldots, O(N)$ are not independent draws from the same law: here they form a dependent sequence, which is a single draw of a distribution over $\mathcal{O}^N$. In that sense, our data reduces to a single sample.

We let $O^N \sim P_0^N$, where P_0^N denotes the true probability distribution of O^N. The subscript "0" stands for the "truth" throughout the rest of the chapter, denoting the true, unknown features of the distribution of the data. Realizations of a random variable O^N are denoted with lower case letters, o^N. We suppose that P_0^N admits a density p_0^N w.r.t. a dominating measure μ over $\mathcal{O}^N$ that can be written as the product measure $\mu = \times_{t=1}^N (\mu_A \times \mu_Y \times \mu_W)$, with μ_A, μ_Y, and μ_W measures over $\mathcal{A}$, $\mathcal{Y}$, and $\mathcal{W}$. The likelihood under the true data distribution P_0^N of a realization $\bar{o}^N$ of $\bar{O}^N$ can be factorized according to the time ordering of observation nodes as:

$$p_0^N(o^N) = \prod_{t=1}^{N} p_{0,a(t)}(a(t) \mid \bar{o}(t-1)) \times \prod_{t=1}^{N} p_{0,y(t)}(y(t) \mid \bar{o}(t-1), a(t)) \times \prod_{t=0}^{N} p_{0,w(t)}(w(t) \mid \bar{o}(t-1), a(t), y(t)), \tag{3.1}$$

where $a(t) \mapsto p_{0,a(t)}(a(t) \mid \bar{o}(t-1)$, $y(t) \mapsto p_{0,y(t)}(y(t) \mid \bar{o}(t-1), a(t))$, and $w(t) \mapsto p_{0,w(t)}(w(t) \mid \bar{o}(t-1), a(t), y(t))$ are conditional densities w.r.t. the dominating measures μ_A, μ_Y, and μ_W.

Statistical Model

Since O^N represents a single time-series, a dependent process, we observe only a single draw from P_0^N. As a result, we are unable to estimate any part of P_0^N without additional assumptions. In particular, we assume that the conditional distribution of $O(t)$ given $\bar{O}(t-1)$, $P_{O(t)|\bar{O}(t-1)}$, depends on $\bar{O}(t-1)$ through a summary measure $C_o(t) = C_o(\bar{O}(t-1)) \in \mathcal{C}$ of fixed dimension; each $C_o(t)$ might contain t-specific summary of previous measurements of

context, or is of a particular Markov order. For later notational convenience, we denote this conditional distribution $P_{O(t)|\bar{O}(t-1)}$ with $P_{C_o(t)}$. Then, the density $p_{C_o(t)}$ of $P_{C_o(t)}$ with respect to a dominating measure $\mu_{C_o(t)}$ is a conditional density $(o, C_o) \to p_{C_o(t)}(o \mid C_o)$ so that for each value of $C_o(t)$, $\int p_{C_o(t)}(o \mid C_o(t))d\mu_{C_o(t)}(o) = 1$. We extend this notion to all parts of the likelihood as described in subsection (3.1), defining $q_{y(t)}$ as the density for node $Y(t)$ conditional on a fixed dimensional summary $C_y(t)$, with $C_w(t)$ and $C_a(t)$ corresponding to fixed dimensional summaries for $q_{w(t)} = p_{0,w(t)}(w(t) \mid C_w(t))$ and $g_t = p_{0,a(t)}(a(t) \mid C_a(t))$, respectively.

Additionally, we assume that $p_{C_o(t)}$ is parameterized by a common (in time t) function $\theta \in \Theta$, with inputs $(c, o) \to \theta(c, o)$. The conditional distribution $p_{C_o(t)}$ depends on θ only through $\theta(C_o(t), \cdot)$. We write $p_{C_o(t)} = p_{\theta,C_o(t)}$ interchangeably.Let q_y be the common conditional density of $Y(t)$, given $(A(t), C_o(t))$; we make no such assumption on $q_{w(t)}$. Additionally, we make no conditional stationarity assumptions on g_t if randomization probabilities are known, as is the case for an adaptive sequential trial. We define $\bar{Q}(C_o(t), A(t)) = E_{P_{C_o(t)}}(Y(t) \mid C_o(t), A(t))$ to be the conditional mean of $Y(t)$ given $C_o(t)$ and $A(t)$. As such, we have that $\bar{Q}(C_y(t)) = \bar{Q}(C_o(t), A(t)) = \int y q_y(y \mid C_o(t), A(t))d\mu_y(o)$, and $\bar{Q}$ is a common function across time t; we put no restrictions on $\bar{Q}$. We suppress dependence of the conditional density $q_{w(t)}$ in future reference, as this factor plays no role in estimation. In particular, $q_{w(t)}$ does not affect the efficient influence curve of the target parameter, allowing us to act as if $q_{w(t)}$ is known. Finally, we define $\theta = (g, \bar{Q})$.

Let $p_{\theta,C_o(t)}$ and p_θ^N be the density for $O(t)$ given $C_o(t)$ and O^N, implied by θ. This defines a statistical model $\mathcal{M}^N = \{P_\theta^N : \theta\}$ where P_θ^N is the probability measure for the time-series implied by $p_{\theta,C_o(t)}$. Additionally, we define a statistical model of distributions of $O(t)$ at time t, conditional on realized summary $C_o(t)$. In particular, let $\mathcal{M}(C_o(t)) = \{P_{\theta,C_o(t)} : \theta\}$ be the model for $P_{C_o(t)}$ for a given $C_o(t)$ implied by $\mathcal{M}^N$. Note that, by setup, both $\mathcal{M}^N$ and $\mathcal{M}(C_o(t))$ contain their truth P_0 and $P_{C_o(t)}$, respectively. Similarly to the likelihood expression in sub section (3.1), we can factorize the likelihood under the above defined statistical model according to time ordering as:

$$p_\theta(o^N) = \prod_{t=1}^{N} g_t(a(t) \mid C_a(t)) \prod_{t=1}^{N} q_y(y(t) \mid C_y(t)) \prod_{t=0}^{N} q_{w(t)}(w(t) \mid C_w(t)). \tag{3.2}$$

Causal Target Parameter and Identification

Structural equation model and causal target parameter

By specifying a structural equations model (SEM; equivalently, structural causal model), we assume that each component of the observed time-specific data structure is a function of an observed, fixed-dimensional history and an unmeasured exogenous error term [93]. We

encode the time-ordering of the variables using the following SEM:

$$\begin{aligned} W(0) &= f_{W(0)}(U_W(0)), \\ A(t) &= f_{A(t)}(C_A(t), U_A(t)), \quad t = 1, \ldots, N, \\ Y(t) &= f_{Y(t)}(C_Y(t), U_Y(t)), \quad t = 1, \ldots, N, \\ W(t) &= f_{W(t)}(C_W(t), U_W(t)), \quad t = 1, \ldots, N, \end{aligned} \tag{3.3}$$

where $(f_A(t) : t = 1, \ldots, N)$, $(f_Y(t) : t = 1, \ldots, N)$ and $(f_W(t) : t = 0, \ldots, N)$ are unspecified, deterministic functions and $U = (U_W(0), \ldots, U_A(1), \ldots, U_Y(1), \ldots, U_Y(N))$ is a vector of exogenous errors.

We denote $\mathcal{M}^F$ the set of all probability distributions P^F over the domain of (O, U) that are compatible with the NPSEM defined above. Let P_0^F be the true probability distribution of (O, U), which we assume to belong to $\mathcal{M}^F$; we denote $\mathcal{M}^F$ as the *causal model*. The causal model $\mathcal{M}^F$ encodes all the knowledge about the data-generating process, and implies a model for the distribution of the counterfactual random variables; as such, causal effects are defined in terms of hypothetical interventions on the SEM. Consider a treatment rule $C_o(t) \to d(C_o(t)) \in \{0, 1\}$, that maps the observed, fixed dimensional history $C_o(t)$ into a treatment decision for $A(t)$. We introduce a counterfactual random variable $O^{N,d}$, defined by substituting the equation for node A at time t in the SEM with the intervention d:

$$\begin{aligned} W^d(0) &= f_{W(0)}(U_W(0)) \\ A^d(t) &= d(C_A(t)), \qquad\qquad t = 1, \ldots, N \\ Y^d(t) &= f_{Y(t)}(C_Y(t), U_Y(t)), \quad t = 1, \ldots, N \\ W^d(t) &= f_{W(t)}(C_W(t), U_W(t)), \quad t = 1, \ldots, N, \end{aligned} \tag{3.4}$$

We gather all of the nodes of the above modified SEM in the random vector $O^{N,d} := (O^d(t) : t = 1, \ldots, N)$, where $O^d(t) := (A^d(t), Y^d(t), W^d(t))$. The random vector $O^{N,d}$ represents the counterfactual time-series, or counterfactual trajectory the subject of interest would have had, had each treatment assignment $A(t)$, for $t = 1, \ldots, N$, had been carried out following the treatment rule d.

We now formally define time-series causal parameters. First, we introduce a time- and context-specific causal model. Let $\mathcal{M}^F(C_o(t))$ be the set of conditional probability distributions $P^F_{C_o(t)}$ over the domain of $(O(t), U_A(t), U_Y(t), U_W(t))$ compatible with the nonparametric structural equation model (3.3) imposing that $C_A(t) = C_o(t) = c$:

$$\begin{aligned} A_c(t) =& f_{A(t)}(c, U_A(t)) \\ Y_c(t) =& f_{Y(t)}(c_Y(c, A(t)), U_Y(t)) \\ W_c(t) =& f_{W(t)}(c_W(c, A(t), Y(t)), U_W(t)). \end{aligned} \tag{3.5}$$

Let $O_c^d(t)$ be the counterfactual observation at time t, obtained by substituting the $A(t)$ equation in the above set of equations with the deterministic intervention d:

$$\begin{aligned} A_c^d(t) =& d(c) \\ Y_c^d(t) =& f_{Y(t)}(c_Y(c, A(t)), U_Y(t)) \\ W_c^d(t) =& f_{W(t)}(c_W(c, A(t), Y(t)), U_W(t)). \end{aligned} \tag{3.6}$$

We define our causal parameter of interest as

$$\Psi^{F,d}_{C_o(t)}(P^F_{C_o(t)}) := E[Y^d_{C_o(t)}], \tag{3.7}$$

which is the expectation of the counterfactual random variable Y^d, generated by the above modified SEM. It corresponds to starting at $c = C_o(t)$, the current context, and assigning treatment following d. Our causal target parameter is the mean outcome we would have obtained after one time-step, if, starting at time t from the observed context $C_o(t)$, we had carried out intervention d.

Identification of the causal target and defining the statistical target

Once we have defined our causal target parameter, the natural question that arises is how to identify it from the observed data distribution. We can identify the distribution of the d-specific time series $O^{N,d}$, and also of the $(d, C_o(t))$-specific observation $O^d_{C_o(t)}$, from the observed data via the G-computation formula - under the sequential randomization and positivity assumptions, which we state below.

Assumption 15 (Sequential randomization)**.** *For every t, $Y^d(t) \perp\!\!\!\perp A(t) \mid C_o(t)$ (and $Y^d_{C_o(t)}(t) \perp\!\!\!\perp A(t) \mid C_o(t)$).*

Assumption 16 (Positivity)**.** *It holds that under the treatment mechanism $g_{0,t}$, each treatment value $a \in \{0, 1\}$ has a positive probability of being assigned, under every possible treatment history:*

$$g_{0,t}(a \mid c) > 0, \forall t \geq 1,\ a \in \{0, 1\} \text{ and every } c \in \mathcal{C} \text{ such that } P_0[C_o(t) = c] > 0. \tag{3.8}$$

Note that under the setting of the present article, as we suppose that $A(t)$ is assigned at random conditional on $C_o(t)$ by the experimenter, assumption 15 concerning the sequential randomization automatically holds. Under identification assumptions 15 and 16, we can write our causal parameter $\Psi^{F,d}_{C_o(t)}(P^F_0)$ as a feature of the data-generating distribution:

$$\Psi^{F,d}_{C_o(t)}(P^F_0) = \Psi^d_{C_o(t)}(P_0) := E_{P_0}\left[Y(t)\frac{g^*(A(t) \mid C_o(t))}{g_{0,t}(A(t) \mid C_o(t))} \mid C_o(t)\right], \tag{3.9}$$

which for a deterministic d can be written as

$$\Psi^{F,d}_{C_o(t)}(P^F_{0,C_o(t)}) = \Psi_{C_o(t)}(P_{0,C_o(t)}) := E_{P_{0,C_o(t)}}\left[Y(t) \mid A(t) = d(C_o(t)), C_o(t)\right]. \tag{3.10}$$

We note that we also could have expressed the target parameter as $\Psi_{C_o(t)}(P_0)$, where it is implied that $\Psi_{C_o(t)}(P_0)$ depends on P_0 only through the true conditional distribution of $O(t)$ given $C_o(t)$. For every P, we remind that $P_{C_o(t)}$ denotes the distribution of $O(t)$ given $C_o(t)$, and let $\mathcal{M}(C_o(t))$ be the set of such distributions corresponding to $P \in \mathcal{M}$. At each time-point t, given a $C_o(t)$, we define a target parameter $\Psi_{C_o(t)} : \mathcal{M}(C_o(t)) \to \mathbb{R}$ that is pathwise differentiable with canonical gradient $D^*_{C_o(t)}(P_{C_o(t)})(o)$ at $P_{C_o(t)}$ in $\mathcal{M}(C_o(t))$. As described in Section 3.2, we have that $\Psi_{C_o(t)}(P_{C_o(t)}) = \Psi_{C_o(t)}(\theta)$, where $\Psi_{C_o(t)}(\theta)$ depends on θ only though its section $\theta(C_o(t), \cdot)$. We denote the collection of $C_o(t)$-specific canonical gradients as $(c, o) \to D^*(P_{C_o(t)})(c, o)$, so that we can write them uniformly as a function of the observed components; with that, we have that $D^*_{C_o(t)}(P_{C_o(t)})(o) = D^*_{C_o(t)}(\theta)(o) = D^*(\theta)(c_o(t), o)$. As is custom for canonical gradients, for a given $C_o(t)$, $D^*(\theta)$ is a function of the observed data with conditional mean zero with respect to $P_{C_o(t)}$.

Finally, we propose a class of statistical target parameters $\bar{\Psi}(\theta)$ defined as the average over time of $C_o(t)$-specific counterfactual means under the treatment rule. In particular, the target parameter on $\mathcal{M}^N$, $\Psi^N : \mathcal{M}^N \to \mathbb{R}$ of the data distribution $P^N \in \mathcal{M}^N$ is defined as:

$$\bar{\Psi}(\theta) = \frac{1}{N} \sum_{t=1}^{N} \Psi_{C_o(t)}(\theta). \tag{3.11}$$

The statistical target parameter $\bar{\Psi}(\theta)$ is data-dependent, as it is defined as an average over time of parameters of the conditional distribution of $O(t)$ given the observed realization of $C_o(t)$; as such, it depends on $(C_o(1), \ldots, C_o(N))$. In practice, $\bar{\Psi}(\theta)$ is an average of the means under optimal treatment decisions over all observed contexts over time. As an average of $C_o(t)$-specific causal effects with a double robust efficient influence curve $D^*_{C_o(t)}(\theta)(o)$, it follows we can estimate $\bar{\Psi}(\theta)$ in a double robust manner as well, as we further emphasize in the following section.

Canonical gradient and first order expansion of the target parameter

In the following theorem we provide the canonical gradient of our target parameter that admits a first order expansion with a double-robust second order term.

Theorem 8 (Canonical gradient and first order expansion)**.** *Under the strong positivity assumption, the target parameter mapping $\Psi_{C_0(t)} : \mathcal{M}(C_o(t)) \to \mathbb{R}$ is pathwise differentiable w.r.t. $\mathcal{M}(C_o(t))$, with a canonical gradient w.r.t. $\mathcal{M}(C_o(t))$ given by*

$$D^*_{C_o(t)}(\theta)(o) = \frac{g^*_t(a \mid C_o(t))}{g_t(a \mid C_o(t))} \left(y - \bar{Q}(a, C_o(t))\right), \tag{3.12}$$

where $A(t) = a$ and $Y(t) = y$. Furthermore $\Psi_{C_o(t)}(\bar{Q})$ admits the following first order expansion:

$$\Psi_{C_o(t)}(\bar{Q}) - \Psi_{C_o(t)}(\bar{Q}_0) = -P_{0,C_o(t)} D^*_{C_o(t)}(\theta) + R(\bar{Q}, \bar{Q}_0, g_t, g_{0,t}), \tag{3.13}$$

where R is a second order remainder that is doubly-robust, with $R(\bar{Q}, \bar{Q}_0, g_t, g_{0,t}) = 0$ if either $\bar{Q} = \bar{Q}_0$ or $g_t = g_{0,t}$.

Previous works on statistical parameters defined over a single time series model [131, 65] consider what we refer to as *marginal* parameters. Unlike the conditional parameters we consider here, the efficient influence function of marginal parameters is not double-robust in the usual sense; that is, robust w.r.t. a pair of variation independent nuisance parameters. More importantly, knowing or consistently estimating the treatment mechanism does not guarantee consistency of the causal effect for parameters described by [131] and [65].

Optimal rule

Now that we have identified the context-specific counterfactual outcome under d as a parameter of the observed data distribution P_0^N, we can identify the optimal treatment rule. The optimal treatment rule is a priori a causal object defined as a function of P_0^F, and a parameter of the observed data generating distribution P_0^N. Under the identification assumptions, we can identify the optimal rule from the observed data distribution as follows. Fix arbitrarily $\bar{Q} \in \bar{\mathcal{Q}}$. To alleviate notation, we further introduce the blip function under the true data generating distribution as:

$$B_0(C_o(t)) \equiv \bar{Q}_0(C_o(t), A(t) = 1) - \bar{Q}_0(C_o(t), A(t) = 0). \tag{3.14}$$

Intuitively, if $B_0(C_o(t)) > 0$, assigning treatment $A(t) = 1$ is more beneficial (in terms of optimizing $Y(t)$) than $A(t) = 0$ for time point t under the current context $C_o(t)$. If $B_0(C_o(t)) < 0$, we can optimize the t-specific outcome by assigning the subject treatment $A(t) = 0$ instead. The true optimal rule for the purpose of optimizing the mean of the next (short-term) outcome $Y(t)$, for binary treatment, is then given by:

$$d_0(C_o(t)) \equiv \mathbb{I}(B_0(C_o(t)) > 0). \tag{3.15}$$

As defined in Equation (3.15), $d_0(C_o(t))$ is a typical treatment rule that maps observed fixed dimensional summary deterministically into one treatment; a stochastic treatment rule does so randomly [77, 75, 19].

3.3 Optimal Rule and the Sampling Scheme

In an adaptive sequential trial, the process of generating $A(t)$ is controlled by the experimenter. As such, one can simultaneously learn and start assigning treatment according to the best current estimate of the optimal treatment rule, with varying exploration-exploitation objectives. In this section we describe different strategies for estimating the optimal treatment rule, as well as propose different sampling schemes for assigning treatment.

Estimating the Optimal Treatment Rule

First, we consider estimating the optimal treatment rule based on a parametric working model. As described previously, consider a treatment rule $C_o(t) \rightarrow d(C_o(t)) \in \{0, 1\}$ that maps the history $C_o(t)$ into a treatment decision for $A(t)$. We define a parametric working model for q_y indexed by parameter ϕ such that $\{q_{y,\phi} : \phi\}$. Notice that under the specified working model, we have that:

$$\bar{Q}_\phi(C_o(t), a) = E(Y(t) \mid C_o(t), A(t) = a) = \int y q_{y,\phi}(y \mid C_o(t), a) d\mu_y(y).$$

The true conditional treatment effect, $B_0(C_o(t))$, can then be expressed as

$$B_\phi(C_o(t)) = \bar{Q}_\phi(C_o(t), 1) - \bar{Q}_\phi(C_o(t), 0)$$

under the parametric working model. Recall that the optimal treatment rule for $A(t)$ for the purpose of maximizing $Y(t)$ is given by:

$$d_0(C_o(t)) = \mathbb{I}(B_0(C_o(t)) > 0).$$

Under the parametric working model, we note that the optimal treatment rule can be represented as:

$$d_\phi(C_o(t)) = \mathbb{I}(B_\phi(C_o(t)) > 0).$$

Let ϕ_{t-1} to be the maximum likelihood estimate of the true ϕ_0 based on the most current history, $\bar{O}(t-1)$, and according to the working model $q_{y,\phi}$. We could define the fixed dimensional history $C_o(t)$ such that for each time point t, ϕ_{t-1} is included in the relevant history $C_o(t)$ for $O(t)$. The current estimate of the rule is then defined as:

$$d_{\phi_{t-1}}(C_o(t)) = \mathbb{I}(B_{\phi_{t-1}}(C_o(t)) > 0).$$

If the parametric model is very flexible, $B_{\phi_{t-1}}$ might be a good approximation of the true conditional treatment effect $B_0(C_o(t))$. In that case, $d_{\phi_{t-1}}(C_o(t))$ is a good approximation of the optimal rule $d_0(C_o(t))$. Nevertheless, we argue that ϕ_{t-1} will converge to ϕ_0 defined by a Kullback-Leibler projection of the true $q_{y,0}$ onto the working model $\{q_{y,\phi} : \phi\}$. Consequently, the rule $d_{\phi_{t-1}}(C_o(t))$ will converge to a fixed $\mathbb{I}(B_0(C_o(t)) > 0)$ as t converges to infinity.

Instead of considering a parametric working model, we explore estimation of the optimal treatment rule based on more flexible, possibly nonparametric approaches drawn from the machine learning literature. As in the previous subsection, we define $B_{\bar{Q}_{t-1}}(C_o(t))$ to be an estimator of the true blip function, $B_0(C_o(t))$, based on the most recent observations up to time t, $\bar{O}(t-1)$. In particular, we consider estimators studied in previous work, including Online Super-Learner of $\bar{Q}_0$ which provides convenient computational and statistical properties for dense time-series data described elsewhere [135, 7, 79]. Additionally, we might consider ensemble machine learning methods that target B_0 directly [77]. As mentioned in the previous section, we can view $B_{\bar{Q}_{t-1}}(C_o(t))$ as just another univariate covariate extracted from the past, and include it in our definition of $C_o(t)$. If $B_{\bar{Q}_{t-1}}$ is consistent for B_0, then the rule $d_{\bar{Q}_{t-1}}(C_o(t))$ based on $B_{\bar{Q}_{t-1}}$ will converge to the optimal rule $\mathbb{I}(B_0(C_o(t)) > 0)$, as shown in previous work [77, 19].

Defining the Sampling Scheme

In the following subsection, we describe two sampling schemes that define $g^N = \{g_t : t = 1, \cdots, N\}$ precisely. Both rely on estimating parts of the likelihood based on the time-points collected so far for the single subject studied. The t-dependent current estimate of $\bar{Q}_0$ and B_0 are then further utilized to assign the next treatment, collect the next corresponding block of data, and estimate the target parameter of interest. Following the empirical process literature, we sometimes write $P_N f$ to be the empirical average of function f, and $Pf = \mathbb{E}_P f(O)$.

Stochastic Optimal Treatment Rules

Let $\bar{Q}_{t-1}$ denote the time t estimate of $\bar{Q}_0$ based on the time-series points collected so far, $\bar{O}(t-1)$. For a small number of samples, $d_{\bar{Q}_{t-1}}(C_o(t))$ might not be a good estimate of $d_0(C_o(t))$. As such, assigning treatment deterministically based on the current estimate of the rule could be ill-advised. In addition, without exploration (enforced via a deterministic rule), we cannot guarantee consistency of the optimal rule estimator. In light of that, we define $\{c_t\}_{t\geq 1}$ and $\{e_t\}_{t\geq 1}$ as user-defined, non-increasing sequences such that $c_1 \leq \frac{1}{2}$, $\lim_t c_t \equiv c_\infty > 0$ and $\lim_t e_t \equiv e_\infty > 0$. More specifically, we let $\{e_t\}_{t\geq 1}$ define the level of random perturbation around the current estimate $d_{\bar{Q}_{t-1}}(C_o(t))$ of the optimal rule. We define $\{c_t\}_{t\geq 1}$ as the probability of failure, so choosing $c_1 = \cdots = c_t = 0.5$ would yield a balanced stochastic treatment rule. In particular, we define a design that ensures that, under any context and with a positive probability c_t, we pick the treatment uniformly at random. This positive probability c_t is what is often referred to as the *exploration rate* in the bandit and reinforcement learning literature [120]. For every $t \geq 1$, we could have the following function G_t over $[-1, 1]$:

$$G_t(x) = c_t \mathbb{I}[x < -e_t] + (1 - c_t)\mathbb{I}[x \geq e_t] + \left(-\frac{1/2 - c_t}{2e_t^3}x^3 + \frac{1/2 - c_t}{2e_t/3}x + \frac{1}{2}\right)\mathbb{I}[-e_t \leq x \leq e_t],$$

where $G_t(x)$ is used to derive a stochastic treatment rule from an estimated blip function, such that

$$g_t(1 \mid C_o(t)) = G_t(B_{\bar{Q}_{t-1}}(C_o(t))).$$

Note that G_t is a smooth approximation to $x \to \mathbb{I}[x \geq 0]$ bounded away from 0 and 1, mimicking the optimal treatment rule as an indicator of the true blip function. With that in mind, any other non-decreasing k_n-Lipschitz function with $F_t(x) = c_t$ for $x < -e_t$ and $F_t(x) = 1 - c_t$ for $x \geq e_t$ would approximate the optimal treatment rule as well. The definitions of G_t and g_t prompt the following lemma, which illustrates the ability of the sampling scheme to learn form the collected data, while still exploring:

Lemma 3. *Let $t \geq 1$. Then we have that:*

$$\inf_{c_o(t)} g_t(d(c_o(t)) \mid c_o(t)) \geq \frac{1}{2}$$
$$\inf_{c_o(t)} g_t(1 - d(c_o(t)) \mid c_o(t)) \geq c_t.$$

Note that under Lemma 3, the positivity assumption needed for the identification result is met. Finally, we reiterate that the stochastic treatment rule $g_t(1 \mid C_o(t))$ approximates $d(C_o(t))$ in the following sense:

$$|g_t(1 \mid C_o(t)) - d(C_o(t))| \leq c_\infty \mathbb{I}[|B(C_o(t)) \geq e_\infty|] + \frac{1}{2}\mathbb{I}[|B(C_o(t)) < e_\infty|].$$

If c_∞ and e_∞ are small and $|B(C_o(t)) \geq e_\infty|$, then drawing treatment assignment from a smooth approximation of $d(C_o(t))$ is not much different than $d(C_o(t))$, with little impact on the mean value of the reward.

Target sequential sampling with Highly Adaptive Lasso

Alternatively, one could allocate randomization probabilities based on the tails of an estimate of the blip function, $B(C_o(t))$. In particular, we present a sampling scheme that utilizes the Highly Adaptive Lasso (HAL) estimator for obtaining the bounds around the estimate of the true blip function. The Highly Adaptive Lasso is a nonparametric regression estimator that does not rely on local smoothness assumptions [6, 130]. Briefly, for the class of functions that are right-hand continuous with left-hand limits and a finite variation norm, HAL is an MLE which can be computed based on L_1-penalized regression. As such, it is similar to standard lasso regression function in its implementation, except that the relationship between the predictors and the outcome is described by data-dependent basis functions instead of a parametric model. For a thorough description of the Highly Adaptive Lasso estimator, we refer the reader to [6] and [130].

We propose to use HAL to estimate $B_0(C_o)$, which implies an estimator for the optimal rule $d_0(C_o) = \mathbb{I}(B_0(C_o) > 0)$. We define a quadratic loss function as follows:

$$L_B(\theta)(o, C_o) = (D_1(\theta)(O) - B(C_o))^2,$$

which is indexed by $\theta = (g, \bar{Q})$ required to evaluate $D_1(\theta)(O)$. This influence function has the property that $E_0(D_1(\theta)|C_o) = B_0(C_o)$ if either $\bar{Q} = \bar{Q}_0$ or $g = g_0$, under positivity. As such, $L_B(\theta_0)$ is a double robust and efficient loss function for the true risk in the sense that $P_n L_B(\theta)$ is a double robust locally efficient estimator of the true risk under regularity conditions. As a double robust and efficient loss, the true risk of the loss function $L_B(\theta)(o, C_o)$ equals $P_0(B_0 - B)^2(C_o)$ up until a constant if either $D_1(\theta) = D_1(\bar{Q}_0, g)$ or $D_1(\theta) = D_1(\bar{Q}, g_0)$.

Let $E(D_1(\theta)|C_o) = \psi^{\text{blip}}$, with $\psi^{\text{blip}} \in D[0, \tau]$, the Banach space of d-variate cadlag functions. Define $C_{o,s} = \{C_{o,j} : j \in s\}$ for a given subset $s \subset \{1, ..., d\}$. For $\psi^{\text{blip}} \in D[0, \tau]$,

we define the s^{th} section of ψ^{blip} as $\psi_s^{\text{blip}}(c_o) = \psi^{\text{blip}}(c_{o,1}\mathbb{I}(1 \in s), \ldots, c_{o,d}\mathbb{I}(d \in s))$, where c_o denotes all possibilities of C_o. We assume the variation norm of ψ^{blip} is finite:

$$\|\psi^{\text{blip}}\|_v = \psi^{\text{blip}}(0) + \sum_{s \subset \{1,\ldots,d\}} \int_{0_s}^{\tau_s} |\psi_s^{\text{blip}}(du)| < M.$$

The HAL estimator represents ψ^{blip} as

$$\begin{aligned}\psi^{\text{blip}}(c_o) &= \psi^{\text{blip}}(0) + \sum_{s \subset \{1,\ldots,d\}} \int_{0_s}^{\tau_s} \psi_s^{\text{blip}}(du) \\ &= \psi^{\text{blip}}(0) + \sum_{s \subset \{1,\ldots,d\}} \int_{0_s}^{\tau_s} \mathbb{I}(u \leq c_{o,s}) \psi_s^{\text{blip}}(du),\end{aligned}$$

which uses a discrete measure ψ_m^{blip} with m support points to approximate this representation. For each subset s, at time $t = N$, we select as support points the N observed values $\tilde{c}_{o,s}(t)$, $t = 1, \ldots, N$, of the context $C_{o,s}(t)$. Then, for each subset s, we have a discrete approximation of ψ_s^{blip} with support defined by the actual N observations and point-masses $d_{\psi_{m,s,t}^{\text{blip}}}$, the pointmass assigned by ψ_m^{blip} to point $\tilde{c}_{o,s}(t)$, $t = 1, \ldots, N$. This approximation consists of a linear combination of basis functions $c_o \to \phi_{s,t}(c_o) = \mathbb{I}(c_{o,s} \geq \tilde{c}_{o,s}(t))$ with corresponding coefficients $d_{\psi_{m,s,t}^{\text{blip}}}$ summed over s and $t = 1, \ldots, N$. The minimization of the empirical risk $P_n L_B(\theta)(o, C_o)$ of this estimator, ψ_n^{blip}, corresponds to lasso regression with predictors $\phi_{s,t}$ across all subsets $s \subset \{1, \ldots, d\}$ and for $t = 1, \ldots, N$. That is, for

$$\psi_\beta^{\text{blip}} = \beta_0 + \sum_{s \subset \{1,\ldots,d\}} \sum_{t=1}^{N} \beta_{s,t} \phi_{s,t}$$

and corresponding subspace $\Psi_{n,M} = \{\psi_\beta : \beta, \beta_0 + \sum_{s \subset \{1,\ldots,d\}} \sum_{t=1}^{N} |\beta_{s,t}| < M\}$,

$$\beta_n = \text{argmin}_{\beta, \beta_0 + \sum_{s \subset \{1,\ldots,d\}} \sum_{t=1}^{N} |\beta_{s,t}| < M} P_n L_{B_\beta}(\theta).$$

The linear combination of basis function with non-zero coefficients in the HAL MLE represent a working model. We can use this data adaptively chosen parametric working model to obtain approximate (non-formal) inference for the blip function. For example, we could use the delta-method to obtain a Wald-type confidence interval for the blip function, recognizing that β_n is an MLE for this working model. Alternatively, we use the nonparametric bootstrap, fixing the model that was selected by HAL (to maintain $L1$-norm), and running lasso with the selected model for each bootstrap. We denote the resulting confidence interval bounds around the HAL MLE of the blip as $\pm\text{CI}(\psi_n^{\text{blip}})$, and propose using these bounds around the HAL MLE of the blip - that is, we would let $\pm\text{CI}(\psi_n^{\text{blip}})$ replace $\pm e_t$ in $G_t(x)$. Incorporating the Highly Adaptive Lasso blip estimate into the sampling scheme encourages exploitation of the known uncertainty in the blip estimates so far, allowing for more efficient use of the exploration step then the procedure described in subsection 3.3.

3.4 Targeted Maximum Likelihood Estimator

In the following, we build a Targeted Maximum Likelihood Estimator (TMLE) for the target parameter, $\bar{\Psi}(\theta)$ [140, 138, 137]. TML estimation is a multistep procedure, where one first obtains an estimate of the relevant parts of the data-generating distribution using machine learning algorithms and appropriate cross-validation [136, 7, 79]. The second stage updates the initial fit in a step targeted towards making an optimal bias-variance trade-off for $\bar{\Psi}(\theta)$, instead of the whole density.

Let $L(\bar{Q})(O(t), C_o(t))$ be a loss function for $\bar{Q}_0$ where $L(\bar{Q}) : \mathcal{O} \times \mathcal{C} \to \mathbb{R}$; for notational simplicity, we can also write $L(\bar{Q})$, with dependence on $(O(t), C_o(t))$ implied. In particular, we define $L(\bar{Q})$ as the quasi negative log-likelihood loss function,

$$L(\bar{Q}) = -[Y(t) \log \bar{Q}(C_o(t), A(t)) + (1 - Y(t)) \log(1 - \bar{Q}(C_o(t), A(t)))],$$

where the true $\bar{Q}_0$ minimizes the risk under the true conditional density $P_{0,C_0(t)}$:

$$P_{0,C_0(t)} L(\bar{Q}_0)(O(t), C_o(t)) = \min_{\bar{Q}} P_{0,C_o(t)} L(\bar{Q})(O(t), C_o(t)).$$

Let $\bar{Q}_N$ be an initial estimator of $\bar{Q}_0$, obtained via Online Super Learner and cross-validation suited for dependent data, such as the rolling-window or recursive-origin scheme [8, 7, 79]. For a $\bar{Q}_N$ in the statistical model, we define a parameteric working model $\{\bar{Q}_{N,\epsilon} : \epsilon\}$ through $\bar{Q}_N$ with finite-dimensional parameter ϵ; note that $\bar{Q}_{N,\epsilon=0} = \bar{Q}_N$. We define a parametric family of fluctuations of the initial estimator $\bar{Q}_N$ of $\bar{Q}_0$ along with the loss function, $L(\bar{Q})$, so that the linear combination of the components of the derivative of the loss evaluated at $\epsilon = 0$ span the efficient influence curve at the initial estimator:

$$\left\langle \frac{d}{d\epsilon} L(\bar{Q}_{N,\epsilon}) \Big|_{\epsilon=0} \right\rangle \supset D^*_{C_o(t)}(\bar{Q}_N),$$

where we used the notation $\langle S \rangle$ for the linear span of the components of the function S. We note that $\{\bar{Q}_{N,\epsilon} : \epsilon\}$ is known as the local least favorable submodel; one could also define a universal least favorable submodel, where the derivative of the loss evaluated at any ϵ will equal the efficient influence curve at the fluctuated initial estimator $\bar{Q}_{N,\epsilon}$ [134]. We proceed to maximize the log-likelihood over the parametric model:

$$\epsilon_N = \arg\min_{\epsilon} \frac{1}{N} \sum_{t=1}^{N} L(\bar{Q}_{N,\epsilon})(O(t), C_o(t)).$$

In order to perform the update of the conditional expectations, we rely on the logistic fluctuation model,

$$\text{logit}(\bar{Q}_{N,\epsilon}) = \text{logit}(\bar{Q}_N) + \epsilon H,$$

where H denotes the clever covariate specific to the target parameter, $H = \frac{g^*_t(A(t)|C_o(t)}{g_t(A(t)|C_o(t))}$. The TMLE update, denoted as $\bar{Q}^*_N = \bar{Q}_{N,\epsilon_N}$, is the TMLE of $\bar{Q}_0$ which solves the efficient score

equation,

$$\frac{1}{N}\sum_{t=1}^{N} D^*(\bar{Q}_N^*)(O(t), C_o(t)) \approx 0.$$

We define the TMLE as the plug-in estimator $\bar{\Psi}(\bar{Q}_N^*)$, obtained by evaluating $\bar{\Psi}$ at the last update of the estimator of $\bar{Q}_0$.

3.5 Asymptotic normality of the TMLE

Decomposition of the TMLE estimator

Our theoretical analysis relies on the fact that the difference between the TML estimator and the target can be decomposed as the sum of (1) the average of a martignale difference sequence, and (2) a martingale process for which we can show an equicontinuity result. We present formally this decomposition in theorem 9 below.

Theorem 9. *For any $\bar{Q}_1 \in \bar{\mathcal{Q}}$, the difference between the TMLE and its target decomposes as*

$$\bar{\Psi}(\bar{Q}_N^*) - \bar{\Psi}(\bar{Q}_0) = M_{1,N}(\bar{Q}_1) + M_{2,N}(\bar{Q}_N^*, \bar{Q}_1),$$

with

$$M_{1,N}(\bar{Q}_1) = \frac{1}{N}\sum_{t=1}^{N} D^*(\bar{Q}_1)(C_o(t), O(t)) - P_{0,C_o(t)} D^*(\bar{Q}_1),$$

$$M_{2,N}(\bar{Q}_N^*, \bar{Q}_1) = \frac{1}{N}\sum_{t=1}^{N} (\delta_{C_o(t),O(t)} - P_{0,C_o(t)})(D^*(\bar{Q}_N^*) - D^*(\bar{Q}_1)).$$

The term $M_{1,N}(\bar{Q}_1)$ is the average of a martingale difference sequence, and we will analyze it with a classical martingale central limit theorem. The second term is a martingale process indexed by $\bar{Q} \in \bar{\mathcal{Q}}$, evaluated at $\bar{Q} = \bar{Q}_N^*$. We will prove an equicontinuity result under a complexity condition for a process derived from the function class $\{D^*(\bar{Q}) : \bar{Q} \in \bar{\mathcal{Q}}\}$, which will imply that if $\bar{Q}_N^* \xrightarrow{P} \bar{Q}_1 \in \bar{\mathcal{Q}}$ then $M_{2,N}(\bar{Q}_N^*, \bar{Q}_1) = o_P(N^{-1/2})$.

Analysis of the term $M_{1,N}(\bar{Q}_1)$

A set of sufficient conditions for the asymptotic normality of the term $M_{1,N}(\bar{Q}_1)$ is that (a) the terms $D^*(\bar{Q}_1)(C_o(t), O(t))$ remain bounded, and (b) that the average of the conditional variances of $D^*(\bar{Q}_1)(C_o(t), O(t))$ stabilize. A sufficient condition for condition (a) to hold is the following strong version of the positivity assumption.

Assumption 17 (Strong positivity)**.** *There exists $\delta > 0$ such that, for every $t \geq 1$,*

$$g_{0,t}(A(t) \mid C_a(t)) \geq \delta, P_0\text{-a.s.}$$

Assumption 18 (Stabilization of conditional variances)**.** *There exists $\sigma_0^2(\bar{Q}_1) \in (0, \infty)$ such that*

$$\frac{1}{N} \sum_{t=1}^{N} Var_0 \left(D^*(\bar{Q}_1)(C_o(t), O(t)) \mid C_o(t) \right) \xrightarrow{d} \sigma_0^2(\bar{Q}_1).$$

Theorem 10. *Suppose that assumption 17 and assumption 18 hold. Then*

$$\sqrt{N} M_{1,N}(\bar{Q}_1) \xrightarrow{d} \mathcal{N}(0, \sigma_0^2(\bar{Q}_1)).$$

Proof. The result follows directly from various versions of martingale central limit theorems (e.g. theorem 2 in [17]). □

The conditional variances stabilize under (1) mixing and ergodicity conditions for the sequence $(C_o(t))$ of contexts, and if (2) the design $g_{0,t}$ stabilizes asymptotically. We discuss special cases in which these mixing and ergodicity conditions can be checked explicitly in the Appendix. For variance estimation we rely on the empirical variance estimator,

$$\widehat{\sigma}_N^2 := \frac{1}{N} \sum_{t=1}^{N} D^*(\bar{Q}_N^*, g_{0,t})^2 (C_o(t), O(t)),$$

which converges to the asymptotic variance $\sigma_0^2(\bar{Q}_1)$ of $M_{1,N}(\bar{Q}_1)$.

Negligibility of the term $M_{2,N}(\bar{Q}_N^*, \bar{Q}_1)$

In this susbsection, we give a brief overview of the analysis of the term $M_{2,N}(\bar{Q}_N^*, \bar{Q}_1)$. We show that $M_{2,N}(\bar{Q}_N^*, \bar{Q}_1) = o_P(N^{-1/2})$ by proving an equicontinuity result for the process $\{M_{2,N}(\bar{Q}, \bar{Q}_1) : \bar{Q} \in \bar{\mathcal{Q}}\}$. Our equicontinuity result relies on a measure of complexity for the process

$$\Xi_N := \left\{ \left(D^*(\bar{Q}, g_{0,t})(C_o(t), O(t)) - D^*(\bar{Q}_1, g_{0,t})(C_o(t), O(t)) \right)_{t=1}^{N} : \bar{Q} \in \bar{\mathcal{Q}} \right\}, \tag{3.16}$$

which we refer to as *sequential bracketing entropy*, introduced by [128] for the analysis of martingale processes. Further, we denote $N_{[]}(\epsilon, b, \Xi_N, \bar{O}(N))$ as the *sequential bracketing number* of Ξ_N corresponding to brackets of size ϵ. We provide theoretical derivations answering following two important questions (1) how to connect the sequential bracketing entropy of the process Ξ_N to a traditional bracketing entropy measure for the outcome model $\bar{\mathcal{Q}}$, and (2) how to obtain consistency of an estimator $\bar{Q}_N^*$ fitted from sequentially collected data. Answers to both of these questions entail bracketing entropy preservation results. Our equicontinuity result is consequently a sequential equivalent of similar results for i.i.d. settings (e.g. [126]) and similarly relies on a Donsker-like condition.

Assumption 19 (Sequential Donsker condition). *Define the sequential bracketing entropy integral as* $J_{[]}(\epsilon, b, \Xi_N, \bar{O}(N)) := \int_0^\epsilon \sqrt{\log(1 + \mathcal{N}_{[]}(u, b, \Xi_N, \bar{O}(N))} du$. *Suppose that there exists a function* $a : \mathbb{R}^+ \to \mathbb{R}^+$ *that converges to* 0 *as* $\delta \to 0$, *such that*

$$J_{[]}(\epsilon, b, \Xi_N, \bar{O}(N)) \leq a(\delta).$$

Note that a sufficient condition for assumption 19 to hold is that $\log(1+\mathcal{N}_{[]}(u, b, \Xi_N, \bar{O}(N)) \leq C\epsilon^{-p}$, with $p \in (0, 2)$ and $C > 0$ a constant that does not depend on N.

Assumption 20 (L_2 convergence of the outcome model). *It holds that* $\|\bar{Q}_N^* - \bar{Q}_1\|_{2,g^*,h_N} = o_P(1)$, *where* h_N *is the empirical measure* $h_N := N^{-1} \sum_{t=1}^N \delta_{C_o(t)}$.

Theorem 11 (Equicontinuity of the martingale process term). *Consider the process* Ξ_N *defined in equation* (3.16). *Suppose that assumptions 17, 19 and 20 hold. Then* $M_{2,N}(\bar{Q}_N^*, \bar{Q}_1) = o_P(N^{-1/2})$.

Asymptotic normality theorem

As an immediate corollary of Theorems 10 and 11, we have the following asymptotic normality result for our TML estimator.

Theorem 12 (Asymptotic normality of the TMLE). *Suppose that assumptions 17, 18, 19 and 20 hold. Then*

$$\sqrt{N}\left(\bar{\Psi}(\bar{Q}_N^*) - \bar{\Psi}(\bar{Q}_0)\right) \xrightarrow{d} \mathcal{N}(0, \sigma_0^2(\bar{Q}_1)).$$

The empirical variance estimator $\widehat{\sigma}_N^2$ converges in probability to $\sigma_0^2(\bar{Q}_1)$, which implies that

$$\widehat{\sigma}_N^{-1}\sqrt{N}\left(\bar{\Psi}(\bar{Q}_N^*) - \bar{\Psi}(\bar{Q}_0)\right) \xrightarrow{d} \mathcal{N}(0, 1).$$

Therefore, denoting $q_{1-\alpha/2}$ the $1-\alpha/2$-quantile of the standard normal distribution, we have that the confidence interval

$$\left[\bar{\Psi}(\bar{Q}_N^*) - \frac{q_{1-\alpha/2}\widehat{\sigma}_N}{\sqrt{N}}, \bar{\Psi}(\bar{Q}_N^*) + \frac{q_{1-\alpha/2}\widehat{\sigma}_N}{\sqrt{N}}\right]$$

has asymptotic coverage $1 - \alpha$ for the target $\bar{\Psi}(\bar{Q}_0)$.

3.6 Simulations

In this section we present simulation results concerning the adaptive learning of the optimal individualized treatment rule estimated using machine learning methods for a single time-series. We focus on the stochastic sampling scheme described in subsection 3.3, and explore performance of our estimator with different initial sample sizes and consequent sequential

updates. We consider binary outcome and treatment, but note that the results will be comparable for continuous bounded outcome. Finally, unless specified otherwise, we present coverage of the mean under the current estimate of the optimal individualized treatment rule at each update based on 500 Monte Carlo draws. We set the reference treatment mechanism to a balanced design, assigning treatment with probability 0.5 for the data draw used to learn the initial estimate of the optimal individualized treatment rule.

Simulation 1a

We explore a simple dependence setting first, emphasising the connection with i.i.d sequential settings. We data consists of a binary treatment ($A(t) \in \{0,1\}$) and outcome ($Y(t) \in \{0,1\}$). The time-varying covariate $W(t)$ decomposes as $W(t) \equiv (W_1(t), W_2(t))$ with binary W_1 and continuous W_2. The outcome Y at time t is conditionally drawn given $\{A(t), Y(t-1), W_1(t-1)\}$ from a Bernoulli distribution, with success probability defined as $1.5 * A(t) + 0.5 * Y(i-1) - 1.1 * W_1(i-1)$. We generate the initial sample of size $t = 1000$ and $t = 500$ by first drawing a set of four $O(t)$ samples randomly from binomial and normal distributions in order to have a starting point to initiate time dependence. After the first 4 draws, we draw $A(t)$ from a binomial distribution with success probability 0.5, $Y(t)$ from a Bernoulli distribution with success probability dependent on $\{A(t), A(t-1), Y(t-1), W_2(t-1)\}$, followed by $W_1(t)$ conditional on $\{Y(t-1), W_1(t-1), W_2(t-1)\}$ and $W_2(t)$ conditional on $\{A(t-1), Y(t-1), W_1(t-1)\}$. After $t = 1000$ or $t = 500$, we continue to draw $O(t)$ as above, but with $A(t)$ drawn from a stochastic intervention approximating the current estimate $d_{\bar{Q}_{t-1}}$ of the optimal rule $d_{\bar{Q}_0}$. This procedure is repeated until reaching a specified final time point indicating the end of a trial. Our estimator of $\bar{Q}_0$, and thereby the optimal rule d_0, is based on an online super-learner with an ensemble consisting of multiple algorithms, including simple generalized linear models, penalized regressions, HAL and extreme gradient boosting [27]. For cross-validation, we relied on the online cross-validation scheme, also known as the recursive scheme in the time-series literature. The sequences $\{c_t\}_{t\geq 1}$ and $\{e_t\}_{t\geq 1}$ are chosen constant, with $c_\infty = 10\%$ and $e_\infty = 5\%$. The TMLEs are computed at sample sizes a multiple of 200, and no more than 1800 (for initial $t = 1000$) or 1300 (for initial $t = 500$), at which point sampling is stopped. We use the coverage of asymptotic 95% confidence intervals to evaluate the performance of the TMLE in estimating the average across time t of the $d_{\bar{Q}_{t-1}}$-specific mean outcome. The exact data-generating distribution used is as

follows:

$$
\begin{aligned}
A(0:4) &\sim \text{Bern}(0.5) \\
Y(0:4) &\sim \text{Bern}(0.5) \\
W_1(0:4) &\sim \text{Bern}(0.5) \\
W_2(0:4) &\sim \text{Normal}(0,1) \\
A(4:t) &\sim \text{Bern}(0.5) \\
Y(4:t) &\sim \text{Bern}(expit(1.5 * A(i) + 0.5 * Y(i-1) - 1.1 * W_1(i-1))) \\
W_1(4:t) &\sim \text{Bern}(expit(0.5 * W_1(i-1) - 0.5 * Y(i-1) + 0.1 * W_2(i-1))) \\
W_2(4:t) &\sim \text{Normal}(0.6 * A(i-1) + Y(i-1) - W_1(i-1), sd = 1) \\
A(t:1800) &\sim d_{\bar{Q}_{t-1}} \\
Y(t:1800) &\sim \text{Bern}(expit(1.5 * A(i) + 0.5 * Y(i-1) - 1.1 * W_1(i-1))) \\
W_1(t:1800) &\sim \text{Bern}(expit(0.5 * W_1(i-1) - 0.5 * Y(i-1) + 0.1 * W_2(i-1))) \\
W_2(t:1800) &\sim \text{Normal}(0.6 * A(i-1) + Y(i-1) - W_1(i-1), sd = 1).
\end{aligned}
$$

From Table 3.1, we can see that the 95% coverage for the average across time of the counterfactual mean outcome under the current estimate of the optimal dynamic treatment approaches nominal coverage with increasing time-steps, for both $t = 500$ and $t = 1000$ length of the initial time-series. The mean conditional variance stabilizes with increasing time-steps, as illustrated in Table 3.2 and Figure 3.1A, thus satisflying assumption 18 necessary for showing asymptotic normality of the TML estimator.

Simulation 1b

In Simulation 1b, we explore the behavior of our estimator in case of more elaborate dependence. As in Simulation 1a, we only consider binary treatment ($A(t) \in \{0,1\}$) and outcome ($Y(t) \in \{0,1\}$), with binary and continuous time-varying covariates. We set the reference treatment mechanism to a balanced treatment mechanism assigning treatment with probability $P(A(t) = 1) = 0.5$, and generate the initial sample of size $t = (1000, 500)$ by sequentially drawing $W_1(t), W_2(t), A(t), Y(t)$. As before, upon the first $t = 1000$ or $t = 500$ time-points, we continue to draw $O(t)$ with $A(t)$ sampled from a stochastic intervention approximating the current estimate $d_{\bar{Q}_{t-1}}$ of the optimal rule $d_{\bar{Q}_0}$. The estimator of the optimal rule $d_{\bar{Q}_0}$ was based on an ensemble of machine learning algorithms and regression-based algorithms, with honest risk estimate achieved by utilizing online cross-validation scheme with validation set size of 30. The sequences $\{c_t\}_{t\geq 1}$ and $\{e_t\}_{t\geq 1}$ were set to 10% and 5%, respectively. The TMLEs are computed at initial $t = 1000$ or $t = 500$, and consequently at sample sizes being a multiple of 200, and no more than 1800 (or 1300), at which point sampling is stopped.

The exact data-generating distribution used is as follows:

$$
\begin{aligned}
A(0:4), Y(0:4), W_1(0:4) &\sim \text{Bern}(0.5) \\
W_2(0:4) &\sim \text{Normal}(0, 1) \\
A(4:t) &\sim \text{Bern}(0.5) \\
Y(4:t) &\sim \text{Bern}(expit(1.5 * A(i) + 0.5 * Y(i-3) - 1.1 * W_1(i-4))) \\
W_1(4:t) &\sim \text{Bern}(expit(0.5 * W_1(i-1) - 0.5 * Y(i-1) + 0.1 * W_2(i-2))) \\
W_2(4:t) &\sim \text{Normal}(0.6 * A(i-1) + Y(i-1) - W_1(i-2), sd = 1) \\
A(t:1800) &\sim d_{\bar{Q}_{t-1}} \\
Y(t:1800) &\sim \text{Bern}(expit(1.5 * A(i) + 0.5 * Y(i-3) - 1.1 * W_1(i-4))) \\
W_1(t:1800) &\sim \text{Bern}(expit(0.5 * W_1(i-1) - 0.5 * Y(i-1) + 0.1 * W_2(i-2))) \\
W_2(t:1800) &\sim \text{Normal}(0.6 * A(i-1) + Y(i-1) - W_1(i-2), sd = 1).
\end{aligned}
$$

As demonstrated in Table 3.1, the TML estimator approaches 95% coverage with increasing number of time points with more elaborate dependence structure as well. The assumption of stabilization of the mean of conditional variances is shown to be valid in Table 3.2 and Figure 3.1B, allowing for the asymptotic coverage $1 - \alpha$ for the target $\bar{\Psi}(\bar{Q}_0)$.

	t	$\mathbf{Cov}_t$	$\mathbf{Cov}_{t_1}$	$\mathbf{Cov}_{t_2}$	$\mathbf{Cov}_{t_3}$	$\mathbf{Cov}_{t_4}$
Simulation 1a	1000	92.60	94.00	95.20	95.40	95.80
Simulation 1a	500	90.00	93.20	93.80	94.80	94.60
Simulation 1b	1000	92.60	92.60	93.00	93.40	93.80
Simulation 1b	500	89.60	90.20	89.90	90.80	91.40

Table 3.1: The 95% coverage for the average across time of the counterfactual mean outcome under the current estimate of the optimal dynamic treatment at time points t, $t_1 = t + 200$, $t_2 = t + 400$, $t_3 = t + 600$ and $t_4 = t + 800$. The first t time points sample treatment with probability 0.5. The sequences $\{c_n\}_{t \geq 1}$ and $\{e_n\}_{t \geq 1}$ are chosen constant, with $c_\infty = 10\%$ and $e_\infty = 5\%$. TMLEs are computed at $t = \{500, 1000\}$, t_1, t_2, t_3 and t_4, with sequential updates being of size 200. The results are reported over 500 Monte-Carlo draws for Simulations 1a and 1b with initial sample sizes 1000 and 500.

	t	$\mathbf{Var}_t$	$\mathbf{Var}_{t_1}$	$\mathbf{Var}_{t_2}$	$\mathbf{Var}_{t_3}$	$\mathbf{Var}_{t_4}$
Simulation 1a	1000	0.0018	0.0019	0.0017	0.0016	0.0004
Simulation 1a	500	0.0011	0.0024	0.0035	0.0014	0.0011
Simulation 1b	1000	0.0072	0.0075	0.0069	0.0067	0.0018
Simulation 1b	500	0.0199	0.0171	0.0187	0.0152	0.0087

Table 3.2: Variance for the average across time of the counterfactual mean outcome under the current estimate of the optimal dynamic treatment at time points t, $t_1 = t+200$, $t_2 = t+400$, $t_3 = t + 600$ and $t_4 = t + 800$, over 500 Monte-Carlo draws for Simulations 1a and 1b with initial sample sizes 1000 and 500.

3.7 Discussion

In this chapter, we once again consider causal parameters based on observing a single time series with asymptotic results derived over time t. The data setup constitutes a typical longitudinal data structure, where within each t-specific time-block one observes treatment and outcome nodes, and possibly time-dependent covariates in-between treatment nodes. Each t-specific data record $O(t)$ is viewed as its own experiment in the context of the observed history $C_o(t)$, carrying information about a causal effect of the treatment nodes on the next outcome node. A key assumption necessary in order to obtain the presented results is that the relevant history for generating $O(t)$, given the past $\bar{O}(t-1)$, can be summarized by a fixed dimensional summary $C_o(t)$. We note that our conditions allow for $C_o(t)$ to be a function of the whole observed past, allowing us to avoid Markov-order type assumptions that limit dependence on recent, or specifically predefined past. Components of $C_o(t)$ that depend on the whole past, such as an estimate of the optimal treatment rule based on $(O(1), \ldots, O(t-1))$, will typically converge to a fixed function of a recent past - so that the martingale condition on the stabilization of the mean of conditional variances holds.

Due to the dimension reduction assumption, each t-specific experiment corresponds to drawing from a conditional distribution of $O(t)$ given $C_o(t)$. We assume that this conditional distribution is either constant in time or is parametrized by a constant function. As such, we can learn the true mechanism that generates the time-series, even when the model for the mechanism is nonparametric. With the exception of parametric models allowing for maximum likelihood estimation, we emphasize that statistical inference for proposed target parameters of the time-series data generating mechanism is a challenging problem which requires targeted machine learning.

The work of [131] and [65] studies marginal causal parameters, marginalizing over the distribution of $C_o(t)$, defined on the same statistical model as the parameter we consider in this article. In particular, [131] define target parameters and estimation of the counterfactual mean of a future (e.g., long term) outcome under a stochastic intervention on a subset of the treatment nodes, allowing for extensions to single unit causal effects. As such, the target

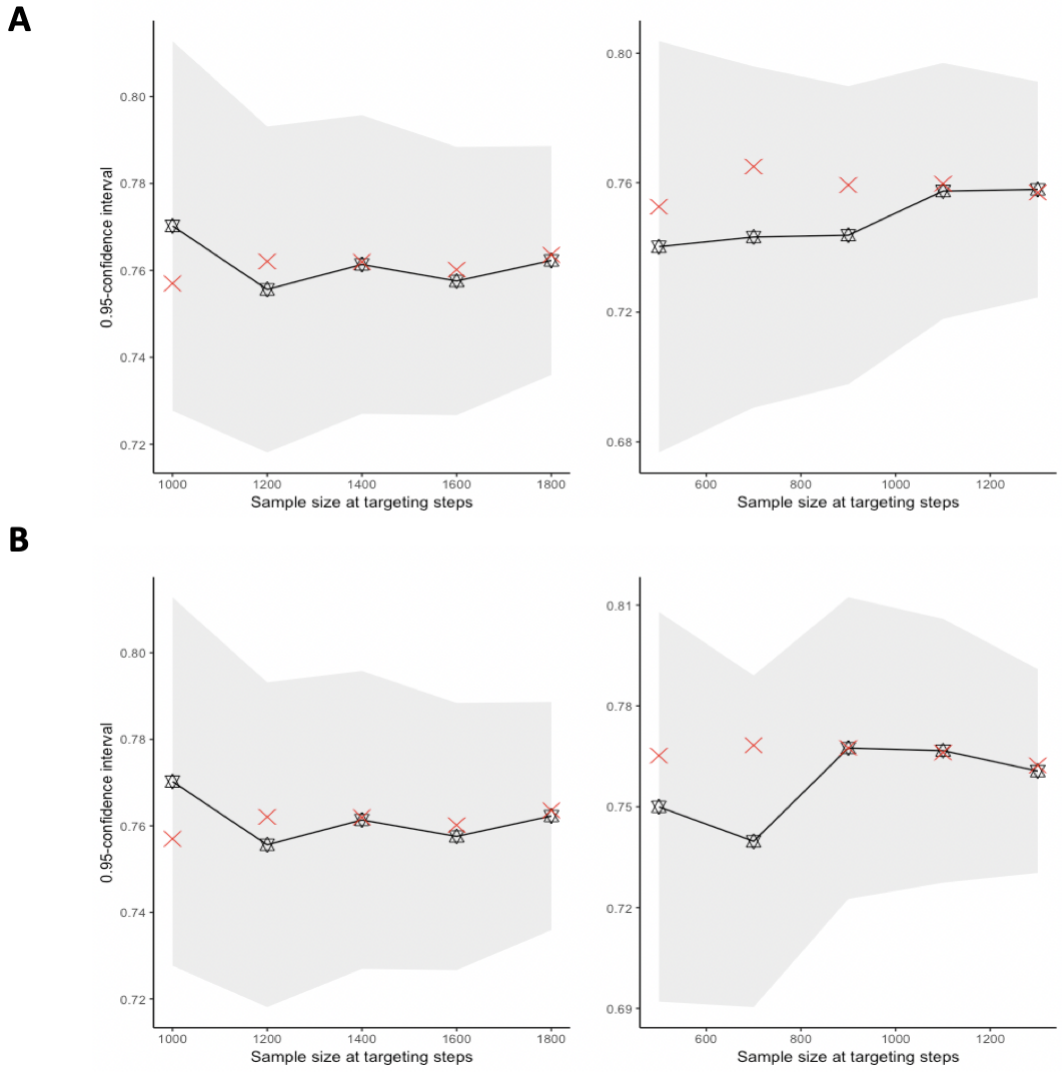


Figure 3.1: Illustration of the data-adaptive inference of the mean reward under the optimal treatment rule with initial sample size $n = 1000$ and $n = 500$ for Simulation 1a and 1b. The red crosses reflect successive values of the data-adaptive true parameter, with stars representing the estimated parameter with the corresponding 95% confidence interval for the data-adaptive parameter.

parameter proposed by [131] addresses the important question regarding the distribution of the outcome at time t, had we intervened on some of the past treatment nodes in a (possibly single) time-series. While important, the TMLE of such target parameters are challenging to implement due to their reliance on the density estimation of the marginal density of $C_o(t)$ averaged across time t. Additionally, we remark that such marginal causal parameters

cannot be robustly estimated if treatment is sequentially randomized, due to the lack of double robustness of the second order remainder.

In this work, we focus on a context-specific target parameter is order to explore robust statistical inference for causal questions based on observing a single time series of a particular unit. We note that for each given $C_o(t)$, any intervention-specific mean outcome $EY_{g^*}(t)$ with g^* being a stochastic intervention w.r.t. the conditional distribution of $P_{C_o(t)}$ (with deterministic rule being a special case), represents a well studied statistical estimation problem based on observing many i.i.d. copies. Even though we do not have repeated observations from the $C_o(t)$-specific distribution at time t, the collection $(C_o(t), O(t))$ across all time points represent the analogue of an i.i.d. data set $(C_o(t), O(t)) \sim_{iid} P_0$, where $C_o(t)$ can be viewed as a baseline covariate for the longitudinal causal inference data structure; we make the connection with the i.i.d. sequential design in one of our simulations. The initial estimation step of the TMLE should still respect the known dependence in construction of the initial estimator, by relying on appropriate estimation techniques developed for dependent data. Similarly, variance estimation can proceed as in the i.i.d case using the relevant i.i.d. efficient influence curve. This insight relies on the fact that the TMLE in this case allows for the same linear approximation as the TMLE for i.i.d. data, with the martingale central limit theorem applied to the linear approximation instead. Since the linear expansion of the time-series TMLE for context-specific parameter is an element of the tangent space of the statistical model, our derived TMLE is asymptotically efficient.

Our motivation for studying the proposed context-specific parameter strives from its important role in precision medicine, in which one wants to tailor the treatment rule to the individual observed over time. In particular, we derive a TMLE which uses only the past data $\bar{O}(t-1)$ of a single unit in order to learn the optimal treatment rule for assigning $A(t)$ to maximize the mean outcome $Y(t)$. Here, we assign the treatment at the next time point $t+1$ according to the current estimate of the optimal rule, allowing for the time-series to learn and apply the optimal treatment rule at the same time. The time-series generated by the described adaptive design within a single unit can be used to estimate, and most importantly provide inference, for the average across all time-points t of the counterfactual mean outcome of $Y(t)$ under the estimate $d(C_o(t))$ of the optimal rule at a relevant time point t. Assuming that the estimate of the optimal rule is consistent, as the number of time-points increases, our target parameter converges to the mean outcome one would have obtained had they carried out the optimal rule from the start. As such, we can effectively learn the optimal rule and simultaneously obtain valid inference for its performance. Interestingly, this does not provide inference relative to, for example, the control that always assigns $A(t) = 0$. This is due to the fact that by assigning treatment $A(t)$ according to a rule, the positivity assumption needed to learn $\frac{1}{N}\sum_t E(Y_{A(t)=0}(t) \mid C_o(t))$ is violated. However, we note that one can safely conclude that one will not be worse than this control rule, even when the control rule is equal to the optimal rule. If one is interested in inference for a contrast based on a single time-series, then we advocate for random assignment between the control and estimate of optimal rule. As such, our proposed methodology still allows to learn the desired contrast.

Finally, we note that while the context-specific parameter enjoys many important statistical and computational advantages as opposed to the marginal target parameter based on a single time-series, the formulation employed in this article is only sensible if one is interested in the causal effect of treatment on a short-term outcome. In particular, if the amount of time necessary to collect outcome $Y(t)$ in $O(t)$ is long, then generating a long time series would take too much time to be practically useful. If one is interested in causal effects on a long term outcome and is willing to forgo utilizing known randomization probabilities for treatment, we advocate for the marginal target parameters as described in previous work by [131] or [65].

3.8 Appendix

Comparison with marginal parameters

We present below two alternative statistical parameters defined on the same statistical model as the parameter we consider in this article, and which were considered in previous works [131, 65]. The parameters are *marginal*, as opposed to context-specific parameters we consider in the present article. The definition of the marginal parameters entails integrating against certain marginal distributions of contexts, as we make explicit below. Let $(O^*(t))_{t=1}^{\infty} \sim P_{Q,g^*}$, with $O^*(t) = (A^*(t), Y^*(t), W^*(t))$. Consider the distribution P_{Q,g^*} over infinite sequences taking values in the infinite cartesian product space $\times_{t=1}^{\infty}\mathcal{O}$, defined from the factors of $P \in \mathcal{M}$ by the following G-computation formula:

$$P_{Q,g^*}\left((o(t))_{t=1}^{\infty}\right) := P_{C_o(1)}(c_o(1)) \prod_{t=1}^{\infty} g^*(a(t) \mid c_o(t)) Q(y(t) \mid c_o(t)) Q_w(w(t) \mid c_o(t)).$$

Marginal parameter by van der Laan et al. 2018 [131]

As a first example of a marginal parameter, [131] consider a class of parameters which includes

$$\Psi_{1,\tau}(P) := E_{Q,g^*}[Y^*(\tau)],$$

for $\tau \geq 1$. Under the causal identifiability assumptions, $\Psi_{1,\tau}(P_0)$ equals the mean outcome we would obtain at time τ, under a counterfactual time series with initial context distribution $P_{0,C_o(1)}$ and intervention g^* (instead of the observed intervention g) at every time point. We note that $P_{0,C_o(1)}$ is the initial, observed data-generating distribution. The canonical gradient of $\Psi_{1,\tau}$ w.r.t. our model $\mathcal{M}$ (where $\mathcal{M}$ assumes $P_{C_o(1)}$ known[1]) is

$$D^*(P)(o^N) := \frac{1}{N}\sum_{t=1}^{N} \bar{D}(Q, \omega, g)(c_o(t), o(t))$$

with

$$\bar{D}(Q,\omega,g)(c_o,o) := \sum_{s=1}^{\tau} \omega_s(c) \frac{g^*(a \mid c_o)}{g(a \mid c_o)} \{E_{Q,g^*}[Y^*(\tau) \mid O^*(s) = o, C_o^*(s) = c_o] - E_{Q,g^*}[Y^*(\tau) \mid A^*(s) = a, C_o^*(s) = c_o]\},$$

[1]If we instead supposed that $P_{C_o(1)}$ is unknown and lies in a certain model $\mathcal{M}_{P_{C_o(1)}}$, the canonical gradient would have one additional component, which would be lying in the tangent space of $\mathcal{M}_{P_{C_o(1)}}$. As far as the conditional parameter of the main text are concerned, this distinction has no effect, as these do not depend on the marginal distribution of contexts and therefore its canonical gradient has no components in the tangent spaces corresponding to the context distributions.

with $\omega_s(c_o) = h_{C_o^*(s)}(c_o)/\bar{h}_N(c_o)$, where

$$\begin{aligned} h_{C_o(s)}(c_o) =& P_{Q,g}[C_o(s) = c_o], \\ \bar{h}_N(c_o) =& \frac{1}{N}\sum_{t=1}^{N} h_{C_o(t)}(c_o), \\ \text{and } h_{C_o^*(s)}(c_o) =& P_{Q,g^*}[C_o^*(s) = c_o] \end{aligned}$$

are the marginal density of context $C_o(s)$ under P, the average thereof over observed time points $t = 1, \ldots, N$, and the marginal density of context $C_o^*(s)$ under P_{Q,g^*}. We note that $\Psi_{1,1}$ is the marginal equivalent of our parameter $\Psi_{C_o(1)}$. Specifically,

$$\Psi_{1,1}(P) = \int dP_{C_o(1)}(c_o(1))\Psi_{c_o(1)}(P).$$

Marginal parameter by Kallus and Uehara, 2019 [65]

Let $\gamma \in (0, 1)$. Kallus and Uehara [65] consider the parameter

$$\begin{aligned} \Psi_2(P) :=& E_{Q,g^*}\left[\sum_{\tau=1}^{\infty} \gamma^\tau Y^*(\tau)\right] \\ =& \sum_{\tau \geq 1} \gamma^\tau \Psi_{1,\tau}(P). \end{aligned}$$

Under the causal identifiability assumptions, $\Psi_2(P_0)$ is the expected total discounted outcome from time point 1 until ∞ that we would get if we carried out intervention g^* forever — starting from initial context distribution $P_{0,C_o(1)}$, as in the observed data generating distribution. The canonical gradient Ψ_2 w.r.t. $\mathcal{M}$ (again, supposing that $\mathcal{M}$ considers $P_{0,C_o(1)}$ known) is

$$D^*(P)(o^N) := \frac{1}{N}\sum_{t=1}^{N} \bar{D}(Q, \omega, g)(c_o(t), o(t)),$$

with

$$\bar{D}(Q, \omega, g)(c_o, o) := \sum_{s=1}^{\infty} \omega_s(c_o) \frac{g^*(a \mid c_o)}{g(a \mid c_o)} \left\{y + \gamma V_{1,Q,g^*}(c_o, o) - V_{2,Q,g^*}(c_o, a)\right\},$$

where ω_s is defined as in the previous example, and

$$\begin{aligned} V_{1,Q,g^*}(c_o, o) :=& E_{Q,g^*}\left[\sum_{\tau \geq 2} \gamma^\tau Y^*(\tau) \mid C_o^*(1) = c_o, O^*(1) = o\right] \\ \text{and } V_{2,Q,g^*}(c_o, o) :=& E_{Q,g^*}\left[\sum_{\tau \geq 1} \gamma^\tau Y^*(\tau) \mid A^*(1) = a, O^*(1) = o\right]. \end{aligned}$$

Robustness properties

In this article we are concerned with adaptive trials where the intervention is controlled by the experimenter, hence g_0 is known; we therefore only consider the case $g = g_0$. Under $g = g_0$, both parameters $\Psi' \in \{\Psi_{1,\tau}, \Psi_2\}$ defined above admit a first order expansion of the form

$$\Psi'(P) - \Psi'(P_0) = -P_0 D^*(P) + R'(Q, Q_0, \omega, \omega_0),$$

where R' is a second-order remainder term such that $R(Q, Q_0, \omega, \omega_0) = 0$ if either $Q = Q_0$ and $\omega = \omega_0$. While this resembles a traditional double-robustness property, as that which holds in the i.i.d. setting for the ATE or in the time series setting for our conditional parameter (as opposed to arbitrary time-series dependence or Markov decision process) it is important to note the following:

1. For $\Psi' \in \{\Psi_{1,\tau}, \Psi_2\}$, knowledge of the treatment mechanism is not sufficient to guarantee that the remainder term is zero; we direct the interested reader to [131] for the exact form of R'.

2. The parameters ω and Q are not variation independent, as appears explicitly from the definition of ω_s. In fact, when estimating ω_s from a single time series, one must a priori rely on an estimator of Q to obtain estimates of ω_s (see [131]). Therefore, if the estimator of Q is inconsistent, the corresponding estimator of ω_s will be inconsistent as well.

Stabilization of conditional variances

Assumption 18 on the stabilization of the conditional variance of the canonical gradient can be checked under mixing conditions on the sequence of context $(C_o(t))$, and under the condition that the design $g_{0,t}$ converges to a fixed design. We state formally below such a set of conditions.

Assumption 21 (Convergence of the marginal law of contexts)**.** *Suppose that the marginal law of contexts converges to a limit law, that is* $C_o(t) \xrightarrow{d} C_\infty$*, for some random variable* C_∞*.*

Definition 4 (ρ-mixing)**.** *Consider a couple of random variables* $(Z_1, Z_2) \sim P$*. The* ρ*-mixing coefficient, or maximum correlation coefficient of* Z_1 *and* Z_2 *is defined as*

$$\rho_P(Z_1, Z_2) := \sup \left\{ Corr(f_1(Z_1), f_2(Z_2)) : f_1 \in L_2(P_{Z_1}), f_2 \in L_2(P_{Z_2}) \right\}.$$

Assumption 22 (ρ-mixing condition)**.** *Suppose that*

$$\sup_{t \geq 1} \sum_{s=1}^{N} \rho(C_o(t), C_o(t+s)) = o(N).$$

Observe that if g is common across time points, the process $(C_o(t))$ is an homogeneous Markov chain. Conditions under which homogeneous Markov chains have marginal law converging to a fixed law and are mixing have been extensively studied. A textbook example, albeit perhaps a bit too contrived for many specifications of the setting of our current chapter, is when the Markov chain has finite state space and the probability of transitioning between any two states from one time point to the next is non-zero. In this case, ergodic theory shows that the transition kernel of the Markov chain admits a so-called invariant law - the marginal laws converge exponentially fast (in total variation distance) to the invariant law, and the mixing coefficients have finite sum. We refer the interested reader to the survey paper by [16] for more general conditions under which Markov chains have convergent marginal laws and are strongly mixing (for various types of mixing coefficients, one of them being ρ-mixing)

Assumption 23 (Design stabilization)**.** *There is a design* g_∞ *such that* $\|g_{0,t} - g_\infty\|_{1, P_{g^*, h_{0,t}}} = o(1)$*, and* $g_\infty \geq \delta$*, for some* $\delta > 0$*.*

We note that, as we will always use assumption 23 along with assumption 17, we will suppose that the constant δ in the statement of both assumptions is the same.

Lemma 4 (Conditional variance stabilization under mixing)**.** *Suppose that assumptions 17, 21 and 22 hold. Then assumption 18 holds.*

Analysis of the martingale process term

We analyze the martingale process $\{M_{2,N}(\bar{Q}, \bar{Q}_1) : \bar{Q} \in \bar{\mathcal{Q}}\}$ under a measure of complexity introduced by [128], which we will refer to in the present work as *sequential bracketing entropy*. We state below the definition of sequential bracketing entropy particularized to our setting.

Definition 5 (Sequential bracketing entropy)**.** *Consider a stochastic process of the form* $\Xi_N := \{(\xi_t(f))_{t=1}^N : f \in \mathcal{F}\}$ *where* $\mathcal{F}$ *is an index set such that, for every* $f \in \mathcal{F}$*,* $t \in [N]$*,* $\xi_t(f)$ *is an* $\bar{O}(t)$*-measurable real valued random variable. We say that a collection of random variables of the form* $\mathcal{B} := \{(\Lambda_t^j, \Upsilon_t^j)_{t=1}^N : j \in [J]\}$ *is an* $(\epsilon, b, \bar{O}(N))$ *bracketing of* Ξ_N *if*

1. *for every* $t \in [N]$*, and* $j \in [J]$*,* $(\Lambda_t^j, \Upsilon_t^j)$ *is* $\bar{O}(t)$*-measurable,*
2. *for every* $f \in \mathcal{F}$*, there exists* $j \in [J]$*, such that, for every* $t \in [J]$*,* $\Lambda_t^j \leq \xi_t(f) \leq \Upsilon_t^j$*,*
3. *for every* $t \in [N]$*,* $j \in [J]$*,* $|\Lambda_t^j - \Upsilon_t^j| \leq b$ *a.s.,*
4. *for every* $j \in [J]$*,*

$$\frac{1}{N}\sum_{t=1}^{N} E\left[(\Upsilon_t^j - \Lambda_t^j)^2 | \bar{O}(t-1)\right] \leq \epsilon^2.$$

We denote $\mathcal{N}_{[]}(\epsilon, b, \Xi_N, \bar{O}(N))$ *the minimal cardinality of an* $(\epsilon, b, \Xi_N, \bar{O}(N))$*-bracketing.*

Applied to our problem, observe that the process $\{M_{2,N}(\bar{Q}, \bar{Q}_1) : \bar{Q} \in \mathcal{Q}\}$ is derived from the process

$$\Xi_N := \left\{ \left((D^*(\bar{Q}) - D^*(\bar{Q}_1))(C_o(t), O(t)) \right)_{t=1}^N : \bar{Q} \in \bar{\mathcal{Q}} \right\}.$$

Natural questions that arise are (1) how to connect the sequential bracketing entropy of the process Ξ_N to a traditional bracketing entropy measure for the outcome model $\bar{\mathcal{Q}}$, and (2) how to obtain consistency of an estimator $\bar{Q}_N^*$ fitted from sequentially collected data. Answers to both of these questions entail bracketing entropy preservation results that we present in the upcoming subsection, 3.8.

We emphasize that the notion of *sequential covering numbers*, and the corresponding *sequential covering entropy* introduced by [101], represent a measure of complexity under which one can control martingale processes and obtain equicontinuity results. One motivation for the development of the notion of sequential covering numbers is that results that hold for i.i.d. empirical processes under traditional covering entropy conditions do not hold for martingale processes. Interestingly, while classical covering number conditions cannot be used to control martingale processes, classical bracketing number bounds can usually be turned into sequential bracketing number bounds. Our choice to state results in terms of

one measure of sequential complexity rather than the other (or both) is motivated by concision purposes, and also by the fact that we know how to bound bracketing entropy of a certain class of statistical models we find realistic in many applications, as we describe in later subsections.

Bracketing preservation results

We formalize the connection between the sequential bracketing entropy of the process Ξ_N to a traditional bracketing entropy measure for the outcome model $\bar{\mathcal{Q}}$ in lemma 5 below. In particular, lemma 5 bounds the sequential bracketing entropy of the canonical gradient process Ξ_N in terms of the bracketing entropy of the outcome model $\bar{\mathcal{Q}}$ w.r.t. a norm defined below.

Lemma 5. *Suppose that assumption 17 holds. Then*

$$\mathcal{N}_{[]}(\epsilon, \Xi_N, \bar{O}(N)) \lesssim N_{[]}(\epsilon, \bar{\mathcal{Q}}, L_2(P_{g^*,h_N})),$$

where $P_{g*,h_N}(a,c) = g^*(a \mid c)h_N(c)$*, with* h_N *being the empirical measure* $h_N := N^{-1}\sum_{t=1}^{N}\delta_{C_o(t)}$.

Proof. Suppose $\mathcal{B} = \{(\lambda_j, \upsilon_j) : j \in [J]\}$ is an ϵ-bracketing in $L_2(P_{g^*,h_N})$ norm of $\bar{\mathcal{Q}}$. Let $\bar{Q} \in \mathcal{Q}$. There exists $j \in [J]$ such that $\lambda_j \leq \bar{Q} \leq \upsilon_j$. Without loss of generality, we can suppose that $0 \leq \lambda_j \leq \upsilon_j \leq 1$, since the bracket $(\lambda_j \vee 0, \upsilon_j \wedge 1)$ brackets the same functions of $\bar{\mathcal{Q}}$ as (λ_j, υ_j), as every element of $\bar{\mathcal{Q}}$ has range in $[0,1]$. We have that

$$D^*(\bar{Q}) - D^*(\bar{Q}_1) = \frac{g^*}{g_{0,t}}(\bar{Q}_1 - \bar{Q}) + \sum_{a=1}^{2} g^*(a \mid \cdot)(\bar{Q} - \bar{Q}_1))(a, \cdot).$$

Denoting

$$\Lambda_t^j := \frac{g^*}{g_{0,t}}(\bar{Q}_1 - \upsilon_j) + \sum_{a=1}^{2} g^*(a \mid \cdot)(\lambda_j - \bar{Q}_1))(a, \cdot),$$
$$\text{and } \Upsilon_t^j := \frac{g^*}{g_{0,t}}(\bar{Q}_1 - \lambda_j) + \sum_{a=1}^{2} g^*(a \mid \cdot)(\upsilon_j - \bar{Q}_1))(a, \cdot),$$

we have that

$$\Lambda_t^j \leq (D^*(\bar{Q}, g_{0,t}) - D^*(\bar{Q}, g_{0,t})(C_o(t), O(t)) \leq \Upsilon_t^j.$$

We now check the size of the sequential bracket $(\Lambda_t^j, \Upsilon_t^j)_{t=1}^N$. We have that

$$
\begin{aligned}
&\frac{1}{N}\sum_{t=1}^{N} E_{Q_0,g}\left[(\Upsilon_t^j - \Lambda_t^j)^2 \mid \bar{O}(t-1)\right] \\
&= \frac{1}{N}\sum_{t=1}^{N} E_{Q_0,g_0}\left[\left\{\frac{g^*}{g_{0,t}}(\upsilon_j - \lambda_j)(A(t), C_o(t)) + \sum_{a=1}^{2}(g^*(\upsilon_j - \lambda_j))(a, C_o(t))\right\}^2 \mid C_o(t)\right] \\
&\leq \frac{2}{N}\sum_{t=1}^{N} E_{Q_0,g_0}\left[\left(\frac{g^*}{g_{0,t}}\right)^2 (\upsilon_j - \lambda_j)^2(A(t), C_o(t)) \mid C_o(t)\right] \\
&+ E_{Q_0,g^*}\left[(\upsilon_j - \lambda_j)(A(t), C_o(t))\right]^2 \\
&\leq \frac{4\delta^{-1}}{N}\sum_{t=1}^{N} E_{Q_0,g^*}\left[(\upsilon_j - \lambda_j)^2(A(t), C_o(t)) \mid C_o(t)\right] \\
&= 4\delta^{-1}\|\upsilon_j - \lambda_j\|^2_{2,P_{g^*,h_N}} \\
&\leq 4\delta^{-1}\epsilon^2,
\end{aligned}
$$

where we have used assumption 17 and Jensen's inequality in the fourth line above. From assumption 17, it is also immediate to check that $|\Upsilon_t^j - \Lambda_t^j| \leq 2\delta^{-1}$. So far, we have proven that one can construct a $(2\delta^{-1/2}\epsilon, 2\delta^{-1}, \bar{O}(N))$ bracketing of Ξ_N from an ϵ-bracketing in $L_2(P_{g^*,h_N})$ norm of $\bar{\mathcal{Q}}$. Treating δ as a constant, this implies that $\log N_{[]}(\epsilon, 2\delta^{-1}, \Xi_N, \bar{O}(N)) \lesssim \log N_{[]}(\epsilon, \bar{\mathcal{Q}}, L_2(P_{g^*,h_N}))$. □

When proving consistency and convergence rate results for the outcome model estimator $\bar{Q}_N^*$, we need bounds on the sequential bracketing entropy of the following martingale process:

$$
\mathcal{L}_N := \left\{\left(\ell_t(\bar{Q})(C_o(t), O(t))\right)_{t=1}^N : \bar{Q} \in \bar{\mathcal{Q}}\right\},
$$

where $\ell_t(\bar{Q})(c, o) := (g^*(a \mid c)/g_{0,t}(a \mid c))(\ell(\bar{Q})(o) - \ell(\bar{Q}_1)(o))$, with $\ell(f)$ denoting a loss function. We refer to $\mathcal{L}_N$ as an *inverse propensity weighted loss process*. Lemma 4 in [10] provides conditions, which hold for most common loss functions, under which the bracketing entropy of the loss class $\{\ell(f)(\bar{Q}) : \bar{Q} \in \bar{\mathcal{Q}}\}$ is dominated up to a constant by the bracketing entropy of $\bar{\mathcal{Q}}$. As a direct corollary of this lemma, we state the following result on the sequential bracketing entropy of the process $\mathcal{L}_N$; we refer to [10] for examples of common settings where assumption 24 is satisfied.

Assumption 24. *The loss function can be written as $\ell(\bar{Q})(c, a, y) = \widetilde{\ell}(\bar{Q}(c, a), y)$, where $\widetilde{\ell}$ satisfies the following conditions:*

- *for all f, c, a, $y \mapsto \widetilde{\ell}(\bar{Q}(c, a), y)$ is unimodal,*
- *for any y, $u \mapsto \widetilde{\ell}(u, y)$ is L-Lispchitz, for $L = O(1)$.*

Lemma 6 (Sequential bracketing entropy of loss process)**.** *Suppose that assumptions 24 and 17 hold. Then*

$$\mathcal{N}_{[]}(\epsilon, \mathcal{L}_N, \bar{O}(N)) \lesssim N_{[]}(\epsilon, \bar{\mathcal{Q}}, L_2(P_{g^*,h_N})).$$

Convergence rate of sequentially fitted outcome model estimators

In this subsection, we give convergence guarantees for outcome model estimators $\bar{Q}_N$, and their targeted counterpart $\bar{Q}_N^*$, fitted on sequentially collected data. We first give convergence rate guarantees for empirical risk minizers $\bar{Q}_N$ over a class $\bar{\mathcal{Q}}$, in terms of the bracketing entropy in $L_2(P_{g^*,h_N})$-norm of $\bar{\mathcal{Q}}$. As briefly defined in section 3.4, let $\ell = L$ be a loss function for the outcome regression such that, for every $\bar{Q} : \mathcal{C} \times \mathcal{A} \to [0, 1]$, we have that

$$\bar{Q}_0 \in \arg \min_{\bar{Q}\text{-measurable}} P_{Q_0,g^*,h_N} \ell(\bar{Q}).$$

We denote $R_{0,N}(\bar{Q}) := P_{Q_0,g^*,h_N}\ell(\bar{Q})$ as the population risk; we note that this population risk is equal to the average across t of the conditional risks $P_{Q_0 g^*,h_N}\ell(\bar{Q})$ given $C_o(t)$. Let $\bar{Q}^*$ be a minimizer of $R_{0,N}(\bar{Q})$ over $\bar{\mathcal{Q}}$. We further define the empirical risk as

$$\widehat{R}_N(\bar{Q}) := \frac{1}{N} \sum_{t=1}^{N} \frac{g^*}{g_{0,t}}(A(t) \mid C_o(t))\ell(\bar{Q})(C_o(t), O(t)).$$

Note that the empirical risk minimizer over $\bar{\mathcal{Q}}$ is any minimizer over $\bar{\mathcal{Q}}$ of $\widehat{R}_N(\bar{Q})$; as such, we use importance sampling weighting factor $g^*/g_{0,t}$ in front of each term $\ell(\bar{Q})(C_o(t), O(t))$. This choice is motivated by the fact that we want convergence rates guarantees for $\bar{Q}_N$ in $L_2(P_{g^*,h_N})$, as is natural to control the size of the sequential brackets of the canonical gradient process Ξ_N in terms of the size of brackets of $\bar{\mathcal{Q}}$ in $L_2(P_{g^*,h_N})$ norm (see lemma 5). In the following, we state the entropy condition and additional assumptions on the loss function.

Assumption 25 (Entropy of the outcome model)**.** *Suppose that there exists $p > 0$ such that*

$$\log(1 + N_{[]}(\epsilon, \bar{\mathcal{Q}}, L_2(P_{g^*,h_N}))) \leq \epsilon^{-p}.$$

Assumption 26 (Variance bound for the loss)**.** *Suppose that*

$$\|\ell(\bar{Q}) - \ell(\bar{Q}^*)\|^2_{2,\bar{Q}_0,g^*,h_N} \lesssim R_{0,N}(\bar{Q}) - R_{0,N}(\bar{Q}^*)$$

for all $\bar{Q} \in \bar{\mathcal{Q}}$.

Assumption 27 (Excess risk dominates L_2 norm)**.** *Suppose that*

$$\|\bar{Q} - \bar{Q}^*\|_{2,g^*,h_N} \lesssim R_{0,N}(\bar{Q}) - R_{0,N}(\bar{Q}^*).$$

Theorem 13. *Consider an empirical risk minimizer $\bar{Q}_N$ over $\bar{\mathcal{Q}}$, and a population minimizer $\bar{Q}^*$, as defined above. Suppose that assumptions 25, 26, 27, and assumption 24 hold. Then,*

$$\|\bar{Q}_N - \bar{Q}^*\|_{2,g^*,h_N} = \begin{cases} O_P(N^{-\frac{1}{1+p/2}}) & \text{if } p < 2, \\ O_P(N^{-\frac{1}{p}}) & \text{if } p > 2. \end{cases}$$

Proof. Consider the process $\mathcal{L}_N$ defined in subsection 3.8. We define $M_{0,N}(\bar{Q}, \bar{Q}^*)$ and $\widehat{M}_N(\bar{Q}, \bar{Q}^*)$ as population and empirical risk differences

$$M_{0,N}(\bar{Q}, \bar{Q}^*) := R_{0,N}(\bar{Q}) - R_{0,N}(\bar{Q}^*) \qquad \text{and} \qquad \widehat{M}_N(\bar{Q}, \bar{Q}^*) := \widehat{R}_N(\bar{Q}) - \widehat{R}_N(\bar{Q}^*). \quad (3.17)$$

Let

$$\sigma_N^2(\bar{Q}, \bar{Q}^*) := \frac{1}{N} \sum_{t=1}^{N} E\left[\left(\frac{g^*}{g_{0,t}} (A(t) \mid C_o(t))(\ell(\bar{Q}) - \ell(\bar{Q}^*))(C_o(t), O(t)) \right)^2 \mid C_o(t) \right].$$

The quantity $\sigma_N(\bar{Q}, \bar{Q}^*)$ can be seen as a sequential equivalent of an L_2 norm for the process $\{(g^*/g_{0,t})(A(t) \mid C_o(t))(\ell(\bar{Q}) - \ell(\bar{Q}^*))(C_o(t), O(t))\}_{t=1}^N$. From assumption 17, we have that $\sup_{t \geq 1} \|(g^*/g_{0,t})(\ell(\bar{Q}) - \ell(\bar{Q}^*))\|_\infty = O(1)$. From theorem A.4 in [54], with probability at least $1 - 2e^{-x}$, we have that

$$\begin{aligned} &\sup \left\{ M_{0,N}(\bar{Q}, \bar{Q}^*) - \widehat{M}_N(\bar{Q}, \bar{Q}^*) : \bar{Q} \in \bar{\mathcal{Q}},\ \sigma_N(\bar{Q}) \leq r \right\} \\ &\lesssim r^- + \frac{1}{\sqrt{N}} \int_{r^-}^{r} \sqrt{\log(1 + N_{[]}(\epsilon, 1, \mathcal{L}_N, \bar{O}(N)))} d\epsilon \\ &\quad + \frac{1}{N} \log(1 + N_{[]}(r, 1, \mathcal{L}_N, \bar{O}(N))) + r\sqrt{\frac{x}{N}} + \frac{x}{N}. \end{aligned}$$

From assumption 17, we have that

$$\sigma_N(\bar{Q}) \lesssim \|\ell(\bar{Q}) - \ell(\bar{Q}^*)\|_{2,g^*,h_N} \lesssim M_{0,N}(\bar{Q}, \bar{Q}^*).$$

Combined with lemma 6, we have that

$$\begin{aligned} &\sup \left\{ M_{0,N}(\bar{Q}, \bar{Q}^*) - \widehat{M}_N(\bar{Q}, \bar{Q}^*) : \bar{Q} \in \bar{\mathcal{Q}},\ M_{0,N}(\bar{Q}, \bar{Q}^*) \leq r \right\} \\ &\lesssim r^- + \frac{1}{\sqrt{N}} \int_{r^-}^{r} \sqrt{\log(1 + N_{[]}(\epsilon, \bar{\mathcal{Q}}, L_2(P_{g^*,h_N}))} d\epsilon \\ &\quad + \frac{1}{N} \log(1 + N_{[]}(r, \bar{\mathcal{Q}}, L_2(P_{g^*,h_N})) + r\sqrt{\frac{x}{N}} + \frac{x}{N} \end{aligned}$$

with probability at least $1 - 2e^{-x}$. In the following, we treat the cases $p < 2$ and $p > 2$ separately.

Case p > 2. Observe that

$$\begin{aligned}\|\bar{Q}_N - \bar{Q}^*\|_{2,g^*,h_N} \lesssim & M_{0,N}(\bar{Q}_N, \bar{Q}^*) \\ = & M_{0,N}(\bar{Q}_N, \bar{Q}^*) - \widehat{M}_N(\bar{Q}_N, \bar{Q}^*) + \widehat{M}_N(\bar{Q}_N, \bar{Q}^*) \\ \leq & M_{0,N}(\bar{Q}_N, \bar{Q}^*) - \widehat{M}_N(\bar{Q}_N, \bar{Q}^*) \\ \leq & \sup\left\{M_{0,N}(\bar{Q}, \bar{Q}^*) - \widehat{M}_N(\bar{Q}, \bar{Q}^*) : \bar{Q} \in \bar{\mathcal{Q}},\ M_{0,N}(\bar{Q}, \bar{Q}^*) \leq r_0\right\}\end{aligned}$$

where $r_0 := \sup_{\bar{Q}\in\bar{\mathcal{Q}}} M_{0,N}(\bar{Q}, \bar{Q}^*)$. The third line follows from the fact that Q_N minimizes $\widehat{R}_N(\bar{Q})$ over $\bar{\mathcal{Q}}$, wich implies that $\widehat{M}_N(\bar{Q}_N, \bar{Q}^*) \leq 0$. We now use equation (**??**) to bound the last line of the inequality. From assumption 17, we know that $r_0 = O(1)$. Using the entropy bound from assumption 25 and minimizing the right hand side of (**??**) w.r.t. r^-, we obtain that, with probability at least $1 - 2e^{-x}$,

$$\|\bar{Q}_N - \bar{Q}^*\|^2_{2,g^*,h_N} \lesssim N^{-2/p} + \frac{x}{\sqrt{N}} + \frac{x}{N},$$

which, by picking x appropriately, then implies that $\|\bar{Q}_N - \bar{Q}^*\|_{2,g^*,h_N} = O_P(N^{-1/p})$.

Case p < 2. Starting from the bound (**??**), via some algebra and by taking an integral, we obtain

$$\begin{aligned}& E_{P_0}\left[\sup\left\{M_{0,N}(\bar{Q}, \bar{Q}^*) - \widehat{M}_N(\bar{Q}, \bar{Q}^*) : \bar{Q} \in \bar{\mathcal{Q}},\ M_{0,N}(\bar{Q}, \bar{Q}^*) \leq r\right\}\right] \\ \lesssim\ & r^- + \frac{1}{\sqrt{N}}\left(r + \int_{r^-}^{r}\sqrt{\log(1 + N_{[]}(\epsilon, \bar{\mathcal{Q}}, L_2(P_{g^*,h_N}))}d\epsilon\right) \\ +\ & \frac{1}{N}\left(r + \log(1 + N_{[]}(r, \bar{\mathcal{Q}}, L_2(P_{g^*,h_N}))\right).\end{aligned}$$

Let $r^- = 0$. By using the entropy bound from assumption 25, we obtain that

$$\begin{aligned}& E_{P_0}\left[\sup\left\{M_{0,N}(\bar{Q}, \bar{Q}^*) - \widehat{M}_N(\bar{Q}, \bar{Q}^*) : \bar{Q} \in \bar{\mathcal{Q}},\ M_{0,N}(\bar{Q}, \bar{Q}^*) \leq r\right\}\right] \\ \lesssim\ & \frac{1}{\sqrt{N}} r^{1-p/2}\left(1 + \frac{r^{1-p/2}}{r^2\sqrt{N}}\right).\end{aligned}$$

Theorem 3.4.1 in [126] then implies that

$$M_{0,N}(\bar{Q}_N, \bar{Q}^*) = O_P(N^{-\frac{2}{1+p/2}}),\ \text{and therefore}\ \|\bar{Q}_N - \bar{Q}^*\|_{2,g^*,h_N} = O_P(N^{-\frac{1}{1+p/2}}).$$

□

Outcome model classes

Now that we know how to characterize the sequential bracketing entropy of Ξ_N and $\mathcal{L}_N$ in terms of the bracketing entropy w.r.t. the norm $L_2(P_{Q_0,h_{C,N}})$ of the outcome model $\mathcal{Q}$, we look at specific function classes $\mathcal{Q}$ for which we know how to bound the latter.

Holder classes $H(\beta, M)$ over $\mathcal{C} \times \mathcal{O}$

Consider functions over a certain domain $\mathcal{X}$; in our setting we note that $\mathcal{X} = \mathcal{C} \times \mathcal{O}$. Suppose that $\dim(\mathcal{X}) = d$. We denote $H(\beta, M)$ the class of functions over a certain domain $\mathcal{X}$, such that, for any $x, y \in \mathcal{X}$, and any non-negative integers $\beta_1, \ldots, \beta_d$ such that $\beta_1 + \ldots + \beta_d = \lfloor \beta \rfloor$,

$$\left| \frac{\partial^{\lfloor \beta \rfloor} f}{\partial x_1^{\beta_1} \ldots \partial x_d^{\beta_d}}(x) - \frac{\partial^{\lfloor \beta \rfloor} f}{\partial x_1^{\beta_1} \ldots \partial x_d^{\beta_d}}(y) \right| \leq M \|x - y\|.$$

The bracketing entropy w.r.t. the uniform norm $\|\cdot\|$ of such a class satisfies

$$\log N_{[]}(\epsilon, H(\beta, M), \|\cdot\|) \lesssim \epsilon^{-d/\beta}.$$

For more detail, we refer the interested reader to, for example, chapter 2.7 in [126]. As such, our Donsker condition 19 is satisfied for $\beta > d/2$. Nevertheless, we caution that assuming that the outcome model lies in a Holder class of differentiability order $\beta > d/2$ might be an overly restrictive assumption.

HAL class

A class of functions that is much richer that the previous Holder classes is the class of cadlag functions with bounded sectional variation norm — also referred to as Hardy-Krause variation. We refer to this class as the Highly Adaptive Lasso class (HAL class), as it is the class in which the estimator, introduced in [130], takes values. The Highly Adaptive Lasso class is particularly attractive in i.i.d. settings for various reasons, which we enumerate next. In particular, (1) unlike Holder classes, it doesn't make local smoothness assumptions. Rather it only restricts a global measure of irregularity, the sectional variation norm, thereby allowing for functions to be differentially smooth/variable depending on the area of the domain. (2) Emprical risk minimzers over the HAL class were shown to be competitive with the best supervised machine learning algorithms, including Gradient Boosting Machines and Random Forests. (3) We know how to bound both the uniform metric entropy and the bracketing entropy of these classes of functions. These bounds show that the corresponding entropy integrals are bounded, which imply that the HAL class is Donsker. In particular, [10] provide a bound on the bracketing entropy w.r.t. $L_r(P)$, for $r \in [1, \infty)$, for probability distribution that have bounded Radon-Nikodym derivative w.r.t. the Lebesgue measure, that is $dP/d\mu \leq C$. [10] use this bracketing entropy bound to prove the rate of convergence $O(N^{-1/3}(\log N)^{2d-1})$.

Unfortunately, to bound the sequential bracketing entropies of Ξ_N and of $\mathcal{L}_N$ we would need a bracketing entropy bound w.r.t. $L_2(P_{Q_0,h_{C,N}})$, which, owing to the fact that $h_{C,N}$ is a discrete measure, does not have bounded Radon-Nikodym derivative w.r.t. the Lebesgue measure over $\mathcal{C} \times \mathcal{O}$. Under the assumption 21 on the convergence of the marginals of $(C_o(t))$ to a limit law (we shall denote it h_∞), we have that $h_{C,N} \xrightarrow{d} h_\infty$, which can reasonably be a continuous measure dominated by the Lebesgue measure. By convergence in distribution of $h_{C,N}t$ to h_∞, we have at least that the size of brackets w.r.t. $h_{C,N}$ converges to the size of brackets under h_∞. If this convergence were uniform over bracketings of $\bar{\mathcal{Q}}$, and that $dh_\infty/d\mu \leq C$, then we would have that $N_{[]}(\epsilon, \bar{\mathcal{Q}}, L_2(P_{Q_0,h_{C,N}})) \lesssim N_{[]}(\epsilon, \bar{\mathcal{Q}}, L_2(\mu))$. Proving the uniformity over bracket seems to be a relatively tough theoretical endeavor, and we leave it to future research.

A modified HAL class

Given the difficulty in bounding $N_{[]}(\epsilon, \bar{\mathcal{Q}}, L_2(P_{Q_0,g^*,h_N}))$ for the HAL, class, we consider a modified HAL class in the case where $\mathcal{C}$ is discrete, that is $\mathcal{C} = \{c_1, \ldots, c_J\}$. We define the modified class as the set of functions $f : \mathcal{C} \times \mathcal{O} \to \mathbb{R}$ such that, for every $c \in \mathcal{C}$, $o \mapsto f(c, o)$ is cadlag with sectional variation norm smaller than M_1. It is straightforward to show that the bracketing entropy of such a class $\mathcal{F}$ is bounded as follows:

$$\log N_{[]}(\epsilon, \mathcal{F}, L_2(P_{Q_0,g^*,h_N})) \lesssim |\mathcal{C}|\epsilon^{-1}(\log(1/\epsilon))^{2(\dim(\mathcal{O})-1)}.$$

Chapter 4

Adaptive Sequential Design with Network and Time Dependence

Infectious disease surveillance via resource-constrained test allocation has become an increasingly important topic during the COVID-19 pandemic. Even as vaccination rates increase for COVID-19, widespread testing informs public health policy by (1) reducing transmission via identifying and isolating cases, and (2) tracking the outbreak dynamics. However, infectious disease surveillance presents unique technical challenges. For instance, the true outcome of interest we wish to minimize for outbreak control — one's positive infectious status, is a latent variable. In addition, unlike the usual i.i.d. settings, the presence of both network and temporal dependence reduces the data to a single observation with dependent components. Finally, the current literature advocates primarily for simple *rule-based* testing strategies (e.g., symptom based, contact tracing, travel history), without taking into account individual risk. In this chapter, we study an adaptive sequential design involving n individuals over a period of τ time-steps, allowing for unspecified dependence among individuals and across time. Our causal target parameter is the mean latent outcome we would have obtained after one time-step, if, starting at time t given the observed past, we had carried out a stochastic intervention that maximizes the outcome under a resource constraint. With that, we propose an Online Super Learner for adaptive sequential surveillance that learns the optimal choice of tests strategies over time, adapting to the current state of the outbreak. Relying on a series of working models, the proposed method decides whether to learn across samples, through time, or both — based on the underlying (unknown) structure in the data at each time point of the disease trajectory. In addition, we present an identification result for the latent outcome in terms of the observed data, and propose a *risk-based* strategy as one of the candidate testing schemes that assigns tests based on the current risk of being infected. We demonstrate the superior performance of the proposed strategy over commonly implemented testing schemes in a simulation modeling a residential university environment during the COVID-19 pandemic.

4.1 Introduction

Most higher education institutions faced a difficult decision during the COVID-19 pandemic: reopen and conduct in-person instruction, or face financial challenges and unpropitious social impacts associated with continued closure. The spread of SARS-CoV-2 in a residential college is particularly hazardous for the broader community due to a large percent of younger, potentially asymptomatic individuals, higher likelihood of shared accommodation, and abundant social contacts [83]. In the absence of pertinent prior experience, most institutions turned to simulation models and sequential testing in order to track, and contain, the spread of COVID-19. A rich literature on different modeling techniques emerged as a consequence — resulting in variations of compartmental models, contact networks and agent- or individual-based models [49, 102, 91, 82, 73, 86, 44, 21]. The interest in effective and safe reopening strategy for an university campus extended across continents [124, 57], campus size [145, 4] and urban settings [53]. Other groups resorted to empirical proximity networks of college students in order to simulate and study the spread of the virus [52].

We differentiate between *rule-based* testing, broadly defined as simple deterministic rules (e.g., based on symptoms, location, timing or network) and *risk-based* testing (e.g., testing individuals at an estimated higher risk of infection). Available literature on testing strategies for a residential campus focused mostly on *rule-based* testing: symptom tracking and contact tracing, as well as scheduled on-campus screening with varying frequency [49, 15, 82, 91, 4, 102, 142, 113, 21, 96, 73, 44]. In general, the predominant infectious disease testing recommendation made by the World Health Organization suggests assigning tests to individuals having (i) symptoms consistent with COVID-19, (ii) contact with confirmed or suspected COVID-19 cases and (iii) evidence of recent travel history [90]. Alternate suggestions advocate for fast and frequent random population testing [70] and scheduled screening with repeated tests [143]. While contact tracing via efficient tracking system can be advantageous, its implementation is costly and often not comprehensive enough as the spread of infectious disease advances [42]. Other simple *rule-based* strategies tend to miss asymptomatic infections (e.g., symptom-based), or require significant financial burden and compliance for a large and heterogeneous population (e.g., frequent random testing). In addition, most of the suggested *rule-based* strategies are not clear on how to distribute tests across different prioritization groups.

Other concentrated efforts consist of finding optimal testing strategies that inform epidemic dynamics [22] and help reduce disease spread [11, 64, 48, 31]. In particular, [64] focus on optimal allocations designed as a combination of group and segmented testing; segments of the population based on occupation, age and geographical location are given testing priority. Both [11] and [48] advocate for contextual bandits as a possible approach to the optimal testing allocation, with [11] additionally suggesting an utility-based active learning solution. On the other hand, [31] develop a probabilistic framework accounting for resource limitations, imperfect testing and the need for prioritizing higher risk patient populations. However, all of the proposed strategies impose strong modeling assumptions — either on the type of dependence allowed (assuming homogeneous Markov Decision Process), model-

ing conditional probabilities necessary to estimate the number of positive tests, or assuming which strata of the population constitutes at-risk profile.

In this work, we propose an adaptive sequential design for a setting with network and temporal dependence where the goal is to optimize a short term outcome. The statistical problem is handled within a fully nonparametric model, respecting the true (unknown) dependence structure. While the proposed method is very general, it is particularly suited for infectious disease surveillance and control. We consider a longitudinal structure following n individuals over a trajectory until time τ. At each time point t for sample i, one observes the exposure variable $A_i(t)$ (e.g., indicator of testing), outcome $Y_i(t)$ (e.g., health status) and other time-varying covariates in $L_i(t)$ (e.g., network structure, location, symptoms). For an infectious disease surveillance, a decision maker/experimenter is in charge of assigning a test $A_i(t)$ to sample i at time t, then collecting a vector of measurements $L_i(t)$ for the same individual, including the outcome. The exposure of interest is defined as a known stochastic intervention, where each treatment denotes a specific testing design (e.g., *rule-based* or *risk-based* testing, etc). We study a setting in which the same decision maker can also adapt treatment assignment over time in response to past observations. Structuring the test allocation problem as an adaptive sequential design is paramount in order for the testing strategy to be able to adapt as the infectious disease trajectory changes and other variants become dominant.

In a setting where goal is surveillance and control, it is natural to define performance of a treatment rule in terms of a short-term average over samples. Our causal target parameter is defined as the mean outcome we would have obtained after one time-step, if, starting at time t given the observed past, we had carried out a stochastic intervention g_t^*. The main goal is to optimize the next time-point outcome under g_t^*, at each t, under a possible resource constraint. Alternatively, one can also seek to optimize the short-term outcome under stochastic intervention as an average over time, therefore targeting the entire trajectory. The history-adjusted optimal choice for a single time point intervention then defines a new adaptive design over time, which we denote the *Online Super Learner (SL) for adaptive sequential surveillance*. The regret minimization objective of the proposed design ensures that we assign tests such that as many infectious individuals as possible are subsequently caught. As the design is adaptive, it learns the optimal choice of test strategies over time, responding to the current state of the epidemic.

The proposed adaptive sequential design has crucial advantages over competing methods which make it particularly propitious in the infectious disease context. Key strength comes from not having to make strong (conditional) independence assumptions, or modeling network and time dependence. Instead of imposing unrealistic assumptions on the statistical model, the proposed method selects among adaptive designs with a short term performance Online Super Learner [7, 79]. As such, it imposes an honest benchmark to choose the best performing estimate for the sake of the adaptive design performance. The necessary parts of the design (e.g., conditional expectation of outcome given the past) is estimated via an Online Super Learner which relies on working models for dependence structure, letting appropriate cross-validation choose the correct model at time t. Therefore, the proposed method decides

whether to learn across samples, through time, or both, based on the underlying (unknown) structure in the data. This is in contrast to previously described adaptive sequential designs, which rely on conditional independence assumptions (across time or samples) in order to deal with unknown dependence [78, 9]. Secondly, as the true infectious status is unknown, the proposed target parameter is defined in terms of a latent outcome. In this work, we show that the average of true latent infectious status at time t can be identified as the average of observed outcomes. As such, the statistical target parameter is defined in terms of the observed outcome, delineated as a function of the stochastic intervention we implement.

In order to illustrate utility of the proposed adaptive sequential design, we simulate a hypothetical residential campus during the SARS-CoV-2 infection. The modeling parameters - including campus available resources, class schedule, on-campus housing and expected population size - reflect environment at the University of California (Berkeley) in the Fall of 2020. While we strive to model a specific environment, settings and simulations can be easily modified to reflect any residential campus and infectious disease. Extensive simulations demonstrate superior performance of the proposed adaptive design as opposed to all considered *rule-based* schemes (repeat *random* testing, *contact tracing + symptomatic*). We emphasize that the reported simulation results reflect the best case scenarios for the competitor testing strategies: accurate observed network and full symptoms for symptomatic individuals. The advantage of the proposed adaptive design is evident over a variety of scenarios, including varying resource constraints and level of problem difficulty (determined by the percent latent component of the network and individual risk). In addition to considering gold standard testing schemes, we investigate performance of a learned *risk-based* strategy individually and as a candidate stochastic intervention in the proposed Online SL for adaptive surveillance design. Our simulations demonstrate that learned and flexible *risk-based* testing is crucial for an infectious disease with a large asymptomatic population — but might need time to perform well, as we need data to learn who is at risk. An Online SL for adaptive designs that uses both simple *rule-based* and learned *risk-based* testing strategies as candidate designs outperforms all schemes individually, while learning which testing strategy does best in each infectious disease context over time. As such, it tends to pick *rule-based* strategies at the beginning of the infectious disease trajectory, and *risk-based* testing as more data is collected and the risk function is learned.

The article structure is as follows. In Section 4.2 we formally define the general formulation of the statistical estimation problem, consisting of specifying notation, likelihood, and the nonparametric model. In subsection 4.2, we describe all the relevant working models, including assumptions underlying each. We define the target parameter, causal assumptions, and provide identification results in subsection 4.2. In Section 4.3, we proceed to describe the proposed adaptive design, denoted as the Online Super Learner for adaptive sequential surveillance. Section 4.3 includes various proposed selectors, aimed at learning the optimal testing strategy for the sake of adaptive design performance. We provide details on the agent-based model used for simulations, as well as how each testing strategy considered can be described as a stochastic intervention, in Section 4.4. Section 4.5 contains simulation results based on the proposed agent-based model for moderate size residential campus. We

conclude with a short discussion in Section 4.6.

4.2 Statistical Formulation of the Problem

Data and the Causal Model

Consider a random variable denoted as O_i for $i = 1, \cdots, n$, where O_i is a sample i trajectory. For each individual i, we define the following longitudinal data structure where

$$O_i = (L_i(0), A_i(1), L_i(1), \ldots, A_i(\tau), L_i(\tau)),$$

corresponding to observations from time $t = 0$ to the final time point $t = \tau$. Within time point t, we arbitrarily order data points by increasing sample index i, such that

$$(A_1(t), \ldots, A_n(t), L_1(t), \ldots, L_n(t))$$

reflects the unit ordering. We further decompose sample i trajectory into baseline and time-varying parts. In particular, we define $O_i(0) = L_i(0)$ as a vector of baseline covariates which, by definition, are initiated at $t = 0$. For an infectious disease surveillance, $L_i(0)$ includes baseline infectious status, as well as other covariates (e.g., demographic information, initial network structure). The time-varying part of sample i trajectory decomposes as $O_i(t) = (A_i(t), L_i(t))$, for $t = 1, \ldots, \tau$; it includes the treatment status occurring before the response variable and time-varying covariates, all indexed by time t. In particular, we let $A_i(t)$ denote the exposure variable, corresponding to a time t indicator of being tested in an infectious disease surveillance design. We define $L_i(t)$ as a vector of time-varying covariates, with the first component being the response variable — infectious status for sample i at time t. In addition to outcome, $L_i(t)$ also possibly tracks the risk profile of unit i, as well as information on other units $\{1, \ldots, n\} \setminus \{i\}$ that belong to the network of sample i. The network of individual i contained in $L_i(t)$ is denoted as $F_i(t)$, which reflects all the samples connected to unit i at time t. In particular, we allow $|F_i(t)|$ to vary in i, but assume that this number is bounded by some known global constant K that does not depend on n. Finally, we emphasize that the true infectious status for each sample and at each time point is typically not observed. Hence, we define the true latent outcome, notably the infectious status for sample i at time t, as $Y_i^l(t)$. The observed outcome for sample i at time point t is denoted as $Y_i(t)$, where $L_i(t) = (Y_i(t), \ldots, F_i(t), \ldots)$.

For n observed trajectories, we write $O = O^{\tau,n} = \{O_i\}_{i=1}^n$, where O is a simplified notation that does not make dependence on τ and n explicit. Under this notation, data observed throughout the course of the trial is O, with $O(t) = \{O_i(t)\}_{i=1}^n$ being the collection of n time t-specific points. Similarly, let $L(t)$ and $A(t)$ denote n dimensional time-specific vectors, effectively including time t-specific information across all n collected samples; with that, we have that $L(t) = (L_1(t), \ldots, L_n(t))$ and $A(t) = (A_1(t), \ldots, A_n(t))$. Further, we write $Pa(O(t)) = \bar{O}(t-1) = (O(0), \ldots, O(t-1))$ to represent history of all samples up

to time t. The complete histories until node $L(t)$ and $A(t)$ are denoted as $Pa(L(t)) = (\bar{O}(t-1), A(t))$ and $Pa(A(t)) = (\bar{O}(t-1))$, which are time t histories of all n samples. We also let time and unit-specific histories $Pa(A_i(t))$ and $Pa(L_i(t))$ denote all observations that come before $A_i(t)$ and $L_i(t)$, according to both time and sample ordering. In particular, let $Pa(L_i(t)) = (Pa(L(t)), L_1(t), \cdots, L_{i-1}(t))$ and $Pa(A_i(t)) = (Pa(A(t)), A_1(t), \cdots, A_{i-1}(t))$, where $i-1$ denotes sequential samples until sample i. Consequently, we let $Pa(O_i(t)) = \bar{O}_i(t) = (\bar{O}(t-1), O_1(t), \cdots, O_{i-1}(t))$, where $Pa(O_i(t))$ includes all history until time $t-1$ and t-specific samples until individual i.

Statistical Model

Let $\mathcal{M}$ denote the *statistical model* for the probability distribution of the data that is non-parametric, beyond possible knowledge of the treatment mechanism. The more we know, or are willing to assume about the experiment that produces the data, the smaller the model. Let $P_0 \in \mathcal{M}$ denote the true probability distribution of O, such that $O \sim P_0$, and let P denote any probability distribution where $P \in \mathcal{M}$. We let p_o denote the density of P_0 with respect to (w.r.t) a dominating measure μ. The likelihood of o can be factorized according to the time-ordering as follows:

$$\begin{aligned} p_0(o) &= \prod_{i=1}^{n} p_{0,l(0)}(l_i(0)) \prod_{t=1}^{\tau} p_{0,a_i(t)}(a_i(t)|Pa(a_i(t)))p_{0,l_i(t)}(l_i(t)|Pa(l_i(t))) \\ &= \prod_{i=1}^{n} p_{0,l(0)}(l_i(0)) \prod_{t=1}^{\tau} g_{0,i,t}(a_i(t) \mid Pa(a_i(t)))q_{0,i,t}(l_i(t) \mid Pa(l_i(t))), \end{aligned} \tag{4.1}$$

where we let $a_i(t) \mapsto p_{0,a_i(t)}(a_i(t) \mid Pa(a_i(t)))$ and $l_i(t) \mapsto p_{0,l_i(t)}(l_i(t) \mid Pa(l_i(t)))$ denote conditional densities w.r.t. the dominating measures μ_A and μ_L, respectively. We use shorthand notation for conditional densities and distributions of the relevant nodes. In particular, we write $q_{0,i,t}$ as the true (i,t)-specific conditional density of $L_i(t)$ based on the observed past until time t, $Pa(L_i(t))$. The corresponding true conditional distribution of $L_i(t)$ conditional on $Pa(L_i(t))$ is written as $Q_{0,i,t}$. At time t, $g_{0,i,t}$ reflects the true probability of drawing the testing indicator $A_i(t)$ conditional on the past until time t, $Pa(A_i(t))$. In our randomized experimental setting, $g_{0,i,t}$ is known and in control of the experimenter for most testing allocations; practically, it denotes a particular sampling and testing design implemented for sample i at time t. Due to the data ordering, we make the following two remarks

Remark 1 *For every $t \in [\tau]$ and $i \in [n]$, $A(t) = (A_1(t), \ldots, A_n(t))$ are independent conditional on $Pa(A(t))$.*

Remark 2 *For every $t \in [\tau]$ and $i \in [n]$, $Y(t) = (Y_1(t), \ldots, Y_n(t))$ are independent conditional on $Pa(L(t))$.*

Both Remark (1) and (2) follow from the time and sample ordering: testing is allocated based on all of the observed past $\bar{O}(t-1)$ (and not influenced by other tests at time t), and

observed outcome is a direct consequence of tested individuals and observed past. We define $\bar{Q}_{0,i,t}(Pa(L(t))) = \bar{Q}_{0,i,t}(A(t), Pa(O(t))) \equiv \mathbb{E}_0[Y(t)|A(t), Pa(O(t))]$ as the true conditional expectation of $Y(t)$ given the observed past. In order to emphasize dependence on the treatment mechanism $g_{i,t}$, we might also write $\bar{Q}^{g_{i,t}}_{0,i,t}(Pa(L(t)))$ as the conditional expectation of $Y(t)$ given the observed past under $g_{i,t}$. Finally, recall that $Pa(O(t)) = \bar{O}(t-1)$ denotes observed history of $O(t)$ until time t. With that in mind, we write $P_{O_i(t)|\bar{O}(t-1)}$ (shorthand, $P_{\bar{O}(t-1)}$) as the time t conditional distribution of $O_i(t)$ given the observed past $\bar{O}(t-1)$.

Note that the decomposition presented in likelihood expression (4.1) places no restrictions on the type of dependence possible. Therefore, the data reduces to a dependent observation with temporal and network dependence, and we observe only a single draw from P_0. In order to learn relevant parts of the data generating distribution, we would have to put some restrictions on the statistical model $\mathcal{M}$. In the following, we discuss several possible working models that enable us to learn parts of the likelihood, without assuming any of them explicitly. Via the proposed working models, one can distinguish between different types of dependence we are willing to assume, depending on whether we can learn through time (therefore assuming some level of conditional stationarity), learn through the number of individuals (therefore assuming independence of samples given a known network), or both. We emphasize that one of the strengths of the proposed method is that it does not impose any direct assumptions on the statistical model $\mathcal{M}$; we let the data decide on the appropriate working model at each time point. In the following, we describe all considered working models, and motivation behind each.

Working Models

We start by restricting the complexity of dependence allowed by supposing that each $L_i(t)$ can depend on the past only through a fixed dimensional summary measure of history, instead of the entire observed history. As such, we assume that $q_{0,i,t}(L_i(t) \mid Pa(L(t)))$ depends on the past only through a fixed dimensional summary measure $C_{L_i}(t)$, where $C_{L_i}(t)$ is a function of the observed history. Therefore, for every $t \in [\tau]$ and $i \in [N]$, $L_i(t)$ is independent of its past conditional on $C_{L_i}(t)$ and $q_{i,t}(L_i(t) \mid Pa(L(t))) = q_{i,t}(L_i(t) \mid C_{L_i}(t))$. For some applications, the summary measure might cover a limited history, and the dependent process has a finite memory allowing us to learn through time. A particular example of summary measures are fixed dimensional extractions from the complete history, such that $C_{L_i}(t) = h_{L_i}(Pa(L(t))) \in \mathbb{R}^k$ is a (k)-dimensional extraction of the form $C_{L_i}(t) = \{(L_j(s), A_j(s+1)) : s = t-1, t-2, \cdots, t-k, j \in [n]\}$. For other applications, the fixed dimensional summary measure might be a function of the sample i's network; as such, we might have that conditional probability of $L_i(t)$ depends only on the history of j samples, where $j \in F_i(t) \cup i$. In the case of both time and network dependence, $C_{L_i}(t)$ could be a function of both sample i's network and previous past time-points where $C_{L_i}(t) = \{(L_j(s), A_j(s+1)) : s = t-1, t-2, \cdots, t-k, j \in F_i(t) \cup i\}$; then $C_{L_i}(t)$ is a summary measure of the history over last k steps of a set $F_i(t)$ of at most K friends. We note that, if $F_i(t) = \emptyset$, our formulation reduces to an i.i.d. setting across samples. In order to formally present our target parameter under a working model, we make

the following assumption on the decomposition of the fixed dimensional summary measure, as stated below.

Assumption 28 (Decomposition of the fixed dimensional summary)**.** *For every $t \in [\tau]$ and $i \in [n]$, the fixed dimensional summary measure $C_{L_i}(t)$ can be written as*

$$C_{L_i}(t) = (A_i(t), C_{A_i}(t)),$$

where $C_{A_i}(t) = h_{A_i}(Pa(A(t))) = h_{A_i}(Pa(O(t)))$.

Assumption 29 (Conditional independence given a summary measure)**.** *For every $t \in [\tau]$ and $i \in [n]$,*

$$q_{i,t}(l_i(t) \mid c_{L_i}(t)) = q_{i,t}(l_i(t) \mid Pa(l(t)))$$

where $c_{L_i}(t)$ is the observed fixed dimensional summary of the past until time t.

The following key assumption is a modeling assumption on the conditional density of $L_i(t)$ given the observed past. Consistent with Assumption 4 in [9], we might assume that the conditional distribution of $L_i(t)$ given the observed fixed dimensional summary of the history is a constant function across samples and time. As such, there exists a common in i and t conditional density q such that $q_{i,t} = q$. Drawing from the reinforcement learning literature, this assumption is analogous to the homogeneity assumption for the Markov Decision Process [1]. Under Assumption 29 and allowing for a common in i and t conditional density of $L_i(t)$ given the history, we can rewrite the likelihood from equation (4.1) as:

$$p(o) = \prod_{i=1}^{n} p_{l(0)}(l_i(0)) \prod_{t=1}^{\tau} g_{0,i,t}(a_i(t) \mid Pa(a(t)))q(l_i(t) \mid c_{L_i}(t)). \tag{4.2}$$

Note that, since $g_{0,i,t}$ is known, we don't need to put any restrictions on the treatment mechanism given the past. We emphasize that assuming common in i and t conditional density q still allows for a rich network and time-dependent structure given $C_{L_i}(t)$. The proposed formulation lets us learn and measure factors that result in changes over time and network, captured with varying $C_{L_i}(t)$. For example, we could have that $C_{L_i}(t) = h_{L_i}(Pa(L(t))) \in \mathbb{R}^{k \times j}$ is a (k, j)-dimensional extraction of the form $C_{L_i}(t) = \{(L_j(s), A_j(s+1)) : s = t-1, t-2, \cdots, t-k, \ \ j \in F_i(t) \cup i\}$ where $F_i(t) < K$ and K is not a function of n. As such, our working model covers finite memory time dependence and network structure where each individual has a limited number of contacts, both of which could possibly vary as the trajectory advances. Alternatively, the proposed working model could cover dependence structures described by summary measures of the time series pattern (e.g.: moving average, finite memory, features related to STL decomposition of the series, spectral entropy, Hurst coefficient) and summary measures of the current state of the network (e.g.: current state of the epidemic, percent isolated, percent wearing masks).

Overall, in the adaptive sequential surveillance design, such modeling assumptions equate to conditional stationarity of the outcome mechanism over the entire trajectory (common in

time) and for each sample (common across samples) given a fixed dimensional summary of the past. Instead of assuming a common conditional density of $L_i(t)$, we might alternatively only need to assume a common conditional expectation. Then, under the common in (i,t)-working model, we have that $\bar{Q}_{i,t} = \bar{Q}$ instead of the full $q = q_{i,t}$. We write assumptions on the conditional expectation of $Y(t)$ given the observed past under the common in (i,t) model as Assumption 30. With that, Assumption 28, 29 and 30 constitute working model $\mathcal{M}^{tn}$.

Assumption 30 (Common in i and t conditional expectation of the outcome)**.** *There exists a common across samples (i) and time (t) conditional expectation of $Y_i(t)$ given the observed past, such that $\bar{Q}_{i,t} = \bar{Q}$ for every $t \in [\tau]$ and $i \in [n]$*

$$\bar{Q}_{i,t}(A_i(t), C_{A_i}(t)) = \bar{Q}(A_i(t), C_{A_i}(t)).$$

Definition 6 (Working model $\mathcal{M}^{tn}$)**.** *We define a working model $\mathcal{M}^{tn}$ as the set of distributions P over the domain $\mathcal{O}$ that satisfy Assumptions 28, 29 and 30.*

Alternatively, we may assume that the conditional expectation of $Y_i(t)$ given the past is a t-common mechanism given the history, allowing for a possibly very dense network structure which might be observed during a highly contagious epidemic. Conditional on the observed fixed dimensional summary $C_{L_i}(t)$, $q_{i,t}$ is a common in t density smooth enough to be learned through time. This working model assumption is analogous to models previously described in the time-series literature, extended to multiple trajectories [131, 78, 79]. We can rewrite the likelihood from equation (4.1) under common-in-t density as follows

$$p(o) = \prod_{i=1}^{n} p_{l(0)}(l_i(0)) \prod_{t=1}^{\tau} g_{0,i,t}(a_i(t) \mid Pa(a(t))) q_i(l_i(t) \mid c_{L_i}(t)). \tag{4.3}$$

In terms of the conditional expectation, we emphasize that the functional form of $\bar{Q}_{i,t} = \bar{Q}_i$ is unspecified, with the only assumption being that $\bar{Q}_i$ is common in time conditional on a fixed dimensional summary. In the current application, such modeling assumptions would equate to conditional stationarity of the expected outcome over the entire trajectory (common in time), but not common across samples. We denote the working model described by the Assumptions 28, 29 and 31 as $\mathcal{M}^t$.

Assumption 31 (Common in t conditional expectation of the outcome)**.** *There exists a common across time (t) conditional expectation of $Y_i(t)$ given the observed past, such that $\bar{Q}_{i,t} = \bar{Q}_i$ for every $t \in [\tau]$ and $i \in [n]$*

$$\bar{Q}_{i,t}(A_i(t), C_{A_i}(t)) = \bar{Q}_i(A_i(t), C_{A_i}(t)).$$

Definition 7 (Working model $\mathcal{M}^t$)**.** *We define a working model $\mathcal{M}^t$ as the set of distributions P over the domain $\mathcal{O}$ that satisfy Assumptions 28, 29 and 31.*

Instead of learning across time, one might instead rely on asymptotics in the number of individuals. An important ingredient of this modeling approach is to assume that any dependence of unit i can be fully described by a function of the known network over time. Following work by [116], let $F_i(t) \leq K$ denote the network for sample i at time t. Then, there is a common in i density, q_t, allowing for possibly very long and elaborate time-dependence. Similarly, there is a common-in-i expectation conditional on a fixed dimensional summary measure $C_{L_i}(t)$. In contrast to decomposition presented in 4.1, likelihood under common-in-i density is written as follows

$$p(o) = \prod_{i=1}^{n} p_{l(0)}(l_i(0)) \prod_{t=1}^{\tau} g_{0,i,t}(a_i(t) \mid Pa(a(t))) q_t(l_i(t) \mid c_{L_i}(t)). \tag{4.4}$$

Under no conditional stationarity assumption, one could use the recent estimates of $\bar{Q}_t$ in order to optimize the next sampling mechanism w.r.t the status of the epidemic few time points in the future. This implies that it is possible to learn the common-in-i expectation $\bar{Q}_t$ from a draw O as $n \to \infty$, resulting in a well-defined statistical estimation problem. For the adaptive surveillance problem, this formulation allows us to learn across samples, as dynamics of the trajectory is not stationary over time but possibly evolving. Assumptions 28, 29 and 32 constitute the working model $\mathcal{M}^n$.

Assumption 32 (Common in i conditional expectation of the outcome)**.** *There exists a common across samples (i) conditional expectation of $Y_i(t)$ given the observed past, such that $\bar{Q}_{i,t} = \bar{Q}_t$ for every $t \in [\tau]$ and $i \in [n]$*

$$\bar{Q}_{i,t}(A_i(t), C_{A_i}(t)) = \bar{Q}_t(A_i(t), C_{A_i}(t)).$$

Definition 8 (Working model $\mathcal{M}^n$)**.** *We define a working model $\mathcal{M}^n$ as the set of distributions P over the domain $\mathcal{O}$ that satisfy Assumptions 28, 29 and 32.*

Target Parameters

In the following, we describe a counterfactual scenario in which the initial treatment mechanism is replaced by user-defined conditional distributions, and define the corresponding target parameter of interest. Our main aim is to describe an adaptive sequential surveillance design for infectious disease under unknown network and time dependence. This entails defining the time t-specific testing strategy which optimizes the short term outcome among a set of proposed testing schemes. The optimal testing strategy maximizes the number of positive cases caught, with the target parameter under a resource constraint being of particular interest in practice. Instead of focusing only on the time t-parameter, we also define an average over the entire trajectory as a target parameter of interest. In the following Sections, we describe a new testing allocation scheme based on the current risk of infection, termed the *risk-based* strategy.

Structural Equations Model

In the previous section, we discuss the distribution of the observed data. Given a data-set, we can estimate parameters of this distribution, resulting in statistical parameters. However, without more structure, statistical parameters do not have a causal interpretation. In order to translate the scientific question of interest into a formal causal quantity, we must first specify a structural equations model (SEM; equivalently, structural causal model (SCM)) [93].

By specifying a SEM, we assume that each component of the data structure is a function of the observed endogenous variables and an unmeasured exogenous error term [93]. We encode the time-ordering of the variables using the following SEM for each t:

$$\begin{aligned} L_i(0) &= z_{L_i(0)}(U_{L_i}(0)) \\ A_i(t) &= z_{A_i(t)}(\bar{O}(t-1), U_{A_i}(t)) \\ L_i(t) &= z_{L_i(t)}((\bar{O}(t-1), A_i(t)), U_{L_i}(t)), \end{aligned} \tag{4.5}$$

where $U := (U_A, U_L)$ with $U_A := (U_{A_i}(t) : t \in [\tau], i \in [n])$ and $U_L := (U_{L_i}(t) : t \in [\tau], i \in [n])$. The unmeasured exogenous variables are sampled from P_U, such that $U \sim P_U$. Given an input (U, O), structural equations $z_{A_i(t)}$ and $z_{L_i(t)}$ for each time $t \in [\tau]$ and sample $i \in [n]$ deterministically assign a value to each of the nodes. While we have a specification of $z_{A_i(t)}$ in a randomized trial, the structural equations $z_{L_i(t)}$ do not restrict the functional form of the causal relationships for any t or i. The SEM defines a collection of distributions (U, O) representing the full data model, here defined in terms of U and observed data O. Let P_0^F denote the true probability distribution of (U, O); in the remainder of the article, we will use the subscript "0" to indicate true probability distributions or components thereof. Here we emphasize that any distribution P^F on the domain of the full data fully determines a corresponding distribution P on the domain of the observed data. Finally, we denote the model for P_0^F as $\mathcal{M}^F$, known as the *causal model*.

We can also define the *time- and history- specific causal model*. Let $\mathcal{M}_t^F(\bar{o}(t-1))$ denote the set of conditional probability distributions $P^F_{\bar{O}(t-1)}$, which condition on the observed history by time t, $\bar{o}(t-1)$. In particular, $\mathcal{M}_t^F(\bar{o}(t-1))$ is compatible with the structural equations model (4.5) by imposing $\bar{O}(t-1) = \bar{o}(t-1)$:

$$\begin{aligned} L_i(0) &= z_{L_i(0)}(U_{L_i}(0)) \\ A_i(t) &= z_{A_i(t)}(\bar{o}(t-1), U_{A_i}(t)) \\ L_i(t) &= z_{L_i(t)}((\bar{o}(t-1), A_i(t)), U_{L_i}(t)). \end{aligned} \tag{4.6}$$

Target Parameter on the SEM and Identifiability

The causal model allows us to define counterfactual random variables as functions of (U, O) corresponding with arbitrary interventions. In particular, we can replace data generating distribution for the treatment mechanism by user-specified conditional distributions; such

non-degenerate choices of intervention distributions are referred to as stochastic interventions [30]. Let $g^*_{i,t}$ denote a stochastic intervention at time t identified as a conditional distribution of $A^*_i(t)$ given the observed past. We write $g^*_t = \{g^*_{1,t}, \ldots, g^*_{n,t}\}$ for all n interventions at time t. With that, $O^*(t)$ is the counterfactual full data generated from the SEM described in (4.6) by replacing the equation associated with the exposure node by the counterfactual intervention $g^*_{i,t}$ at time t,

$$\begin{aligned} L_i(0) &= z_{L_i(0)}(U_{L_i}(0)) \\ A_{i,g^*_{i,t}}(t) &\sim g^*_{i,t}(\cdot|\bar{o}(t-1)) \\ L_{i,g^*_{i,t}}(t) &= z_{l_i(t)}((\bar{o}(t-1), A_{i,g^*_{i,t}}(t)), U_{L_i}(t)). \end{aligned} \tag{4.7}$$

We write $(U(t), O^*(t))$ as the full post-intervention data at time t, with the post-intervention distribution denoted as $P^{F*}_{\bar{O}(t-1)}$. Consequently, the counterfactual latent outcome under $g^*_{i,t}$ is written as $Y^l_{i,g^*_{i,t}}(t)$ for the sample i at time t. We define our causal parameter of interest as

$$\Psi^F_{t,g^*_t}(P^F_{\bar{O}(t-1)}) = \mathbb{E}_{P^{F*}_{\bar{O}(t-1)}}\left[\frac{1}{n}\sum_{i=1}^{n} Y^l_{i,g^*_{i,t}}(t)\right], \tag{4.8}$$

which is the expectation of the counterfactual random variable $Y^l_{i,g^*_{i,t}}(t)$ generated by the modified SEM as written in equation (4.7). Our causal target parameter is the mean latent outcome we would have obtained after one time-step, if, starting at time t given the observed past, we had carried out intervention g^*_t.

By defining the causal quantity of interest in terms of stochastic interventions on the SEM and providing a link between the causal model and the observed data, we lay the groundwork for addressing identifiability through P_0. In order to express $\Psi^F_{t,g^*_t}(P^F_{\bar{O}(t-1)})$ as a parameter of the distribution $P_{\bar{O}(t-1)}$ of the observed data O, we add two key assumptions on the SEM: the sequential randomization assumption (Assumption 33) and the positivity assumption (Assumption 34) in the following.

Assumption 33 (Sequential Randomization)**.** *For any $t \in [\tau]$ and $i \in [n]$,*

$$A_i(t) \perp\!\!\!\perp Y^l_{i,g^*_{i,t}}(t) \mid Pa(A(t)) \quad \textit{and} \quad A_{i,g^*_{i,t}}(t) \perp\!\!\!\perp Y^l_{i,g^*_{i,t}}(t) \mid Pa(A(t)).$$

Assumption 34 (Positivity)**.** *For any $t \in [\tau]$ and $i \in [n]$ with $P(Pa(A(t)) = Pa(a(t))) > 0$,*

$$g_o(A_i(t) \mid Pa(A(t)) = Pa(a(t))) > 0.$$

Theorem 14. *Assume assumptions 33 and 34 hold. Under consistency, we denote the time t value under the stochastic intervention g^*_t as*

$$\begin{aligned} \Psi^F_{t,g^*_t}(P^F_{\bar{O}(t-1)}) = \Psi_{t,g^*_t}(P_{\bar{O}(t-1)}) &= \int_a \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_P[Y^l_i(t) \mid A_i(t) = a, \bar{o}(t-1)] g^*_{i,t}(a \mid \bar{o}(t-1)) d\mu_a(a) \\ &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{\bar{Q}_{i,t},g^*_{i,t}}[Y_i(t) \mid \bar{o}(t-1)] \end{aligned}$$

where the observed outcome is defined as $Y_i(t) = A_i(t)Y_i^l(t)/g_{i,t}(A_i(t) \mid \bar{O}(t-1))$ *and* $\psi_{t,g_t^*} = \Psi_{t,g_t^*}(P_{\bar{O}(t-1)})$.

Proof. We allocate the derivation to the Appendix section. □

Note that Theorem 14 identifies the causal parameter in terms of both the latent and observed outcome at each time point. As we can identify the causal target parameter in terms of the observed O, we can proceed with the estimation step in the following sections. As per Theorem 14, the statistical target parameter is denoted as

$$\psi_{t,g_t^*} = \Psi_{t,g_t^*}(P_{\bar{O}(t-1)}) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{\bar{Q}_{i,t},g_{i,t}^*}[Y_i(t) \mid \bar{O}(t-1)]. \tag{4.9}$$

Instead of focusing on just the time t-target $\Psi_{t,g_t^*}(P_{\bar{O}(t-1)})$, we can additionally define a time- and sample- specific target parameter

$$\psi_{i,t,g_{i,t}^*} = \Psi_{i,t,g_{i,t}^*}(P_{\bar{O}(t-1)}) = \mathbb{E}_{\bar{Q}_{i,t},g_{i,t}^*}[Y_i(t) \mid \bar{O}(t-1)], \tag{4.10}$$

where $\psi_{t,g_t^*} = 1/n\sum_{i=1}^{n}\psi_{i,t,g_{i,t}^*}$. Of even more interest is the target parameter defined as an average of observed outcomes over the length of the entire trajectory, denoted as

$$\psi = \frac{1}{\tau}\sum_{t=1}^{\tau}\Psi_{t,g_t^*}(P_{\bar{O}(t-1)}). \tag{4.11}$$

We refer to all three in the following sections, with a particular focus on parameters in Equation (4.9) and (4.11).

Finally, as testing resources are typically limited during a highly contagious infectious disease, we assume a fixed testing capacity at each time-point until the end of the epidemic. As such, it is necessary to provide an optimal allocation of the available resources, analogous to the resource constrained optimal individualized treatment literature [75]. Suppose that the number of available tests are limited at each time point t, so that at most $k \in (0,1)$ proportion of the population can get tested. Our ultimate interest might be in optimizing Equation (4.11) under a resource constraint, meaning that we want to optimize the true number of infected individuals by the end of the trajectory. The more positive cases we can detect (and isolate) at each t under the k testing constraint, the fewer incidence of downstream transmission can occur — resulting in a greater infection control. In the following, we focus on the target parameters presented in Equation (4.9) and (4.11) under a possible k resource constraint.

4.3 Online Super Learner for Adaptive Surveillance

As defined in Section 4.2, our goal is to optimize the time t parameter

$$\begin{aligned}\psi_{t,g_t^*} = \Psi_{t,g_t^*}(P_{\bar{O}(t-1)}) &= \frac{1}{n}\sum_{i=1}^{n} \psi_{i,t,g_{i,t}^*} \\ &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{\bar{Q}_{i,t},g_{i,t}^*}[Y_i(t) \mid \bar{O}(t-1)],\end{aligned}$$

or the full trajectory target parameter

$$\psi = \frac{1}{\tau}\sum_{t=1}^{\tau} \Psi_{t,g_t^*}(P_{\bar{O}(t-1)}),$$

under a possible resource constraint. We reiterate here that ψ_{t,g_t^*} is (t, g_t^*)- specific and $\psi_{i,t,g_{i,t}^*}$ is $(i, t, g_{i,t}^*)$- specific. With a slight abuse of notation, we write $\Psi_{t,g_t^*}(P_{\bar{O}(t-1)})$ as $\Psi_{t,g_t^*}(\bar{Q}_t)$. Let $\{g_{t,1}^*, \ldots, g_{t,S}^*\} \in \mathcal{G}$ denote a collection of S user-specified stochastic interventions for all samples $i \in [n]$ at t. Note that all considered testing schemes are an element of a space $\mathcal{G}$, which consists of a finite number of testing strategies considered at each time point. Therefore, $g_{t,s}^*$ is a s-specific conditional distribution of $A(t)$ given the observed past $\bar{O}(t-1)$ at time t. For the s-specific stochastic intervention, it then follows that

$$\psi_{t,g_{t,s}^*} = \Psi_{t,g_{t,s}^*}(\bar{Q}_t) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{\bar{Q}_{i,t},g_{i,t,s}^*}[Y_i(t) \mid \bar{O}(t-1)],$$

and we have a separate $\psi_{t,g_{t,s}^*} = \Psi_{t,g_{t,s}^*}(\bar{Q}_t)$ for each $s \in \{1, \ldots, S\}$ at t. As infectious disease progression evolves over time, we want the proposed adaptive sequential design to be able to respond to the current state of the epidemic. At the beginning of the disease trajectory, catching the few infected individuals and testing their proximate network might be enough to control the spread. However, as the contagion reaches the state of an epidemic, identifying individuals at high risk might be crucial in order to establish control. While one of the s-specific stochastic interventions might be optimal at the beginning of the trajectory, another one might be optimal at later points. The enforced adaptive sequential surveillance should evolve and adapt over time in response to the current state of the infectious disease progression. The problem then becomes — how to do we pick among stochastic interventions in $\mathcal{G}$, over the entire trajectory, while not imposing assumptions on the statistical model $\mathcal{M}$? In the following, we describe an *Online Super Learner for adaptive sequential surveillance* which uses different selectors to pick the optimal stochastic intervention s at time t.

Loss-based selector

We can define an adaptive design as an online algorithm that at each time point t fits a conditional distribution of treatment given the past observations. As such, it's an online

mapping of past data into a conditional distribution $g_t^*(A(t) \mid \bar{O}(t-1))$, while learning over time how to adapt g_t^* in order to optimize the short term outcome. With that it mind, we can formulate the problem at hand within the loss-based estimation paradigm [34, 132, 141, 133]. In the following, we proceed to define key concepts necessary for establishing an Online Super Learner — including a valid loss, risk, cross-validation scheme, and the discrete Super Learner [136, 7, 79].

To start, let $P_{n,t}$ denote the empirical distribution of n time-series collected until time t. We define the estimator mapping, $\hat{\Psi}_{t,g_t^*}$, as a function from the empirical distribution to the parameter space. In particular, let $P_{n,t} \mapsto \hat{\Psi}_{t,g_t^*}$ represent a mapping from $P_{n,t}$ (based on n time-series collected until time t) into a function $\hat{\Psi}_{t,g_t^*}(P_{n,t})$. Then, $\hat{\Psi}_{t,g_t^*}(P_{n,t})(\bar{O}(t-1))$ denotes the target function evaluated at the observed past. Similarly, the estimator mapping $\hat{\bar{Q}}_{i,t}$ is defined as a function of the empirical distribution. We can write $\hat{\bar{Q}}_{i,t}(P_{n,t})(A_i(t), \bar{O}(t-1))$ as the predicted outcome for unit i of the estimator $\hat{\bar{Q}}_{i,t}(P_{n,t})$ at time t, based on $(A_i(t), \bar{O}(t-1))$. In Section 4.7, elaborate more on the loss-based parameter definition and estimation of the conditional expectation of the outcome given the past under working models described in Section 4.2.

Let $C(i,m)$ denote the time m- and unit i-specific collection $C(i,m) = (Y_i(m), A_i(m), \bar{O}(m-1))$; similarly, we write $C(m) = (Y(m), A(m), \bar{O}(m-1))$ as the time m- specific record. Let $L(\hat{\Psi}_{t,g_t}(P_{n,t}))(C(m))$ define a loss function for the time- specific target, such that $L(\hat{\Psi}_{t,g_t}(P_{n,t})) : \mathcal{C} \to \mathbb{R}$. By construction, a valid loss for a given parameter of interest is defined as a function whose true conditional mean is minimized by the true value of the target. For instance, for the time- specific target we then have that

$$P_{0,\bar{O}(t-1)} L(\hat{\Psi}_{t,g_{0,t}}(\bar{Q}_{0,t}))(C(t)) = \min_{\hat{\Psi}_{t,g_t}(\bar{Q}_t)} P_{0,\bar{O}(t-1)} L(\hat{\Psi}_{t,g_t}(\bar{Q}_t))(C(t)).$$

For a binary outcome, we can further define $L(\hat{\Psi}_{t,g_t}(P_{n,t}))$ as the inverse weighted mean squared error function (MSE), which is the loss we are trying to minimize

$$\frac{1}{n} \sum_{i=1}^{n} \frac{1}{g_{i,t}(A_i(t) \mid \bar{O}_i(t-1))} \left(Y_{i,g_{i,t}}(t) - \hat{\Psi}_{i,t,g_{i,t}}(P_{n,t})(\bar{O}(t-1)) \right)^2.$$

Consequently, the true risk is defined as the expected value of the loss evaluated w.r.t the true distribution. As such, it establishes the true measure of performance for the target parameter with respect to the specified loss — however, it is unattainable quantity, as the truth is unknown. In order to obtain an unbiased estimate of the true risk, we instead resort to its cross-validated estimate. Let $P_{n,t}^0$ denote the empirical distribution of the training sample until time t, with $P_{n,t}^1$ the corresponding empirical distribution of the validation set for any cross-validation scheme (CV). In general, we use different cross-validation schemes to evaluate how well an estimator trained on specific samples' past is able to predict an outcome for samples in the future, reflected in different empirical distributions $P_{n,t}^0$ and $P_{n,t}^1$.

For an infectious disease, we might expect its trajectory to vary over time, but have a similar profile across close time points. Therefore, we let $P_{n,t}^0$ be the empirical distribution

of all the data until time t, with $P_{n,t}^1$ consisting of samples at the next time step $t+1$. The cross-validated risk over all times then corresponds to

$$R_{CV,m} = \frac{1}{\tau - m} \sum_{m=t+1}^{\tau} L(\hat{\Psi}_{m-1,g_{m-1}}(P_{n,m-1}^0))(C(m)).$$

Let $\hat{\Psi}_{t,g_{t,s}^*}(P_{n,t})$ denote the estimator of the target parameter under design $g_{t,s}^*$, where we have a separate $\hat{\Psi}_{t,g_{t,s}^*}(P_{n,t})$ for each $s \in S$. We can evaluate the performance of each stochastic intervention $g_{t,s}^*$ using the loss-based framework. The proposed evaluation therefore proceeds as follows: with each new $C(t+1)$, evaluate the loss $L(\hat{\Psi}_{t,g_{t,s}^*}(P_{n,t}^0))(C(t+1))$ for each $s \in S$; add this loss to the current estimate of the online CV risk; update each online estimator $\hat{\Psi}_{t,g_{t,s}^*}$ into $\hat{\Psi}_{t+1,g_{t+1,s}^*}$ using $C(t+1)$. Upon observing the next batch of data, $C(t+2)$, the process is repeated. The Online CV risk gives us estimated performance of the adaptive design over time. We can use the full online CV risk, or an average over a more recent window in order to pick among proposed designs at each time point. We define the discrete SL design s_t as the design which minimizes the online CV risk:

$$\begin{aligned} s_t &= \min_s \frac{1}{\tau_w - m} \sum_{m=t+1}^{\tau_w} L(\hat{\Psi}_{m,g_{m,s}^*}(P_{n,m}))(C(m)) \\ &= \min_s \frac{1}{\tau_w - m} \sum_{m=t+1}^{\tau_w} \frac{1}{n} \sum_{i=1}^{n} \frac{1}{g_{i,m,s}^*(A_i(t) \mid \bar{O}_i(t-1))} \\ &\quad \left(Y_{i,g_{i,m,s}^*}(t) - \hat{\Psi}_{i,m,g_{i,m,s}^*}(P_{n,m})(\bar{O}(m-1))\right)^2, \end{aligned}$$

where τ_w is a future time point based on the window size w.

TMLE- and TMLE-CI-based selector

Thinking further in terms of an adaptive sequential design being an online algorithm, we could use the past data in order to fit the likelihood of O. At each time step t, we run a simulation under a different design $g_{t,s}^*$, and select the one that optimizes the short term mean outcome. The loss-based selector in the previous section optimizes over a window of recent losses (e.g., inverse weighted MSE over a window of time points). We could instead optimize for the MSE such that we also have inference for the target parameter — allowing us to pick a design by taking into account uncertainty in the point estimate as well. This motivates a new selector, based on the Targeted Minimum Loss Estimation (TMLE) [139, 138, 137]. In order to derive a TMLE, we utilize one of the working model outlined in Section 4.2. The estimated mean outcome under each of the s designs is a TMLE based on the working model, optimized for MSE with inference. Here we emphasize that reliance on working models is a necessary step in order to obtain a ranking of designs based on the TMLE, but our proposed method does not rely on assumptions imposed by the working models.

For the TMLE-based selector, we want to obtain a TMLE of each s-specific conditional mean outcome $\psi_{t,g^*_{t,s}}$. The standard TMLE, as originally defined by [139], first computes an initial estimator of $\bar{Q}_{0,i,t}$. In general, the functional form of $\bar{Q}_{0,i,t}$ is unknown, with arbitrary dependence structure. In order to avoid unnecessary assumptions, we resort to data-adaptive predictive methods such as the Online Super Learner under various working models; as such, we allow for flexibility in the specification of the functional form and dependence structure. Consistent estimation of $\bar{Q}_{0,i,t}$ is key for achieving asymptotic efficiency of the target parameter [138, 137]. We denote the initial estimator of the conditional mean outcome given the past as $\hat{\bar{Q}}_{i,t}(P^0_{n,t})$, trained on the training data available until time t, $P^0_{n,t}$.

The initial estimator of the conditional mean outcome given the past is then updated in such a way that the efficient influence function (EIF) estimating equation is zero when computed at the updated estimate. The TML estimators defined in this way generally require optimizing a loss function iteratively for the likelihood of the observed data. Achieving a solution to the EIF estimating equation guarantees, under regularity assumptions, that the estimator enjoys optimality properties such as double robustness and local efficiency [139, 138, 137]. We solve the estimating equation by fitting the following logistic model

$$\text{logit } \hat{\bar{Q}}_{t,\epsilon}(A(t), C_{A_i}(t)) = \text{logit } \hat{\bar{Q}}_t(A(t), C_{A_i}(t)) + \epsilon,$$

with weights defined as $w_t = g^*_t(A(t) \mid C_{A_i}(t))/g_t(A(t) \mid C_{A_i}(t))$. We emphasize that $g_t(A(t) \mid C_{A_i}(t))$ denotes the treatment mechanism generating the data so far. The estimate of ϵ is written as ϵ_t, with the updated initial estimator of $\bar{Q}_{0,i,t}$ evaluated at $(A_i(t), C_{A_i}(t))$ denoted as $\hat{\bar{Q}}^*_{i,t}(A_i(t), C_{A_i}(t)) = \hat{\bar{Q}}_{i,t,\epsilon_t}(A_i(t), C_{A_i}(t))$. The targeted estimate $\hat{\bar{Q}}^*_{i,t}(A_i(t), C_{A_i}(t))$ then solves the following EIF estimating equation,

$$\frac{1}{n}\sum_{i=1}^{n} \frac{g^*_{t,i}(A_i(t) \mid C_{A_i}(t))}{g_{t,i}(A_i(t) \mid C_{A_i}(t))}(Y_i(t) - \hat{\bar{Q}}^*_{i,t}(A_i(t), C_{A_i}(t))) = 0.$$

The TMLE of the s-specific stochastic intervention is defined as the plug-in estimator under the targeted estimate $\hat{\bar{Q}}^*_{i,t}$ and $g^*_{i,t,s}$,

$$\Psi_{t,g^*_{t,s}}(\hat{\bar{Q}}^*_t) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{\hat{\bar{Q}}^*_{i,t}, g^*_{i,t,s}}[Y_i(t) \mid C_{A_i}(t)].$$

In the following, we refer to Equation (4.10) denoting the time- and sample- specific target parameter in order to more easily define the canonical gradient and the first order expansion. The desired target parameter and the subsequent analysis is then defined as an average over samples, under the working model $\mathcal{M}^{tn}(\bar{O}(t-1))$.

Lemma 7 (Canonical gradient and first order expansion)**.** *Under the positivity assumption, target parameter mapping $\Psi_{i,t,g_{i,t}} : \mathcal{M}^{tn}(\bar{O}(t-1)) \to \mathbb{R}$ is pathwise differentiable with respect to $\mathcal{M}^{tn}(\bar{O}(t-1))$ and has a canonical gradient defined as*

$$\phi_{P_{\bar{O}(t-1)}} = D^*_{\bar{O}(t-1)}(\bar{Q}_{i,t})(o) = \frac{g^*_{t,i}(A_i(t)|C_{A_i}(t))}{g_{t,i}(A_i(t)|C_{A_i}(t))}(Y_i(t) - \bar{Q}_{i,t}(A_i(t), C_{A_i}(t))).$$

The time- and sample- specific parameter admits the following first order expansion:

$$\Psi_{i,t,g_{i,t}}(\bar{Q}_{i,t}) - \Psi_{i,t,g_{0,i,t}}(\bar{Q}_{0,i,t}) = -P_{0,\bar{O}(t-1)} D^*_{\bar{O}(t-1)}(\bar{Q}_{i,t}) + R(\bar{Q}_{i,t}, \bar{Q}_{0,i,t}, g_{i,t}, g_{0,i,t}),$$

where R is a second order remainder that is doubly-robust, with $R(\bar{Q}_{i,t}, \bar{Q}_{0,i,t}, g_{i,t}, g_{0,i,t}) = 0$ if either $\bar{Q}_{i,t} = \bar{Q}_{0,i,t}$ or $g_{i,t} = g_{0,i,t}$.

Since we are in a randomized trial and the treatment mechanism is known, the second order remainder in Lemma 7 is zero. All further theoretical analysis relies on the fact that the difference between the TML estimator and the target can be decomposed as the average of a martingale difference sequence, as shown in Theorem 15.

Theorem 15 (Asymptotic Normality of the t-specific TMLE)**.** *Under the working model $\mathcal{M}^{tn}(\bar{O}(t-1))$, the difference between the TMLE and its target decomposes as*

$$\begin{aligned}
\Psi_{t,g_t^*}(\hat{\bar{Q}}_t^*) - \Psi_{t,g_{0,t}}(\bar{Q}_{0,t}) &= \frac{1}{n}\sum_{i=1}^{n} D^*_{\bar{O}(t-1)}(\hat{\bar{Q}}^*_{i,t})(O_i(t)) - E_{\bar{Q}_{0,t},g_{0,t}} D^*_{\bar{O}(t-1)}(\hat{\bar{Q}}^*_{i,t}) \\
&= \frac{1}{n}\sum_{i=1}^{n} \frac{g^*_{t,i}(A_i(t)|C_{A_i}(t))}{g_{t,i}(A_i(t)|C_{A_i}(t))}(Y_i(t) - \hat{\bar{Q}}^*_{i,t}(A_i(t), C_{A_i}(t))) \\
&\quad - E_{\bar{Q}_{0,t},g_{0,t}}\left[\frac{g^*_{t,i}(A_i(t)|C_{A_i}(t))}{g_{t,i}(A_i(t)|C_{A_i}(t))}(Y_i(t) - \hat{\bar{Q}}^*_{i,t}(A_i(t), C_{A_i}(t)))\right]
\end{aligned}$$

Under weak conditions we have that

$$\Psi_{t,g_t^*}(\hat{\bar{Q}}_t^*) - \Psi_{t,g_{0,t}}(\bar{Q}_{0,t}) \xrightarrow{d} \mathcal{N}(0, \sigma_t^2),$$

where σ_t^2 is the asymptotic variance.

Theorem 16 (Asymptotic Normality of the TMLE)**.** *Under working model $\mathcal{M}^{tn}(\bar{O}(t-1))$ and weak conditions, we have that*

$$\psi - \psi_0 \xrightarrow{d} \mathcal{N}(0, \sigma^2),$$

where $\psi = \frac{1}{\tau}\sum_{t=1}^{\tau} \Psi_{t,g_t^*}(\hat{\bar{Q}}_t^*)$, $\psi_0 = \frac{1}{\tau}\sum_{t=1}^{\tau} \Psi_{t,g_{0,t}}(\bar{Q}_{0,t})$, $\hat{\bar{Q}}_t^*$ is the TMLE and σ^2 is the asymptotic variance.

Relevant derivations and conditions are presented in previous chapters. Note that under Theorem 15, each adaptive design $g^*_{t,s}$ has its corresponding asymptotic variance $\sigma^2_{t,s}$. We can estimate $\sigma^2_{t,s}$ using the empirical variance estimator as follows

$$\hat{\sigma}^2_{t,s} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{g^*_{t,i}(A_i(t)|C_{A_i}(t))}{g_{t,i}(A_i(t)|C_{A_i}(t))}(Y_i(t) - \hat{\bar{Q}}^*_{i,t}(A_i(t), C_{A_i}(t)))\right)^2,$$

corresponding to the variance of the EIF. Therefore, each adaptive design $g^*_{t,s}$ has a confidence interval for its overall mean outcome given by

$$\frac{1}{n}\sum_{i=1}^{n}\int_a \hat{\bar{Q}}^*_{i,t}(a, C_{A_i}(t))g^*_{i,t,s}(a \mid C_{A_i}(t)) \pm 1.96\frac{\hat{\sigma}_{t,s}}{\sqrt{n}}.$$

We can define two different selectors based on the TML estimator. First, denoted as the the TMLE-based selector, chooses the design s_t among $s \in S$ that maximizes the point TMLE estimate of the target parameter, such that

$$s_t = \max_s \Psi_{t,g^*_{t,s}}(\hat{\bar{Q}}^*_t).$$

Alternatively, we can take advantage of the asymptotic normality of the TMLE under working model $\mathcal{M}^{tn}(\bar{O}(t-1)$. The second selector, denoted TMLE-CI-based selector, maximizes the lower bound of the confidence interval — picking the design s_t with the highest minimum value of the confidence interval

$$s_t = \max_s [\Psi_{t,g^*_{t,s}}(\hat{\bar{Q}}^*_t) - 1.96\sigma_{t,s}/\sqrt{n}].$$

Either way, s_t corresponds to the discrete Online Super Learner selector, and at time point $t+1$ we use the design g^*_{t,s_t} in order to assign tests to all subjects.

4.4 Agent-based model of the university campus

In the following, we illustrate utility of the proposed Online Super Learner for adaptive sequential surveillance by simulating an environment which models the University of California, Berkeley in the Fall of 2020. In addition, we elaborate on how all testing strategies (*risk-based*, *symptomatic*, *contact tracing*) can be seen as stochastic interventions with different sampling strategies. In Section 4.5 we show the performance of the proposed adaptive design for surveillance using the closed sample population and infectious disease dynamics described below.

Model Description

We develop a dynamic, agent-based model of transmission of SARS-CoV-2 on a residential university campus. The model parameters reflect transmission dynamics among students and faculty on a medium-size public university with on- and off-campus living arrangements, class number and size compatible with the University of California, Berkeley. In particular, we present an agent-based model for a setting with 20,000 university-affiliated individuals with varying age, social network, class size, risk-level and living accommodations in a U.S. college town. While we parameterized the model according to U.C. Berkeley, our simulations can be easily modified to reflect any residential higher-education institution. In the following subsections, we describe the core features, dynamics and assumptions underlying the agent-based model.

Population and Network Structure

We focus on a residential campus community as the target population, broadly reflecting students and faculty/staff at a medium size university. During the pandemic, most universities implemented a wide array of measures aimed at limiting transmission on campus — including mandatory vaccination, mask-wearing, social distancing, enhanced cleaning protocols, increased availability of sanitizing products and cancelling large social gatherings. While each of these interventions has benefits, we focus solely on the optimal testing strategy. By not including other interventions which can alleviate the spread of the infectious disease, we aim to investigate performance of the proposed method under the worst-case scenario. We don't model any breaks or high-risk events during the semester, such as holiday travel. Most importantly, we assume the modeled population is constant, where individuals can only be removed due to isolation, but no new individuals can join. While we model the campus population as a closed community, infections induced by interactions with individuals outside the campus population are allowed - thus providing a steady stream of new cases.

We assume the campus population is divided into distinct groups with different collective behavior, dictated by their covariates. For example, we model three sub-populations — students living on-campus, students living off-campus, and faculty/staff, as shown in Table 4.1. The distinct groups are differentiated by baseline covariates, underlying risk of infection, and different degrees of interaction within one's network. The age distribution is modeled in order to reflect a predominately younger campus population, with 6 age categories (< 18, $18-28$, $29-38$, $39-48$, $49-68$, > 68) sampled with probabilities $(0, 1, 0.5, 0.2, 0.07, 0.07, 0.06)$, respectively. The baseline risk of infection is sampled from a *beta(1, n/36000)* distribution, reflecting dependence on the size of the observed population. The modeled network structure consists of several components, including off-campus living for both students and faculty, on-campus housing, in person classes, and random exposure. Each network type has an unique probability of transmission within a graph, with highest being for students in a communal housing. Below, we describe each component of the network structure in more details.

1. **Off-campus housing for students:** We model off-campus housing separately for students and faculty. Students of similar age category are grouped in off-campus housing units, with household sample size drawn from $(1-8)$ range from a *negative binomial* distribution with $\mu = 2$. While we concentrate on the university campus population, this is not exclusively represented in the observed housing structure population, with students being vastly dispersed in the region. With that in mind, we assign housemates randomly to a total of $n * 0.01$ off-campus houses of varying household size. The probability of transmission among the same household is 0.03.

2. **Off-campus housing for faculty and staff:** We model off-campus housing for faculty/staff similar to the student off-campus accommodations. Individuals older that 28 years old are candidates for faculty/staff housing. We sample the size of a faculty/staff household from $(0-2)$ range from a *negative binomial* distribution with $\mu = 0.5$. We assign housemates randomly with no age preference (accounting for

possible family ties). We distribute sampled household individuals to a total of $n*0.005$ faculty/staff houses. The probability of transmission among the same household is 0.03, as in off-campus living arrangements for students.

3. **On-campus housing:** The probability of a student being part of on-campus housing available at U.C. Berkeley is derived based on the 25 communal living buildings with 8908 students at maximum occupancy. During the regular school year, 14% of the student population is living in a university provided dorms and Greek housing. In the Summer of 2020, approximately 5% were projected to return to communal housing in the Fall of 2020, with an average occupancy of 2 students per room. We sample the size of a household from a *negative binomial* distribution with $\mu = 2$, and allocate people to the 25 available communal buildings with equal probability. We assume that the risk of transmission from the community to off-campus students is lower than the risk of transmission from the community to on-campus students, based on the evidence from a campus outbreak of H1N1 in 2009 [50]. We further assume that on-campus students living in congregate settings are considered at a higher risk for transmission. If a student is part of the communal housing (dorms or Greek housing), the risk of infection increases by 0.01 in addition to the regular housing risk.

4. **In-person classes:** We modeled in-person, online and hybrid available classes in Fall of 2020 based on the suggested schedule announced by the U.C. Berkeley administration in the Summer of 2020. This consisted of 111 hybrid and 98 flexible classes, and up to 314 in-person lectures with a maximum of 25 students per class. This is in accordance with the wide-spread university policy to strive for majority online classes, with few small-in-size in-person lectures and staggered class times in order to decrease student contact. We sample the in-person class size from a *negative binomial* distribution with a minimum of 15 and maximum of 25 students per class. We further assume 18% of total student and faculty/staff population would return to U.C. Berkeley for in-person classes and university-based responsibilities. The probability of transmission for an in-person attendance among class members and people that frequent the same classroom/building is 0.01.

5. **Random:** Finally, we account for random exposure from people being in close contact during their regular day-to-day activities (e.g., taking public transportation). We model random exposure as the number of people outside one's network that come in close contact with the individual in question. We sample the number of nodes in a random graph from the *negative binomial* distribution with a minimum of 3 and maximum of 25 encounters per day. This allocation is dependent on individual risk, and corresponds to the latent part of the network structure. The probability of transmission in a random graph is 0.005. In Appendix Section 4.7 we investigate how the proposed method responds to increased individual risk.

Individual Disease Progression

We assume that the disease evolution always progresses through set stages in each infected individual, as exemplified in Figure 4.4 and Table 4.2. We separate the overall student and faculty/staff populations into the following compartments at each time step: susceptible (S), exposed (E), detectable infectious (It), symptomatic infectious (Is), asymptomatic infectious (Ia), recovered (R) and isolated (I). The probability of a new infection, conditional on being susceptible, depends on the hazard at the current time-step t. In particular, the hazard function takes into account the individual's current stage, time spent in the state, one's full network, individual risk and covariates, and the current state of the epidemic; the more infectious people and the more advanced the epidemic, the greater the risk of a new infection. If an individual is never infected, they remain susceptible until the end of the observation period. Infected individuals advance to the next compartment, or remain in their current one, stochastically at each time step. The transition probabilities and average length of stay for each compartment are modeled based on the literature available in Summer of 2020, and described in more details in the following. We initiate the infectious disease trajectory with 8 exposed, 2 temporarily infectious and 2 symptomatic infectious cases of SARS-CoV-2 infection at the start of data collection.

1. **Susceptible (S):** Except for the seeded samples, all individuals start as susceptible. The S stage has no time to next state, as all units remain susceptible until infection or end of the semester.

2. **Exposed (E):** If infected, a susceptible individual transitions to an exposed status. This stage is not yet infectious due to a low viral load, and it is not detectable via testing. Exposed units spend 4.5 days on average as exposed, with the number of days in the E compartment sampled from $gamma(9, 2)$ distribution.

3. **Detectable Infectious (It):** From the E compartment, exposed individual transitions to a temporary state It. Each individual spends 1 day on average as It, with the number of days in the detectable infectious compartment sampled from the $gamma(1, 1)$ distribution. During this transition period, individuals are not symptomatic, but are infectious. The compartment It is also the first stage of the disease trajectory at which the infection can be detected via a test. We model the ability to infect others as reduced while It, but increasing with each time step with a peak at the transition to the next compartment. Since a significant fraction of COVID-19 patients are asymptomatic, especially within the younger population, we divided the next infectious compartment into symptomatic and asymptomatic with distinct transition probabilities and duration of the infectious state. Determination whether one is asymptomatic or symptomatic is a Bernoulli trial with success probability depending on the age of the exposed individual. For instance, all samples within the age gap < 18 to 28 were symptomatic with success probability 0.4, $29 - 48$ with probability 0.6, and the oldest members of the campus probability were symptomatic with probability 0.8.

4. **Symptomatic (Is):** Infected individual can transition to a symptomatic class following the temporarily infectious state. One spends 13 days on average as a symptomatic patient, with the number of days in the Is stage sampled from the $gamma(13, 1)$ distribution. The symptomatic compartment encompasses any COVID-19 symptom, or combination of, that warrants a test; we assume a patient that is symptomatic exhibits symptoms the entire time while Is. Consequently, a test can be requested and administered at any time point during the symptomatic stage, not just at the onset of symptoms. As a Is, a person is infectious for the duration of the state; however, their infectiousness decreases the longer they are in the current stage.

5. **Asymptomatic (Ia):** Infected individual can also transition to an asymptomatic infectious state following the It compartment. If asymptomatic, one spends 7.5 days on average as Ia, with the number of days in the current stage sampled from the $gamma(7.5, 1)$ distribution. While there is still emerging research in this area, we assume relative infectiousness for the asymptomatic class to be less than for the symptomatic class; the reduction in transmission probability is assumed at 39% [94]. We let asymptomatic individuals remain asymptomatic for the duration of the infection, assuming no transition from Ia to Is compartment. As the campus population is predominately young, the percentage of asymptomatic individual is higher than in the general population. This makes detecting active infections a more difficult task for the campus community.

6. **Recovered (R):** An individual recovers by going through the full disease cycle, $(S, E, It, Is/Ia, R)$. As the campus population is predominately young and healthy, we don't model death as an outcome. Instead, all individuals eventually reach a terminal state R. We assume recovered individuals obtain at least a temporary immunity that lasts the length of the semester. Therefore, recovered individuals do not become susceptible again. Studies are still ongoing regarding the duration, if any, of temporary immunity [148].

7. **Isolated (I):** Finally, an infected member of the campus population can be identified via testing at any point of the disease trajectory, except for the E and R state. Positive test results at time t lead to possible contact tracing at $t + 1$ for the complete known network. Individuals administered a test and diagnosed positive for COVID-19 are isolated. Quarantine conditions are modeled on a continuum via the isolation factor, in an attempt to mimic realistic conditions. In particular, if the isolation factor is 0, quarantine is modeled as isolation with complete reduction in one's contact rate for the duration of the infection. For higher isolation factor, detected infectious individual is still able to infect others at a reduced rate. If caught, the infected individual goes though all the usual disease stages with a varying isolation level, until it reaches a terminal R state. As mentioned previously, isolated, and eventually recovered patients, do not become susceptible again.

Table 4.1: Simulated university population during COVID-19 pandemic

Parameter	**Population**	**On-campus housing**	**In-person class**
Value (total or %)	20,000	5%	18%

Table 4.2: Simulation model parameters for the COVID-19 pandemic

Parameter	**Symbol**	**Average (days)**	**Range (Q1,Q3)**	**Distribution**
Latent Period	**E**	4.5	(3.4, 5.4)	$gamma(9,2)$
Detectable and Infectious	**It**	1	(0.7, 1.4)	$gamma(1,1)$
Asymptomatic	**Ia**	7.5	(5.5, 9.1)	$gamma(7.5,1)$
Symptomatic	**Is**	13	(10.4, 15.2)	$gamma(13,1)$

Testing Strategies as Stochastic Interventions

We can describe a wide range of testing schemes by a stochastic intervention g_t^*. The familiar case of static interventions — defined by setting $A_i(t)$ to a value a in its support $\mathcal{A}$ — can be recovered by choosing degenerate candidate distributions which place all mass on just a single value [118, 30]. In particular, we can conceptualize a testing strategy as an intervention that assigns all individuals testing allocations $\{A_1(t), \ldots, A_n(t)\} \in \{0,1\}^n$ at time t in a two-step procedure. Operationally, we delineate between the probability of receiving a test based on the available history, and how the said probabilities are used in order to output a final testing decision. For example, g_t^* could assign tests based on the ranking of the probability of receiving a test given the past, or based on the probability itself. Alternatively, g_t^* could be a deterministic intervention conditional on the observed past (e.g., *rule-based*, static interventions). Let f denote a function that takes the probability of being assigned a test given the past, and outputs either a stochastic or deterministic rule as to how such probabilities are to be used to assign a test. We formally define the stochastic intervention g_t^* as a function that maps every f to $g_t^*(f) : (Pa(A(t))) \mapsto g_t^*(f)(1 \mid Pa(A(t))) = g_t^*(f)(1 \mid \bar{O}(t-1))$. If f is an identity function, and each sample is to receive a test just based on its conditional probability, we simply write $g_t^*(f)(1 \mid Pa(A(t))) = g_t^*(1 \mid Pa(A(t)))$.

In the following, we define several testing strategies used to pick k percent of the population to be tested — from a new *risk-based* testing scheme to commonly allocated strategies, such as *symptomatic* testing and *contact tracing*. In connection to our general stochastic intervention framework, we emphasize that the testing allocation could be a static rule depending on just the current knowledge of the network, recovering the *contact tracing* testing scheme. On the other hand, a static rule depending on the reported symptoms results in *symptomatic* testing. Instead of relying on simple rules, g_t^* could depend on the current estimate of $\bar{Q}_{0,i,t}$, in which case we can incorporate the current risk of being infected as part of the test allocation strategy. Further, we can then sample based on the current estimate of $\bar{Q}_{0,i,t}$, or pick the top k percent ranked samples. We compare all proposed testing strategies

to benchmarks where we either know the true unknown status of each individual or the true $\bar{Q}_{0,i,t}$. We denote as "realistic" all testing strategies that can be implemented in the general population during an epidemic.

Testing allocation functions

Let f denote any function that takes as input the conditional probability of being assigned a test given the observed past, and outputs either a stochastic or deterministic rule as to how such probabilities are to be used to assign a test. The function f, together with the collected past, defines a stochastic intervention g_t^* at time t. We study two such functions: $f = f_S$ and $f = f_R$, defining sample and rank functions, respectively. We define f_S as an identity function such that

$$g_t^*(f_S)(1 \mid \bar{O}(t-1)) \equiv g_t^*(1 \mid \bar{O}(t-1)) = p^*_{a_i(t)}(A_i(t) \mid \bar{O}(t-1)). \tag{4.12}$$

With that, $g_t^*(f_S)(1 \mid \bar{O}(t-1))$ is a stochastic intervention which assigns tests according to the probability of being tested given the past. On the other hand, f_R ranks the current estimate of the conditional probability of being tested and allocates k percent of tests to the top ranked samples. In particular, let S_P denote the survival function of $g_t^*(1 \mid \bar{O}(t-1))$ such that $c \mapsto P(g_t^*(1 \mid \bar{O}(t-1)) > c)$. Then, we can define $c^* \equiv \inf\{c : S_P(c) \leq k\}$ as the cutoff at which at most k percent of individuals get tested based on S_P. We define the rank-based stochastic intervention as

$$g_t^*(f_R)(1 \mid \bar{O}(t-1)) \equiv \mathbb{I}(g_t^*(1 \mid \bar{O}(t-1)) > c^*), \tag{4.13}$$

which allocates tests to the top k percent of individuals with the highest ranked probability of treatment given the observed history. In the following subsections, we describe and compare a range of testing allocation schemes based on $g_t^*(f)(1 \mid \bar{O}(t-1))$ for each $i \in [n]$ and $t \in [\tau]$, with $f = f_S$ and $f = f_R$.

Realistic Testing Strategies

The "realistic" testing strategies include test allocations often described in the literature, including *symptomatic* and *contact tracing*. In addition, it includes the new *risk-based* strategy based on the current estimate of the conditional expectation of $Y_i(t)$ given the observed past. In the following, we assume $L(t)$ includes covariates that describe the current presence of symptoms associated with the infectious disease in question ($L^{\text{symp}}(t)$) and the time t network of each individual ($F(t)$).

1. **Symptomatic Testing**
   One of the most commonly described test allocation strategy in the literature is *symptomatic* testing. Briefly, *symptomatic* testing entails giving a test to k percent of

individuals with reported symptoms, where $L_i^{\text{symp}}(t) = 1$ describes a symptomatic patient at time t. We can denote such testing strategy as the following deterministic intervention g_t^* at t:

$$g_t^*(f)(1 \mid \bar{O}(t-1)) = \begin{cases} 1, & L_i^{\text{symp}}(t) = 1 \\ 0, & L_i^{\text{symp}}(t) = 0, \end{cases}$$

where $E[g_t^*(f)(1 \mid \bar{O}(t-1))] \leq k$.

2. **Contact Tracing**
Another commonly described and implemented testing strategy is based on *contact tracing*. Here, each individual with a current positive test has their entire known network tested for the same cause as well, while respecting the testing resource constraint. The sample i's network consists of family, friends, colleges and all other individuals who came in close contact with an infected individual within a specified time frame. Typically the more comprehensive one's network, the more effective contact tracing is as a testing strategy. We write *contact tracing* as the following deterministic intervention at time t:
$$g_t^*(f)(1 \mid \bar{O}(t-1)) = \begin{cases} 1, & F_i(t) = 1 \\ 0, & F_i(t) = 0. \end{cases}$$

3. **Random**
Random testing corresponds to assigning tests with equal probability to k proportion of available individuals. Here, no information on the samples, or the current state of the epidemic trajectory, is used to assign tests. The stochastic intervention g_t^* corresponds to assigning uniform testing weights
$$g_t^*(f)(1 \mid \bar{O}(t-1)) = 1/n,$$
at each time $t \in [\tau]$ and for all samples $i \in [n]$.

4. **Risk-Based Testing**
Instead of relying solely on the current symptoms, known network of each patient or the combination of both — we can instead incorporate an estimate of the current risk of being infected into the testing scheme. In particular, we use the current estimate of $\bar{Q}_{0,i,t}$ fit on the training set and available covariates in order to assign tests at the next time step. We can define the *risk-based* testing scheme as the following intervention at time t:
$$g_t^*(f)(1 \mid \bar{O}(t-1)) \equiv f(\hat{\bar{Q}}_{i,t}(Pa(L(t)))$$
which assigns tests based on the current estimate of one's risk of being infected, given their observed past.

Benchmark Testing Strategies

In this subsection, we define several benchmark testing strategies used to evaluate performance of above described "realistic" testing schemes. As previously noted, most benchmark allocations are impossible to implement in practice due to ethical concerns, dependence on unknown statistical quantities, or latent variables.

1. **No Testing**
   The first benchmark strategy includes the natural progression of the epidemic when no individuals are isolated, and no tests are allocated. As such, we define g_t^* at each time $t \in [\tau]$ and for all samples $i \in [n]$ as $g_t^*(f)(1 \mid \bar{O}(t-1)) = 0$.

2. **True Risk**
   The true risk benchmark entails using the true risk function in order to assign tests at each time point $t \in [\tau]$, instead of the current estimate of $\bar{Q}_{0,i,t}$. The true risk scheme is based on unknown components of the process, as we don't know $\bar{Q}_{0,i,t}$. While not possible to implement in practice (unless we a priory know $\bar{Q}_{0,i,t}$), it is a useful benchmark for *risk-based* testing strategies, setting an upper limit on their performance over time. We can denote the true risk testing benchmark as the following intervention at time t:
   $$g_t^*(f)(1 \mid \bar{O}(t-1)) \equiv f(\bar{Q}_{0,i,t}(Pa(L(t)))$$
   which assigns tests based on the true risk of being infected at time t, given the observed past.

3. **Perfect (True Status)**
   The true status benchmark entails testing k proportion of individuals with current active infection. As $Y^l(t)$ is a latent variable, it is not observed. While impossible to implement in practice, the true status benchmark serves as an upper bound on the effectiveness of testing for infection control. We define it as the following deterministic intervention g_t^* at time t:
   $$g_t^*(f)(1 \mid \bar{O}(t-1)) = \begin{cases} 1, & Y_i^l(t) = 1 \\ 0, & Y_i^l(t) = 0. \end{cases}$$

4.5 Simulations

In the following we report simulation results testing the performance of the Online Super Learner for adaptive sequential surveillance in a closed population. We compare several different testing strategies using the agent-based model described in Section 4.4, which simulates transmission of SARS-CoV-2 on a residential university campus of moderate size. We focus on various stochastic interventions reflecting different testing strategies and design selectors including TMLE-, TMLE-CI- and loss-based, while assessing the state of the infection

spread during the length of an academic semester. In particular, we compare performance of the proposed Online Super Learner for adaptive surveillance using different selectors with commonly implemented *rule-based* testing strategies — including *symptomatic*, *contact tracing* and *random* testing. A simple *risk-based* strategy with $\bar{Q}_{0,i,t}$ learned using a generalized linear model is also used as a comparison (denoted as *risk-based* with glm). All designs are further compared to benchmarks, including no testing and when true infectious status is known ("perfect"), corresponding respectively to the lower and upper bounds of performance for any intervention.

All simulations results represent averages over 500 Monte Carlo draws and trajectory of $t = 120$ time points. While the size of the population is set to $n = 20,000$, we investigate performance of the proposed methodology under various resource constraints, testing $1\%-4\%$ of the total population at each time point (corresponding to $k = \{200, 400, 600, 800\}$). In the Appendix Section 4.7 we also consider different levels of outside transmission, reflected by the risk scale parameter; higher the risk scale score, higher the role of the latent parts of the network and individual risk on transmission dynamics. All simulations are initiated with 8 exposed, 2 temporarily infectious and 2 symptomatic cases of COVID-19. We purposely focus on the scenario where simple rule-based strategies might do well (knowing the network of the few infected individuals), and it is difficult to learn one's risk due to a limited number of infections. We also want to mimic a new start of a semester in an environment with a stable number of daily infections, as otherwise the in-person instruction might be omitted. While we only present results with the $(E = 8, It = 2, Is = 2, Ia = 0)$ configuration, other random seeds result in a similar design performance and ranking.

Testing Performance

We evaluate testing performance by the cumulative incidence curve at each time point and the cumulative percent of infected individuals by $t = 120$ (final cumulative incidence). The best performing testing strategy keeps the infection rate low over time, and achieves the lowest cumulative incidence at the end of observation. Testing performance is a function of testing strategy, number of available tests, and the number of currently infected individuals. As such, we evaluate multiple different testing designs (from simple *rule-based* and *risk-based*, to Online SL for adaptive surveillance with the TMLE-, TMLE-CI- and loss-based selector) under different resource constrains ($k = \{200, 400, 600, 800\}$) across the entire trajectory of the infectious disease progression.

The average cumulative incidence curves at each time point for all considered designs and available resources are shown in Figure 4.1. For instance, *random* strategy performs as good as no testing under $k = 200$, but gets better as more tests become available. Nevertheless, it is always outperformed by competing testing strategies in our simulations, no matter the number of available tests or starting conditions. The *symptomatic + contact* and the *risk-based* strategy with `glm` perform much better than *random* and no testing, with barely any overlap of trajectories. The Online SL for adaptive surveillance has the lowest cumulative incidence compared to competitors across all times and all selectors. Differences between

individual selectors occur at the beginning and end of the trajectory. As can be seen in Figure 4.1, loss-based approach starts with lower cumulative incidence, but gets outperformed by TMLE-based strategies towards the end of the trajectory. While testing only $1\% - 4\%$ of the total population at each time point, no design achieves performance of the oracle that knows the true infectious status. However, as more resources are allocated, cumulative incidence curves for all strategies get flatter, especially for the Online SL for adaptive surveillance.

The average cumulative incidence by time point $t = 120$ is shown in Figure 4.2; we refer to it as the average final cumulative incidence. The Online SL for adaptive surveillance with TMLE-CI selector outperforms all competing designs across all simulation setups — with the most stark difference at $k = 800$ (CI: $(1.8\%, 2.1\%)$ vs. $(2.4\%, 2.9\%)$ for the second best design, TMLE-based selector). As expected, the performance of the TMLE-CI selector gets better as a function of more available tests (CI with $k = 200$: $(14.9\%, 16.2\%)$; CI with $k = 400$: $(7.1\%, 8.0\%)$; CI with $k = 600$: $(3.6\%, 4.2\%)$; CI with $k = 800$: $(1.8\%, 2.1\%)$). As such, even with testing only 4% of the campus population, we can achieve good control of the infectious disease spread in our simulations. Compared to simple *rule-based* and *risk-based* competitors, Online SL for adaptive surveillance achieves much lower cumulative incidence by $t = 120$, across all proposed selectors. On average and across all resource constraints, Online SL with TMLE-CI selector has 1.6 lower cumulative incidence then implementing just *risk-based* testing strategy with glm. Compared with the *symptomatic + contact* scheme, which might be considered standard testing practice, mean final cumulative incidence was reduced from 23.8% to 15.5% at $k = 200$, and from 3.6% to 1.9% at $k = 800$ with the TMLE-CI selector. Figure 4.2 also shows that TMLE-based and loss-based selector have similar performance under higher k values (CI with $k = 200$: $(16.2\%, 17.4\%)$ vs. $(17.2\%, 18.6\%)$; CI with $k = 400$: $(8.1\%, 9.1\%)$ vs. $(9.6\%, 10.8\%)$; CI with $k = 600$: $(4.5\%, 5.4\%)$ vs. $(4.3\%, 5.0\%)$; CI with $k = 800$: $(2.4\%, 2.9\%)$ vs. $(2.5\%, 3.0\%)$). One reason for similar performance could be that the advantage of smooth transitions across designs achieved by weights in the TMLE-based selector is offset by averaging loss over a recent window (size 5 in simulations) in the loss-based selector. Ultimately, both TMLE- and loss-based strategies perform worse than the TMLE-CI selector in terms of final cumulative incidence — this could be explained by the fact that neither of the two selectors take into account uncertainty in the point estimates. Finally, as also observed in Figure 4.1, all designs perform better than no testing (except for *random* at $k = 200$) and worse than the oracle which knows the true infectious status at all values of k.

Finally, we investigate design performance under different levels of outside transmission governed by the risk scale parameter. In our agent-based model, higher values of the risk parameter correspond to higher weighted latent parts of the network and individual risk on transmission dynamics. Intuitively, higher the risk parameter, more difficult of a problem catching infectious individuals becomes. In Appendix Section 4.7 we show the cumulative incidence curve at each time point (Figure 4.4) and the final cumulative incidence by $t = 120$ (Figure 4.5). Similarly to previously reported results for the risk scale parameter (value 0.5), the Online SL for adaptive surveillance with TMLE-CI-based selector outperforms all other designs over a grid of risk parameters (values considered: 0.4, 0.6, 0.7). In fact, as the

problem becomes harder, TMLE-CI-based selector seems to gain even more advantage over the second best design (CI for risk parameter 0.4: (2.1%,2.5%) vs. (2.8%,2.9%); CI for risk parameter 0.5: (3.6%,4.2%) vs. (4.3%,5.0%); CI for risk parameter 0.6: (5.6%,6.5%) vs. (8.4%,9.9%); CI for risk parameter 0.7: (10.5%,12.1%) vs. (16.2%,18.1%)). In comparison to the standard testing practice, *symptomatic + contact*, Online SL for adaptive surveillance with TMLE-CI-based selector vastly outperforms across all risk scale parameters (CI for risk parameter 0.4: (2.1%,2.5%) vs. (3.5%,4.1%); CI for risk parameter 0.5: (3.6%,4.2%) vs. (6.8%,7.7%); CI for risk parameter 0.6: (5.6%,6.5%) vs. (12.1%,13.8%); CI for risk parameter 0.7: (10.5%,12.1%) vs. (21.9%,23.4%)). Comparison with the *risk-based* testing scheme remains parallel to *symptomatic + contact*, with high advantages over *random* testing as well, as shown in Figure 4.5. Finally, as expected, higher risk scale parameter corresponds to higher average final cumulative incidence under any design. Even for the TMLE-CI-based selector, as the risk scale jumps to 0.7, average final cumulative incidence at $k = 600$ is 11.3% vs. 3.9% at value 0.5. Nevertheless, no matter the level of difficulty of the problem, the ranking of the designs in terms of testing performance remains the same.

Picked Designs

Designs used as candidates in the Online SL for adaptive surveillance include a combination of *rule-based* and *risk-based* strategies. Figure 4.3 demonstrates picked designs (discrete Super Learners) over time and 500 simulations using the TMLE-CI-based selector. In particular, one of the candidates used is the canonical *symptomatic + contact* testing strategy with a known network. As shown in Figure 4.3, *symptomatic + contact* design is often picked at the beginning of the trajectory, while there is not a lot of information collected; as more data becomes available, it is picked less. All the risk-based strategies include Online Super Learners of $\bar{Q}_{0,i,t}$ with candidate algorithms which reflect working models described in Section 4.2. In particular, some of the candidates train on the full training history collected, windows of past points (window sizes used: $(7, 10, 14)$), and exponential weights of the form $(1 - \text{rate})^{\text{lag}}$ where rate is $(0.01, 0.05, 0.1)$. As shown in Figure 4.3, designs trained on full history are often picked at the beginning of the trajectory. As time progresses, candidates trained on particular extractions of the past win for most resource constraints, converging to a stable allocation of picked designs. This could be explained by the fact that, as the trajectory progresses, distant past becomes irrelevant in the current state of epidemic due to the fast evolving nature of an infectious disease. Once the selector learns that designs trained on more recent past does well, it starts picking them as discrete SL more consistently. Other working models not shown in Figure 4.3 include various network components, corresponding to different working models for the network dependence.

In addition to various working models for $\bar{Q}_{0,i,t}$, Online SL for adaptive surveillance can also pick among different sampling schemes. As described in Section 4.4, one can allocate tests to the top ranked samples or sample based on the estimated risk. As shown in Figure 4.3 each risk-based strategy was a candidate design based on both sampling and top sampling strategy; hence, if the full training data was used to estimate $\bar{Q}_{0,i,t}$, tests were allocated both

based on the top ranked samples and sample proportional to the estimated $\bar{Q}_{0,i,t}$. In terms of sampling schemes, top based strategy seems to outperform sampling across all simulation scenarios considered. This could be explained by the fact that the sample strategy introduces more exploration than necessary for this problem, especially as we achieve a good estimate of $\bar{Q}_{0,i,t}$ with more data over time. While a deterministic sampling strategy with a good estimate of $\bar{Q}_{0,i,t}$ is preferential in our simulations, it is important to keep the sampling option as a candidate design for cases when more exploration is necessary.

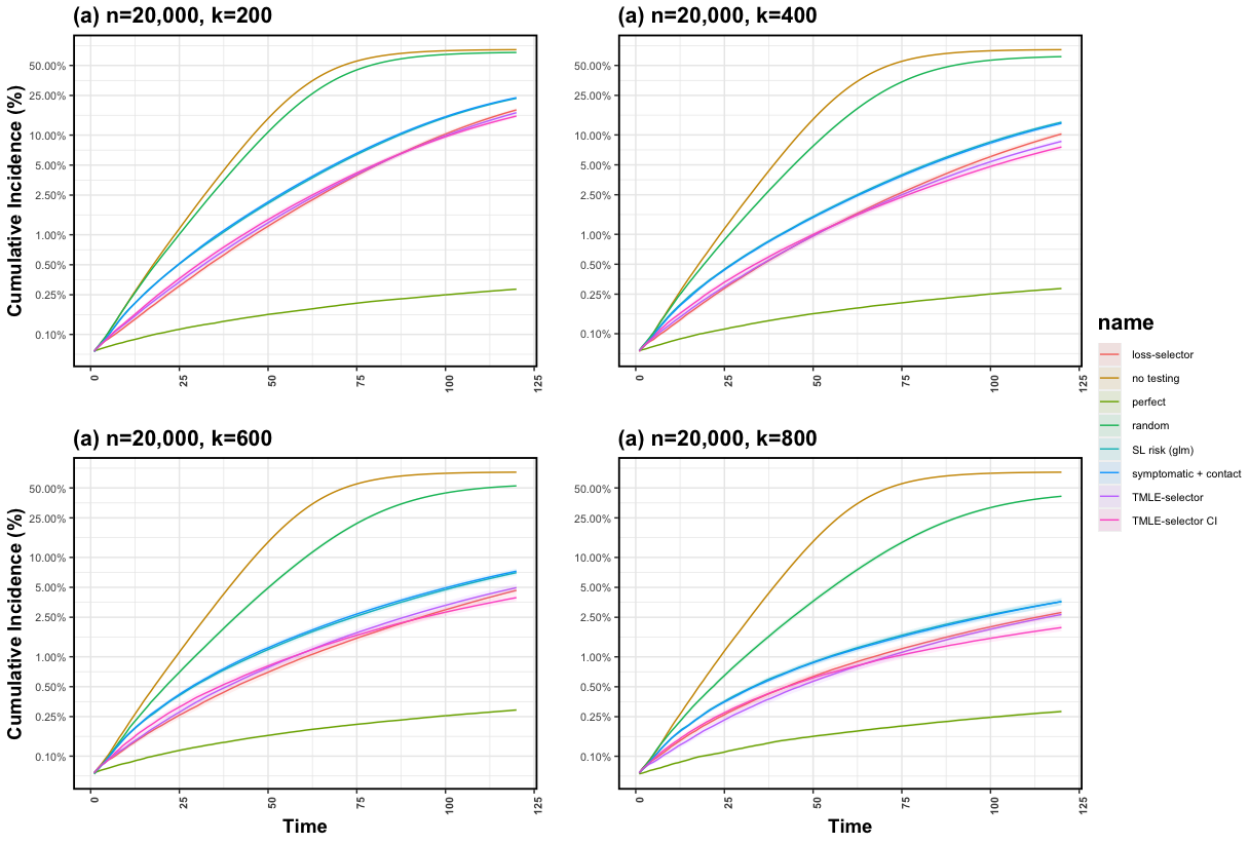


Figure 4.1: Average cumulative incidence at each time point over 500 simulations with $n = 20,000$ sample size and capacity $k = \{200, 400, 600, 800\}$ using TMLE-based, TMLE-CI-based and loss-based selectors of the testing strategy. We compare different proposed selectors to *symptomatic + contact*, *random* and *glm risk-based* testing, with *perfect* as the upper and *no testing* as lower bound on performance.

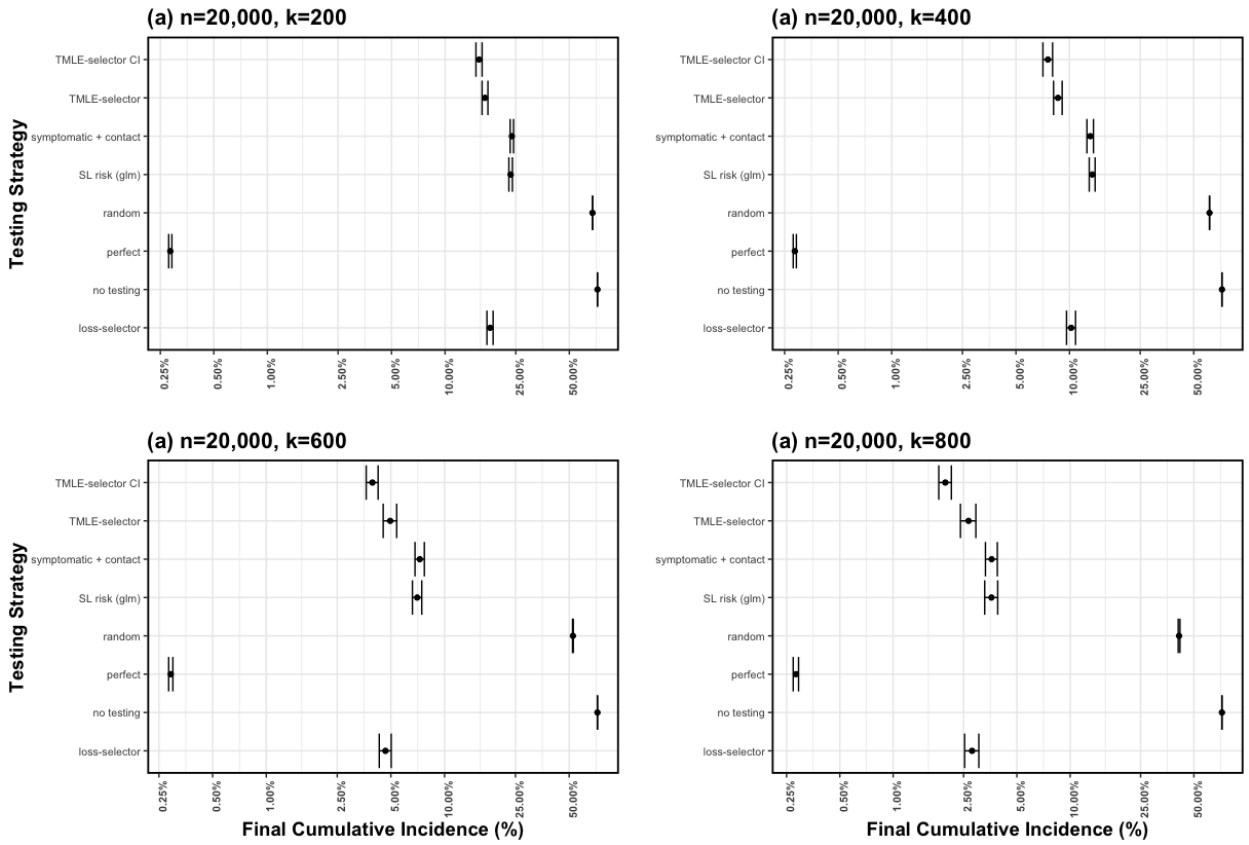


Figure 4.2: Average final cumulative incidence at $t = 120$ over 500 simulations, with $n = 20,000$ sample size and capacity $k = \{200, 400, 600, 800\}$ using TMLE-based, TMLE-CI-based and loss-based selectors of the testing strategy. We compare different proposed selectors to *symptomatic + contact*, *random* and *glm risk-based* testing, with *perfect* as the upper and *no testing* as lower bound on performance.

4.6 Discussion

In this work, we develop an Online Super Learner for the adaptive sequential design. Our proposed method is especially suited for infectious disease surveillance and control, or any adaptive sequential problem with unknown dependence within a fully nonparameteric model. The data setup constitutes a typical longitudinal structure of n individuals over a period of τ time points. Within each t-specific time block, one observes the exposure variable (e.g., indicator of testing), time-varying covariates (e.g., network structure, health status) and outcome (e.g., infectious status) for all n individuals.

Our causal target parameter is the mean outcome we would have obtained after one time-step, if, starting at time t given the observed past, we had carried out a stochastic

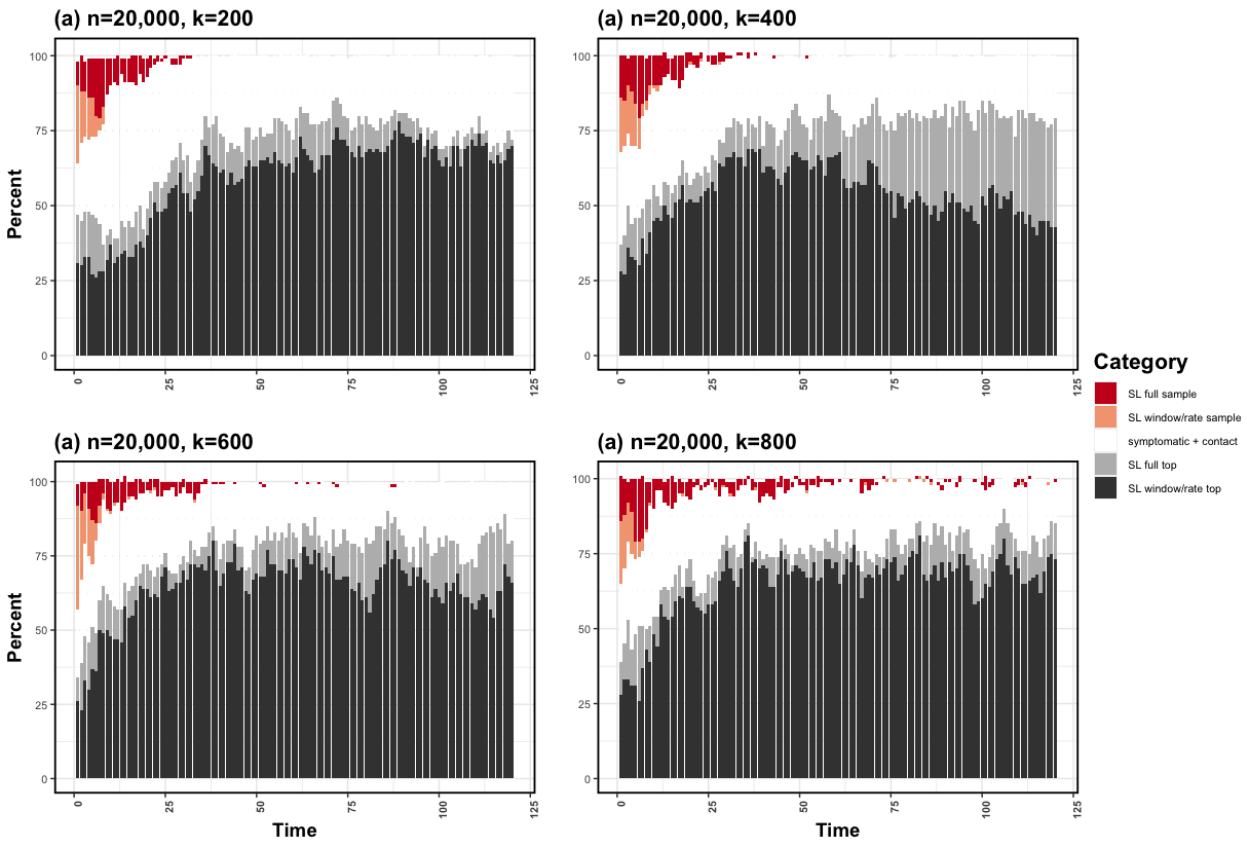


Figure 4.3: Percent design over the full trajectory and 500 simulations with $n = 20,000$ sample size and capacity $k = \{200, 400, 600, 800\}$ using TMLE-CI-based selector of the testing strategy. Candidate designs include *symptomatic + contact* and various *risk-based* designs where the Super Learner is either trained on the full past, exponentially weighted past, or a window of $t = 14$ days. All *risk-based* testing strategies also consider different sampling schemes, including picking the top samples for testing or randomly sampling based on the estimated risk.

intervention g_t^*. The main goal is to optimize the next time-point outcome under g_t^* at each t, or as an average of short term outcomes over time, under a possible resource constraint. As such, the history-adjusted optimal choice for a single time point intervention defines the adaptive design over time. In the setting of an infectious disease outbreak, we define exposures of interest as user-defined stochastic interventions, where each g_t^* denotes a specific testing design (*symptomatic*, *random*, *contact tracing*, *risk-based* testing, etc). The proposed Online Super Learner for adaptive sequential surveillance then learns the optimal choice of test strategies over time, adapting to the current state of the epidemic. The optimal testing allocation aims to maximize the number of caught infectious individuals — resulting in infectious disease monitoring, prompt isolation, and prevention of further spread.

The infectious disease setup, however, presents unique technical challenges. For instance, our target parameter is defined in terms of a latent outcome, as the true infectious status is often unknown. In this work we present an identification result for the causal target parameter in terms of the observed outcome, defined as a function of the stochastic intervention — as we only observe the true status for individuals we test. In addition, unlike the usual i.i.d. settings, infectious disease propagation induces both network and temporal dependence. The adaptive sequential designs described in the literature are typically asymptotic in the number of subjects enrolled in the trial [19], or in the number of time points [78]. Unlike the usual settings, the presence of both network and temporal dependence reduces data to a single observation. Instead of imposing unrealistic assumptions on the statistical model $\mathcal{M}$, we rely on working models in order estimate the conditional mean function $\bar{Q}_{0,i,t}$ and an honest benchmark to choose the best performing estimate for the sake of the adaptive design performance. Therefore, the proposed method decides whether to learn across samples, through time, or both, based on the underlying (unknown) structure in the data at each time point of the disease trajectory.

As part of the Online SL for adaptive sequential surveillance, we propose a discrete Super Learner for selecting among different adaptive designs. Namely at each time t, we evaluate the performance of a choice g_t^* by proportion of infected individuals we catch. We might use as criterion for an adaptive design its average loss over a recent time window (loss-based selector), the TMLE estimate under g_t^* and working model (TMLE-based), or the maximum lower confidence interval under the stochastic intervention (TMLE-CI-based). All of the evaluation strategies aim to provide smooth transitions from discrete SL at time t to the one at $t+1$, by either averaging over a window of recent performances or weighting by the ratio of current and proposed probability of treatment given the past. Therefore, the Online SL for adaptive sequential design is also an adaptive design itself, and the discrete Super Learner evolves and changes as a function of the underlying dynamics over time. The key strength of the proposed method is that it does not depend on a strong statistical model, or imposes unrealistic assumptions. Instead, it relies on different working models to estimate $\bar{Q}_{0,i,t}$, and selects among adaptive designs with a short term performance Online Super Learner. As such, the proposed adaptive design avoids stationarity and independence assumptions, which are unrealistic in an infectious disease setup.

In addition to proposing a new adaptive sequential design suitable for studying infectious

disease, we also develop an agent-based model for a moderate size campus during an epidemic. While we strive to model the environment analogue to the University of California (Berkeley), all the settings and simulations can be easily modified to reflect any residential campus and infectious disease. We made all the code and simulations public for the interested reader, including the proposed adaptive design. Within the simulation framework defined by the agent-based model, the Online SL for adaptive sequential surveillance outperforms all considered gold standard testing schemes (*random*, *contact tracing + symptomatic*) including the *risk-based* testing alone. The advantages of the proposed adaptive design are evident over a variety of scenarios, including varying resource constraints and level of problem difficulty (determined by the percent latent component of the network and individual risk). The reported simulation results reflect the best case scenarios for the competitor testing strategies - perfect accuracy of the allocated tests, accurate observed network, and full symptoms for symptomatic individuals. As response to COVID-19 pandemic evolved, most universities started to require mandatory vaccination, mask-wearing, social distancing, enhanced cleaning protocols, increased availability of sanitizing products and no large social gatherings. In future work, we plan to investigate performance of the proposed Online SL for adaptive surveillance in addition to a wide array of other transmission-limiting measures.

Our proposed method can be extended in various ways. Instead of picking a single selector as done in our simulations, we could instead have an Online SL for adaptive surveillance where each candidate is one of the described methods. As such, it would be possible to pick among TMLE-, TMLE-CI- and loss-based selectors at any time point t; this would be particularly advantageous at the beginning of the trajectory, when loss-based methods seems to perform best, but TMLE-CI-based selector minimizes the final cumulative incidence. In addition, we could extend the proposed methodology to consider a convex combination of different designs, instead of focusing on the discrete Online SL. While this extension could provide better performance (in terms of optimizing our target parameter), it might be more difficult to interpret and implement in practice. In addition, while the Online SL for adaptive surveillance outperforms all considered testing schemes, it does not provides inference for our main target parameter. If we were willing to make additional assumptions, or known more about the data generating process — for example, assume a known network structure over time or conduct detailed surveillance as done in some countries — we could analyze the TMLE of our target parameter under one of the working models discussed in Section 4.2. Alternatively, one could data-adaptively learn the underlying true model by giving up certain statistical properties of the estimator, such as regularity. We explore all of these interesting avenues and extensions in future work.

4.7 Appendix

Identifiability Results

Theorem 1 *Assume assumptions 33 and 34 hold. Under consistency, we denote the time t value under the stochastic intervention g_t^* as*

$$\Psi^F_{t,g_t^*}(P^F_{\bar{O}(t-1)}) = \Psi_{t,g_t^*}(P_{\bar{O}(t-1)}) = \int_a \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_P[Y_i^l(t) \mid A_i(t) = a, \bar{o}(t-1)] g_t^*(a \mid \bar{o}(t-1)) d\mu_a(a)$$
$$= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{\bar{Q}_{i,t},g^*_{i,t}}[Y_i(t) \mid \bar{o}(t-1)]$$

where the observed outcome is defined as $Y_i(t) = A_i(t)Y_i^l(t)/g_{i,t}(A_i(t) \mid \bar{O}(t-1))$ and $\psi_{t,g_t^*} = \Psi_{t,g_t^*}(P_{\bar{O}(t-1)})$.

Proof. First, we identify the causal parameter in terms of the conditional distribution of the observed data $P_{\bar{O}(t-1)}$ and latent outcome $Y^l(t)$. Under Assumptions 33 (A6) and 34 (A7), jointly with consistency (A8), we can denote value at t under the stochastic intervention g_t^* as

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} Y^l_{i,g^*_{i,t}}(t)\right] = \mathbb{E}[\bar{Y}^l_{g_t^*}(t)] \tag{4.14}$$
$$\stackrel{def}{=} \int_a \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[Y^l_{i,g^*_{i,t}}(t) = y^l \mid A_{i,g^*_{i,t}}(t) = a, \bar{o}(t-1)] g^*_{i,t}(a \mid \bar{o}(t-1)) d\mu(a)$$
$$\stackrel{def}{=} \int_a \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[Y^l_{i,a}(t) = y^l \mid A_{i,g^*_{i,t}}(t) = a, \bar{o}(t-1)] g^*_{i,t}(a \mid \bar{o}(t-1)) d\mu(a)$$
$$\stackrel{A6}{=} \int_a \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[Y^l_{i,a}(t) = y^l \mid \bar{o}(t-1)] g^*_{i,t}(a \mid \bar{o}(t-1)) d\mu(a)$$
$$\stackrel{A8}{=} \int_a \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[Y^l_{i}(t) = y^l \mid A_i(t) = a, \bar{o}(t-1)] g^*_{i,t}(a \mid \bar{o}(t-1)) d\mu(a)$$
$$\stackrel{def}{=} \int_a \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[Y^l_{i}(t) = y^l \mid A_i(t) = a, \bar{o}(t-1)] g^*_{i,t}(a \mid \bar{o}(t-1)) d\mu(a).$$

Note that, for conditional expectations to be well defined, Assumption 34 must hold. The last equality in equation (4.14) gives us the identification result in terms of the conditional expectation of the latent outcome. We proceed to define the observed outcome as

$$Y_i(t) = A_i(t)Y_i^l(t)/g_{i,t}(A_i(t) \mid \bar{O}(t-1)), \tag{4.15}$$

with the conditional expectation of the observed outcome as follows

$$\begin{aligned}
\mathbb{E}[\bar{Y}(t) \mid \bar{o}(t-1)] &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\frac{A_i(t)}{g_{i,t}(A_i(t) \mid \bar{o}(t-1))}Y_i^l(t) \mid \bar{o}(t-1)\right] \\
&= \frac{1}{n}\sum_{i=1}^{n}\int_y \frac{y}{g_{i,t}(A_i(t) \mid \bar{o}(t-1))}P(y \mid 1, \bar{o}(t-1))g_{i,t}(A_i(t) \mid \bar{o}(t-1))d\mu(y) \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{g_{i,t}}[Y_i^l(t) \mid 1, \bar{O}(t-1) = \bar{o}(t-1)].
\end{aligned}$$

Therefore, the conditional expectation of the observed outcome defined as in equation (4.15) is equal to the conditional expectation of the latent outcome $Y_i^l(t)$. We can write the final identification results as

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}Y^l_{i,g^*_{i,t}}(t)\right] &= \int_a \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[Y_i^l(t) = y^l \mid A_i(t) = a, \bar{o}(t-1)]g_t^*(a \mid \bar{o}(t-1))d\mu(a) \\
&= \int_a \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[\frac{A_i(t)y^l}{g_{i,t}(A_i(t) \mid \bar{o}(t-1))} \mid a, \bar{o}(t-1)]g^*_{i,t}(a \mid \bar{o}(t-1))d\mu(a) \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{\bar{Q}_{i,t},g^*_{i,t}}[Y_i(t) \mid \bar{o}(t-1)].
\end{aligned}$$

□

Comment on the Oracle Adaptive Design

The oracle for the proposed adaptive design aims to maximize the overall number of detected cases at each time step under a resource constraint, such that

$$\arg\max_{g_t^* \in \mathcal{G}}(\Psi_{t,g_t^*}(P_{\bar{O}(t-1)})) \text{ subject to } E[g_t^*(1 \mid Pa(A(t)))] \leq k \text{ for all } t. \tag{4.16}$$

The proposed Online SL for adaptive surveillance aims to approximate and learn the oracle design in Equation (4.16). Since we are optimizing a short term outcome, this equates to optimizing the mean over time parameter under a resource constraint at each t. The optimal strategy g_t^{opt} then corresponds to the intervention g_t^* that maximizes Equation (4.16) at t under the k percent constraint. Under the t-optimal testing allocation, contagious individuals can be quickly isolated from the general population with the ultimate goal of minimizing transmission at future time-points by prompt detection of active, circulating infections. Consequently, maximizing the number of caught infected individuals at each time step results in minimizing the total number of circulating infectious by time τ. As such, the short term outcome serves as a *surrogate outcome* for the ultimate goal of minimizing the number of active infections at the end of the observed trajectory.

Online Super Learner under Working Models

In the following, we outline the Online Super Learner algorithm under flexible working models described in Subsection 4.2. This work builds on ideas presented by [7] and [79] under various working models describing the possible underlying dependence structure. The proposed Online Super Learner is used to estimate $\bar{Q}_{0,i,t}$, the true conditional expectation of $Y_i(t)$ given the observed past.

Loss-based Parameter Definition and Estimation

Let $\mathcal{M}_Y$ denote the working model for $\bar{Q}_{0,i,t}$, consisting of all functions $\bar{Q}_{i,t}$ that map $\mathcal{A} \times \mathcal{C}_{\mathcal{A}}$ to $[0,1]$. The working model $\mathcal{M}_Y$ reflects a large class of candidate working models described in subsection 4.2. With that, we define $\mathcal{M}_Y$ as a collection of all conditional expectations with some possible dependence structure across time and/or network that could have given rise to the observed data.More specifically, for all $\bar{Q}_{i,t} \in \mathcal{M}_Y$, the conditional expectation of the outcome depends on the past only through a fixed dimensional summary measure $C_{L_i}(t)$, with $\mathcal{M}_Y$ satisfying Assumption 29. Under the decomposition of the fixed dimensional summary outlined in Assumption 28, $C_{L_i}(t) = (A_i(t), C_{A_i}(t))$ where $C_{A_i}(t) = h_{A_i}(Pa(A(t))) = h_{A_i}(Pa(O(t)))$. In addition, any $\bar{Q}_{i,t} \in \mathcal{M}_Y$ could be common in both samples and times (i,t), in time (t), or across samples (i) — thereby satisfying some combination of Assumption 30, 31 or 32, respectively. By specifying $\bar{Q}_{i,t} \in \mathcal{M}_Y$, we are implying there is some structure to the dependent process, as described by one of the working models in Section 4.2; we are, however, not specifying the exact structure.

Let $\bar{Q}_{i,t} \in \mathcal{M}_Y$. Under the working model $\mathcal{M}_Y$, we write the risk-based target parameter $\psi_{0,t}^{\text{risk}}$ corresponding to the *risk-based* testing strategy as

$$\psi_{0,t}^{\text{risk}}(A_i(t), C_{A_i}(t)) = \Psi_t^{\text{risk}}(P_0)(A_i(t), C_{A_i}(t)) \equiv \bar{Q}_{0,i,t}(A_i(t), C_{A_i}(t)),$$

denoting the true conditional expectation of $Y_i(t)$ given the fixed dimensional summary of the observed past. Note that, in addition to defining the *risk-based* testing strategy, estimating $\bar{Q}_{0,i,t}$ is an integral part of the Online SL for adaptive sequential surveillance (both for the loss- and TMLE- based selectors). In the following, we will refer mostly to the risk-based target parameter, but the analysis follows for all applications of $\bar{Q}_{0,i,t}$. Let L denote a valid loss function for $\Psi_t^{\text{risk}}(P_0)$, and $C(i,m)$ the time m- and unit i-specific record $C(i,m) = (Y_i(m), A_i(m), \bar{O}(m-1))$. A valid loss L is defined as a function whose true conditional expectation is minimized by the true value of the target parameter; here, the minimizer is therefore $\bar{Q}_{0,i,t}$. Further, let L be a function that maps every $\Psi_t^{\text{risk}}(P)$ to $L(\Psi_t^{\text{risk}}(P)) : C(i,t) \mapsto L(\Psi_t^{\text{risk}}(P))(C(i,t))$. As our parameter of interest is a conditional mean, we could use the square error to define the loss, resulting in

$$L(\Psi_t^{\text{risk}}(P))(C(i,t)) = w(i,t)(Y_i(t) - \Psi_t^{\text{risk}}(P)(A_i(t), C_{A_i}(t)))^2,$$

for sample i and time t, where $w(i,t)$ is the subject and time specific weight. Our accent on appropriate loss functions strives from their multiple use within our framework — as a

theoretical criterion for comparing the estimator and the truth, and as a way to compare multiple estimators of the target parameter [132, 34, 133, 141]. The loss function L and the working model $\mathcal{M}_Y$ will be used to estimate $\bar{Q}_{0,i,t}$.

We define the true risk, $R(P_0, \Psi_t^{\text{risk}}(P)) = R(P_0, \psi_t^{\text{risk}})$, as the expected value of the loss w.r.t the true probability distribution over all samples and times. As $\psi_{0,t}^{\text{risk}} = \Psi_t^{\text{risk}}(P_0)$, we note that $\psi_{0,t}^{\text{risk}}$ is the minimizer of the true risk over all evaluated ψ_t^{risk} in the parameter space, such that $\psi_{0,t}^{\text{risk}} = \arg\min_{\psi_t^{\text{risk}}} R(P_0, \psi_t^{\text{risk}})$. Therefore, the true risk establishes the true measure of performance for $\Psi_t^{\text{risk}}(P)$, with $\Psi_t^{\text{risk}}(P_0)$ denoting the minimum. Further, we define the estimator mapping, $\hat{\Psi}_t^{\text{risk}}$, as a function from the empirical distribution to the parameter space

We resort to appropriate CV for dependent data in order to obtain an unbiased estimate of the true risk. To derive a general representation for cross-validation, we define a time t specific split vector B_t, where t indicates the final time-point of the currently available data where for all $1 \leq i \leq n$, $B_t(i, \cdot) \in \{-1, 0, 1\}^t$. Let v be a particular v-fold, where v range from 1 to V. A realization of B_t defines a particular split of the learning set into corresponding three disjoint subsets,

$$B_t^v(i, s, \cdot) = \begin{cases} -1, & C(i,m) \text{ not used} \\ 0, & C(i,m) \text{ in the training set} \\ 1, & C(i,m) \text{ in the validation set,} \end{cases}$$

where $B_t^v(i, m)$ reflects the v-fold assignment of, at minimum, unit i at time point m for split B_t^v trained on data until time t. Then, for each t, we define $P_{n,t}^0$ as the empirical distribution of the training sample until time t. Similarly, we let $P_{n,t}^1$ denote the empirical distribution of the validation set. Sets $\mathcal{B}_{t,v}^0$ and $\mathcal{B}_{t,v}^1$ contain all (i, m) indexes in the training and validation sets for fold v, respectively.

Suppose we have K candidate estimators $\hat{\Psi}_{k,t}^{\text{risk}}$, $k = 1, \ldots, K$, that can be applied to $(A_i(t), C_{A_i}(t))$ for $i \in [n]$ and $t \in [\tau]$. For a given problem, a library of prediction algorithms can be proposed. In particular, the candidate estimators for the outcome regression should include different learners corresponding to the underlying working model. The algorithms in the candidate library may range from single time-series learners to networks of time-series, as well as learners that pool data across time, network, or all the data available up to time t. The Online Super Learner library can also include algorithms that put decaying weight of different rates on components of the past, or consider learners indexed by subsets of a network. We utilize working models in the estimation procedure without explicit reliance on any of the described working models in particular; we let the cross-validation procedure determine the underlying structure of the process at each time step. In order to evaluate performance of each $\hat{\Psi}_{k,t}^{\text{risk}}$, we use cross-validation for dependent data to estimate the average loss for each candidate. In particular, each $\hat{\Psi}_{k,t}^{\text{risk}}$ is trained on the training set $P_{n,t}^0$ and results in a predictive function $\hat{\Psi}_{k,t}^{\text{risk}}(P_{n,t}^0)$ for $k = 1, \ldots, K$. We define the online cross-validated risk for each candidate estimator as:

$$R_{t,CV}(P^1_{n,t}, \hat{\Psi}^{\text{risk}}_{k,t}(\cdot)) = \sum_{j=1}^{t} \sum_{v=1}^{V} \sum_{(i,s)\in\mathcal{B}^1_{j,v}} L(\hat{\Psi}^{\text{risk}}_{k,t}(P^0_{n,j}))(C(i,s)),$$

where $R_{t,CV}(P^1_{n,t}, \hat{\Psi}^{\text{risk}}_{k,t}(\cdot))$ is the cumulative performance of $\hat{\Psi}^{\text{risk}}_{k,t}$ trained on the training sets and evaluated on the corresponding validation samples until time t. For instance, while $\hat{\Psi}^{\text{risk}}_{k,t}(P^0_{n,t})$ is trained on the training set $P^0_{n,t}$, its performance will be evaluated over the validation set $P^1_{n,t}$. The online cross-validated risk estimates the following true online cross-validated risk, denoted as $R_{t,CV}(P_0, \hat{\Psi}^{\text{risk}}_{k,t}(\cdot))$ and expressed as

$$R_{t,CV}(P_0, \hat{\Psi}^{\text{risk}}_{k,t}(\cdot)) = \sum_{j=1}^{t} \sum_{v=1}^{V} \sum_{(i,s)\in\mathcal{B}^1_{j,v}} E_0[L(\hat{\Psi}^{\text{risk}}_{k,t}(P^0_{n,j}))(C(i,s))|C_{A_i}(s)].$$

Note that $R_{t,CV}(P_0, \hat{\Psi}^{\text{risk}}_{k,t}(\cdot))$ reflects the true average loss for the candidate estimator with respect to the true conditional distribution. As opposed to the true online cross-validated risk, $R_{t,CV}(P^1_{n,t}, \hat{\Psi}^{\text{risk}}_{k,t}(\cdot))$ gives an empirical measure of performance for each candidate estimator k trained on training data until time t. In light of that, we define the discrete online cross-validation selector as:

$$k_{n,t} = \arg\min_{k=1,\ldots,K} R_{t,CV}(P^1_{n,t}, \hat{\Psi}^{\text{risk}}_{k,t}(\cdot)), \tag{4.17}$$

which is the estimator that minimizes the online cross-validated risk. The discrete (online) Super Learner is the estimator that at each time point uses the estimates from the discrete online cross-validation selector. Since each of the k learners can reflect different working models, the discrete (online) SL picks one of the candidate dependence structures for time point t. We emphasize that the discrete SL can switch from one learner to another as t progresses, in response to accumulating more data and detecting changes in the network and trajectory. Note that, if all the candidate estimators are online estimators, the discrete (online) SL is itself an online estimator.

In order to study performance of an estimator, we construct loss-based dissimilarity measures. In particular, loss-based dissimilarity compares the performance of a particular estimator to the true parameter, defined as

$$d_{0,t}(\hat{\Psi}^{\text{risk}}_{k,t}, \psi^{\text{risk}}_{0,t}) = \sum_{j=1}^{t} \sum_{v=1}^{V} \sum_{(s)\in\mathcal{B}^1_{j,v}} E_0\left[\left(L(\hat{\Psi}^{\text{risk}}_{k,t}(P^0_{n,t})) - L(\psi^{\text{risk}}_{0,t})\right)(C(i,s))\middle| C_{A_i}(s)\right].$$

We define the time t *oracle* selector as the unknown estimator that uses the candidate closest to the truth in terms of the defined dissimilarity measure:

$$\overline{k}_{n,t} = \arg\min_{k=1,\ldots,K} d_{0,t}(\hat{\Psi}^{\text{risk}}_{k,t}(P^0_{n,t}), \psi^{\text{risk}}_{0,t}). \tag{4.18}$$

Due to it being a function of the true conditional mean, the oracle selector cannot be computed in practice. However, we can utilize it as benchmark in order to describe performance of the online cross-validation based estimator. Following the argument given by [7], it follows that the performance of the discrete Online Super Learner is asymptotically equivalent to that of the oracle selector. The result relies on the martingale finite-sample inequality by [54] to show that, as $t \to \infty$,

$$\frac{d_{0,t}(\hat{\Psi}^{\text{risk}}_{k_{n,t},t}(P^0_{n,t}), \psi^{\text{risk}}_{0,t})}{d_{0,t}(\hat{\Psi}^{\text{risk}}_{\overline{k}_{n,t},t}(P^0_{n,t}), \psi^{\text{risk}}_{0,t})} \to_p 1. \tag{4.19}$$

Ensemble of Candidate Estimators

In this section, we consider a more flexible online learner by considering an ensemble of a given set of estimators. As individual learners reflect different candidate working models for time and network dependence, a weighted combination of candidate estimators reflects a mixture of working models. We define $\hat{\Psi}^{\text{risk}}_{\beta,t}$ as an ensemble of K estimators indexed by a finite-dimensional vector of coefficients β, where $\beta = (\beta_1, \ldots, \beta_K)$. For example, $\hat{\Psi}^{\text{risk}}_{\beta,t}$ could represent a convex linear combination:

$$\hat{\Psi}^{\text{risk}}_{\beta,t} = \sum_{k=1}^{K} \beta_k \hat{\Psi}^{\text{risk}}_{k,t},$$

such that $\sum_{k=1}^{K} \beta_k = 1$ with $\beta_k \geq 0$ for all β_k. Let $R_{t,CV}(P^1_{n,t}, \hat{\Psi}^{\text{risk}}_{\beta,t}(\cdot))$ be the online cross-validated risk for $\hat{\Psi}^{\text{risk}}_{\beta,t}$ given by

$$R_{t,CV}(P^1_{n,t}, \hat{\Psi}^{\text{risk}}_{\beta,t}(\cdot)) = \sum_{j=1}^{t} \sum_{v=1}^{V} \sum_{(i,s) \in \mathcal{B}^1_{j,v}} L(\hat{\Psi}^{\text{risk}}_{\beta,t}(P^0_{n,j}))(C(i,s)).$$

We denote $\beta_{n,t}$ as the choice of β that minimizes the online cross-validated risk,

$$\beta_{n,t} = \arg\min_{\beta} R_{t,CV}(P^1_{n,t}, \hat{\Psi}^{\text{risk}}_{\beta,t}(\cdot)). \tag{4.20}$$

Note that $\beta_{n,t}$ itself is not an online estimator, since it involves recomputing the minimum for each t. We can define an oracle selector for this class of estimators as the choice of weights that minimizes the true average of the loss-based dissimilarity:

$$\overline{\beta}_{n,t} = \arg\min_{\beta} d_{0,t}(\hat{\Psi}^{\text{risk}}_{\beta,t}(P^0_{n,t}), \psi^{\text{risk}}_{0,t}). \tag{4.21}$$

The oracle results extend to an ensemble of candidate estimators, and we can show that

$$\frac{d_{0,t}(\hat{\Psi}^{\text{risk}}_{\beta_{n,t},t}(P^0_{n,t}), \psi^{\text{risk}}_{0,t})}{d_{0,t}(\hat{\Psi}^{\text{risk}}_{\overline{\beta}_{n,t},t}(P^0_{n,t}), \psi^{\text{risk}}_{0,t})} \to_p 1 \tag{4.22}$$

as t goes to infinity. As such, the performance of the Online Super Learner is asymptotically equivalent with the optimal ensemble of candidate estimators, which reflect different working models.

Additional Simulations

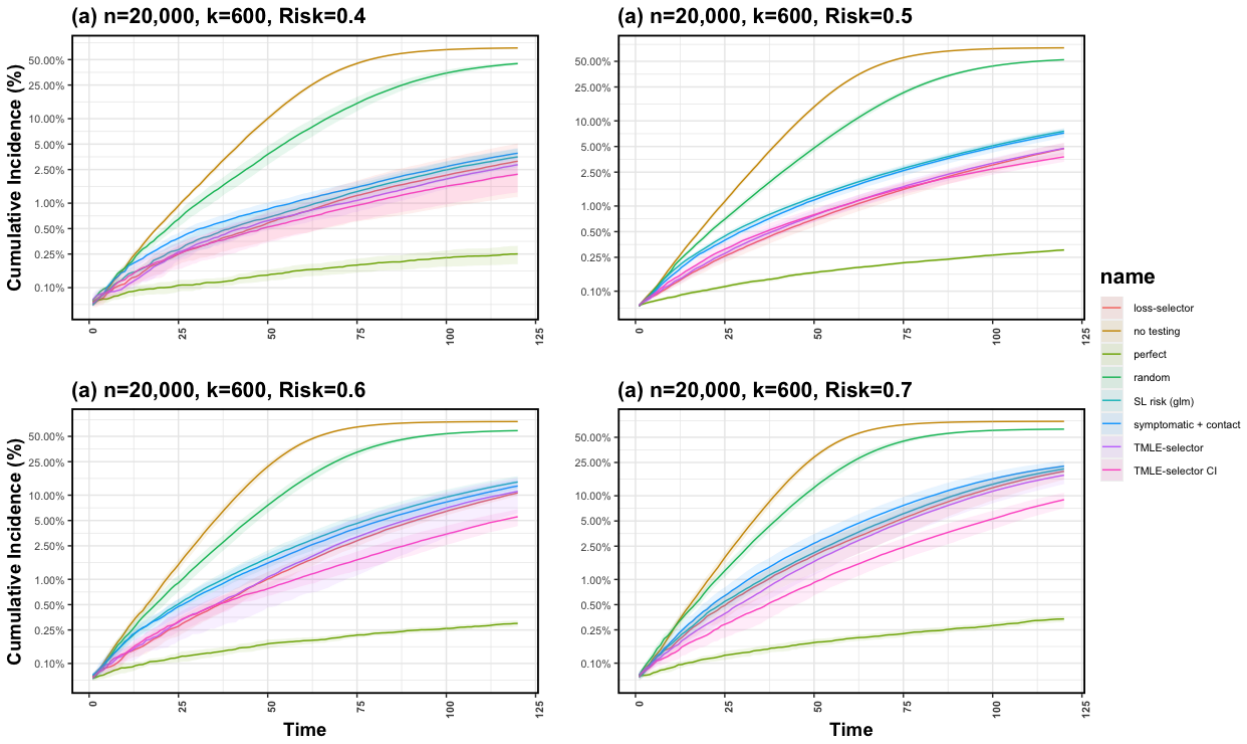


Figure 4.4: Average cumulative incidence at each time point over 500 simulations with $n = 20,000$ sample size and capacity $k = 600$ using TMLE-based, TMLE-CI-based and loss-based selectors of the testing strategy. Performance is evaluated at different risk scale values $\{0.4, 0.5, 0.6, 0.7\}$, where higher risk scale score corresponds to higher individual risk. We compare different proposed selectors to *symptomatic + contact*, *random* and *glm risk-based* testing, with *perfect* as the upper and *no testing* as lower bound on performance.

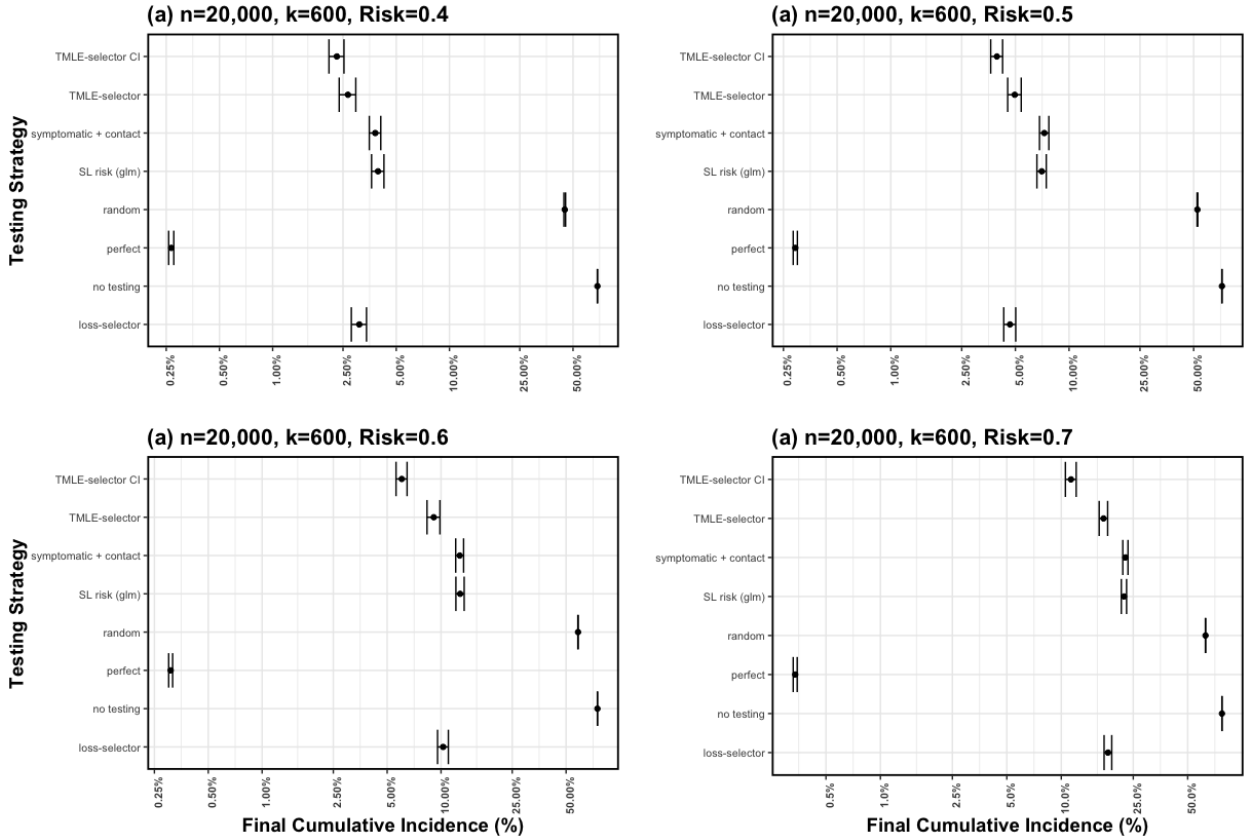


Figure 4.5: Average final cumulative incidence at $t = 120$ over 500 simulations, with $n = 20,000$ sample size and capacity $k = 600$ using TMLE-based, TMLE-CI-based and loss-based selectors of the testing strategy. Performance is evaluated at different risk scale values $\{0.4, 0.5, 0.6, 0.7\}$, where higher risk scale score corresponds to higher individual risk. We compare different proposed selectors to *symptomatic + contact*, *random* and *glm risk-based* testing, with *perfect* as the upper and *no testing* as lower bound on performance.

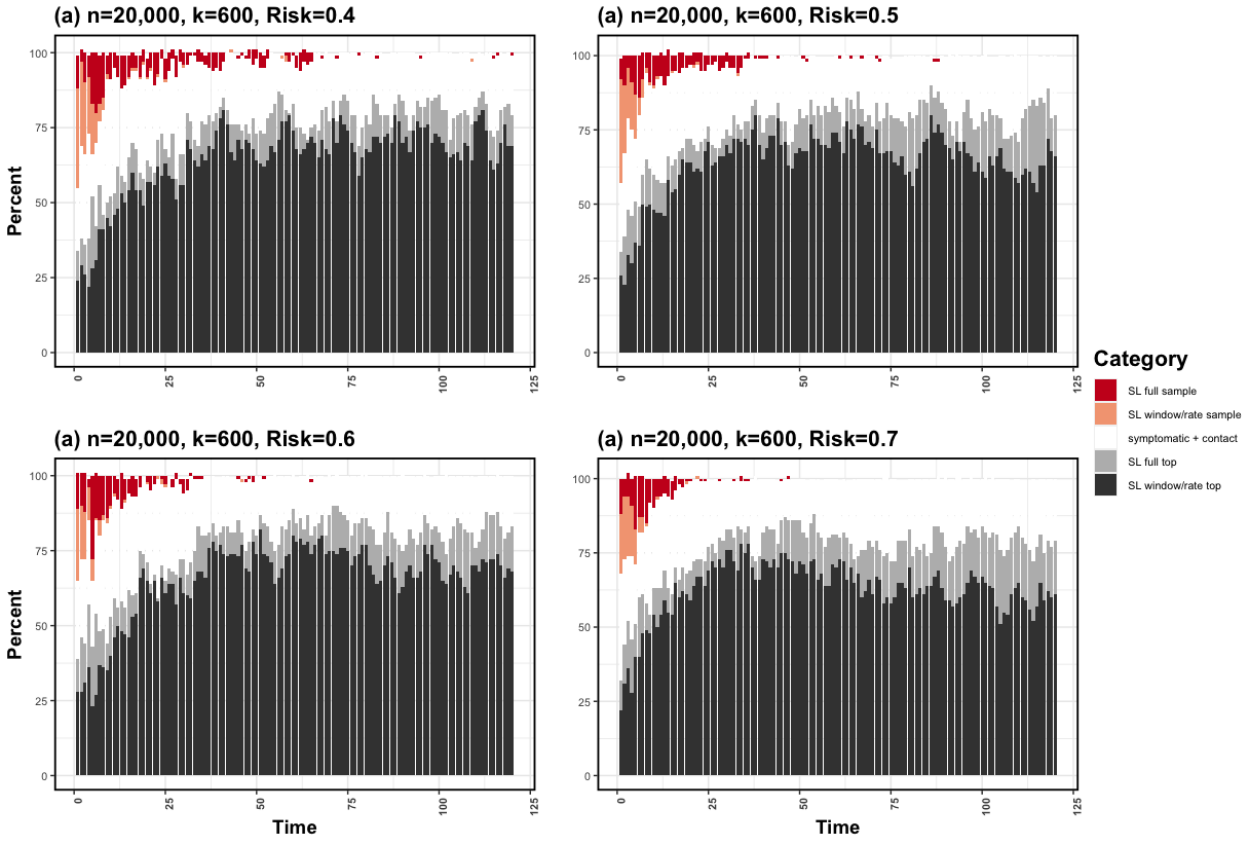


Figure 4.6: Percent design over the full trajectory and 500 simulations with $n = 20,000$ sample size and capacity $k = 600$ using TMLE-CI-based selector of the testing strategy. Performance is evaluated at different risk scale values $\{0.4, 0.5, 0.6, 0.7\}$, where higher risk scale score corresponds to higher individual risk. Candidate designs include *symptomatic + contact* and various *risk-based* designs where the Super Learner is either trained on the full past, exponentially weighted past, or a window of $t = 14$ days. All *risk-based* testing strategies also consider different sampling schemes, including picking the top samples for testing or randomly sampling based on the estimated risk.

Chapter 5

Regularized Targeted Learning in Reinforcement Learning

In this chapter, we study the problem of off-policy evaluation (OPE) in Reinforcement Learning (RL), where the aim is to estimate the performance of a new policy given historical data that may have been generated by a different policy (or policies). In particular, we introduce a novel doubly-robust estimator for the OPE problem in RL, based on the Targeted Maximum Likelihood Estimation principle from the statistical causal inference literature. We also introduce several variance reduction techniques that lead to impressive performance gains in off-policy evaluation. We show empirically that our estimator uniformly wins over existing off-policy evaluation methods across multiple RL environments and various levels of model misspecification. Finally, we further the existing theoretical analysis of estimators for the RL off-policy estimation problem by showing their $O_P(1/\sqrt{n})$ rate of convergence and characterizing their asymptotic distribution.

5.1 Introduction

Off-policy evaluation (OPE) is an increasingly important problem in reinforcement learning. Works on OPE address the pressing issue of evaluating the performance of a novel policy in a setting where actual enforcement might be too costly, infeasible, or even hazardous. This situation arises in many fields, including medicine, finance, advertising, and education, to name a few [88, 95, 121, 59]. The OPE problem can be treated as a counterfactual quantity estimation problem, as we inquire about the mean reward we would have accrued, had we, contrary to fact, implemented the policy π_e at the time of data-collection. Estimating and inferring such counterfactual quantities is a well studied problem in statistical causal inference, and has led to many methodological developments. This chapter aims to further earlier efforts by [33] in bridging the gap between the reinforcement learning and causal inference fields.

There are roughly two predominant classes of approaches to off-policy value evaluation

in RL [63]. The first is the *direct method* (DM), analogous to the *G-computation* procedure in causal inference [106, 107]. The direct method first fits a model of the system's dynamics and then uses the learned fit in order to estimate the mean reward of the target policy (evaluation policy). The estimators produced by this approach usually exhibit low variance, but suffer from high bias when the model fit is misspecified or the sample size is small relative to the complexity of the function class of the model [81]. The second major avenue for off-policy value evaluation is *importance sampling* methods, also termed *inverse propensity score* methods in statistical causal inference [110]. Importance sampling (IS) attempts to correct the mismatch between the distributions produced by the behavior and target policies [98, 97]. The IS estimators are unbiased under mild conditions, but their variance tends to be large when the evaluation and behavior policies differ significantly [39]. Their variance also grows exponentially with the horizon, rendering IS approach [39] impractical for many RL settings. A third class of estimators, *Doubly Robust* (DR) estimators, obtained by combining a DM estimator and an IS estimator, are becoming standard in OPE [39, 63, 123]. These originate from the statistics literature [109, 108, 5, 140, 138, 137], and were introduced in the RL literature by Dudik, Langford, and Li [33]. Combining a DM and an IS estimator under the form of a DR estimator leads to lower bias than DM alone, and lower variance than IS alone.

The contribution of this chapter to OPE in RL is multifold. First, it proposes an adaptation of a doubly robust estimator from statistical causal inference, the Longitudinal Targeted Maximum Likelihood Estimator (LTMLE) to the OPE in RL setting. We show that our adapted estimator converges at rate $O_P(1/\sqrt{n})$ to the true policy value. Deriving the LTMLE requires us to identify a mathematical object known in semiparametric statistics as the *efficient influence function* (EIF) of the estimand (policy value). To the best of our knowledge, this work is the first one to explicitly derive the EIF of the policy value for the OPE problem in RL. Knowledge of the EIF allows us to prove that both our estimator (the LTMLE) and recently proposed DR estimators [63, 123] are optimal in the sense that they achieve the generalized Cramer-Rao lower bound. Second, it introduces an idea from statistics to make better use of the data than prior OPE works [63, 123]. Most OPE papers, at least in theory, use sample splitting: the Q-function is fitted on a split of the data, while the DR estimator is obtained by evaluating the fitted Q-function on another split. This chapter proposes a cross-validation-based technique that allows to average the Q-function over the entire sample, leading to a constant-factor gain in risk. Finally, and most importantly for practice, it proposes several regularization techniques for the LTMLE estimators, out of which some, but not all, apply to other DR estimators. Using the MAGIC ensemble method from Thomas and Brunskill [123], we construct an estimator that combines various regularized LTMLEs. We call our final estimator RLTMLE (TMLE for RL). The simulations reported demonstrate that RLTMLE outperforms all considered competing off-policy methods, uniformly across multiple RL environments and levels of model misspecification.

5.2 Statistical Formulation of the Problem

Markov Decision Process and Target Parameter

Consider a Markov Decision Process (MDP) defined as a tuple $(\mathcal{S}, \mathcal{A}, \mathcal{R}, P_1, P, \gamma)$, where $\mathcal{S}$ and $\mathcal{A}$ are the state and action spaces, and $\gamma \in (0, 1]$ is a discount factor. A trajectory H is a succession of states S_t, actions A_t and rewards R_t, observed from $t = 1$ to the horizon $t = T$: $H = (S_1, A_1, R_1, ..., S_T, A_T, R_T)$. For all $(s, a, r, s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{R} \times \mathcal{S}$, $P(s', r|s, a)$ is the probability of collecting reward r and transitioning to state s', conditional on starting in state s and taking action a, and $P_1(s)$ is the probability that the initial is s. A policy π is a sequence of conditional distributions $(\pi_1, \pi_2, ...)$ that stochastically map a state to an action: for all t, $A_t|S_t \sim \pi_t$. Suppose we are given n i.i.d. T-step trajectories of the MDP, $D = (H_1, ..., H_n)$, collected under the behavior policy $\pi_b = (\pi_{b,1}, ..., \pi_{b,T})$. We assume all trajectories have the same initial state s_1, allowing for the data-generating mechanism to be fully characterized by (P, π_b).

The goal of OPE is to estimate the average cumulative discounted reward we would have obtained by carrying out the target policy π_e instead of policy π_b. That is, we want to estimate the following counterfactual quantity:

$$V_1^{\pi_e}(s_1) := E_{P,\pi_e}\left[\sum_{t=1}^{T} \gamma^t R_t | S_1 = s_1\right]. \tag{5.1}$$

Assumption 35 (Absolute continuity)**.** $\forall s, a \in \mathcal{S} \times \mathcal{A}$, *if* $\pi_b(a|s) = 0$, *then* $\pi_e(a|s) = 0$.

Under assumption 35 and the Markov assumption of the MDP model, $V_1^{\pi_e}(s_1)$ can be written as an expectation under the data-generating mechanism (P, π_b):

$$V_1^{\pi_e}(s_1) = E_{P,\pi_b}\left[\prod_{t=1}^{T} \frac{\pi_{e,t}(A_t|S_t)}{\pi_{b,t}(A_t|S_t)} \sum_{t=1}^{T} \gamma^t R_t \Big| S_1 = s_1\right]. \tag{5.2}$$

For $t = 1, ..., T$, define $\bar{R}_{t:T} := \sum_{\tau=t}^{T} \gamma^{\tau-t} R_\tau$ as the total reward from step t to step T. For all $1 \le t_1 \le t_2 \le T$, define $\rho_{t_1:t_2} := \prod_{\tau=t_1}^{t_2} \pi_{e,\tau}(A_\tau|S_\tau)/\pi_{b,\tau}(A_\tau|S_\tau)$. For all $t = 1, ..., T$, we will use the shortcut notation $\rho_t := \rho_{1:t}$. We use the convention that $\rho_0 = 0$. Denote $\bar{R}_{t:T}^{(i)}$, $\rho_t^{(i)}$, $\rho_{t_1:t_2}^{(i)}$ the corresponding quantities for a sample trajectory H_i. Consistently with (5.1) and (5.2), we define, for any $t = 1, ..., T$, and $s \in \mathcal{S}$, the value function (or reward-to-go) from time point t and state s, as

$$\begin{aligned} V_t^{\pi_e}(s) :&= E_{P,\pi_e}[\bar{R}_{t:T}|S_t = s] \\ &= E_{P,\pi_b}\left[\rho_{t:T}\bar{R}_{t:T}|S_t = s\right]. \end{aligned} \tag{5.3}$$

For every $t = 1, ..., T$, $s \in \mathcal{S}$, $a \in \mathcal{A}$, we further define the action-value function from time step t as

$$\begin{aligned} Q_t^{\pi_e}(s, a) &:= E_{P,\pi_e}\left[\bar{R}_{t:T}|S_t = s, A_t = a\right] \\ &= E_{P,\pi_b}\left[\rho_{t:T}\bar{R}_{t:T}|S_t = s, A_t = a\right]. \end{aligned} \tag{5.4}$$

5.3 Current state-of-the art approach

Our method can be seen as building upon and improving on Thomas and Brunskill [123]. We believe it helps understanding our contribution to first briefly describe their estimators. For a detailed review of OPE methods, we refer the interested reader to the vast and excellent literature on the topic [98, 122, 63, 39].

Weighted Doubly Robust Estimator

Jiang et al [63] were the first authors to propose a doubly robust estimator for off-policy evaluation in the MDP setting. Building on their work, Thomas et al [123] build a stabilized version termed Weighted Doubly Robust (WDR) estimator. The stabilized importance sampling weight for observation i at time step t is then defined as $w_t^{(i)} = \rho_t^{(i)} / \sum_{i=1}^n \rho_t^{(i)}$, with the final WDR estimator written as

$$WDR := \sum_{i=1}^{n} \left\{ \frac{1}{n} V_1^{\pi_e}(S_1^{(i)}) + \sum_{t=1}^{T} \gamma^t w_t^{(i)} \left[R_t^{(i)} - Q_t^{\pi_e}(S_t^{(i)}, A_t^{(i)}) + \gamma V_{t+1}^{\pi_e}(S_{t+1}^{(i)}) \right] \right\}. \tag{5.5}$$

MAGIC

While WDR has low bias and converges at a rate $O_P(1/\sqrt{n})$ to the truth, its reliance on importance weights can make it highly variable. As a result, in some settings, especially if model misspecification is not too prevalent, DM estimators can beat WDR [123]. This motivates the construction of an estimator that interpolates between DM and WDR, so as to benefit from both. Thomas et al. [123] propose *partial importance sampling* estimators, which correspond to cutting off the sum in (5.5) for terms with index $t \geq j$ corresponding to some $0 \leq j \leq T$. Formally, they define their partial importance sampling estimator as the average $g_j := \sum_{i=1}^n g_j^{(i)}$ of the so-called *off-policy j-step return*, that they define, for each trajectory i, as

$$g_i^{(j)} := \sum_{t=1}^{j} \underbrace{\gamma^t w_t^i R_t^{(i)}}_{\text{a}} + \underbrace{\gamma^{j+1} w_j^i V_{j+1}^{\pi_e}(S_{j+1}^i)}_{\text{b}} - \sum_{t=1}^{j} \underbrace{\gamma^t [w_t^i Q_t^{\pi_e}(S_t^{(i)}, A_t^{(i)}) - w_{t-1}^i V_t^{\pi_e}(S_t^{(i)})]}_{\text{c}}. \tag{5.6}$$

Here we emphasize that g_0 is equal to the DM estimator, and that the last component, (c), represents the combined control variate for the importance sampling (a) and model based

term (b). Hence, as j increases, we expect bias to decrease, at the expense of an increase in variance.

The final estimator by Thomas et al. [123] is a convex combination of the partial importance sampling estimators g_j. Ideally, we would like this convex combination to minimize mean squared error (MSE): that is, we would like to use as estimator $(\mathbf{x}^*)^\top \mathbf{g}$, with $\mathbf{g} = (g_0, ..., g_T)$, where

$$\begin{aligned} \mathbf{x}^* &= \arg \min_{\substack{0 \le \mathbf{x} \le 1 \\ \sum_{j=0}^T x_j = 1}} \mathrm{MSE}(\mathbf{x}^\top \mathbf{g}, V_1^{\pi_e}) \\ &= \arg \min_{\substack{0 \le \mathbf{x} \le 1 \\ \sum_{j=0}^T x_j = 1}} \left\{ \mathrm{Bias}^2(\mathbf{x}^\top \mathbf{g}, V_1^{\pi_e}) + \mathrm{Var}(\mathbf{x}^\top \mathbf{g}) \right\}. \end{aligned} \tag{5.7}$$

As we do not have access to the true variance and bias, Thomas et al. [123] propose to use as estimator $\hat{\mathbf{x}}^\top \mathbf{g}$, where $\hat{\mathbf{x}}$ is a minimizer, over the convex weights simplex, of an estimate of the MSE. The covariance matrix of $\mathbf{g}$, which we will denote

High level description

Our proposed estimator extends the longitudinal Targeted Maximum Likelihood Estimation (TMLE) methodology, initially developed in the statistics causal inference literature, to the MDP setting [140, 129, 138, 137]. In order to build intuition on our estimator, we start with a high-level description. Targeted Maximum Likelihood Estimation is a general framework that allows to construct efficient nonparametric estimators of low-dimensional characteristics of the data-generating distribution, given machine learning based estimators of high-dimensional characteristics. Let us illustrate on an example what these low-dimensional and high-dimensional characteristics can be. Suppose we want to estimate an average treatment effect (ATE), where we have pre-treatment covariates X, a treatment T and an outcome Y, with $(X, T, Y) \sim P$. In this situation, the low-dimensional characteristic is the ATE $E_P[E_P[Y|T = 1, X] - E_P[Y|T = 0, X]]$, while the high-dimensional characteristics of P are the outcome regression function $x, a \mapsto E_P[Y|A = a, X = x]$ and the propensity score function $x \mapsto E_P[T|X = x]$.

Suppose we are provided with n i.i.d. trajectories, $D = (H_1, ..., H_n)$. First, we generate two splits of the sample: for some $0 < p < 1$, let $D^{(0)} = (H_1, ..., H_{(1-p)n})$ and $D^{(1)} = (H_{(1-p)n+1}, ..., H_n)$. We use $D^{(0)}$ to fit estimators $\hat{Q}_1^{\pi_e}, \cdots, \hat{Q}_T^{\pi_e}$ of the action value functions $Q_1^{\pi_e}, \cdots, Q_T^{\pi_e}$, and call $\hat{Q}_1^{\pi_e}, \cdots, \hat{Q}_T^{\pi_e}$ the *initial estimators*. Such estimators can be obtained, for instance, by fitting a model of the dynamics of the MDP (or by SARSA) among other methods [120]. Estimators fitted in such a way tend to exhibit low variance but often suffer from misspecification bias. As mentioned in section 5.3, doubly-robust estimators take initial estimates as input and evaluate on $D^{(1)}$. An average of a certain function of the evaluated input produces an unbiased estimator of $V_1^{\pi_e}(s_1)$. These doubly-robust estimators rely on the addition of terms weighted by the importance sampling (IS) ratios $\rho_{i:t}^{(i)}$, $i = 1, \cdots, n$,

$t = 1, \cdots, n$. The TMLE methodology takes another route: for each t, it defines, on top of the initial estimator fit, a parametric model — which we will call a *second-stage parametric model*. The second-stage parametric model $\hat{Q}_t^{\pi_e}$ achieves bias reduction by fitting a maximum likelihood over it, on the sample split $D^{(1)}$.

Formal presentation of the simplified algorithm

To formally describe our algorithm, it suffices to define the second-stage parametric models and describe the loss used for the fit. For all $x \in \mathbb{R}$, we define $\sigma(x) = 1/(1+e^{-x})$ as the logistic function, and we denote σ^{-1} as its inverse. Observe that bounding the range of rewards where $\forall t, R_t \in [r_{min}, r_{max}]$, implies that $\forall t$ and $\forall (s,a) \in \mathcal{S} \times \mathcal{A}$, $Q_t(s,a) \in [-\Delta_t, \Delta_t]$ with $\Delta_t := \sum_{\tau=t}^{T} \gamma^{\tau-t} \max(r_{max}, |r_{min}|)$. We further denote $\tilde{Q}_t^{\pi_e}(s,a) := (\hat{Q}_t^{\pi_e} + \Delta_t)/(2\Delta_t)$ as the normalized initial estimator. In addition, $\forall \delta \in (0, 1/2)$ and $\forall (s,a)$, we define the following thresholded version of $\tilde{Q}_t^{\pi_e}$:

$$\tilde{Q}_t^{\pi_e,\delta}(s,a) := \begin{cases} 1-\delta & \text{if } \tilde{Q}_t^{\pi_e}(s,a) > 1-\delta, \\ \tilde{Q}_t^{\pi_e}(s,a) & \text{if } \tilde{Q}_t^{\pi_e}(s,a) \in [\delta, 1-\delta], \\ \delta & \text{if } \tilde{Q}_t^{\pi_e}(s,a) < \delta. \end{cases} \tag{5.8}$$

For all $\epsilon \in \mathbb{R}$, we can now define the normalized version of our second-stage parametric model as:

$$\tilde{Q}_t^{\pi_e,\delta}(\epsilon)(s,a) := \sigma(\sigma^{-1}(\tilde{Q}_t^{\pi_e,\delta}(s,a)) + \epsilon). \tag{5.9}$$

Finally, we denote $\hat{Q}_t^{\pi_e,\delta}(\epsilon) = 2\Delta_t(\tilde{Q}_t^{\pi_e,\delta}(\epsilon) - 1/2)$ as the rescaled version of $\tilde{Q}_t^{\pi_e,\delta}(\epsilon)$.

The normalization, thresholding and rescaling steps in the definition of the parametric second-stage model ensure that (1) $\tilde{Q}_t^{\pi_e,\delta}(\epsilon) \in [\delta, 1-\delta] \subset (0,1)$ for all ϵ, and that (2) $\hat{Q}_t^{\pi_e,\delta}(\epsilon)$ always stays in the allowed range of rewards $[-\Delta_t, \Delta_t]$. The definition of $\tilde{Q}_t^{\pi_e,\delta}(\epsilon)$ as a logistic transform of ϵ that lies in $(0,1)$ makes the fitting of ϵ possible through maximum likelihood for a logistic likelihood. For $t = T$, since $Q_T^{\pi_e}(s,a) = E_{P,\pi_b}[\rho_{1:T} R_T | S_T = s, A_T = a]$, it is natural to consider the log likelihood,

$$\begin{aligned} \mathcal{R}_{n,T}^{\delta}(\epsilon) = \frac{1}{n} \sum_{i=1}^{n} \rho_{1:T}^{(i)} \Bigg(& \tilde{U}_T^{(i)} \log(\tilde{Q}_T^{\pi_e,\delta}(\epsilon)(S_T^{(i)}, A_T^{(i)})) \\ & + (1 - \tilde{U}_T^{(i)}) \log(1 - \tilde{Q}_T^{\pi_e,\delta}(\epsilon)(S_T^{(i)}, A_T^{(i)})) \Bigg), \end{aligned} \tag{5.10}$$

where $\tilde{U}_T^{(i)} := (R_T^{(i)} + \Delta_T)/(2\Delta_T)$ is the normalized reward at time T. Normalization of the reward is necessary since we are using logistic regression to optimize ϵ, and to keep the definition of $\tilde{U}_T^{(i)}$ and $\tilde{Q}_T^{\pi_e,\delta}(s,a)$ consistent. The thresholding step that defines $\tilde{Q}_t^{\delta}(s,a)$ prevents the log likelihood from taking on non-finite values. In order to make the bias introduced by thresholding vanish as the sample size grows, we use a vanishing sequence $\delta_n \downarrow 0$ of thresholding values.

Let $\epsilon_{n,T}$ be the minimizer over $\mathbb{R}$ of the log likelihood $\mathcal{R}^{\delta}_{n,t}$ for step T. We fit the second-stage models for $t = T-1, ..., 1$ by backward recursion, a procedure which we describe in more detail in this paragraph. Start with observing that for all $t = 1, ..., T$, and for all $(s, a) \in \mathcal{S} \times \mathcal{A}$, $Q^{\pi_e}_t(s, a) = E_{\pi_b}[\rho_{1:t}(R_t + \gamma V^{\pi_e}_{t+1}(S_{t+1}))|S_t = s, A_t = a]$. This motivates defining, as outcome of the rescaled logistic regression model for time step t, the normalized reward-to-go:

$$\tilde{U}^{(i)}_{t,n} := (R^{(i)}_t + \gamma \hat{V}^{\pi_e}_{t+1}(\epsilon_{n,t+1})(S^{(i)}_{t+1}) + \Delta_t)/(2\Delta_t). \tag{5.11}$$

Define $\hat{V}^{\pi_e}_t(\epsilon)$ as the value function corresponding to the action-value function $\hat{Q}^{\pi_e,\delta_n}_t(\epsilon)$, that is, for all $s \in \mathcal{S}$, set $\hat{V}^{\pi_e}_t(\epsilon)(s) = \sum_{a' \in \mathcal{A}} \pi_e(a'|s)\hat{Q}^{\pi_e,\delta_n}(\epsilon)(s, a')$. We define the second-stage model log likelihood for each $t = T-1, ..., 1$ as

$$\begin{aligned}\mathcal{R}^{\delta}_{t,n}(\epsilon) = \frac{1}{n}\sum_{i=1}^{n} \rho^{(i)}_{1:t}\Bigg(&\tilde{U}^{(i)}_t \log(\tilde{Q}^{\pi_e,\delta}_t(\epsilon)(S^{(i)}_t, A^{(i)}_t)) \\ &+ (1 - \tilde{U}^{(i)}_t)\log(1 - \tilde{Q}^{\pi_e,\delta}_t(\epsilon)(S^{(i)}_t, A^{(i)}_t))\Bigg).\end{aligned} \tag{5.12}$$

The fact that the outcome in the second-stage logistic model at time step t depends on the second-stage model fit at time step $t+1$ is why we have to proceed backwards in time. It is also the reason why we say this procedure is a *backward recursion*. Finally, once all of the T second-stage models have been fitted, we define the LTMLE estimator of $V^{\pi_e}_1(s_1)$ as follows:

$$\hat{V}^{\pi_e,LTMLE}_1(s_1) := \hat{V}^{\pi_e}_1(\epsilon_{n,1})(s_1). \tag{5.13}$$

The idea of backward recursion originates from *sequential regression*, first described by [5]. We present the pseudo-code of the procedure as Algorithm 2.

Algorithm 2 Longitudinal TMLE for MDPs

Input: Logged data split $D^{(1)}$, target policy π_e, initial estimators $\hat{Q}^{\pi_e}_1, ..., \hat{Q}^{\pi_e}_T$, discount factor γ.
Set $\Delta_T = 0$ and $\hat{V}^{\pi_e}_{T+1} = 0$.

for $t = T$ ***to*** 1 **do**
Set $\Delta_t = \max_{t,i} |R_t| + \gamma\Delta_t$.
Set $\tilde{U}_t = (R_t + \gamma\hat{V}^{\pi_e}_{t+1} + \Delta_t)/2\Delta_t$.
Set $\tilde{Q}^{\pi_e,\delta_n}_t = \text{threshold}(\delta_n, (\hat{Q}^{\pi_e}_t + \Delta_t)/2\Delta_t)$.
Compute $\epsilon_{n,t} = \arg\min_{\epsilon} \mathcal{R}^{\delta_n}_{n,t}(\epsilon)$.
Set $\hat{Q}^{\pi_e,\delta_n}_t = 2\Delta_t(\tilde{Q}^{\pi_e,\delta_n}_t - 0.5)$.
Set, for all $s \in \mathcal{S}$,

$$\hat{V}^{\pi_e}_t(s) = \sum_{a' \in \mathcal{A}} \pi_e(a'|s)\hat{Q}^{\pi_e,\delta_n}_t(s, a').$$

end

return $\hat{V}^{\pi_e}_1(\epsilon_{n,1})(s_1)$.

Convergence Rate and Asymptotic Distribution

It might at first appear surprising that fitting the second-stage models, which amounts to simply fitting the intercept of a logistic regression model, suffices to fully remove the bias. We nevertheless prove that it does so in Theorem 17 under mild assumptions. Theorem 17 requires Assumption 35 stated in section 5.2, and Assumptions 2-4 stated below. We can also characterize the asymptotic distribution and the asymptotic variance of the LTMLE estimator. In particular, we can show that, provided that $\hat{Q}^{\pi_e}$ is consistent, our estimator attains the generalized Cramer-Rao bound and is therefore *locally efficient*. We also argue that it is asymptotically equivalent with the doubly robust estimator presented before in the literature [123, 63].

Assumption 36. *For all $t = 1, ..., T$, $r_t \in [r_{min}, r_{max}]$ almost surely.*

Assumption 37. *For all $t = 1, ..., T$, the initial estimator $\hat{Q}^{\pi_e}_{t,n}$ converges in probability to some limit $Q_{t,\infty} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, that is $\|\hat{Q}^{\pi_e}_{t,n} - Q_{t,\infty}\|_{P,2} = o_P(1)$.*

Assumption 38. *For all $t = 1, ..., T$, let $Q_{t,\infty}$ be the limit as defined in Assumption 37. Assume there exists a (small) positive constant $\eta \in (0, 1/2)$ such that $\forall t$ and $\forall (s, a) \in \mathcal{S} \times \mathcal{A}$, $Q_{t,\infty}(s, a) \in [\eta, 1 - \eta]$.*

Assumption 39. *Suppose there exists a finite positive constant M such that $\forall t$, $\rho_{1:t} \leq M$ almost surely.*

Theorem 17. *Suppose Assumptions 36, 37, 38, and 39 hold. Then the LTMLE estimator has bias $o(1/\sqrt{n})$, that is*

$$E_{P,\pi_b}[\hat{V}_1^{\pi_e, LTMLE}(s_1)] - V_1^{\pi_e}(s_1) = o(1/\sqrt{n}).$$

In addition, the LTMLE estimator converges in probability at rate $\sqrt{n}$, that is

$$\hat{V}_1^{\pi_e, LTMLE}(s_1) - V_1^{\pi_e}(s_1) = O_P(1/\sqrt{n}). \tag{5.14}$$

5.4 RLTMLE

In the following, we (1) present regularizations which can be applied to LTMLE, and (2) describe our "final estimator", denoted RLTMLE (standing for *LTMLE for RL*). In particular, RLTMLE consists of a convex combination of regularized LTMLE estimators. The weights in the RLTMLE convex combination are obtained following a variant of the ensembling procedure of the MAGIC estimator, presented earlier [123].

Regularization and base estimators

We introduce three ready-to-use regularization techniques which allow variance stabilization of the LTMLE estimator. The first two have a clear WDR analogue, while the third one only applies to LTMLE.

1. **Weight softening:** For $\alpha \in [0,1]$, $x \in \mathbb{R}^d$, we define weight softening as $\texttt{soften}(x,\alpha) := (x_k^\alpha / \sum_{l=1}^d x_l^\alpha : k = 1,...,d)$. The LTMLE algorithm corresponding to softening level α is obtained by replacing, in the second-stage log likelihoods (5.10) and (5.12), the IS ratios $(\rho_{1:t}^{(i)} : i = 1,...,n)$ by $\texttt{soften}((\rho_{1:t}^{(i)} : i = 1,...,n), \alpha)$. The same operation can be applied to the importance weights of the WDR estimator.

2. **Partial horizon:** The LTMLE with partial horizon $\tau < T$ is obtained by setting to zero the coefficients $\epsilon_{n,\tau_1}, ..., \epsilon_{n,T}$ before fitting the other second-stage coefficients. This enforces importance sampling ratios $\rho_{1:t}$ for $t \geq j$ to have no impact on the estimator. The WDR equivalent is to use the τ-step return g_τ.

3. **Penalization:** The penalized LTMLE is obtained by adding a penalty $\lambda|\epsilon_{n,t}|$ for some $\lambda \geq 0$ to the the log-likelihoods (5.10) and (5.12) of the second-stage models.

The three regularizations can be applied simultaneously, as well as individually. A regularized LTMLE estimator can therefore be indexed by a triple (α, τ, λ), where α, τ and λ denote the level of softening, the partial horizon, and the level of likelihood penalization.

Ensemble estimator

Our final proposed estimator is an ensemble of a pool of regularized LTMLE estimators, which we denote $g_1, ..., g_K$. In particular, the ensembled estimator corresponds to a sequence of triples $(\alpha_1, \tau_1, \lambda_1), ..., (\alpha_K, \tau_K, \lambda_K)$ of regularization levels. We set g_K to be the unregularized LTMLE, that is we set $(\alpha_K, \tau_K, \lambda_K) = (1, T, 0)$. We ensemble the regularized LTMLE estimators $g_1, ..., g_K$ by taking a convex combination of them that minimizes an estimate of MSE; the ensembling step closely follows that of the MAGIC procedure. We propose two variants of it, which we call RLTMLE 1 and RLTMLE 2, differing in how we estimate the covariance matrix Ω_n (defined in section 5.3) of base estimators $g_1, ..., g_K$.

For RLTMLE1, covariance estimation relies on asymptotic properties of the LTMLE estimator. In particular, the difference between a regularized LTMLE estimator with regularization parameters (α, τ, λ), and its asymptotic limit is given by $n^{-1} \sum_{i=1}^n \mathrm{EIF}(\hat{Q}, \alpha, \tau, \lambda)(H_i) + o_P(n^{-1/2})$, where EIF is the efficient influence function whose expression is given by

$$\begin{aligned} &\mathrm{EIF}(\hat{Q}^{\pi_e}, \alpha, \lambda, \tau)(h) \\ &= \sum_{t=1}^{T} \gamma^t \rho_t \times \left(r_t + \gamma \hat{V}_{t+1}^{\pi_e}(\epsilon_{n,t+1})(s_{t+1}) - \hat{Q}_t^{\pi_e}(\epsilon_{n,t})(s_t, a_t) \right). \end{aligned} \tag{5.15}$$

Here, for all t, $\epsilon_{n,t}$ is the maximizer of the regularized version of the log-likelihood (5.12) (where ρ_t is replaced with soften(ρ_t, α) and penalized by $\lambda|\epsilon|$). We denote the EIF corresponding to estimator g_k as $\text{EIF}_k(h) = \text{EIF}(\hat{Q}, \alpha_k, \lambda_k, \tau_k)(h)$. The estimate of the covariance matrix Ω_n is then the empirical covariance matrix $\hat{\Omega}_n$ of $(\text{EIF}_1(H), ..., \text{EIF}_K(H))$.

On the other hand, for RLTMLE2, an estimate of the covariance matrix Ω_n of the base estimators $g = (g_1, ..., g_K)$ is obtained by computing bootstrapped values $g^{(1)}, ..., g^{(B)}$ of g. In particular, we generate a large enough number of bootstrap samples B, and compute the empirical covariance $\hat{\Omega}_n$ matrix of $g^{(1)}, ..., g^{(B)}$. In Algorithm 3 we present the pseudo-code description of RLTMLE2, which is our most performant algorithm.

Bias Estimation

For bias estimation, we follow closely the method proposed by Thomas et al. [123]. In particular, for $k = 1, ..., K$, we denote by $b_{n,k}$ the bias of estimator g_K, and $b_n := (b_{n,1}, ..., b_{n,K})$. Further, let CI$(\alpha)$ denote the α-percentile bootstrap confidence interval for the LTMLE estimator. For both RLTMLE 1 and RLTMLE 2, and for each $k = 1, ..., K$, we estimate the bias $b_{n,k}$ with $\hat{b}_{n,k} := \text{dist}(g_k, \text{CI}(\alpha))$. Finally, we denote $\hat{b}_n := (\hat{b}_{n,1}, ..., \hat{b}_{n,K})$.

Algorithm 3 RLTMLE 2

Input:
Logged data split $D^{(1)}$,
target policy π_e,
initial estimator $\hat{Q}^{\pi_e} := (\hat{Q}_1^{\pi_e}, ..., \hat{Q}_T^{\pi_e})$,
discount factor γ,
triples of regularization levels $(\alpha_1, \tau_1, \lambda_1), ..., (\alpha_K, \tau_K, \lambda_K)$,
number of bootstrap samples B.

for $b = 1$ ***to*** B **do**
Sample with replacement from $D^{(1)}$ a bootstrap sample $D^{*,(b)}$.
for $k = 1$ ***to*** K **do**
Compute $g_k^{(b)}$ with algorithm 2 using inputs $D^{*,(b)}$, $\hat{Q}^{\pi_e}$, π_e, γ and $(\alpha_k, \tau_k, \lambda_k)$.
end
end

for $k = 1$ ***to*** K **do**
Compute g_k with algorithm 2 using inputs $D^{(1)}$, $\hat{Q}^{\pi_e}$, π_e, γ and $(\alpha_k, \tau_k, \lambda_k)$.
for $l = 1$ ***to*** K **do**
$\hat{\Omega}_{k,l} \leftarrow n^{-1} \sum_{b=1}^{B} g_k^{(b)} g_l^{(b)} - \left(n^{-1} \sum_{b=1}^{B} g_k^{(b)}\right) \left(n^{-1} \sum_{b=1}^{B} g_l^{(b)}\right)$.
end
$\text{CI}(\alpha) \leftarrow \left[\text{percentile}(\{g_k^{(b)} : b\}, \alpha), \text{percentile}(\{g_k^{(b)} : b\}, 1-\alpha)\right]$.
$\hat{b}_{n,k} \leftarrow \text{distance}(g_k, \text{CI}(\alpha))$.
end

$$\hat{x} \leftarrow \arg \min_{\substack{0 \le x \le 1 \\ x^\top 1 = 1}} \frac{1}{n} x^\top \hat{\Omega}_n x + (x^\top \hat{b}_n)^2.$$

return $\hat{x}^\top g$.

5.5 Simulations

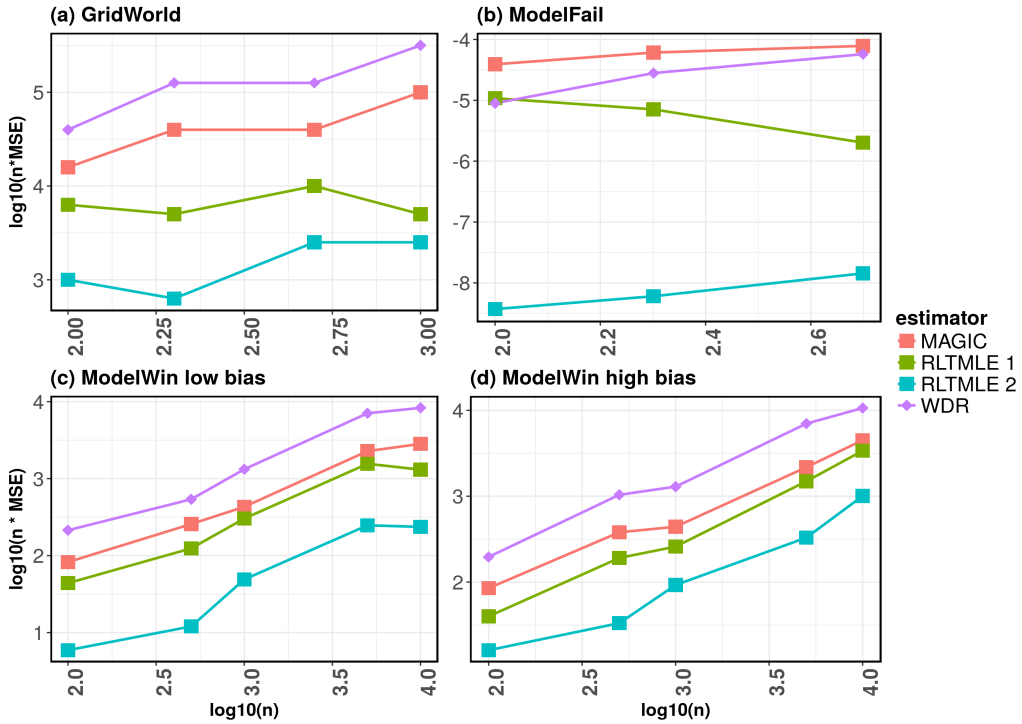


Figure 5.1: Empirical results for three different environments and varying level of model misspecification. **(a)** GridWorld MSE across varying sample size $n = (100, 200, 500, 1000)$ and bias equivalent to $b_0 = 0.005 * \text{Normal}(0, 1)$ over 71 trials; **(b)** ModelFail MSE across varying sample size $n = (100, 200, 500, 1000)$ and bias equivalent to $b_0 = 0.005 * \text{Normal}(0, 1)$ over 71 trials; **(c)** ModelWin MSE across varying sample size $n = (100, 500, 1000, 5000, 10000)$ and bias equivalent to $b_0 = 0.005 * \text{Normal}(0, 1)$ over 63 trials; **(d)** ModelWin MSE across varying sample size $n = (100, 500, 1000, 5000, 10000)$ and bias equivalent to $b_0 = 0.05 * \text{Normal}(0, 1)$ over 63 trials.

In this section, we demonstrate effectiveness of RLTMLE by comparing it with other state-of-the-art methods used for OPE problem in various RL benchmark environments. We used three main domains, often described in the OPE literature. We implement the same behavior and evaluation policies as in previous work [123, 39].

1. **ModelFail**: a partially observable, deterministic domain with $T = 3$. Here the approximate model is incorrect, even asymptotically, due to three of the four states appearing identical to the agent.

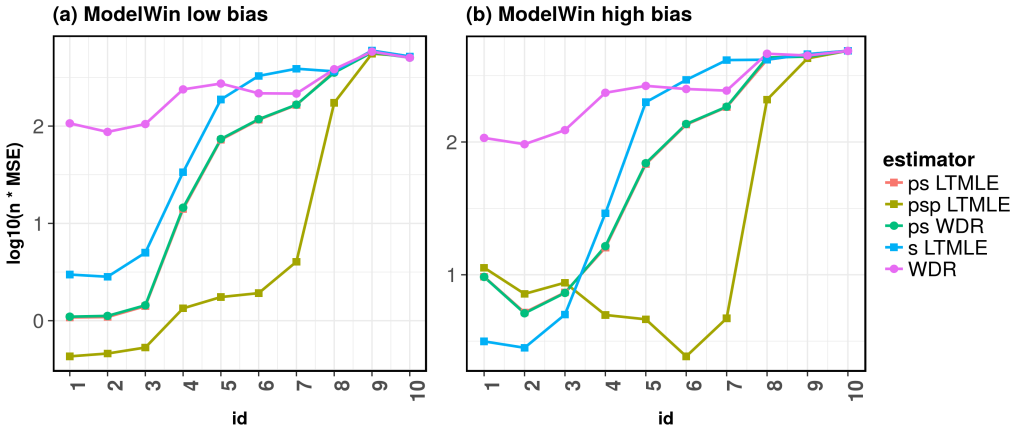


Figure 5.2: Comparison of WDR and LTMLE base estimators across various regularization methods in ModelWin at low ($b_0 = 0.005 * \text{Normal}(0,1)$) and high ($b_0 = 0.05 * \text{Normal}(0,1)$) model misspecification. Regularized base estimators include ps LTMLE (partial, softened LTMLE), ps WDR (partial, softened WDR), psp LTMLE (partial, softened, penalized LTMLE), s LTMLE (softened LTMLE) and WDR (no regularization). The x-axis indicates the id of the k^{th} estimator, corresponding to $(\alpha_k, \lambda_k, \tau_k)$. **(a)** ModelWin MSE for sample size $n = 1000$ and low bias over 315 trials; **(b)** ModelWin MSE for sample size $n = 1000$ and high bias over 315 trials.

2. **ModelWin**: a stochastic MDP with $T = 10$, where the approximate model can perfectly represent the MDP.

3. **GridWorld**: a 4×4 grid used for evaluating OPE methods, with an episode ending at $T = 100$ or when a final state ($s16$) is reached.

We omit benefits of RLTMLE over IS, PDIS (per-decision IS), WIS (weighted IS), CWPDIS (consistent weighted per-decision IS) and DR (doubly robust) estimators due to the extensive empirical studies performed by Thomas et al. [123]. Instead, we compare our estimator to WDR and MAGIC, as they demonstrate improved performance over all simulations in benchmark RL environments considered [123].

In evaluating our estimator, we also explore how various degree of model misspecification and sample size can affect the performance of considered methods. We start with small amount of bias, $b_0 = 0.005 * \text{Normal}(0,1)$, where most estimators should do well. Consequently, we increase model misspecification to $b_0 = 0.05 * \text{Normal}(0,1)$ at the same sample size, and consider the performance of all estimators. In addition, we test sensitivity to the number of episodes in D with $n = \{100, 200, 500, 1000)$ for GridWorld and ModelFail, and $n = \{100, 500, 1000, 5000, 10000)$ for ModelWin. In addition, we consider the benefits of adding few regularization techniques as opposed to all three described in subsection 5.4. In particular, we concentrate on RLTMLE with only weight softening and partial LTMLE

(RLTMLE 1) as opposed to using penalized LTMLE as well (RLTMLE 2). The goal of these experiments was to demonstrate the improved performance of our estimator when fully exploiting all the variance reduction techniques in a clever way. The MSE across varying sample size and model misspecification for GridWorld, ModelFail and ModelWin can be found in Figure 5.1. We can see that RLTMLE 2 outperforms all other estimators for all RL environments and varying levels of model misspecification.

Finally, we compare WDR and LTMLE base estimators augmented with various regularization methods before the ensemble step in Figure 5.5. In particular, for ModelWin, we look at the MSE of $\hat{V}_1^{\pi_e,j}(\epsilon_{n,1})(s_1)$ and g_k for each k, where the k^{th} estimator corresponds to regularization $(\alpha_k, \lambda_k, \tau_k)$. Regularized base estimators considered include ps LTMLE (partial, softened LTMLE), ps WDR (partial, softened WDR), psp LTMLE (partial, softened, penalized LTMLE), s LTMLE (softened LTMLE) and WDR (no regularization). We note the vast improvement of WDR just by adding weight softening across all base estimators, evident for both low and high model misspecification setting. For the low bias environment of ModelWin, psp LTMLE (RLTMLE 2) uniformly outperforms all competitors for all k. High bias setting loses to s LTMLE for low k, but still outperforms majority of the time, including having the best ensemble MSE. While uniform win over all k is not necessary, we note that this behavior stems from the fact that for $k < 3$, $(\alpha_k, \lambda_k, \tau_k)$ used had very small τ_k and α_k. As such, with no strong debiasing effect of LTMLE, minimizing variance becomes more effective with respect to minimizing MSE.

5.6 Discussion

In this chapter, we propose a new doubly robust estimator for off-policy value evaluation in reinforcement learning. In particular, we present a convex combination of regularized LTMLE estimators which aim at minimizing the MSE. We showed that our estimator is consistent and asymptotically optimal, achieving the Cramer-Rao lower bound. We prove the $O_P(1/\sqrt{n})$ rate of convergence of our estimator, and characterize its asymptotic distribution. The LTMLE is guaranteed to lie in the allowed rewards domain, both for discrete and continuous state, and is amenable to several regularization techniques. Finally, our experiments demonstrate uniform win of RLTMLE over all considered off-policy methods across multiple RL environments and various levels of model misspecification.

The RLTMLE enjoys multiple distinguishing features that contribute to its finite sample performance. First, its base estimator is a substitution estimator, therefore it inherently respects the reward domain for the RL problem. While this is true for DR if states and actions are discrete, our estimator by design produces estimates that lie in the allowed reward domain for both discrete and continuous state space. Our estimator also allows for clever usage of importance weights, instead of explicitly summing over IS terms. This property strives from using LTMLE as a base estimator, where stabilized IS ratios can be used as weights of the observations in the log likelihood of the second-stage models. This is an important feature of RLTMLE, that greatly contributes to its stability without introducing

bias. Finally, LTMLE is amenable to many regularization methods, with RLTMLE enjoying a rich family of regularized base estimators. Our experiments show impressive performance gains from utilizing variance reduction techniques for both RLTMLE and WDR. Finally, our method does not refit the entire reward-to-go model for each new target policy as the More Robust Doubly Robust estimator, demonstrating some practical advantages. Since refitting the reward-to-go model can be quite computationally expensive, our estimator might be beneficial in situations where one wants to scan through many candidate target policies.

5.7 Appendix

Simplified sample-splitting-based algorithm

In this section, we present the theoretical analysis for the algorithm derived in section 5.3. We first outline the steps of the proof in the proof sketch below. We then state the four main lemmas on which the proof relies, and present the formal proof.

The first fact underpinning the proof is that for any of candidate action-value $Q' = (Q'_1, ..., Q'_T)$ and corresponding value functions $V' = (V'_1, ..., V'_T)$, the difference between the candidate and the true value function at time point $t = 1$ can be decomposed as follows:

$$V'_1(s_1) - V(s_1) = -\int D(Q')(h) dP^{\pi_b}(h), \tag{5.16}$$

where $D(Q')(h) = \sum_{t=1}^{T} D_t(Q')(h)$, with $D_t(Q')(h) = \rho_{1:t}(h)(r_t + \gamma V'_{t+1}(s_{t+1}) - Q'_t(s_t, a_t))$. This is formally stated in lemma 8 below. For non-random functions Q' and V' note that the RHS of (5.16) is equal to $-E_{P,\pi_b}[D(Q')]$. The second fact the proof relies on is that the estimators $\hat{Q}(\epsilon_n)$ resulting from the fitting of the parametric second stages verify the following equation:

$$\frac{1}{n}\sum_{i=1}^{n} D(\hat{Q}(\epsilon_n))(H_i) = 0. \tag{5.17}$$

This is formally stated in lemma 9 below. The argument in the proof of lemma 9 can be simply summarized as follows. For each t, $D_t(\hat{Q}(\epsilon_{n,t}))$ is the score function of the log likelihood of the second-stage logistic model for time point t. The third fact we use in the proof is that ϵ_n converges in probability to some limit ϵ_∞. Heuristically, the reason why this is the case is that, due to the convergence of $\hat{Q}_n$ to Q_∞, the log likelihoods of the second stage models converge to a limit, which in turns implies that their arg min ϵ_n converges to the arg min of their limit. We make this rigorous in lemma 10 below.

Using the first two facts stated above, we obtain, by adding up equations (5.16) and (5.17), that the difference between our estimator $\hat{V}_1^{LTMLE}(\epsilon_n)(s_1)$ and the truth $V_1(s_1)$ is

$$\begin{aligned}&\hat{V}_1^{LTMLE}(\epsilon_n)(s_1) - V_1(s_1)\\ &= \frac{1}{n}\sum_{i=1}^{n} D(\hat{Q}(\epsilon_n))(H_i) - \int D(\hat{Q}(\epsilon_n))(h) dP^{\pi_b}(h).\end{aligned}$$

Using the third fact stated above, that ϵ_n converges to some ϵ_∞, motivates rewriting the

above display as

$$\begin{aligned}
&\hat{V}_1^{LTMLE}(\epsilon_n)(s_1) - V_1(s_1) \\
&= \frac{1}{n}\sum_{i=1}^{n} D(\hat{Q}(\epsilon_\infty))(H_i) - \int D(\hat{Q}(\epsilon_\infty))(h) dP^{\pi_b}(h) \\
&+ \frac{1}{n}\sum_{i=1}^{n} D(\hat{Q}(\epsilon_n))(H_i) - D(\hat{Q}(\epsilon_\infty))(H_i) \\
&- \int D(\hat{Q}(\epsilon_\infty))(h) - D(\hat{Q}(\epsilon_n))(h) dP^{\pi_b}(h).
\end{aligned}$$

Denote $\mathcal{T}$ as the sample split on which the initial estimators are fitted. Since $h \mapsto D(\hat{Q}(\epsilon_\infty))(h)$ is a non-random function conditional on $\mathcal{T}$, we have that

$$\int D(\hat{Q}(\epsilon_\infty))(h) dP^{\pi_b}(h) = E_{P,\pi_b}[D(\hat{Q}(\epsilon_\infty))|\mathcal{T}].$$

Therefore, applying the Central Limit theorem conditional on $\mathcal{T}$ gives us that the first line of the RHS in the above display is asymptotically normally distributed and is of order $O_P(1/\sqrt{n})$. As we will show in the formal proof, this also holds after marginilazing w.r.t. $\mathcal{T}$. The term formed by the second and third lines in the RHS of the above display can be shown to be $o_P(1/\sqrt{n})$. This is formally stated in the lemma below.

Lemma 8 (First order expansion)**.** *Consider $Q' = (Q'_1, ..., Q'_T)$ a candidate vector of action-value functions $\mathcal{S} \times \mathcal{A} \to \mathbb{R}$ for polict π_e, and let $V' = (V'_1, ..., V'_T)$ the corresponding vector of state-value functions under π_e, that is, for all t, $s \in \mathcal{S}$, $V'_t(s) = \sum_{a' \in \mathcal{A}} \pi_e(a'|s) Q'_t(s, a')$. Denote $Q = (Q_1, ..., Q_T)$ and $V = (V_1, ..., V_T)$ the true action-value and state value functions under π_e. For all t, for all $h \in \mathcal{H}$, denote $\rho'_{1:t}(h)$ an importance sampling ratio for time point t and trajectory h, not necessarily equal to the true importance sampling ratio. Denote $\rho = (\rho_1, ..., \rho_T)$ and $\rho' = (\rho'_1, ..., \rho'_T)$. We have that*

$$V'_1(s_1) - V_1(s_1) = -\int D(\rho', Q')(h) dP^{\pi_b}(h) - \int Rem(\rho, \rho', Q, Q')(h) dP^{\pi_b}(h),$$

$D(\rho', Q')(h) = \sum_{t=1}^{T} D_t(\rho', Q')(h)$ and $Rem(\rho, \rho', Q, Q')(h) = \sum_{t=1}^{T} Rem_t(\rho, \rho', Q, Q')(h)$ with

$$D_t(\rho', Q')(h) = \gamma^{t-1} \rho'_{1:t}(h) \left(r_t + \gamma V'_{t+1}(s_{t+1}) - Q't(s_t, a_t) \right),$$

and

$$\begin{aligned}
&Rem_t(\rho, \rho', Q, Q')(h) \\
&= \gamma^{t-1} \big(\rho_{1:t}(h) - \rho'_{1:t}(h) \big) \big(Q_t(s_t, a_t) - Q'_t(s_t, a_t) \\
&\qquad + (V_{t+1}(s_{t+1}) - V'_{t+1}(s_{t+1})) \big).
\end{aligned}$$

From the expression in the RHS of the above display, it is immediately clear that

$$Rem_t(\rho, \rho', Q, Q')(h) = 0$$

if $\rho = \rho'$ *or* $Q = Q'$.

The lemma below shows that the maximum likelihood fits $\epsilon_{n,t}$ of the second-stage parametric models solve a certain equation, termed score equation in statistics.

Lemma 9 (Score equation)**.** *Consider the simplified LTMLE algorithm described in section 5.3. For each* $t = 1, ..., T$*, the maximum likelihood fit* $\epsilon_{n,t}$ *satisfies*

$$\sum_{i=1}^{n} D_t(\rho_{1:t}, \hat{Q}(\epsilon_{n,t}))(H_i) = 0.$$

The following lemma shows that the vector $\epsilon_n = (\epsilon_{n,1}, ..., \epsilon_{n,T})$ of the maximum likelihood fits of the second stage models converges in probability to a limit.

Lemma 10 (Convergence of ϵ_n)**.** *Under assumptions 36, 37, 38 and 39, there exists* $\epsilon_\infty \in \mathcal{R}^T$ *such that*

$$\epsilon_n - \epsilon_\infty = o_P(1).$$

Lemma 11 (Equicontinuity)**.** *Denote, for all* $h \in \mathcal{H}$, $\epsilon \in \mathbb{R}$, Q' *and* ρ'

$$g_\epsilon(Q', \rho')(h) = D(Q'(\epsilon), \rho')(h),$$

where Q' *and* ρ' *are possibly random. Suppose* $H_1, ..., H_n$ *are i.i.d. trajectories drawn from* P^{π_b}*. Suppose further that* $H_1, ..., H_n$ *are independent from the potentially random functions* Q' *and* ρ'*. Suppose* $\epsilon'_n \xrightarrow{P} \epsilon'_\infty$ *for some* ϵ'_∞*. Then*

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^{n} g_{\epsilon'_n}(Q', \rho')(H_i) - \int g_{\epsilon'_n}(Q', \rho')(h) dP^{\pi_b}(h) \\ &- \frac{1}{n}\sum_{i=1}^{n} g_{\epsilon'_\infty}(Q', \rho')(H_i) - \int g_{\epsilon'_\infty}(Q', \rho')(h) dP^{\pi_b}(h) = o_P\left(\frac{1}{\sqrt{n}}\right).\end{aligned}$$

Proof. From lemma 8,

$$\hat{V}_1^{TMLE}(s_1) - V_1(s_1) = -P^{\pi_b} D(\hat{Q}_n(\epsilon_n), \rho).$$

Since from lemma 9 we have $P_n(D(\hat{Q}_n(\epsilon_n), \rho) = 0$, we can add this latter identity to the above display, which yields

$$\begin{aligned}\hat{V}_1^{TMLE}(s_1) - V_1(s_1) =& (P_n - P^{\pi_b}) D(\hat{Q}_n(\epsilon_n), \rho) \\ =& (P_n - P^{\pi_b}) D(Q_\infty(\epsilon_\infty), \rho) \\ &+ (P_n - P^{\pi_b})(D(Q_\infty(\epsilon_\infty), \rho) - D(\hat{Q}_n(\epsilon_n), \rho)).\end{aligned} \tag{5.18}$$

From the Central Limit theorem applied conditionally on $\mathcal{T}$, $\sqrt{n}((P_n - P^{\pi_b})D(Q_\infty(\epsilon_\infty), \rho)) \xrightarrow{d} \mathcal{N}(0, \sigma^2(Q_\infty(\epsilon_\infty))$, with

$$\sigma^2(Q_\infty(\epsilon_\infty)) := Var_{P^{\pi_b}}(D(Q_\infty(\epsilon_\infty), \rho)).$$

Using dominated convergence on the c.d.f. on the LHS, $\sqrt{n}((P_n - P^{\pi_b})D(Q_\infty(\epsilon_\infty), \rho)) \xrightarrow{d} \mathcal{N}(0, \sigma^2(Q_\infty(\epsilon_\infty))$ also holds true unconditionally. □

Proof of lemma 8

Proof. Let $H \sim P^{\pi_e}$. If Q', V' are random functions, further suppose, without loss of generality, that H is independent of Q' and V'. Denote $\mathcal{G}$ a σ-field such that Q', V' are $\mathcal{G}$-measurable.

Step 1. Observe that

$$P^{\pi_b}D(Q', \rho') = P^{\pi_b}D(Q', \rho) + P^{pi_b}(D(Q', \rho') - D(Q', \rho))$$

Step 2: First order term. Observe that

$$P^{\pi_b}D(Q', \rho) = E_{P^{\pi_b}}[D(Q', \rho)(H)|\mathcal{G}].$$

For all $t \geq 1, ..., T$, denote $\mathcal{F}_t$ the σ-field induced by $S_1, A_1, R_1, ..., S_t, A_t, R_t$. Observe that

$$\begin{aligned}
&E_{P^{\pi_b}}[D_t(Q', \rho)(H)|S_t, A_t, \mathcal{F}_{t-1}, \mathcal{G}] \\
&= \gamma^{t-1} E_{P^{\pi_b}}[\rho_{1:t}(R_t + \gamma V'_{t+1}(S_{t+1}) - Q'_t(S_t, A_t))|S_t, A_t, \mathcal{F}_{t-1}, \mathcal{G}] \\
&= \gamma^{t-1}\rho_{1:t} E_P[R_t + \gamma V_{t+1}(S_{t+1}) - Q'_t(S_t, A_t)|S_t, A_t, \mathcal{G}] \\
&\quad + \gamma^t \rho_{1:t} E_P[(V'_{t+1}(S_{t+1}) - V_{t+1}(S_{t+1}))|S_t, A_t, \mathcal{G}].
\end{aligned}$$

Recall that by definition of Q, we have that $E_P[R_t + \gamma V_{t+1}(S_{t+1})|S_t, A_t] = Q_t(S_t, A_t)$. Inserting this in the last line of the above display yields

$$\begin{aligned}
&E_{P^{\pi_b}}[D_t(Q', \rho)(H)|S_t, A_t, \mathcal{F}_{t-1}, \mathcal{G}] = \\
&\gamma^{t-1}\rho_{1:t}(Q_t(S_t, A_t) - Q'_t(S_t, A_t)) \\
&\quad + \gamma^t \rho_{1:t} E_P[V'_{t+1}(S_{t+1}) - V_{t+1}(S_{t+1})|S_t, A_t, \mathcal{G}].
\end{aligned} \tag{5.19}$$

We take the expectation conditional on S_t, $\mathcal{F}_{t-1}$, $\mathcal{G}$ of the first term in the right-hand side of the above display:

$$\begin{aligned}
&E_{P^{\pi_b}}[\gamma^{t-1}\rho_{1:t}(Q_t(S_t, A_t) - Q'_t(S_t, A_t))|S_t, \mathcal{F}_{t-1}, \mathcal{G}] \\
&= \gamma^{t-1}\rho_{1:t-1} E_{P,\pi_b}[\rho_t(Q_t(S_t, A_t) - Q'_t(S_t, A_t)|S_t, \mathcal{G}] \\
&= \gamma^{t-1}\rho_{1:t-1} E_{P,\pi_e}[(Q_t(S_t, A_t) - Q'_t(S_t, A_t)|S_t, \mathcal{G}] \\
&= \gamma^{t-1}\rho_{1:t-1}(V_t(S_t) - V'_t(S_t)).
\end{aligned} \tag{5.20}$$

The second equality above uses that, for all $\mathcal{G}$-measurable function f, $E_{P,\pi_b}[\rho_t f(S_t, A_t)|S_t, \mathcal{G}] = E_{P,\pi_e}[f(S_t, A_t)|S_t, \mathcal{G}]$. The third equality follows from the relationship between the value function and the action value function. Using the law of iterated expectations, and identities (5.20) and (5.21), we have that

$$
\begin{aligned}
&E_{P,\pi_b}[D_t(Q', \rho)(H)|\mathcal{G}] \qquad (5.21)\\
&= E_{P,\pi_b}[E_{P,\pi_b}[D_t(Q', \rho)(H)|S_t, A_t, \mathcal{F}_{t-1}, \mathcal{G}]|\mathcal{G}]\\
&= E_{P,\pi_b}[\gamma^{t-1}\rho_{1:t}(Q_t(S_t, A_t) - Q'_t(S_t, A_t))|\mathcal{G}]\\
&\quad + E_{P,\pi_b}[\gamma^t \rho_{1:t} E_P[V'_{t+1}(S_{t+1}) - V_{t+1}(S_{t+1})|S_t, A_t, \mathcal{G}]|\mathcal{G}]\\
&= E_{P,\pi_b}[E_{P,\pi_b}[\gamma^{t-1}\rho_{1:t}(Q_t(S_t, A_t) - Q'_t(S_t, A_t))|S_t, \mathcal{F}_{t-1}, \mathcal{G}]|\mathcal{G}]\\
&\quad + E_{P,\pi_b}[\gamma^t \rho_{1:t}(V'_{t+1}(S_{t+1}) - V_{t+1}(S_{t+1}))|\mathcal{G}]\\
&= E_{P,\pi_b}[\gamma^{t-1}\rho_{1:t-1}(V_t(S_t) - V'_t(S_t))|\mathcal{G}]\\
&\quad + E_{P,\pi_b}[\gamma^t \rho_{1:t}(V'_{t+1}(S_{t+1}) - V_{t+1}(S_{t+1}))|\mathcal{G}]
\end{aligned}
$$

Using the above expression in the definition of $D(Q', V')$ yields

$$
\begin{aligned}
E_{P,\pi_b}[D(Q', \rho)(H)|\mathcal{G}] =& \sum_{t=1}^{T} E_{P,\pi_b}[\gamma^t \rho_{1:t}(V'_{t+1}(S_{t+1}) - V'_{t+1}(S_{t+1})) \qquad (5.22)\\
&\qquad - \gamma^{t-1}\rho_{1:t-1}(V'_t(S_t) - V_t(S_t))|\mathcal{G}]\\
=& E_{P,\pi_b}[\gamma^T \rho_{1:T+1}(V'_{T+1}(S_{T+1}) - V_{T+1}(S_{T+1}))\\
&\quad - \rho_{1:0}(V'_1(s_1) - V_1(s_1))|\mathcal{G}]\\
=& -(V'_1(s_1) - V_1(s_1)),
\end{aligned}
$$

where we have used that by convention $V'_{T+1}(S_{T+1}) = V_{T+1}(S_{T+1}) = 0$ and $\rho_{1:0} = 1$.

Step 3: remainder term. Similarly, we show that

$$P^{\pi_b}(D'(Q', \rho) - D(Q', \rho)) = Rem_t(Q, Q', \rho, \rho').$$

□

Proof of lemma 9

We only present a proof sketch in this subsection of the Chapter, with the full proof allocated to the article. The result essentially follows from the following two facts:

1. The score of the logistic likelihood of the second stage model at t is $P_n D_t(\hat{Q}.\rho)$.
2. The maximum likelihood fit solves the empirical score equation.

Proof of lemma 10

The convergence of $\hat{Q}$ to Q_∞ implies the pointwise convergence of the log likelihood risk $\mathcal{R}_{n,t}$ to some asymptotic risk $\mathcal{R}_{\infty,t}$. The fact that $Q_{\infty,t} \in [\delta, 1-\delta] \subset (0,1)$ (in other words, that $Q_{\infty,t}$ is bounded away from 0 and 1) implies that the asymptotic log likelihood risk $\mathcal{R}_{\infty,t}$ is strongly convex; this implies it has a unique minimizer $\epsilon_{\infty,t}$. We then show in the formal proof that since $\mathcal{R}_{n,t}$ are a sequence of convex functions that converge pointwise in probability to a strongly convex function minimized by $\epsilon_{\infty,t}$, the sequence of their minimizers $\epsilon_{n,t}$ converges in probability to $\epsilon_{\infty,t}$

Proof of lemma 11

The proof of lemma 11 relies on the following three technical lemmas. Recall the following definition: for all Q' ρ', $h \in \mathcal{H}$, $\epsilon \in \mathbb{R}$,

$$g_\epsilon(Q', \rho')(h) = D(Q'(\epsilon), \rho')(h).$$

Lemma 12. *Assume that $0 \leq \rho'_{1:t}(H) \leq M$ almost surely for all $t = 1, ..., T$. Under assumption 36 on the range of the rewards, for all $\epsilon \in \mathbb{R}^T$ we have that*

$$\|g_\epsilon(Q', \rho')\|_{L_\infty(P^{\pi_b})} \leq 3MT,$$

and for all $\epsilon_1, \epsilon_2 \in \mathbb{R}^T$

$$\|g_{\epsilon_1}(Q', \rho') - g_{\epsilon_2}(Q', \rho')\|_{L_\infty(P^{\pi_b})} \leq 2MT\|\epsilon_1 - \epsilon_2\|_\infty. \tag{5.23}$$

For any $\epsilon_0 \in \mathbb{R}$, and any $\xi > 0$, define the class of functions

$$\mathcal{G}(Q', \rho')(\epsilon_0, \xi) := \{g_\epsilon(Q', \rho') - g_{\epsilon_0}(Q', \rho') : \|\epsilon - \epsilon_0\|_\infty \leq \xi\}.$$

Lemma 13. *For any $\alpha > 0$ and any probability distribution Λ on $\mathcal{H}$ with $L = 2MT$,*

$$N(\alpha, L_2(\Lambda), \mathcal{G}(\epsilon_0, \xi)) \leq \left(\frac{2\xi L}{\alpha}\right)^T.$$

Proof. Consider the set

$$\left\{\left(\epsilon_{0,1} + i_1\frac{\alpha}{L}, ..., \epsilon_{0,T} + i_T\frac{\alpha}{L}\right) : \forall t = 1, ...T,\ i_t \in \mathbb{Z} \cap \left[-\frac{\xi L}{\alpha}, \frac{\xi L}{\alpha}\right]\right\}. \tag{5.24}$$

Observe that for any $f_\epsilon := g_\epsilon(Q', \rho') - g_{\epsilon_0}(Q', \rho') \in \mathcal{G}(Q', \rho')(\epsilon_0, \xi)$, there exists an $f_{\epsilon'} := g_{\epsilon'}(Q', \rho') - g_{\epsilon_0}(Q', \rho')$ in the set above such that $\|\epsilon - \epsilon'\|_\infty \leq \alpha/L$. From the second claim in lemma 12, for all $h \in \mathcal{H}$, $|f_{\epsilon'}(h) - f_\epsilon(h)| \leq \alpha$. Therefore, for any probability distribution Λ over $\mathcal{H}$,

$$\|f_{\epsilon'} - f_\epsilon\|_{L_2(\Lambda)} = \left(\int (f_{\epsilon'}(h) - f_\epsilon(h))^2 d\Lambda(h)\right)^{1/2} \leq \alpha.$$

Therefore the set defined above is an α-cover of $\mathcal{G}(\epsilon_0, \xi))$ for the norm $L_2(\Lambda)$. Since this set has at most $(2\epsilon_L/\alpha)^T$ elements, this proves that

$$N(\alpha, L_2(\Lambda), \mathcal{G}(\epsilon_0, \xi)) \leq \left(\frac{2\xi L}{\alpha}\right)^T.$$

□

The covering numbers characterized in lemma 13 are the basis for another measure of geometric complexity of a class of function, the uniform entropy integral, whose definition we recall below (see also [126]).

Definition 9 (Uniform entropy integral)**.** *Consider a class of functions $\mathcal{X} \to \mathbb{R}$. Let $F : \mathcal{X} \to \mathbb{R}$ be an envelope function for $\mathcal{F}$, that is a function such that for all $x \in \mathcal{X}$, $|f(x)| \leq F(x)$. The uniform entropy integral of $\mathcal{F}$, w.r.t. the envelope function F and L_2 norm is defined, for all $\beta > 0$ as*

$$J_F(\beta, \mathcal{F}, L_2) := \int_0^\beta \sup_\Lambda \sqrt{\log(1 + N(\alpha \|F\|_{\Lambda,2}, L_2(\Lambda), \mathcal{F})} d\alpha,$$

where the supremum is over all discrete probability distributions on $\mathcal{X}$.

Lemma 14. *Let $\beta > 0$, and denote $L = 2MT$. The function $F_\xi : h \mapsto L\xi$ is an envelope function for $\mathcal{G}(\epsilon_0, \xi)$. The uniform entropy integral of $\mathcal{G}(\epsilon_0, \xi)$ w.r.t. the envelope function F_ξ and for the L_2 norm is upper bounded as follows:*

$$J_{F_\xi}(\beta, \mathcal{G}(\epsilon_0, \xi), L_2) = O\left(T\beta\sqrt{\log(1/\beta)}\right).$$

Proof. For every probability distribution Λ on $\mathcal{H}$, $\|F_\xi\|_{\Lambda,2} = L\xi$. From lemma 13, we have that

$$N(\alpha \|F_\xi\|_{2,\Lambda}, L_2(\Lambda), \mathcal{G}(\epsilon_0, \xi)) \leq (2/\alpha)^T.$$

Therefore,

$$J_{F_\xi}(\beta, \mathcal{G}(\epsilon_0, \xi), L_2) \leq \int_0^\beta \sqrt{\log(1 + (2/\alpha)^T)} d\alpha = O\left(T\beta\sqrt{\log(1/\beta)}\right),$$

where the second equality above follows from an integration by parts. □

Finally, we prove the lemma 11. The proof relies on a classical result in empirical process theory. We first introduce the relevant definitions and the relevant result before stating the proof of our lemma.

Definition 10 (Empirical process and empirical process notation)**.** *Consider $\mathcal{X}, \Sigma, P')$ a probability space and let $X_1, ..., X_n$ be n i.i.d. draws from P'. Let $\mathcal{F}$ be a class of functions $\mathcal{X} \to \mathbb{R}$. For all $f \in \mathcal{F}$, define the so-called "empirical process notation"*

$$P'f := \int f(h) dP'(h).$$

Denote $P_n := n^{-1} \sum_{i=1}^n \delta_{X_i}$ the empirical probability distribution associated to the sample $X_1, ..., X_n$. Observe that using the empirical process notation defined above, we have that $P_n f = n^{-1} \sum_{i=1}^n f(X_i)$. The stochastic process

$$\{(P_n - P')f : f \in \mathcal{F}\}$$

is termed the empirical process associated to P' and n indexed by $\mathcal{F}$.

Lemma 15 (Pollard's maximal inequality, vdV-Wellner 1996 2.14.1)**.** *Consider $(\mathcal{X}, \Sigma, P')$ a probability space and let $X_1, ..., X_n$ be n i.i.d. draws from P'. Let $\mathcal{F}$ be a class of functions $\mathcal{X} \to \mathbb{R}$. Let $\mathcal{F}$ be a class of functions $\mathcal{X} \to \mathbb{R}$ with envelope function F. Then*

$$E_{P'}[\sup_{f \in \mathcal{F}} \sqrt{n}|(P_n - P')f|] \lesssim J_F(1, \mathcal{F}, L_2) \|F\|_{L_2(P')}.$$

Proof of lemma 11. Recasting the claim of lemma 8 in terms of empirical process notations, we want to show that

$$\sqrt{n}(P_n - P^{\pi_b})(g_{\epsilon_n}(Q', \rho') - g_{\epsilon_\infty}(Q', \rho')) = o_P(1).$$

Let $\kappa > 0$, $\gamma \in (0, 1/2)$. Define, for all $\xi > 0$, the following two events:

$$\mathcal{E}_1(\xi) := \{\|\epsilon_n - \epsilon_\infty\|_\infty \leq \xi\}$$
$$\mathcal{E}_2(\xi) := \left\{ \sup_{\epsilon : \|\epsilon - \epsilon_\infty\|_\infty \leq \xi} \sqrt{n}|(P_n - P^{\pi_b})(g_\epsilon(Q', \rho') - g_{\epsilon_\infty}(Q', \rho'))| \leq \kappa \right\}.$$

The function $F_\xi : h \mapsto \xi L$ is an envelope function for $\mathcal{G}(\epsilon_0, \xi)$. By Markov's inequality and lemma 15 applied with the uniform entropy integral bound given in lemma 14, we have that

$$\begin{aligned} 1 - P^{\pi_b}[\mathcal{E}_2(\xi)] =& P^{\pi_b}\left[\sqrt{n}|(P_n - P^{\pi_b})(g_{\epsilon_n}(Q', \rho') - g_{\epsilon_\infty}(Q', \rho'))| \geq \kappa\right] \\ \leq& \kappa^{-1} E_{P^{\pi_b}}\left[\sqrt{n}|(P_n - P^{\pi_b})(g_{\epsilon_n}(Q', \rho') - g_{\epsilon_\infty}(Q', \rho'))|\right] \\ \leq& \kappa^{-1} J_F(1, \mathcal{G}(\epsilon_0, \xi), L_2) \|F_\xi\|_{2,\Lambda} \\ \leq& K \kappa^{-1} \xi L, \end{aligned}$$

for some constant K. Set $\xi = \kappa\gamma/(2KL)$. Then, from the above display, $P^{\pi_b}[\mathcal{E}_2(\kappa\gamma/(2KL))] \geq 1 - \gamma/2$.

Since $\epsilon_n \xrightarrow{P} \epsilon_\infty$, there exists n_0 such that for all $n \geq n_0$, $P^{\pi_b}[\mathcal{E}_1(\kappa\gamma/(2KL))] \geq 1 - \gamma/2$. Observe that if $\mathcal{E}_1(\kappa\gamma/(2KL)) \cap \mathcal{E}_2(\kappa\gamma/(2KL))$ is realized, then

$$\sqrt{n}|(P_n - P^{\pi_b})(g_{\epsilon_n}(Q', \rho') - g_{\epsilon_\infty}(Q', \rho'))| \leq \kappa.$$

Using a union bound, we have that, for all $n \geq n_0$,

$$\begin{aligned}
&P^{\pi_b}\left[\sqrt{n}|(P_n - P^{\pi_b})(g_{\epsilon_n}(Q', \rho') - g_{\epsilon_\infty}(Q', \rho'))| \leq \kappa\right] \\
&\geq 1 - (1 - P^{\pi_b}[\mathcal{E}_1(\kappa\gamma/(2KL))]) - (1 - P^{\pi_b}[\mathcal{E}_2(\kappa\gamma/(2KL))]) \\
&\geq 1 - \gamma.
\end{aligned}$$

Recapitulating the above, we have proven that for all $\kappa > 0$, $\gamma \in (0, 1/2)$, there exists n_0 such that for all $n \geq n_0$,

$$P^{\pi_b}\left[\sqrt{n}|(P_n - P^{\pi_b})(g_{\epsilon_n}(Q', \rho') - g_{\epsilon_\infty}(Q', \rho'))| \leq \kappa\right] \geq 1 - \gamma.$$

In other words, we have thus proven that

$$\sqrt{n}|(P_n - P^{\pi_b})(g_{\epsilon_n}(Q', \rho') - g_{\epsilon_\infty}(Q', \rho'))| = o_P(1).$$

which concludes the proof. □

Experiment Details

In this subsection, we provide full details of our experiments and utilized domains. In particular, we provide detailed descriptions of discrete-state domains ModelWin, ModelFail and Gridworld.

ModelWin

The ModelWin environment was constructed in order to simulate situations in which the approximate model of the MDP will converge quickly to the truth. On the other hand, importance-sampling based methods might suffer from high variance.

The ModelWin MDP consists of 3 states, and the agent always begins at state s_1. At s_1, the agent stochastically picks between two actions, a_1 and a_2. Under action a_1, the agent transitions to s_2 with probability 0.4 and s_3 with probability 0.6. On the other hand, under action a_2 the agent does the opposite- it transitions to s_2 and s_3 with probability 0.6 and 0.4, respectively. Under both actions, if the agent transitions to s_2, it gets a positive reward of +1. Consequently s_1 to s_3 transitions are penalized with -1 reward. In states s_3 and s_2, both actions a_1 and a_2 will take the agent back to s_1 with probability 1 and no reward. The horizon is set to $T = 20$.

The considered behavior policy takes action a_1 from s_1 with probability 0.73, and action a_2 with probability 0.27. The evaluation policy has the opposite behavior. Note that both the behavior and evaluation policies select actions uniformly at random while in states s_1 and s_2.

ModelFail

Unlike the ModelWin domain, the agent does not observe the true underlying states of the MDP in ModelFail. The purpose of this domain is to test environments are not known perfectly, and where the approximate model will fail to converge to the true MDP. ModelFail attempts to mimic partial observability, common in real applications.

The actual MDP consists of 4 states, 3 states and a final absorbing state, however the agent is not able to distinguish between them. The agent always starts at the same state, s_1, where it has two actions available. With actions a_1 it transitions into the upper state (s_2), whereas with action a_2 it goes to the lower state (s_3). No matter which state the agent transitioned to, both s_2 and s_3 lead to the terminal absorbing state s_4. However, s_2 to s_4 transition carries reward +1, whereas s_3 to s_4 leads to reward of -1. The horizon is $T = 2$.

The considered behavior policy takes action a_1 with probability 0.88, and action a_2 with probability 0.22. The evaluation policy has the opposite behavior.

Gridworld

The last discrete-state environment used is a 4×4 gridworld domain with 4 actions (up, down, left, right) developed by [122]. As emphasized by [123], this is a domain specifically developed for evaluation of OPE estimators. However, due to its deterministic nature, it will favor model-based approaches.

The horizon for GridWorld is $T = 100$, after which the episode ends unless the terminal state of s_{12} is reached before T. The reward is always -1, expect at states s_8 where it is +1, s_{12} with +10, and s_6 where the agent is penalized with -10 reward.

We used two different polices for the gridworld, as described in [122]. In particular, policy π_1 selects each of the 4 actions with equal probability regardless of the observation. Intuitively this policy takes a long time to reach the goal, and potentially often visits the state with the maximum negative reward. In addition, we also considered the near-optimal+ policy π_5, which exemplifies a near-deterministic near-optimal policy that moves quickly to s_8 with reward +1, without visiting s_6 with -10 reward. At s_8 it chooses action down with high probability, collecting as many positive rewards as possible until the time limit runs out. Once it eventually chooses the right action, it moves almost deterministically to s_{12} where it collects its final reward and end the episode.

Bibliography

[1] O. Alagoz et al. "Markov decision processes: a tool for sequential decision making under uncertainty". In: *Med Decis Making* 30.4 (2010), pp. 474–483.

[2] N. Altieri et al. "Curating a COVID-19 data repository and forecasting county-level death counts in the United States". In: *arXiv preprint arXiv:2005.07882* (2020).

[3] O. Anava et al. "Online Learning for Time Series Prediction". In: *Proceedings of the 26th Annual Conference on Learning Theory*. Ed. by Shai Shalev-Shwartz and Ingo Steinwart. Vol. 30. Proceedings of Machine Learning Research. Princeton, NJ, USA: PMLR, Dec. 2013, pp. 172–184.

[4] R. Bahl et al. "Modeling COVID-19 spread in small colleges". In: *PLoS One* 16.8 (2021), e0255654.

[5] H. Bang and J.M. Robins. "Doubly robust estimation in missing data and causal inference models". In: *Biometrics* 61.4 (Dec. 2005), pp. 962–973.

[6] D. Benkeser and M.J. van der Laan. "The Highly Adaptive Lasso Estimator". In: *Proc Int Conf Data Sci Adv Anal* 2016 (2016), pp. 689–696.

[7] D. Benkeser et al. "Online cross-validation-based ensemble learning". In: *Statistics in Medicine* 37.2 (2018), pp. 249–260. ISSN: 1097-0258. DOI: 10.1002/sim.7320.

[8] C. Bergmeir and J.M. Benítez. "On the use of cross-validation for time series predictor evaluation". In: *Information Sciences* 191 (2012). Data Mining for Software Trustworthiness, pp. 192–213. ISSN: 0020-0255. DOI: https://doi.org/10.1016/j.ins.2011.12.028.

[9] A. Bibaut et al. *Sequential causal inference in a single world of connected units*. 2021. arXiv: 2101.07380 [math.ST].

[10] A.F. Bibaut and M.J. van der Laan. *Fast rates for empirical risk minimization over càdlàg functions with bounded sectional variation norm*. 2019. arXiv: 1907.09244 [math.ST].

[11] A. Biswas et al. *COVID-19: Strategies for Allocation of Test Kits*. 2020. arXiv: 2004.01740 [cs.AI].

[12] I. Bojinov and N. Shephard. "Time Series Experiments and Causal Estimands: Exact Randomization Tests and Trading". In: *Journal of the American Statistical Association* 0.0 (2019), pp. 1–36. DOI: 10.1080/01621459.2018.1527225.

[13] N. Bolger and J.P. Laurenceau. *Intensive Longitudinal Methods: An Introduction to Diary and Experience Sampling Research*. Methodology in the social sciences. Guilford Publications, 2013. ISBN: 9781462506781.

[14] A. Boruvka et al. "Assessing Time-Varying Causal Effect Moderation in Mobile Health". In: *Journal of the American Statistical Association* 113.523 (2018), pp. 1112–1121. DOI: 10.1080/01621459.2017.1305274.

[15] E.H. Bradley, M.W. An, and E. Fox. "Reopening Colleges During the Coronavirus Disease 2019 (COVID-19) Pandemic-One Size Does Not Fit All". In: *JAMA Netw Open* 3.7 (July 2020), e2017838.

[16] R.C. Bradley. "Basic Properties of Strong Mixing Conditions. A Survey and Some Open Questions". In: *Probab. Surveys* 2 (2005), pp. 107–144.

[17] B.M. Brown. "Martingale Central Limit Theorems". In: *Ann. Math. Statist.* 42.1 (Feb. 1971), pp. 59–66. DOI: 10.1214/aoms/1177693494.

[18] B. Chakraborty and E. Moodie. *Statistical Methods for Dynamic Treatment Regimes*. Springer Publishing Company, Incorporated, 2013. ISBN: 978-1-4614-7427-2.

[19] A. Chambaz, W. Zheng, and M.J. van der Laan. "Targeted sequential design for targeted learning inference of the optimal treatment rule and its mean reward". In: *Ann Stat* 45.6 (2017), pp. 2537–2564.

[20] B. Chan et al. "Generalizable deep temporal models for predicting episodes of sudden hypotension in critically ill patients: a personalized approach". In: *Sci Rep* 10.1 (July 2020), p. 11480.

[21] J.T. Chang, F.W. Crawford, and E.H. Kaplan. "Repeat SARS-CoV-2 testing models for residential college populations". In: *Health Care Manag Sci* 24.2 (June 2021), pp. 305–318.

[22] M. Chatzimanolakis et al. "Optimal allocation of limited test resources for the quantification of COVID-19 infections". In: *Swiss Med Wkly* 150 (Dec. 2020), w20445.

[23] T. Chen and C. Guestrin. "XGBoost: A Scalable Tree Boosting System". In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. San Francisco, California, USA: ACM, 2016, pp. 785–794. ISBN: 978-1-4503-4232-2. DOI: 10.1145/2939672.2939785.

[24] H.A. Chipman, E.I. George, and R.E. McCulloch. "BART: Bayesian additive regression trees". In: ().

[25] J.R. Coyle and N.S. Hejazi. "origami: A Generalized Framework for Cross-Validation in R". In: *The Journal of Open Source Software* 3.21 (Jan. 2018). DOI: 10.21105/joss.00512. URL: https://doi.org/10.21105/joss.00512.

[26] J.R. Coyle et al. *origami: A Generalized Framework for Cross-Validation in R*. R package version 1.0.5. 2021. DOI: 10.5281/zenodo.1155901. URL: https://zenodo.org/record/1155901.

[27] J.R. Coyle et al. *sl3: Modern Pipelines for Machine Learning and Super Learning*. R package version 1.4.2. 2021. DOI: 10.5281/zenodo.1342293. URL: https://doi.org/10.5281/zenodo.1342293.

[28] Jeremy R Coyle et al. *hal9001: The scalable highly adaptive lasso*. R package version 0.4.2. 2022. DOI: 10.5281/zenodo.3558313. URL: https://doi.org/10.5281/zenodo.3558313.

[29] W. Dempsey et al. "Randomised trials for the Fitbit generation". In: *Signif (Oxf)* 12.6 (Dec. 2015), pp. 20–23.

[30] I. Diaz and M.J. van Der Laan. "Population intervention causal effects based on stochastic interventions". In: *Biometrics* 68.2 (June 2012), pp. 541–549.

[31] J. Du et al. "Optimal diagnostic test allocation strategy during the COVID-19 pandemic and beyond". In: *Statistics in Medicine* n/a.n/a (2021). DOI: https://doi.org/10.1002/sim.9238.

[32] D. Dua and C. Graff. *UCI Machine Learning Repository*. 2017. URL: http://archive.ics.uci.edu/ml.

[33] M. Dudik, J. Langford, and L. Li. "Doubly Robust Policy Evaluation and Learning". In: *Proceedings of the 28th International Conference on International Conference on Machine Learning*. ICML'11. Bellevue, Washington, USA: Omnipress, 2011, pp. 1097–1104. ISBN: 978-1-4503-0619-5. URL: http://dl.acm.org/citation.cfm?id=3104482.3104620.

[34] S. Dudoit and M.J. van der Laan. "Asymptotics of cross-validated risk estimation in estimator selection and performance assessment". In: *Statistical Methodology* 2.2 (2005), pp. 131–154. ISSN: 1572-3127. DOI: https://doi.org/10.1016/j.stamet.2005.02.003.

[35] P.L. Dulin, V.M. Gonzalez, and K. Campbell. "Results of a pilot test of a self-administered smartphone-based treatment system for alcohol use disorders: usability and early outcomes". In: *Subst Abus* 35.2 (2014), pp. 168–175.

[36] G. Ecoto, A. Bibaut, and A. Chambaz. "One-step ahead sequential Super Learning from short times series of many slightly dependent data, and anticipating the cost of natural disasters". In: *arXiv preprint arXiv:2107.13291* (2021).

[37] E. Ertin et al. "AutoSense: unobtrusively wearable sensor suite for inferring the onset, causality, and consequences of stress in the field". In: *SenSys*. 2011.

[38] A. fabio di Narzo et al. *tsDyn: Nonlinear Time Series Models with Regime Switching*. R package version 10-1.2. 2020. URL: https://CRAN.R-project.org/package=tsDyn.

[39] M. Farajtabar, Y. Chow, and M. Ghavamzadeh. "More Robust Doubly Robust Off-policy Evaluation". In: *CoRR* abs/1802.03493 (2018). arXiv: 1802.03493. URL: http://arxiv.org/abs/1802.03493.

[40] C. Free et al. "The effectiveness of mobile-health technology-based health behaviour change or disease management interventions for health care consumers: a systematic review". In: *PLoS Med.* 10.1 (2013), e1001362.

[41] J. Friedman, T. Hastie, and R. Tibshirani. "Regularization Paths for Generalized Linear Models via Coordinate Descent". In: *Journal of Statistical Software* 33.1 (2010), pp. 1–22. URL: https://www.jstatsoft.org/v33/i01/.

[42] B.J. Gardner and A.M. Kilpatrick. "Contact tracing efficiency, transmission heterogeneity, and accelerating COVID-19 epidemics". In: *PLoS Comput Biol* 17.6 (June 2021), e1009122.

[43] A. Gelman and Y.S. Su. *arm: Data Analysis Using Regression and Multilevel Hierarchical Models*. R package version 1.11-2. 2020. URL: https://CRAN.R-project.org/package=arm.

[44] N. Ghaffarzadegan. "Simulation-based what-if analysis for controlling the spread of Covid-19 in universities". In: *PLoS One* 16.2 (2021), e0246323.

[45] R.D. Gill, M.J. van der Laan, and J.A. Wellner. "Inefficient estimators of the bivariate survival function for three models." In: *Annales de l'Institut Henri Poincare* 31 (1995), pp. 545–597.

[46] M. Gilliland. "The value added by machine learning approaches in forecasting". In: *International Journal of Forecasting* 36.1 (2020). M4 Competition, pp. 161–166. ISSN: 0169-2070. DOI: https://doi.org/10.1016/j.ijforecast.2019.04.016.

[47] A.L. Goldberger et al. "PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals". In: *circulation* 101.23 (2000), e215–e220.

[48] G.S. Gonsalves et al. "Maximizing the Efficiency of Active Case Finding for SARS-CoV-2 Using Bandit Algorithms". In: *Medical Decision Making* 41.8 (2021). PMID: 34120510, pp. 970–977. DOI: 10.1177/0272989X211021603. eprint: https://doi.org/10.1177/0272989X211021603.

[49] P.T. Gressman and J.R. Peck. "Simulating COVID-19 in a university environment". In: *Math Biosci* 328 (Oct. 2020), p. 108436.

[50] A. Guh et al. "Transmission of 2009 pandemic influenza A (H1N1) at a Public University–Delaware, April-May 2009". In: *Clin Infect Dis* 52 Suppl 1 (Jan. 2011), S131–137.

[51] E.L. Hamaker et al. "At the Frontiers of Modeling Intensive Longitudinal Data: Dynamic Structural Equation Models for the Affective Measurements from the COGITO Study". In: *Multivariate Behavioral Research* 53.6 (2018). PMID: 29624092, pp. 820–841. DOI: 10.1080/00273171.2018.1446819.

[52] H.L. Hambridge, R. Kahn, and J.P. Onnela. "Examining SARS-CoV-2 Interventions in Residental Colleges Using an Emprical Network". In: *medRxiv* (Apr. 2021).

[53] D.H. Hamer et al. "Assessment of a COVID-19 Control Plan on an Urban University Campus During a Second Wave of the Pandemic". In: *JAMA Netw Open* 4.6 (June 2021), e2116425.

[54] R. van Handel. "On the minimal penalty for Markov order estimation". In: *Probability Theory and Related Fields* 150.3-4 (Apr. 2010), pp. 709–738. ISSN: 1432-2064. DOI: 10.1007/s00440-010-0290-y.

[55] K.E. Heron and J.M. Smyth. "Ecological momentary interventions: incorporating mobile technology into psychosocial and health behaviour treatments". In: *Br J Health Psychol* 15.Pt 1 (Feb. 2010), pp. 1–39.

[56] M. Hibon and T. Evgeniou. "To combine or not to combine: selecting among forecasts and their combinations". In: *International journal of forecasting* 21.1 (2005), pp. 15–24.

[57] E.M. Hill et al. "Modelling SARS-CoV-2 transmission in a UK university setting". In: *Epidemics* 36 (June 2021), p. 100476.

[58] S.C.H Hoi et al. "Online Learning: A Comprehensive Survey". In: *arXiv* (2018). eprint: 1802.02871 (cs.LG).

[59] W. Hoiles and M. Van Der Schaar. "Bounded Off-policy Evaluation with Missing Data for Course Recommendation and Curriculum Design". In: *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*. ICML'16. New York, NY, USA: JMLR.org, 2016, pp. 1596–1604. URL: http://dl.acm.org/citation.cfm?id=3045390.3045559.

[60] R. Hyndman et al. *forecast: Forecasting functions for time series and linear models*. R package version 8.15. 2021. URL: https://pkg.robjhyndman.com/forecast/.

[61] R. Istepanian and T. Al-Anzi. "m-Health 2.0: New perspectives on mobile health, machine learning and big data analytics". In: *Methods* 151 (Dec. 2018), pp. 34–40.

[62] R. Istepanian and B. Woodward. *M-Health: Fundamentals and Applications: Fundamentals and Applications*. John Wiley-IEEE, 2017. ISBN: 9781118496985.

[63] N. Jiang and L. Li. "Doubly Robust Off-policy Value Evaluation for Reinforcement Learning". In: *arXiv e-prints*, arXiv:1511.03722 (Nov. 2015), arXiv:1511.03722. arXiv: 1511.03722 [cs.LG].

[64] J. Jonnerby et al. *Maximising the Benefits of an Acutely Limited Number of COVID-19 Tests*. 2020. arXiv: 2004.13650 [q-bio.PE].

[65] Nathan Kallus and Masatoshi Uehara. *Efficiently Breaking the Curse of Horizon in Off-Policy Evaluation with Double Reinforcement Learning*. 2019. arXiv: 1909.05850 [stat.ML].

[66] Y. Kang, R.J. Hyndman, and F. Li. "GRATIS: GeneRAting TIme Series with diverse and controllable characteristics". In: *Statistical Analysis and Data Mining: The ASA Data Science Journal* (May 2020). ISSN: 1932-1872. DOI: 10.1002/sam.11461. URL: http://dx.doi.org/10.1002/sam.11461.

[67] P. Klasnja et al. "Efficacy of Contextually Tailored Suggestions for Physical Activity: A Micro-randomized Optimization Trial of HeartSteps". In: *Ann Behav Med* 53.6 (May 2019), pp. 573–582.

[68] P. Klasnja et al. "Microrandomized trials: An experimental design for developing just-in-time adaptive interventions". In: *Health Psychol* 34S (Dec. 2015), pp. 1220–1228.

[69] S. Kumar et al. "Mobile health technology evaluation: the mHealth evidence workshop". In: *Am J Prev Med* 45.2 (Aug. 2013), pp. 228–236.

[70] D.B. Larremore et al. "Test sensitivity is secondary to frequency and turnaround time for COVID-19 screening". In: *Science Advances* 7.1 (2021), eabd5393. DOI: 10.1126/sciadv.abd5393. eprint: https://www.science.org/doi/pdf/10.1126/sciadv.abd5393. URL: https://www.science.org/doi/abs/10.1126/sciadv.abd5393.

[71] N.D. Le, R.D. Martin, and A.E. Raftery. "Modeling Flat Stretches, Bursts Outliers in Time Series Using Mixture Transition Distribution Models". In: *Journal of the American Statistical Association* 91.436 (1996), pp. 1504–1515. DOI: 10.1080/01621459.1996.10476718. eprint: https://doi.org/10.1080/01621459.1996.10476718.

[72] C.S. Lee and A.Y. Lee. "Clinical applications of continual learning machine learning". In: *The Lancet Digital Health* 2.6 (June 2020), e279–e281.

[73] B. Lopman et al. "A modeling study to inform screening and testing interventions for the control of SARS-CoV-2 on university campuses". In: *Sci Rep* 11.1 (Mar. 2021), p. 5900.

[74] D.J. Luckett et al. "Estimating Dynamic Treatment Regimes in Mobile Health Using V-Learning". In: *Journal of the American Statistical Association* 0.0 (2019), pp. 1–34. DOI: 10.1080/01621459.2018.1537919.

[75] A.R. Luedtke and M.J. van der Laan. "Optimal Individualized Treatments in Resource-Limited Settings". In: *The International Journal of Biostatistics* 12.1 (2016), pp. 283–303. DOI: 10.1515/ijb-2015-0007.

[76] A.R. Luedtke and M.J. van der Laan. "Statistical inference for the mean outcome under a possibly non-unique optimal treatment strategy". In: *Ann Stat* 44.2 (Apr. 2016), pp. 713–742.

[77] A.R. Luedtke and M.J. van der Laan. "Super-Learning of an Optimal Dynamic Treatment Rule". In: *The International Journal of Biostatistics* 12.1 (2016), pp. 305–332. DOI: 10.1515/ijb-2015-0052.

[78] I. Malenica, A. Bibaut, and M.J. van der Laan. *Adaptive Sequential Design for a Single Time-Series*. 2021. arXiv: 2102.00102 [math.ST].

[79] I. Malenica et al. *Personalized Online Machine Learning*. 2021. arXiv: 2109.10452 [stat.ML].

[80] D. Malvey and D.J. Slovensky. *mHealth: Transforming Healthcare*. Springer Publishing Company, Incorporated, 2014. ISBN: 1489974563, 9781489974563.

[81] S. Mannor et al. "Bias and Variance Approximation in Value Function Estimates". In: *Management Science* 53.2 (2007), pp. 308–322. DOI: 10.1287/mnsc.1060.0614. eprint: https://doi.org/10.1287/mnsc.1060.0614. URL: https://doi.org/10.1287/mnsc.1060.0614.

[82] N. Martin, R.T. Schooley, and V. De Gruttola. "Modelling testing frequencies required for early detection of a SARS-CoV-2 outbreak on a university campus". In: *medRxiv* (June 2020).

[83] N.J. Matheson et al. "Mass testing of university students for covid-19". In: *BMJ* 375 (Oct. 2021), n2388.

[84] K.E. Muessig et al. "Mobile phone applications for the care and prevention of HIV and other sexually transmitted diseases: a review". In: *J. Med. Internet Res.* 15.1 (Jan. 2013), e1.

[85] G. Muhammad et al. "A Facial-Expression Monitoring System for Improved Healthcare in Smart Cities". In: *IEEE Access* 5 (2017), pp. 10871–10881. ISSN: 2169-3536. DOI: 10.1109/ACCESS.2017.2712788.

[86] K. Muller and P. Muller. "Mathematical modelling of the spread of COVID-19 on a university campus". In: *Infect Dis Model* (Aug. 2021).

[87] S.A. Murphy. "Optimal dynamic treatment regimes". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 65.2 (2003), pp. 331–355. DOI: 10.1111/1467-9868.00389.

[88] S.A. Murphy, M.J. van der Laan, and J.M. Robins. "Marginal Mean Models for Dynamic Regimes". In: *J Am Stat Assoc* 96.456 (Dec. 2001), pp. 1410–1423.

[89] D.M. Nathan. "The diabetes control and complications trial/epidemiology of diabetes interventions and complications study at 30 years: overview". In: *Diabetes Care* 37.1 (2014), pp. 9–16.

[90] World Health Organization. *Laboratory testing strategy recommendations for COVID-19: interim guidance, 21 March 2020*. Technical documents. 2020, 6 p.

[91] A.D. Paltiel, A. Zheng, and R.P. Walensky. "COVID-19 screening strategies that permit the safe re-opening of college campuses". In: *medRxiv* (July 2020).

[92] M. Pawlikowski and A. Chorowska. "Weighted ensemble of statistical models". In: *International Journal of Forecasting* 36.1 (2020). M4 Competition, pp. 93–97. ISSN: 0169-2070. DOI: https://doi.org/10.1016/j.ijforecast.2019.03.019.

[93] J. Pearl. *Causality: Models, Reasoning and Inference*. 2nd. New York, NY, USA: Cambridge University Press, 2009. ISBN: 052189560X, 9780521895606.

[94] T.A. Perkins et al. "Estimating unobserved SARS-CoV-2 infections in the United States". In: *Proceedings of the National Academy of Sciences* 117.36 (2020), pp. 22597–22602. ISSN: 0027-8424. DOI: 10.1073/pnas.2005476117. eprint: https://www.pnas.org/content/117/36/22597.full.pdf. URL: https://www.pnas.org/content/117/36/22597.

[95] M. Petersen et al. "Targeted Maximum Likelihood Estimation for Dynamic and Static Longitudinal Marginal Structural Working Models". In: *Journal of Causal Inference* 2.2 (2014). PMCID: PMC4405134, pp. 147–185.

[96] S.F. Poole et al. "A holistic approach for suppression of COVID-19 spread in workplaces and universities". In: *PLoS One* 16.8 (2021), e0254798.

[97] D. Precup. "Temporal abstraction in reinforcement learning". PhD thesis. University of Massachusetts Amherst, 2000.

[98] D. Precup, Richard S. Sutton, and S.P. Singh. "Eligibility Traces for Off-Policy Policy Evaluation". In: *Proceedings of the Seventeenth International Conference on Machine Learning*. ICML '00. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2000, pp. 759–766. ISBN: 1-55860-707-2. URL: http://dl.acm.org/citation.cfm?id=645529.658134.

[99] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2020. URL: https://www.R-project.org/.

[100] M. Rabbi et al. "Toward Increasing Engagement in Substance Use Data Collection: Development of the Substance Abuse Research Assistant App and Protocol for a Microrandomized Trial Using Adolescents and Emerging Adults". In: *JMIR Res Protoc* 7.7 (July 2018), e166.

[101] A. Rakhlin, K. Sridharan, and A. Tewari. "Sequential complexities and uniform martingale laws of large numbers". In: *Probability Theory and Related Fields* 161 (2014), pp. 111–153.

[102] L. Rennert et al. "Surveillance-based informative testing for detection and containment of SARS-CoV-2 outbreaks on a public university campus: an observational and modelling study". In: *Lancet Child Adolesc Health* 5.6 (June 2021), pp. 428–436.

[103] J. Rissanen. "Order estimation by accumulated prediction errors". In: *Journal of Applied Probability* 23.A (1986), pp. 55–61. DOI: 10.2307/3214342.

[104] J. Robins. "A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect". In: *Mathematical Modelling* 7.9 (1986), pp. 1393–1512. ISSN: 0270-0255. DOI: https://doi.org/10.1016/0270-0255(86)90088-6. URL: https://www.sciencedirect.com/science/article/pii/0270025586900886.

[105] J.M. Robins. "Optimal Structural Nested Models for Optimal Sequential Decisions". In: *Proceedings of the Second Seattle Symposium in Biostatistics: Analysis of Correlated Data*. New York, NY: Springer New York, 2004, pp. 189–326. ISBN: 978-1-4419-9076-1. DOI: 10.1007/978-1-4419-9076-1_11.

[106] J.M. Robins, S. Greenland, and F.C. Hu. "Estimation of the Causal Effect of a Time-Varying Exposure on the Marginal Mean of a Repeated Binary Outcome". In: *Journal of the American Statistical Association* 94.447 (1999), pp. 687–700. DOI: 10.1080/01621459.1999.10474168.

[107] J.M. Robins, M. A. Hernan, and B. Brumback. "Marginal structural models and causal inference in epidemiology". In: *Epidemiology* 11.5 (Sept. 2000), pp. 550–560.

[108] J.M. Robins and A. Rotnitzky. "Semiparametric Efficiency in Multivariate Regression Models with Missing Data". In: *Journal of the American Statistical Association* 90.429 (1995), pp. 122–129. ISSN: 01621459. URL: http://www.jstor.org/stable/2291135.

[109] J.M. Robins, A. Rotnitzky, and L.P. Zhao. "Estimation of Regression Coefficients When Some Regressors Are Not Always Observed". In: *Journal of the American Statistical Association* 89.427 (1994), pp. 846–866. ISSN: 01621459. URL: http://www.jstor.org/stable/2290910.

[110] P.R. Rosenbaum and D.B. Rubin. "The central role of the propensity score in observational studies for causal effects". In: *Biometrika* 70.1 (1983), pp. 41–55. DOI: 10.1093/biomet/70.1.41. eprint: /oup/backfile/content_public/journal/biomet/70/1/10.1093/biomet/70.1.41/2/70-1-41.pdf. URL: http://dx.doi.org/10.1093/biomet/70.1.41.

[111] M Saeed et al. "Multiparameter Intelligent Monitoring in Intensive Care II (MIMIC-II): a public-access intensive care unit database". In: *Critical care medicine* 39.5 (2011), p. 952.

[112] Alejandro Schuler and Mark van der Laan. *The Selectively Adaptive Lasso*. 2022. DOI: 10.48550/ARXIV.2205.10697. URL: https://arxiv.org/abs/2205.10697.

[113] O. Schultes et al. "COVID-19 Testing and Case Rates and Social Contact Among Residential College Students in Connecticut During the 2020-2021 Academic Year". In: *JAMA Netw Open* 4.12 (Dec. 2021), e2140602.

[114] D. Shaub. "Fast and accurate yearly time series forecasting with forecast combinations". In: *International Journal of Forecasting* 36.1 (2020). M4 Competition, pp. 116–120. ISSN: 0169-2070. DOI: https://doi.org/10.1016/j.ijforecast.2019.03.032.

[115] S. Smyl. "A hybrid method of exponential smoothing and recurrent neural networks for time series forecasting". In: *International Journal of Forecasting* 36.1 (2020). M4 Competition, pp. 75–85. ISSN: 0169-2070. DOI: https://doi.org/10.1016/j.ijforecast.2019.03.017.

[116] O. Sofrygin and M.J. van der Laan. "Semi-Parametric Estimation and Inference for the Mean Outcome of the Single Time-Point Intervention in a Causally Connected Population". In: *J Causal Inference* 5.1 (Mar. 2017).

[117] S.R. Steinhubl, E.D. Muse, and E.J. Topol. "Can Mobile Health Technologies Transform Health Care?" In: *JAMA* 310.22 (Dec. 2013), pp. 2395–2396. ISSN: 0098-7484. DOI: 10.1001/jama.2013.281078.

[118] J.H Stock. "Nonparametric policy analysis". In: *Journal of the American Statistical Association* 84.406 (1989), pp. 567–575.

[119] A. Stone et al. *The Science of Real-Time Data Capture: Self-Reports in Health Research*. Oxford University Press, 2007. ISBN: 9780195178715.

[120] R.S. Sutton and A.G. Barto. *Introduction to Reinforcement Learning*. 1st. Cambridge, MA, USA: MIT Press, 1998. ISBN: 0262193981.

[121] G. Theocharous, P.S. Thomas, and M. Ghavamzadeh. "Personalized Ad Recommendation Systems for Life-time Value Optimization with Guarantees". In: *Proceedings of the 24th International Conference on Artificial Intelligence*. IJCAI'15. Buenos Aires, Argentina: AAAI Press, 2015, pp. 1806–1812. ISBN: 978-1-57735-738-4. URL: http://dl.acm.org/citation.cfm?id=2832415.2832500.

[122] P. Thomas. "Safe Reinforcement Learning". PhD thesis. University of Massachusetts Amherst, 2015.

[123] P. Thomas and E. Brunskill. "Data-Efficient Off-Policy Policy Evaluation for Reinforcement Learning". In: *Proceedings of The 33rd International Conference on Machine Learning*. Ed. by Maria Florina Balcan and Kilian Q. Weinberger. Vol. 48. Proceedings of Machine Learning Research. New York, New York, USA: PMLR, June 2016, pp. 2139–2148.

[124] J. Tuells et al. "Seroprevalence Study and Cross-Sectional Survey on COVID-19 for a Plan to Reopen the University of Alicante (Spain)". In: *Int J Environ Res Public Health* 18.4 (Feb. 2021).

[125] N.S. Tyler et al. "An artificial intelligence decision support system for the management of type 1 diabetes". In: *Nat Metab* 2.7 (July 2020), pp. 612–619.

[126] A. van der Vaart and J. Wellner. *Weak Convergence and Empirical Processes*. Springer-Verlag New York, Mar. 2013. ISBN: 9781475725452.

[127] A.W. van der Vaart. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1998.

[128] S. van de Geer. *Empirical Processes in M-Estimation*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2000.

[129] M.J van der Laan and S. Gruber. *Targeted Minimum Loss Based Estimation of an Intervention Specific Mean Outcome*. Tech. rep. U.C. Berkeley Division of Biostatistics Working Paper Series, 2011.

[130] M.J. van der Laan. "A Generally Efficient Targeted Minimum Loss Based Estimator based on the Highly Adaptive Lasso". In: *Int J Biostat* 13.2 (Oct. 2017).

[131] M.J. van der Laan, A. Chambaz, and S. Lendle. "Online Targeted Learning for Time Series". In: *Targeted Learning in Data Science: Causal Inference for Complex Longitudinal Studies*. Cham: Springer International Publishing, 2018, pp. 317–346. ISBN: 978-3-319-65304-4. DOI: 10.1007/978-3-319-65304-4_19.

[132] M.J. van der Laan and S. Dudoit. *Unified Cross-Validation Methodology For Selection Among Estimators and a General Cross-Validated Adaptive Epsilon-Net Estimator: Finite Sample Oracle Inequalities and Examples*. 2003. eprint: WorkingPaper130.

[133] M.J. van der Laan, S. Dudoit, and A.W. van der Vaart. "The cross-validated adaptive epsilon-net estimator". In: *Statistics & Risk Modeling* 24.3 (Dec. 2006), pp. 1–23.

[134] M.J. van der Laan and S. Gruber. *One-Step Targeted Minimum Loss-based Estimation Based on Universal Least Favorable One-Dimensional Submodels.* Tech. rep. Working Paper 347. U.C. Berkeley Division of Biostatistics Working Paper Series, Mar. 2016.

[135] M.J. van der Laan and S.D. Lendle. *Online Targeted Learning*. Tech. rep. Working Paper 330. U.C. Berkeley Division of Biostatistics Working Paper Series, Sept. 2014.

[136] M.J. van der Laan, E. Polley, and A. Hubbard. "Super learner". In: *Stat Appl Genet Mol Biol* 6 (2007), Article25.

[137] M.J. van der Laan and Sherri Rose. *Targeted Learning in Data Science: Causal Inference for Complex Longitudinal Studies*. Springer Science and Business Media, 2018.

[138] M.J. van der Laan and Sherri Rose. *Targeted Learning: Causal Inference for Observational and Experimental Data (Springer Series in Statistics)*. Springer, 2011.

[139] M.J. van Der Laan and D. Rubin. "Targeted Maximum Likelihood Learning". In: *The International Journal of Biostatistics* 2.1 (2006). DOI: doi:10.2202/1557-4679.1043. URL: https://doi.org/10.2202/1557-4679.1043.

[140] M.J. van der Laan and D. Rubin. *Targeted Maximum Likelihood Learning*. Tech. rep. Working Paper 213. U.C. Berkeley Division of Biostatistics Working Paper Series, Oct. 2006.

[141] A.W. van der Vaart, S. Dudoit, and M.J. van der Laan. "Oracle inequalities for multi-fold cross validation". In: *Statistics & Risk Modeling* 24.3 (Dec. 2006), pp. 1–21.

[142] N.A. Vander Schaaf et al. "Routine, Cost-Effective SARS-CoV-2 Surveillance Testing Using Pooled Saliva Limits Viral Spread on a Residential College Campus". In: *Microbiol Spectr* 9.2 (Oct. 2021), e0108921.

[143] H.T. Walke, M.A. Honein, and R.R. Redfield. "Preventing and Responding to COVID-19 on College Campuses". In: *JAMA* 324.17 (Nov. 2020), pp. 1727–1728. ISSN: 0098-7484. DOI: 10.1001/jama.2020.20027. URL: https://doi.org/10.1001/jama.2020.20027.

[144] Th. Walls and J. Schafer. *Models for intensive longitudinal data*. Methodology in the social sciences. Oxford University Press, 2006. ISBN: 9780195173444.

[145] K.A. Weeden and B. Cornwell. "The Small-World Network of College Classes: Implications for Epidemic Spread on a University Campus". In: *Sociological Science* 7.9 (2020), pp. 222–241. ISSN: 2330-6696. DOI: 10.15195/v7.a9. URL: http://dx.doi.org/10.15195/v7.a9.

[146] C.S. Wong and W.K. Li. "On a Mixture Autoregressive Model". In: *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* 62.1 (2000), pp. 95–115. ISSN: 13697412, 14679868. URL: http://www.jstor.org/stable/2680680.

[147] M.N. Wright and A. Ziegler. "ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R". In: *Journal of Statistical Software* 77.1 (2017), pp. 1–17. DOI: 10.18637/jss.v077.i01.

[148] J. Wu et al. "SARS-CoV-2 infection induces sustained humoral immune responses in convalescent patients following symptomatic COVID-19". In: *Nat Commun* 12.1 (Mar. 2021), p. 1813.

[149] M.W. Zhang et al. "The alcohol tracker application: an initial evaluation of user preferences". In: *BMJ Innov* 2.1 (Jan. 2016), pp. 8–13.