Mark J. van der Laan*

# Causal Inference for a Population of Causally Connected Units

**Abstract:** Suppose that we observe a population of causally connected units. On each unit at each time-point on a grid we observe a set of other units the unit is potentially connected with, and a unit-specific longitudinal data structure consisting of baseline and time-dependent covariates, a time-dependent treatment, and a final outcome of interest. The target quantity of interest is defined as the mean outcome for this group of units if the exposures of the units would be probabilistically assigned according to a known specified mechanism, where the latter is called a stochastic intervention. Causal effects of interest are defined as contrasts of the mean of the unit-specific outcomes under different stochastic interventions one wishes to evaluate. This covers a large range of estimation problems from independent units, independent clusters of units, and a single cluster of units in which each unit has a limited number of connections to other units. The allowed dependence includes treatment allocation in response to data on multiple units and so called causal interference as special cases. We present a few motivating classes of examples, propose a structural causal model, define the desired causal quantities, address the identification of these quantities from the observed data, and define maximum likelihood based estimators based on cross-validation. In particular, we present maximum likelihood based super-learning for this network data. Nonetheless, such smoothed/regularized maximum likelihood estimators are not targeted and will thereby be overly bias w.r.t. the target parameter, and, as a consequence, generally not result in asymptotically normally distributed estimators of the statistical target parameter.

To formally develop estimation theory, we focus on the simpler case in which the longitudinal data structure is a point-treatment data structure. We formulate a novel targeted maximum likelihood estimator of this estimand and show that the double robustness of the efficient influence curve implies that the bias of the targeted minimum loss-based estimation (TMLE) will be a second-order term involving squared differences of two nuisance parameters. In particular, the TMLE will be consistent if either one of these nuisance parameters is consistently estimated. Due to the causal dependencies between units, the data set may correspond with the realization of a single experiment, so that establishing a (e.g. normal) limit distribution for the targeted maximum likelihood estimators, and corresponding statistical inference, is a challenging topic. We prove two formal theorems establishing the asymptotic normality using advances in weak-convergence theory. We conclude with a discussion and refer to an accompanying technical report for extensions to general longitudinal data structures.

**Keywords:** networks; causal inference; targeted maximum likelihood estimation; stochastic intervention; efficient influence curve

***Corresponding author: Mark J. van der Laan,** University of California – Berkeley, Berkeley, CA, USA, E-mail: laan@berkeley.edu

## 1 Introduction and motivation

Most of the literature on causal inference has focussed on assessing the causal effect of a single or multiple time-point intervention on some outcome based on observing $n$ longitudinal data structures on $n$

independent units that are not causally connected. For literature reviews, we refer to a number of books on this topic: Rubin [1], Pearl [2], van der Laan and Robins [3], Tsiatis [4], Hernán and Robins [5], and van der Laan and Rose [6].

Such a causal effect is defined as an expectation of the effect of the intervention assigned to the unit on the unit's outcome, and causal effects of the intervention on other units on the unit's outcome are assumed non-existent. As a consequence, causal models only have to be concerned about the modeling of causal relations between the components of the unit-specific data structure. Statistical inference is based on the assumption that the $n$ data structures can be viewed as $n$ independent realizations of a random variable, so that central limit theorems (CLTs) for sums of independent random variables can be employed. The latter requires that the sample size $n$ is large enough so that statistical inference based on the normal limit distributions is indeed appropriate.

In many applications, one may define the unit as a group of causally connected individuals, often called a community or cluster. It is then assumed that the communities are not causally connected, and that the community-specific data structures can be represented as $n$ independent random variables. One can then define a community-specific outcome and assess the causal effect of the community level intervention/exposure on this community-specific outcome with methods from the causal inference literature. Such causal effects incorporate the total effect of community level intervention, where the effect of the community level exposure on an individual in a community also occurs through other individuals in that same community.

We refer to Halloran and Struchiner [7], Hudgens and Halloran [8], VanderWeele et al. [9], and Tchetgen Tchetgen and VanderWeele [10] for defining different types of causal effects in the presence of causal interference between units. Lacking a general methodological framework, many practical studies assume away interference for the sake of simplicity. The risk of this assumption is practically demonstrated in Sobel [11], who shows that ignoring interference can lead to completely wrong conclusions about the effectiveness of the program. We also refer to Donner and Klar [12], Hayes and Moulton [13], and Campbell et al. [14] for reviews on cluster randomized trials and cluster level observational studies.

In many such community randomized trials or observational studies, the number of communities is very small (e.g. around 10 or so), so that the number of independent units itself is not large enough for statistical inference based on limit distributions. In the extreme, but not uncommon, case, one may observe a single community of causally connected individuals. Can one now still statistically evaluate a causal effect of an intervention assigned at the community level on a community level outcome, such as the average of individual outcomes? This is the very question we aim to address in this article. Clearly, causal models incorporating all units are needed in order to define the desired causal quantity, and identifiability of these causal quantities under (minimal) assumptions need to be established without relying on asymptotics in a number of *independent* units.

An important ingredient of our modeling approach carried out in this article is the incorporation of network information that describes for each unit $i$ (in a finite population of $N$ units) at certain points in time $t$ a set of other units $F_i(t) \subset \{1, \ldots, N\}$ this unit may receive input from. This allows us to pose a structural equation model for this group of units in which the observed data node at time $t$ of a unit $i$ is only causally affected by the observed data on the units in $F_i(t)$, beyond exogenous errors. This group of friends needs to include the actual immediate friends of unit $i$ that directly affect the data at time $t$ of unit $i$, and if one knows the actual immediate friends, then $F_i(t)$ should not include anybody else. Such a structural equation model could be visualized through a so-called causal graph involving all $N$ units, which one might call a network. Our assumptions on the exogenous errors in the structural equation model will correspond with assuming sequential conditional independence of the unit-specific data nodes at time $t$, conditional on the past of all units at time $t$. That is, conditional on the most recent past of all units, including the recent network information, the data on the units at the next time-point are independent across units. The smaller these sets $F_i(t)$ (i.e. friends of $i$ at time $t$) can be selected, the fewer incoming edges for each node in the causal graph, the larger the effective sample size will be for targeting the desired quantity. Even though these causal graphs allows the units to depend on each other in complex ways, if the size of $F_i(t)$ is bounded

universally in $N$ and under our independence assumptions on the exogenous errors, it will follow that the likelihood of the data on all $N$ units allows statistical inference driven by the number of units $N$ instead of driven by the number of communities (e.g. 1). In future work, we will generalize our formal asymptotic results in which $F_i(t)$ is universally bounded, to the case in which the size of $F_i(t)$ can grow with $N$.

To precisely define and solve the estimation problem, we will apply the roadmap for targeted learning of a causal effect (e.g. Refs [2, 6, 15]). We start out with defining a structural causal model [2] that models how each data node is a function of parent data nodes and exogenous variables, and defining the causal quantity of interest in terms of stochastic interventions on the unit-specific treatment nodes. The structural assumptions of the structural causal model could be visualized by a causal graph describing the causal links between the $N$ units and how these links evolve over time, and from that it is clear that this structural causal model describes what one might call a dynamic causal network.

As mentioned above, our structural equation model also makes strong independence assumptions on the exogenous errors, which imply that the unit-specific data nodes at time $t$ are independent across the $N$ units, *conditionally* on the past of all $N$ units. We refer to this assumption as a sequential conditional independence assumption. Thus, it is assumed that any dependence of the unit-specific data nodes at time $t$ can be fully explained by the observed past on all $N$ units. (In our technical report, we weakened this assumption to allow for residual dependence after this adjustment, among units that are causally connected.) As a next step in the roadmap, we then establish the identifiability of the causal quantity from the data distribution under transparent additional (often non-testable) assumptions. This identifiability result allows us to define and commit to a statistical model that contains the true probability distribution of the data, and an estimand (i.e. a target parameter mapping applied to true data distribution) that reduces to this causal quantity if the required causal assumptions hold. The statistical model needs to contain the true data distribution, so that the statistical estimand can be interpreted as a pure statistical target parameter, while under the stated additional causal conditions that were needed to identify the causal effect, it can be interpreted as the causal quantity of interest. This statistical model, and the target parameter mapping that maps data distributions in this statistical model into the parameter values, defines the pure statistical estimation problem. As a next step in the roadmap, we develop targeted estimators of the statistical estimand and develop the theory for statistical inference. To understand the deviation between the estimand and the causal quantity under a variety of violations of these causal assumptions, one may carry out a sensitivity type analysis [16–18, 36], which represents the final step of the roadmap.

Since the statistical model does not assume that the data generating experiment involves the repetition of independent experiments, the development of targeted estimators and inference represents novel and new challenges in estimation and inference that, to the best of our knowledge, have not been addressed by the current literature. TMLE was developed for estimation in semi-parametric models for i.i.d. data [6, 19, 20] and extended to a particular form of dependent treatment/censoring allocation as present in group sequential adaptive designs [19, 21, 22] and community randomized trials [23]. In this article, we need to generalize TMLE to the complex semi-parametric statistical model presented in this article, and we also need to develop corresponding statistical inference.

Our models generalize the models in the causal inference literature for independent units. Even though in this article our causal model models a single group of units, it obviously includes the case that the units can be partitioned in multiple causally independent groups of units. In addition, our models also incorporate group sequential adaptive designs in which treatment allocation to an individual can be based on what has been observed on previously recruited individuals in the trial [19, 21, 22, 24]. Our models also allow that the outcome of an individual is a function of the treatments other individuals received. The latter is referred to as interference in the causal inference literature. Thus the causal models proposed in this article do not only generalize the existing causal models for independent units, but they also generalize causal models that incorporate previously studied causal dependencies between units. Finally, we note that our models and corresponding methodology can also be used to establish a methodology for assessing causal effects of *interventions on the network* on the average of the unit-specific outcomes. For example, one might want to

know how the community level outcome changes if we change the network structure of the community through some intervention, such as increasing the connectivity between certain units in the community. In this case, our treatment nodes need to be defined as properties of the sets $F_i(t)$ so that a change in treatment corresponds with a change in the network structure.

Nonetheless, our assumed universal bounds on the size of $F_i(t)$ in our formal results exclude many realistic and important types of networks, demonstrating that our asymptotic theorems need to be further generalized in order to capture many realistic networks, but that will be beyond the scope of this article.

## 1.1 A bibliographic remark and possible relation to network literature

We acknowledge that our contribution does not really fit well in the current literature on networks, which is much more concerned with properties of the network structure and uses particular types of models and estimands that are often not embedded within a causal model as we have done here (e.g. Ref. [25]). Our contribution is aligned and builds on the current causal inference literature (Neyman–Rubin or Pearl's structural equation models) to define the causal quantity of interest and establish identifiability from observed data. In addition, it builds on the modern literature of targeted learning in semi-parametric models and weak-convergence theory in order to deal with the estimation problem based on dependent data. Nonetheless, we think it is appropriate to define and model networks of units in terms of a structural equation model, so that the impact of interventions on this network of units can be formally defined, and methods for assessing such causal effects can be developed, as we do in this article. Therefore, we suggest and hope that our contributions may become relevant to the literature on networks.

In this article, we focussed on the case that we observe all $N$ units in the population, while we refer to our technical report for generalizing this to sampling a random sample from this population of $N$ units. We also restricted our attention to particular types of causal quantities, namely the counterfactual mean under a stochastic intervention on the unit-specific treatment nodes (and thereby also causal contrasts). The network literature, on the other hand, has been much more focussed on particular types of direct/indirect and peer effects among others (e.g. see Bakshy et al. [26] and Airoldi et al. [27] for estimation of causal peer influence effects, and the above references). We hope to apply our framework and approach to tackle such questions as well in future research.

We refer to Aronow and Samii [28] for an inverse probability of treatment weighted approach for estimation of an average causal effect (ACE) under general interference, relying on the experimental design to generate these required generalized propensity scores. In addition, these authors also provide finite sample positively biased estimators of the true (non-identifiable) conditional variance of this IPTW-estimator, conditioning on the underlying counterfactuals, again relying on knowing the generalized propensity score. In addition, the authors consider asymptotics when one observes multiple independent samples from subpopulations, the number of subpopulations converging to infinity, each sample allowing for their general type of interference.

Their innovative approach relies on defining an exposure model that maps the treatment nodes of the $N$ units and specified characteristics of unit $i$ into a generalized exposure of unit $i$. For example, you might define this generalized exposure as the vector of exposures of the friends of unit $i$, beyond the exposure of unit $i$ itself. It defines for each unit $i$ the counterfactual outcome corresponding with the static intervention that sets this generalized exposure to a certain value, same for each unit $i$, and then defines the counterfactual mean outcome as the expectation of the average of these unit-specific counterfactuals. It inverses probability weights by the conditional probability of this generalized exposure to obtain an unbiased estimator of this expectation of the average of these counterfactual outcomes.

Our model includes the case of observing many independent clusters of units as a special case, but by assuming more general conditional independence assumptions we also allow for asymptotic statistical inference when we only observe one population of interconnected units, we define causal quantities in terms of stochastic interventions on the $N$ unit-specific exposures, we allow for more general dependencies

than interference, and we develop highly efficient estimators that are very different from the above-mentioned IPTW-type estimator, overall making our approach distinct from Aronow and Samii [28].

## 1.2 Organization of article

The organization of this article is as follows.

**Section 2:** We formulate a counterfactual causal model that can be viewed as an analogue of the structural causal model actually used in this article. This section provides a perspective of the contribution of this article in the context of the causal inference literature that relies on the Neyman–Rubin model, demonstrating that in essence it corresponds with allowing for (statistical) dependence between the unit-specific counterfactuals indexed by interventions on the total of $N$ unit-specific exposures, allowing for the unit-specific counterfactuals to be affected by the treatments of other units (i.e. causal interference between the units), and that the treatment assigned to a unit are informed by other units in the population. This section is succinct and is not necessary for understanding the remainder of the article.

**Section 3:** We present our structural causal model that models the data generating process for a population of interconnected units, where changes of the connections over time (i.e. $F_i(t)$) themselves are part of the randomness. Specifically, it represents a model for the distribution of $(O, U) = (O_i, U_i : i = 1, \ldots, N)$, where $O_i$ denotes the observed data on unit $i$, and $U_i$ represents a vector of exogenous errors for the structural equations for unit $i$. This structural causal model allows us to define stochastic interventions denoted with $g^*$ on the collection of unit-specific treatment nodes (contained in $O_i$), and corresponding counterfactual outcomes. The causal quantity, denoted with $E\left(1/N \sum_{i=1}^{N} Y_{i,g^*}\right)$, is defined in terms of the (possibly conditional) expectation of the intervention-specific counterfactual outcomes $Y_{i,g^*}$, and it represents a parameter of the distribution of $(O, U)$. Subsequently, we establish identifiability of the causal quantity from the data distribution $P_0$ of data $O = (O_1, \ldots, O_N)$ on the $N$ units, commit to a statistical model $\mathcal{M}$ for the probability distribution $P_0$ of $O$, define the statistical target parameter mapping $\Psi : \mathcal{M} \to \mathbb{R}$ that defines the estimand $\Psi(P_0)$, where the latter reduces to the causal quantity under the additional assumptions that were needed to establish the identifiability. The statistical estimation problem is now defined by the data $O \sim P_0 \in \mathcal{M}$, the statistical model $\mathcal{M}$ and target parameter $\Psi : \mathcal{M} \to \mathbb{R}$. The parameter $\Psi(P)$ only depends on $P$ through a parameter $Q = Q(P)$. Therefore, we also use the notation $\Psi(Q)$ to denote this target parameter $\Psi(P)$.

**Section 4:** We discuss maximum likelihood estimation (MLE), unified loss-based cross-validation [29–31], and likelihood based super-learning [32, 33] of the relevant factor of $P_0$ (which implies $Q_0$). The resulting smoothed/regularized maximum likelihood substitution estimators $\Psi(Q_N)$ are not targeted and will thereby be overly biased w.r.t. the target parameter $\Psi(Q_0)$, and, as a consequence, generally not result in asymptotically normally distributed estimators of the statistical target parameter. Thus there is a need for targeted learning (targeting the fit toward $\psi_0$) instead of MLE.

**Section 5:** We present heuristic arguments demonstrating that the log-likelihood of $O$ will satisfy a local asymptotic normality condition [34, 35] so that efficiency theory can be applied to pathwise differentiable target parameters of the data distribution. As demonstrated in van der Vaart [35], under local asymptotic normality the normal limit distribution of the MLE (ignoring all regularity conditions that would be needed to establish the asymptotic normality of the MLE) is optimal in the sense of the convolution theorem [34]. In this section, we demonstrate that the variance of the efficient influence curve (i.e. the canonical gradient of the pathwise derivative of the target parameter) corresponds with the asymptotic variance of a maximum likelihood estimator of the target parameter. From this, we learn that our goal should be to construct estimators that are asymptotically normally distributed with variance equal to the standardized variance of the efficient influence curve (and thus asymptotically equivalent with a MLE), while appropriately dealing with the curse of dimensionality through super-learning and TMLE [6, 20].

In the remainder of the article, we focus on the simpler single time-point longitudinal data structure in which $O_i = (W_i, A_i, Y_i)$, where $W_i$ are baseline covariates, $A_i$ is the subsequent treatment assigned to unit $i$, and $Y_i$ is the final outcome of interest measured on unit $i$. This simplification allows us to present a TMLE in closed form and formally analyze this TMLE, while much of what we learn can be generalized to general longitudinal data structures.

**Section 6:** We derive the efficient influence curve, also called the canonical gradient of the pathwise derivative of the statistical target parameter [34, 35]. We also establish that the expectation of the efficient influence curve $D^*(Q, g)$ under misspecified parameters $(Q, g)$ of the data distribution can be represented as $\Psi(Q_0) - \Psi(Q)$ plus a product of differences of $Q$ and $Q_0$ and a specified $h(Q, g)$ and $h(Q_0, g_0)$. This result provides a fundamental ingredient in establishing a first-order expansion of the TMLE under conditions that make these second-order terms negligible relative to the first-order term, while a separate analysis of the first-order term (which is a sum of dependent random variables) establishes the asymptotic normality of the TMLE.

**Section 7:** We present the TMLE for the causal effect of a single time-point intervention on an outcome, controlling for the baseline covariates across the units. This TMLE generalizes the TMLE of the causal effect of a single time-point intervention under causal and statistical independence of the units [6, 36–39]. It is shown that the efficient influence curve satisfies a double robustness property, which implies the double robustness of the TMLE. We also present an estimator defined as a solution of the efficient influence curve based estimating equation: Robins and Rotnitzky [40] and van der Laan and Robins [3]. We propose effective schemes for implementing the TMLE.

**Section 8:** We present a theorem establishing asymptotic normality of this TMLE for the causal effect of a single time-point intervention and discuss statistical inference based on its normal limit distribution. The theorem relies on modern advances in weak convergence of processes as presented in van der Vaart and Wellner [41] and van der Vaart [35]. The proof of the theorem is deferred to the Appendix. The generalization of the formal asymptotics results for this TMLE to the TMLE for general longitudinal data structures is also discussed in the Appendix of our accompanying technical report.

**Section 9:** We present an analogue theorem for this TMLE as an estimator of the intervention-specific mean outcome, conditional on all baseline covariates $W = (W_1, \ldots, W_N)$. This result avoids making any independence assumptions on the distribution of $W$, and the asymptotic variance of the TMLE is reduced.

**Section 10:** We conclude with a summary and some remarks.

We will address the actual implementation of the proposed TMLE and simulation studies in an article in the near future. We refer to our accompanying technical report for various additional results such as weakening of the sequential conditional independence assumption (still heavily restricting the amount of dependence, but allowing that, even conditional on the observed past, a subject can be dependent on maximally $K$ other subjects), and only observing a random sample of the complete population of causally connected units, among others.

## 2 Formulation of estimation problem in terms of Neyman–Rubin model for counterfactuals

The estimation problem defined in the next section in terms of a semi-parametric structural equation model corresponds with the following counterfactual missing data problem formulation also called the Neyman–Rubin causal model [1, 42–46].

Let $X_i^F = (L_{i,a} : a \in \mathcal{A})$ be the full-data structure consisting of all static regimen-specific counterfactuals $L_{i,a}$ for unit $i$, where $a = (a_1, \ldots, a_N)$ represents the static regimens for all $N$ units,

$L_{i,a} = (L_{i,a}(0), L_{i,a}(1), \ldots, L_{i,a}(\tau+1))$ is a time-dependent process up till time $\tau+1$, and $L_{i,a}(t)$ only depends on $a$ through $(\bar{a}_j(t-1) = (a_j(0), \ldots, a_j(t-1)) : j = 1, \ldots, N)$. Note that counterfactuals $L_{i,a}$ are indexed by $a$ and not just the treatment $a_i$ for unit $i$: we refer to Halloran and Struchiner [7], Hudgens and Halloran [8], VanderWeele et al. [9], Tchetgen Tchetgen and VanderWeele [10], Aronow and Samii [28] for discussions of counterfactuals under interference.

Let $P_0^F$ be the probability distribution of $X^F = (X_1^F, \ldots, X_N^F)$ and let $\mathcal{M}^F$ be the full-data model, i.e. the collection of possible distributions of $X^F$. This full-data model will thus incorporate additional assumptions such as that the counterfactuals of unit $i$ only depend on the regimens of a subset of the $N$ individuals and conditional independence assumptions, as presented below. We observe the missing data structure $O = (O_i : i = 1, \ldots, N)$, $O_i = (A, L_i = L_{i,A})$ on the full-data $X^F = (X_i^F : i = 1, \ldots, N)$. We view $O = (O_1, \ldots, O_N)$ as a missing data structure on the full-data $X^F = (X_1^F, \ldots, X_N^F)$ with censoring variable $A$. In other words, $O = \Phi(A, X^F)$ for a specified function $\Phi$. We assume that the conditional density $g_0$ of $A = (A_1, \ldots, A_N)$, given $X^F$, satisfies

$$g_0(A|X^F) = \prod_{t=0}^{\tau} \prod_{i=1}^{N} g_{0,t,i}(A_i(t)|c_{t,i}^A),$$

where $c_{t,i}^A$ is a function of $(\bar{A}_j(t-1), \bar{L}_j(t) : j = 1, \ldots, N)$. Note that this corresponds with assuming that at each time $t$, $A_i(t)$, $i = 1, \ldots, N$, are independent, conditional on the past of the $N$ subjects (i.e. a sequential randomization assumption (SRA). We remind the reader that one definition of coarsening at random [47–49] is that the conditional density $g_0(A|X^F)$ of censoring variable $A$, given full-data $X^F$, w.r.t. an appropriate dominating measure, only depends on $A, X^F$ through the censored data structure $O = \Phi(A, X^F)$. Thus the SRA implies that the missingness mechanism on the full-data $X^F$ satisfies coarsening at random: Note that $g_0(A|X^F) = h_0(O)$ is a measurable function $h_0$ of $O$ so that this assumption indeed implies the coarsening at random assumption.

Due to this coarsening at random assumption, the likelihood of $O$ factorizes in a full-data distribution factor and the joint intervention mechanism $g_0$:

$$P_0(A = a, L = l) = P_{P_0^F}(L_{i,a} = l_i : i = 1, \ldots, N)\Big|_{a=A} g_0(a|X^F).$$

We use the notation $L = (L_i : i = 1, \ldots, N)$, $L_a = (L_{i,a} : i = 1, \ldots, N)$, and $\bar{L}_a(t) = (\bar{L}_{i,a}(t) : i = 1, \ldots, N)$. Note that the full-data distribution factor equals the likelihood of $(L_{i,a} : i = 1, \ldots, N)$ at set regimen $a = (a_1, \ldots, a_N)$ at value $A = (A_1, \ldots, A_N)$ and is thus identified by the full-data distribution $P_0^F$. We could model this full-data distribution factor of the likelihood as follows:

$$\begin{aligned}
P_{P_0^F}(L_{i,a} = l_i : i = 1, \ldots, N)\Big|_{a=A} &= \textstyle\prod_{t=0}^{\tau+1} P_{P_0^F}(L_a(t) = l(t)|\bar{L}_a(t-1) = \bar{l}(t-1)) \\
&= \textstyle\prod_{t=0}^{\tau+1} P_0(L(t) = l(t)|\bar{L}(t-1) = \bar{l}(t-1), \bar{A}(t-1) = \bar{a}(t-1)) \\
&= \textstyle\prod_{t=0}^{\tau+1} \prod_{i=1}^{N} P_0(L_i(t) = l_i(t)|\bar{L}(t-1) = \bar{l}(t-1), \bar{A}(t-1) = \bar{a}(t-1)) \\
&= \textstyle\prod_{t=0}^{\tau+1} \prod_{i=1}^{N} P_0(L_i(t) = l_i(t)|c_{t,i}^L(\bar{l}(t-1), \bar{a}(t-1))) \\
&= \textstyle\prod_{t=0}^{\tau+1} \prod_{i=1}^{N} \bar{Q}_{0,t}(l_i(t)|c_{t,i}^L(\bar{l}(t-1), \bar{a}(t-1))),
\end{aligned}$$

where the second equality assumes coarsening at random, the third equality assumes that $L_i(t), i = 1 \ldots, N$ are conditionally independent, given $\bar{L}(t-1), \bar{A}(t-1)$, the fourth equality assumes that $L_i(t)$ only depends on the past through an $i$-specific fixed (in $N$) dimensional summary measure $c_{t,i}^L$ of $(\bar{L}(t-1), \bar{A}(t-1))$, and the final equality assumes that each $L_i(t)$ is drawn from $\bar{Q}_{0,t}(\cdot|c_{t,i}^L(\bar{L}(t-1), \bar{A}(t-1)))$ for a common $\bar{Q}_{0,t}(\cdot|\cdot)$. These assumptions define the full-data model $\mathcal{M}^F$. Because of these assumptions, the full-data distribution

factor of the distribution of $O$ only depends on $P_0^F$ through $Q = (\bar{Q}_{0,t} : t = 0, \ldots, \tau + 1)$, so that the data distribution if parameterized by $Q, g$ and could thus denoted with $P_{Q,g}$. The statistical model $\mathcal{M}$ is now defined as $\{P_{Q,g} : Q, g\}$, where $Q, g$ are unspecified beyond the specifications presented above.

Our full-data target parameter is a parameter $\Psi^F : \mathcal{M}^F \to \mathbb{R}^d$ defined on the full-data model. The factorization of the likelihood of $O$ due to coarsening at random establishes the identifiability of $\psi_0^F = \Psi^F(P_0^F)$ as a parameter of the distribution $P_0$ of $O$, under the assumption that $\Psi^F(P_0^F)$ only depends on the full-data distribution $P_0^F$ through $Q_0$. As a consequence, we can now define a statistical target parameter $\Psi : \mathcal{M} \to \mathbb{R}^d$ so that $\psi_0^F = \psi_0 = \Psi(P_0)$. We need to construct an estimator $\psi_N$ of $\psi_0$ based on this single draw of $O \sim P_0 \in \mathcal{M}$, and we need to establish a limit distribution of the standardized estimator: $\sqrt{N}(\psi_N - \psi_0) \Rightarrow_d Z$ as $N \to \infty$ for some limit distribution $Z$ (e.g. $N(0, \Sigma)$).

Let $g^*$ be a conditional distribution of a random variable $A_*$, given $X$, satisfying coarsening at random so that $g^*(A_*|X) = h^*(A_*, L_{A_*})$ for some function $h^*$. We refer to this choice $g^*$ as a stochastic intervention which can be used to define a modified version of the data distribution $P$ by replacing $g$ by $g^*$ resulting in the probability distribution

$$P^{g^*}(A_*, L_*) \equiv \prod_{t=0}^{\tau+1} \prod_{i=1}^{N} \bar{Q}_{0,t}(L_{i*}(t)|c_{t,i}^L(\bar{L}_*(t-1), \bar{A}_*(t-1)))h^*(A_*, L_*)$$

whose random variable is denoted with $(A_*, L^{g^*})$, where we will also use the notation $L^*$ for $L^{g^*}$. The latter distribution $P^{g^*}$ is the so-called $G$-computation formula for the post-intervention distribution of $L$ under the stochastic intervention $g^*$ [45] and is a parameter of $P$. Under the causal model including the SRA and a positivity assumption, $P^{g^*}$ would equal the post-intervention distribution $P_{g^*}$ of the (counterfactual) random variable obtained by first drawing the counterfactuals $X^F = (L_a : a)$, then drawing an $A_* \sim g^*(\cdot|X)$, and reporting $(A_*, L_{A_*})$. A possible statistical target parameter is now given by $\Psi(P_0) = E_0 \frac{1}{N} \sum_{i=1}^N Y_i^{g^*}$, as addressed in this article, which equals the full-data parameter $\Psi^F(P_0^F) = E_0 \frac{1}{N} \sum_{i=1}^N Y_{i,g^*}$ under the causal model.

The fact that the counterfactual outcome of subject $i$ can be a function of the treatments of other subjects is referred to as interference in the causal inference literature. In addition, the above formulation allows that treatment allocation for unit $i$ depends on data collected on other units. The above formulation also allows dependence between the counterfactuals between different units. The above formulation can thus be viewed as the causal inference estimation problem when interference, adaptive treatment allocation, and dependence of the counterfactuals of different units is allowed. Our structural equation model defined in next section implies such restrictions on the distribution of the counterfactuals and defines this same particular full-data model $\mathcal{M}^F$.

# 3 Formulation of estimation problem using a structural causal model

For a unit $i$, let $O_i = (L_i(0), A_i(0), \ldots, L_i(\tau), A_i(\tau), Y_i = L_i(\tau + 1))$ be a time-ordered longitudinal observed data structure, where $L_i(0)$ are baseline covariates, $A_i(t)$ denotes an action/treatment/exposure at time $t$, which will play the role of intervention node in the structural equation model below, $L_i(t)$ denotes time-dependent measurements on unit $i$, possibly including an outcome process $Y_i(t)$, and $Y_i$ denotes a final outcome, realized after the final intervention node $A_i(\tau)$. Let $F_i(t)$ be a component of $L_i(t)$ that denotes the set of friends individual $i$ may receive input from at time $t$, $t = 0, \ldots, \tau + 1$. Thus, $F_i(t) \subset \{1, \ldots, N\}$.

If we define $L(t) = (L_i(t) : i = 1, \ldots, N)$, and similarly we define $A(t) = (A_i(t) : i = 1, \ldots, N)$, then the observed data $O = (O_1, \ldots, O_n)$ can be represented by a single time-ordered data structure

$$O = (L(0), A(0), \ldots, L(\tau), A(\tau), Y = L(\tau + 1)).$$

The latter ordering is the only causally relevant ordering, and the ordering of units within a time-point is user supplied but inconsequential. We define $Pa(A(t)) = (\bar{L}(t), \bar{A}(t-1))$ and $Pa(L(t)) = (\bar{L}(t-1), \bar{A}(t-1))$, as the parents of $A(t)$ and $L(t)$, respectively, w.r.t. this ordering. The parents of $A_i(t)$, denoted with $Pa(A_i(t))$, are defined to be equal to $Pa(A(t))$, and the parents of $L_i(t)$, denoted with $Pa(L_i(t))$, are also defined to be equal to $Pa(L(t))$, $t = 0, \ldots, \tau+1$, $i = 1, \ldots, N$.

In order to define causal quantities, we assume that $O$ is generated by a structural equation model of the following type: first generate a collection of exogenous errors $U_N = (U_i : i = 1, \ldots, N)$ across the $N$ units, where the exogenous errors for unit $i$ are given by

$$U_i = (U_{L_i(0)}, U_{A_i(0)}, \ldots, U_{L_i(\tau)}, U_{A_i(\tau)}, U_{Y_i}), \quad i = 1, \ldots, N,$$

and then generate $O$ deterministically by evaluating functions as follows:

$$\begin{aligned}
&L_i(t) = f_{L_i(t)}(Pa(L_i(t)), U_{L_i(t)}) \\
&i = 1, \ldots, N \\
&A_i(t) = f_{A_i(t)}(Pa(A_i(t)), U_{A_i(t)}) \\
&i = 1, \ldots, N \\
&t = 0, \ldots, \tau \\
&Y_i = f_{Y_i}(Pa(Y_i(\tau+1)), U_{Y_i(\tau+1)}) \\
&i = 1, \ldots, N.
\end{aligned}$$

These functions $(f_{L_i(t)} : t = 0, \ldots, \tau+1)$, $(f_{A_i(t)} : t = 0, \ldots, \tau)$ are unspecified at this point, but will be subjected to modeling below.

Since $Pa(L_i(t)) = (\bar{A}(t-1), \bar{L}(t-1))$ and $Pa(A_i(t)) = (\bar{A}(t-1), \bar{L}(t))$, an alternative succinct way to represent this structural equation model is

$$\begin{aligned}
&L(t) = \mathbf{f}_{L(t)}(Pa(L(t)), U_{L(t)}) \\
&A(t) = \mathbf{f}_{A(t)}(Pa(A(t)), U_{A(t)}) \\
&t = 0, \ldots, \tau \\
&Y = L(\tau+1) = \mathbf{f}_Y(Pa(Y), U_Y).
\end{aligned}$$

Recall that set of friends, $F_i(t)$, is a component of $L_i(t)$ and is thus also a random variable defined by this structural equation model, $t = 1, \ldots, \tau$, although we decided to condition on $F_i(0)$ in our formal theorems for the point-treatment data structure $O_i = (L_i(0), A_i(0), Y_i)$ in our later sections, representing the case $\tau = 0$.

**Counterfactuals and stochastic interventions:** This structural equation model for

$$(L(0), A(0), \ldots, L(\tau), A(\tau), Y = L(\tau+1)),$$

allows us to define counterfactuals $Y_d(\tau+1)$ corresponding with a dynamic intervention $d$ on $A$ [46, 50–53]. For example, one could define $A_i(t)$ at time $t$ as a particular deterministic function $d_{i,t}$ of the parents $Pa(A_i(t))$ of subject $i = 1, \ldots, N$. Such an intervention corresponds with replacing the equations for $A(t)$ by this deterministic equation $d_t(Pa(A(t)))$, $t = 0, \ldots, \tau$. More generally, we can replace the equations for $A(t)$ that describe a degenerate distribution for drawing $A(t)$, given $U = u$, and $Pa(A(t))$, by a user-supplied conditional distribution of an $A_*(t)$, given $Pa(A_*(t))$. Such a conditional distribution defines a so-called stochastic intervention: Dawid and Didelez [54], Didelez et al. [55], and Diaz and van der Laan [56].

Let $g^* = (\bar{g}^*_t : t = 0, \ldots, \tau)$ denote our selection of a stochastic intervention identified by a set of conditional distributions of $A_*(t)$, given $Pa(A_*(t))$, $t = 0, \ldots, \tau$. For convenience, we represent the stochastic intervention with equations $A_*(t) = f_{A_*(t)}(Pa(A_*(t)), U_{A_*(t)})$ in terms of random errors $U_{A_*(t)}$. This implies the following modified system of structural equations:

$$
\begin{aligned}
&L_*(t) = \mathbf{f}_{L(t)}(Pa(L_*(t)), U_{L(t)})\\
&A_*(t) = \mathbf{f}_{A_*(t)}(Pa(A_*(t)), U_{A_*(t)})\\
&t = 0, \ldots, \tau\\
&Y_* = L_*(\tau+1) = \mathbf{f}_Y(Pa(Y), U_Y),
\end{aligned}
$$

where $Pa(L_*(t))$ is the same set of variables as $Pa(L(t))$, but with $A, L$ replaced by $A_*, L_*$. Let $Y_{i,g^*}$, or shorthand $Y_{i,*}$, denote the corresponding counterfactual outcome for unit $i$. A causal effect at the unit level could now be defined as a contrast such as $Y_{i,g_1^*} - Y_{i,g_2^*}$ for two interventions $g_1^*$ and $g_2^*$. Note that, for a given $g^*$, $Y_{i,g^*} = Y_{i,g*}(U^*)$ is a deterministic function of the error-term $U^* \equiv (U, U_{A_*})$ that are inputted in the structural equations.

**Post-intervention distribution, and** SRA**:** We assume the SRA on $U$,

$$
A(t) \perp L_{g^*}, \text{ conditional on } Pa(A(t)) \ , \tag{1}
$$

and $U_{A_*} \perp U$. Then, the probability distribution $P_{g^*}$ of $(A_*, L_{g^*})$ is given by the so-called $G$-computation formula [45, 52, 53, 55, 57]

$$
P_{g^*}(A_*, L_*) = \prod_{t=0}^{\tau+1} \prod_{i=1}^{N} P_{L_i(t)}(L_{i,*}(t) | Pa(L_{i,*}(t))) \bar{g}_t^*(A_{i,*}(t) | Pa(A_{i,*}(t))),
$$

where $P_{L_i(t)}$ is the conditional distribution of $L_i(t)$, given $Pa(L_i(t))$, and $Pa(L_{i,*}(t)) = (\bar{L}_*(t-1), \bar{A}_*(t-1))$. We will denote the distribution of $L_{g^*}$ with $P_{L_{g^*}}$. Thus, under this SRA, the post-intervention distribution $P_{g^*}$ is identified from the observed data distribution of $O$ generated by the structural equation model. The distribution of $Y_{i,g^*}$ corresponds now with a marginal distribution of $P_{L_{g^*}}$.

**ACE:** One might now define an ACE as the following target parameter of this distribution of $P_{g^*}$:

$$
E_{P_{g_1^*}}\left\{\frac{1}{N}\sum_{i=1}^{N} Y_{i,g_1^*}\right\} - E_{P_{g_2^*}}\left\{\frac{1}{N}\sum_{i=1}^{N} Y_{i,g_2^*}\right\}.
$$

Let $\bar{Y} = \frac{1}{N}\sum_{i=1}^{N} Y_i$, so that we can also write this causal effect as $E(\bar{Y}_{g_1^*} - \bar{Y}_{g_2^*})$. Since the distribution $P_{g^*}$ is indexed by $N$, the parameter depends on $N$. In particular, the effect of stochastic intervention on a population of $N$ interconnected units will naturally depend on the size $N$ of that population, and the network information $F$: i.e. adding a unit will change the dynamics. As we will do in our point-treatment sections, one might decide to replace these marginal expectations by conditional expectations conditioning on $F_i(0)$, $i = 1, \ldots, N$, or even conditioning on $(L_i(0) : i = 1, \ldots, N)$. We will focus on the causal quantity $\psi^F = E_{P_{g^*}} \bar{Y}_{g^*}$ for a user-supplied stochastic intervention, and our results naturally generalize to causal quantities that a Euclidean valued function of a collection of such intervention-specific means.

**Iterative conditional expectation representation of ACE:** The parameter $E\bar{Y}_{g^*}$ can be represented as an iterative conditional expectation w.r.t. the probability distribution $P_{g^*}$ of $(A_*, L_{g^*})$ [58, 59]:

$$
\begin{aligned}
&\bar{Y} = \frac{1}{N}\sum\nolimits_{i=1}^{N} Y_i(\tau+1)\\
&\bar{Q}_{\tau+1,1}^{g^*} = E(\bar{Y} \mid \bar{L}(\tau), \bar{A}(\tau))\\
&\bar{Q}_{\tau+1}^{g^*} = E_{g_\tau^*}(\bar{Q}_{\tau+1,1}^{g^*} | \bar{L}(\tau), \bar{A}(\tau-1))\\
&\bar{Q}_{\tau,1}^{g^*} = E(\bar{Q}_{\tau+1}^{g^*} | \bar{L}(\tau-1), \bar{A}(\tau-1))\\
&\bar{Q}_{\tau}^{g^*} = E_{g_{\tau-1}^*}(\bar{Q}_{\tau,1}^{g^*} | \bar{L}(\tau-1), \bar{A}(\tau-2))\\
&\text{Iterate}\\
&\bar{Q}_{1,1}^{g^*} = E(\bar{Q}_2^{g^*} | \bar{L}(0), \bar{A}(0))\\
&\bar{Q}_{1}^{g^*} = E_{g_0^*}(\bar{Q}_{1,1}^{g^*} | \bar{L}(0))\\
&\bar{Q}_{0}^{g^*} = E_{L(0)}\bar{Q}_1^{g^*}(L(0)),
\end{aligned}
$$

where $EY_{g^*} = \bar{Q}_0^{g^*}$. Thus, this mapping involves iteratively integrating w.r.t. the observed data distribution of $L(t)$, given its parents, and the conditional intervention distribution $\bar{g}_t^*$ of $A_*(t)$, given $Pa(A_*(t))$, respectively, starting at $t = \tau + 1$, till $t = 0$.

**Dimension reduction and exchangeability assumptions:** The above-stated identifiability of $\psi^F$ is not of interest since we cannot estimate the distribution of $O$ based on a single observation. Therefore, we will need to make much more stringent assumptions that will allow us to learn the distribution of $O$ based on a single draw. One could make such assumptions directly on the distribution of $O$, but below we present these assumptions in terms of assumptions on the structural equations and exogenous errors.

Beyond the assumptions above, we will also assume that for each node $A_i(t)$ and $L_i(t)$, we can define known functions, $Pa(A_i(t)) \to c_{t,i}^A(Pa(A_i(t)))$ and $Pa(L_i(t)) \to c_{t,i}^L(Pa(L_i(t)))$, that map into a Euclidean set with a dimension that does not depend on $N$, and corresponding common (in $i$) functions $f_{L(t)}, f_{A(t)}, f_Y$, so that

$$L_i(t) = f_{L(t)}(c_{t,i}^L(Pa(L_i(t))), U_{L_i(t)}) \tag{2}$$

$$A_i(t) = f_{A(t)}(c_{t,i}^A(Pa(A_i(t))), U_{A_i(t)})$$

$$i = 1, \ldots, N, t = 0, \ldots, \tau$$

$$Y_i = f_Y(c_{\tau+1,i}^Y(Pa(Y_i)), U_{Y_i})$$

$$i = 1, \ldots, N.$$

(As mentioned above, an interesting variation of this structural causal model treats $L(0)$ as given and thus removes that data generating equation.) Examples of such dimension reductions are $c_{t,i}^L(Pa(L_i(t))) = ((\bar{L}_i(t-1), \bar{A}_i(t-1)), (\bar{L}_j(t-1), \bar{A}_j(t-1) : j \in F_i(t-1)))$, i.e. the observed past of unit $i$ itself and the observed past of its current friends, and, similarly, we can define $c_{t,i}^A(Pa(A_i(t))) = ((\bar{L}_i(t), \bar{A}_i(t-1)), (\bar{L}_j(t), \bar{A}_j(t-1) : j \in F_i(t-1)))$. By augmenting these reductions to data on maximally $K$ friends, filling up the empty cells for units with fewer than $K$ friends with a missing value, these dimension reductions have a fixed dimension and include the information on the number of friends. This structural equation model assumes that, across all units $i$, the data on unit $i$ at the next time-point $t$ is a common function of its own past and past of its friends. In our formal asymptotic results for the TMLE based on the point-treatment data structure $O_i = (L_i(0), A_i(0), Y_i(0))$, we assume this particular type of summary measure of maximally $K$ friends in order to enforce enough independence to establish an asymptotic normal limit distribution, but the sequel and the TMLE are defined for any summary measure, and in future work we hope to address the analysis of the TMLE for more general summary measures.

**Independence assumptions on exogenous errors:** Beyond the SRA (1), we make the following (conditional) independence assumptions on the exogenous errors. Firstly, we assume independence assumptions on $U_{L_i(0)}$ (and thereby $L_i(0)$, $i = 1, \ldots, N$) such as that $U_{L_i(0)}$, $i = 1, \ldots, N$, are independent (so that $L_i(0)$, $i = 1, \ldots, N$, are independent), or that $U_{L_i(0)}$ is independent of $U_{L_j(0)}$ if $F_i(0) \cap F_j(0) = \emptyset$. We will estimate the joint distribution of $L(0)$ with the empirical counterpart that puts mass 1 on the actual observed $L(0) = (L_1(0), \ldots, L_N(0))$, and the resulting empirical expectation w.r.t. this empirical distribution in our estimator, i.e. $\bar{Q}_1^{g^*}(L(0))$ in the iterative algorithm above, has to satisfy that $\sqrt{N}(\bar{Q}_1^{g^*}(L(0)) - E_0\bar{Q}_1^{g^*}(L(0)))$ has to converge to a normal distribution. The key assumption for this convergence in distribution is that $L_i(0)$ depends on at most $K$ $L_j(0)$ for a universal $K$. So we will assume a model on the distribution of $L(0)$ that assumes the latter, at minimal.

In addition, for all $t = 0, \ldots, \tau$, conditional on $(\bar{A}(t-1), \bar{L}(t))$, $U_{A_i(t)}$, $i = 1, \ldots, N$ are independent and identically distributed, and for all $t = 1, \ldots, \tau + 1$, conditional on $(\bar{A}(t-1), \bar{L}(t-1))$, $U_{L_i(t)}$, $i = 1, \ldots, N$, are independent and identically distributed. The important implication of the latter assumptions is that, given

the observed past $Pa(L(t))$, for any two units $i$ and $j$ that have the same value for their summaries $c^L_{t,i} = c^L_{t,j}$ as functions of $Pa(L(t))$, we have that $L_i(t)$ and $L_j(t)$ are independent and identically distributed, and similarly, we have this statement for the treatment nodes. This allows us to factorize the likelihood of the observed data as done below, parameterized by common conditional distributions $\bar{Q}_{0,L(t)}$ and $\bar{g}_{0,t}$ that can actually be learned from a single (but growing) $O$ when $N \to \infty$.

**Identifiability: *G*-computation formula for stochastic intervention.** For notational convenience, let $C^L_{t,i} = c^L_t(Pa(L_i(t)))$, and let $C^{L,*}_{t,i}$ be defined accordingly with $A, L$ replaced by $A_*, L_*$. Due to the exchangeability and dimension reduction assumptions, the probability distribution $P_{g^*}$ of $L_{g^*} = (L_{i,g^*} : i = 1, \ldots, N)$ now simplifies:

$$P_{g^*}(L_*, A_*) = P_{L(0)}(L(0)) \prod_{i=1}^{N} \prod_{t=0}^{\tau+1} \bar{Q}_{L(t)}(L_{i,*}(t)|C^{L,*}_{t,i}))\bar{g}^*_t(A_{i,*}(t)|Pa(A_{i,*}(t))) \equiv P^{g^*}(L_*, A_*), \quad (3)$$

where $\bar{Q}_{L(t)}$ are the above defined conditional distributions of $L_i(t)$, given $Pa(L_i(t))$, where, by our assumptions, these $i$-specific conditional densities are constant in $i = 1, \ldots, N$, as functions of $C^L_{t,i}$, $t = 1, \ldots, \tau + 1$. We will also use the notation $Q_{L(t)}$ for the conditional distribution of $L(t)$, given $Pa(L(t))$, which is thus parameterized in terms of $\bar{Q}_{L(t)}$. Similarly, we use the notation $g_{A(t)}$ or $g_t$ to denote the conditional distribution of $A(t)$, given $Pa(A(t))$, which is thus parameterized in terms of $\bar{g}_t$. We introduced the notation $P^{g^*}$ for the right-hand side in eq. (3) which thus represents an expression in terms of the distribution of the data under the assumption that the conditional densities of $L_i(t)$, given $Pa(L_i(t))$, are constant in $i$ as functions of $C^L_{t,i}$, indexed by the choice of stochastic intervention $g^*$, while one needs the causal model and randomization assumption in order to have that the right-hand side actually models the counterfactual post-intervention distribution $P_{g^*}$. This shows that $\psi^F_0 = \Psi(P_0)$ for a mapping $\Psi$ from the distribution $P_0$ of $O$ to the real line. Strictly speaking this does not establish a *desired* identifiability result yet, since we cannot learn $P_0$ based on a single draw $O$. To start with, we need to realize that $P^N_0$, $\psi^{F,N}_0$, and $\psi^N_0$ are indexed by $N$, and we only observed one draw from $P^N_0$. Therefore, we still need to show that we can construct an estimator based on a single draw $O^N$ that is consistent for $\psi^N_0$ as $N \to \infty$. For that purpose, we note that the distribution $P^{g^*}$ is identified by the common conditional distributions $\bar{Q}_{L(t)}$, $t = 1, \ldots, \tau + 1$, and $P_{L(0)}$ with $L(0) = (L_i(0) : i = 1, \ldots, N)$. We can construct consistent estimators of these common conditional distributions $\bar{Q}_{0,L(t)}$ based on MLE that are consistent as $N \to \infty$, which follows from our presentation of estimators and theory. This demonstrates the identifiability of $\bar{Q}_{0,L(t)}$ as $N \to \infty$, $t = 1, \ldots, \tau + 1$. In addition, our target parameter involves an average $E_{L(0)}\bar{Q}^{g^*}_1(L(0))$ w.r.t. $P_{L(0)}$ which can be consistently estimated by its empirical counterpart under our independence assumptions, as discussed above. This demonstrates the desired identifiability of $\psi^{F,N}_0$ from the observed data as $N \to \infty$.

**Likelihood and statistical model:** Let $Q_{L(0)}$ denote the distribution of $L(0)$. By our assumptions, the likelihood of the data

$O = (L(0), A(0), \ldots, L(\tau), A(\tau), Y = L(\tau + 1))$ is given by:

$$P_{Q,g}(O) = Q_{L(0)}(L(0)) \prod_{i=1}^{N} \prod_{t=1}^{\tau+1} \bar{Q}_{L(t)}(L_i(t)|C^L_{t,i})\bar{g}_t(A_i(t)|C^A_{t,i}). \quad (4)$$

We denoted the factors representing the conditional distributions of $L_i(t)$ with $\bar{Q}_{L(t)}$, where these conditional densities at $L_i(t)$, given $Pa(L_i(t))$, are constant in $i$, as functions of $L_i(t)$ and $C^L_{t,i}$. Similarly, we modeled the $g$-factor in terms of common conditional distributions $(\bar{g}_t : t = 0, \ldots, \tau)$. Let $Q = (Q_{L(0)}, \bar{Q}_{L(t)} : t = 1, \ldots, \tau + 1)$ represent the collection of all these factors, and $g = (\bar{g}_t : t = 0, \ldots, \tau)$, so that the distribution of $O$ is parameterized by $(Q, g)$. The conditional distributions $\bar{Q}_{L(t)}(L(t)|C^L_t)$ are unspecified functions of $L(t)$ and $C^L_t$, beyond that for each value of $C^L_t$ it is a conditional density, and $Q_{L(0)}$ satisfies a particular independence model discussed above. Similarly, the conditional distributions $\bar{g}_t$ are unspecified conditional

densities. This defines now a statistical parameterization of the distribution of $O$ in terms of $Q, g$, and a corresponding statistical model

$$\mathcal{M} = \{P_{Q,g} : Q \in \mathcal{Q}, g \in \mathcal{G}\}, \tag{5}$$

where $\mathcal{Q}$ and $\mathcal{G}$ denote the parameter spaces for $Q$ and $g$, respectively. Note that we derived the same likelihood and statistical model based on the Neyman–Rubin model in Section 2: instead of making assumptions on the structural equation model, we assumed coarsening at random, and made assumptions on the full-data distribution factor of the likelihood.

**Statistical target parameter:** Let $L^{g^*}$ denote a random variable with distribution $P^{g^*}$ (eq. 3), defined as a function of the data distribution $P$ of $O$. We define our statistical target parameter as $E\bar{Y}^{g^*}$ which is a function of the intervention-specific distribution $P^{g^*}$, so that it equals the causal quantity $E\bar{Y}_{g^*}$ under the above-stated causal assumptions. Thus

$$E_{P_{Q,g}}\bar{Y}^{g^*} = \Psi(P_{Q,g}) = \Psi(Q) \tag{6}$$

depends on the distribution $P$ of the data $O$ through $Q = (Q_{L(0)}, \bar{Q}_{L(t)} : t = 1, \ldots, \tau+1)$. Note that $Q$ is determined by the distribution of $L(0)$, and the conditional distributions of $L_i(t)$, given $(\bar{A}(t-1), \bar{L}(t-1))$, which, by assumption, equal a common function $\bar{Q}_{L(t)}(L_i(t)|C^L_{t,i})$, $t = 1, \ldots, \tau+1$. As shown above, we can represent this statistical target parameter also as an iterative conditional expectation involving the iterative integration w.r.t. $\bar{Q}_{L(t)}$, $g^*_{A(t-1)}$, starting at $t = \tau+1$ and moving backward till the expectation over $L(0)$:

$$\begin{aligned}
&\bar{Q}_{\tau+2} \equiv \bar{Y}\\
&\bar{Q}_{\tau+1,1} = E_{Q_{\tau+1}}(\bar{Q}_{\tau+2}|\bar{A}(\tau), \bar{L}(\tau))\\
&\bar{Q}_{\tau+1} = E_{g^*_\tau}(\bar{Q}_{\tau+1,1}|\bar{A}(\tau-1), \bar{L}(\tau))\\
&\text{Iterate, } t = \tau, \ldots, 0\\
&\bar{Q}_{t+1,1} = E_{Q_{t+1}}(\bar{Q}_{t+1}|\bar{A}(t), \bar{L}(t))\\
&\bar{Q}_{t+1} = E_{\bar{g}^*_t}(\bar{Q}_{t+1,1}|\bar{A}(t-1), \bar{L}(t))\\
&\bar{Q}_{t=0} = E_{L(0)}\bar{Q}_1\\
&= E\bar{Y}^*
\end{aligned}$$

This representation allows the effective evaluation of $\Psi(Q)$ by first evaluating a conditional expectation w.r.t. conditional distribution of $L(\tau+1)$, and thus w.r.t. $\prod_{i=1}^N \bar{Q}_{L(\tau+1)}(L_i(\tau+1)|C^L_{\tau+1,i})$, then the conditional mean of the previous conditional expectation w.r.t. conditional distribution of $A_*(\tau)$, and iterating this process of taking a conditional expectation w.r.t. $L(t)$ and $A_*(t-1)$ till we end up with a conditional expectation over $A_*(0)$, given $L(0)$, and finally we take the marginal expectation w.r.t. the distribution of $L(0)$. Note that each conditional expectation involves an expectation over vector $(L_i(t) : i = 1, \ldots, N)$ or $(A_{i,*}(t-1) : i = 1, \ldots, N)$ w.r.t. product measure of common conditional distributions $\bar{Q}_{L(t)}(L_i(t)|c^L_{t,i})$ or $g^*_{t-1}(A_{i,*}(t-1)|c^{A_*}_{t-1,i})$, $t = 1, \ldots, \tau+1$.

One can also define an $L(0)$-conditional statistical target parameter as $E(\bar{Y}^{g^*}|L(0))$, which can still be effectively evaluated by the iterative conditional expectations presented above, but one simply removes the final integration over the distribution of $L(0)$.

**Statistical estimation problem:** We have now defined a statistical model $\mathcal{M}$ (eq. 5) for the distribution (eq. 4) of $O$, and a statistical target parameter mapping $\Psi : \mathcal{M} \to \mathbb{R}$ (eq. 6) for which $\Psi(P_{Q,g})$ only depends on $Q$. We will also denote this target parameter with $\Psi(Q)$, with some abuse of notation by letting $\Psi$ represent these two mappings. Given a single draw $O \sim P_{Q_0,g_0}$, we want to estimate $\Psi(Q_0)$. In addition, we want to construct an asymptotically valid confidence interval. Recall that our notation suppressed the dependence on $N$ and $F$ of the data distribution $P_{Q,g}$, statistical model $\mathcal{M}$, and target parameter $\Psi$. In the conditional model for the conditional distribution of $O$, given $L(0)$, we will make the dependence on $L(0)$ of the data distribution $P^{L(0)}_{Q,g}$, $\mathcal{M}^{L(0)}$, and $\Psi^{L(0)}$ explicit.

**Summary:** So we defined a structural causal model (eq. 2), including the stated independence (and i.i.d.) assumptions on the exogenous errors, the dimension reduction assumptions, and the SRA (eq. 1). This resulted in the likelihood (eq. 4) and corresponding statistical model $\mathcal{M}$ (eq. 5) for the distribution $P_0$ of $O$. In addition, these assumptions allowed us to write the causal quantity $\psi_0^F$ as a statistical estimand $\Psi(Q_0)$ (eq. 6): $\psi_0^F = \Psi(Q_0)$, where $Q_0$ can be learned from a single draw $O$ as $N \to \infty$. The pure statistical estimation problem is now defined: $O \sim P_0 \in \mathcal{M}$, and we want to learn $\psi_0 = \Psi(P_0)$ where $\Psi : \mathcal{M} \to \mathbb{R}$. Under the non-testable causal assumptions, beyond the statistical assumption $P_0 \in \mathcal{M}$, we can interpret $\psi_0$ as $\psi_0^F$, but, even without these non-testable assumptions, one might interpret $\psi_0$ (and its contrasts) pure statistically as an effect measure of interest controlling for the observed confounders.

## 3.1 Example

In order to provide the reader with some sense of the type of applications that can be addressed with our model approach, we present a few examples.

Consider a study in which we wish to evaluate the effect of starting HIV treatment early after HIV-infection on the rate of HIV-infection for the population of interest. For that purpose, the study tracks the cohort of individuals for 5 years, and for each individual one obtains baseline characteristics $L_i(0)$, one regularly tests for HIV-infection ($Y_i(t)$), one measures when the individual starts treatment and one measures if the person was lost to follow up ($A_i(t) = (A_{1i}(t), A_{2i}(t))$), one regularly measures biomarkers and other time-dependent characteristics of interest such as condom use ($L_i(t)$), and one regularly measures the set $F_i(t) \subset L_i(t)$ of sexual partners. Let $t = 0, \ldots, \tau + 1$, where the $\tau + 1$th point represents end-point 5 years after baseline. Suppose one is interested in the effect of early HIV treatment ($A_{1i}(t)$) on the proportion $1/N \sum_i Y_i(\tau + 1)$ of HIV-infections at 5 years. One knows that an HIV-infected person that is being treated is much less infectious than a non-treated HIV-infected person, so that early treatment might have a strong beneficial effect on the spread of HIV-infection. One might be interested to estimate the mean outcome $1/N \sum_i E_0 Y_{i,g^*}(\tau + 1)$ under a stochastic intervention $g^*$ on $A_i(t)$, $t = 0, \ldots, \tau$. For example, the stochastic intervention deterministically starts HIV-treatment after the first observed HIV-infection, and it enforces no-right-censoring. This would be an example of a deterministic dynamic intervention. In our model, we may assume that our conditional distributions of $L_i(t)$, $Y_i(t)$, and $A_i(t)$, given the past on all individuals only depends on the individual pasts of the sexual partners of subject $i$, beyond the past of subject $i$ itself. In particular, it is clear that the HIV-infection at time $t$ for individual $i$ is very much a function of the treatment status of its sexual partners.

A simplified version of this example is the case that we only observe on individual $i$ baseline covariates $L_i(0)$ (including baseline HIV-infection status), treatment status $A_i(0)$, and subsequent HIV-infection $Y_i(1)$, for the $N$ individuals. One might now assume that the treatment status of individual $i$ is not only a function of its own baseline characteristics, but also of the baseline characteristics of its sexual partners, and that its outcome status is a function of the baseline characteristics and treatment status of its friends as well as its own.

Similarly, the treatment node could be defined as the indicator of condom use, so that the counterfactual mean outcomes evaluates the effect of condom use on the spread of the HIV-epidemic. One could also think about interventions on $F_i(t)$ itself, such as interventions that decrease the number $|F_i(t)|$ of sexual partners. This corresponds with specifying a conditional distribution of $|F_i(t)|$, given the past, at each time $t$, where such a conditional distribution might be a part of the actual distribution of the set $F_i(t)$, given the past.

It is also of interest to note that a stochastic intervention could only target a random subset of the total set of intervention nodes, $(A_i(t) : i, t)$, by focussing on a subset of individuals and a subset of the time-points. That is, a stochastic intervention could be equal to the actual treatment mechanism that generated the $A_i(t)$ at certain times and for certain individuals while it enforces an intervention elsewhere. For example, resources might only allow one to carry out a limited number of interventions, and one wishes to evaluate different strategies for selecting the nodes for which the intervention will be enforced.

Another example of interest might be one in which taking an anti-depression drug or an intervention is the treatment node, and a depression score at a final time-point is the outcome of interest. Consider a group of individuals that are socially connected and for which a reasonable proportion is subjected to this intervention or drug-treatment. One might expect that drug/intervention node of the friends of individual $i$ affects (indirectly) the psychological health and thereby outcome of individual $i$, so that this would be an example of causal interference. In addition, one expects that drug/intervention node of individual $i$ is affected by the drug/intervention nodes of its friends and other features of its friends, so that this is also an example of adaptive treatment allocation (i.e. treatment for individual $i$ is affected by the past of the friends of individual $i$, beyond the past of individual $i$ itself). Thus, this would be an example where one naturally needs to allow that both the treatment nodes and the outcome nodes of an individual are affected by the observed past of its friends. Clearly, the causal effect of different stochastic interventions on the anti-depression drug/intervention nodes for the individuals in the population will include the peer effects.

# 4 Maximum likelihood estimation, cross-validation, super-learning, and targeted maximum likelihood estimation

We could estimate the distribution of $L(0)$ with the empirical distribution that puts mass 1 on $(L_i(0) : i = 1, \ldots, N)$. This choice also corresponds with a TMLE of the intervention-specific mean outcome $E(\bar{Y}^{g^*}|L(0))$ that conditions on $L(0)$, as we formally show in our later sections for the single time-point data structure. If it is assumed that $(L_i(0) : i = 1, \ldots, N)$ are independent, then we estimate the distribution of $L(0)$ with the NPMLE that maximizes the log-likelihood $\sum_i \log Q_{L_i(0)}(L_i(0))$ over all possible distributions of $L(0)$ that the statistical model $\mathcal{M}$ allows. In particular, if it is known that $L_i(0)$ are i.i.d., then we would estimate the common distribution $Q_0$ of $L_i(0)$ with the empirical distribution that puts mass $1/N$ on $L_i(0)$, $i = 1, \ldots, N$.

Regarding estimation of $\bar{Q}_{0,t} = \bar{Q}_{0,L(t)}$ for $t = 1, \ldots, \tau + 1$, we consider the log-likelihood loss function for $\bar{Q}_t$:

$$L_t(\bar{Q}_t) \equiv -\sum_{i=1}^{N} \log \bar{Q}_t(L_i(t)|C_{i,t}^L).$$

Note that $E_0 L_t(\bar{Q}_t)$ is minimized in $\bar{Q}_t$ by the true $\bar{Q}_{0,t}$, since, conditional on $(\bar{A}(t-1), \bar{L}(t-1))$, the true distribution of $L_i(t)$ is given by $\bar{Q}_{0,t}(\cdot|C_{i,t}^L)$, $i = 1, \ldots, N$. In addition, this expectation $E_0 L_t(\bar{Q}_t)$ is well approximated by $\frac{1}{N}\sum_{i=1}^{N} \log \bar{Q}_t(L_i(t)|C_{i,t}^L)$, since, conditional on $(\bar{A}(t-1), \bar{L}(t-1))$, this is a sum of independent random variables $L_i(t)$, $i = 1, \ldots, N$. The latter allows us to prove convergence of the empirical mean process to the true mean process uniformly in large parameter spaces for $\bar{Q}_t$, using similar techniques as we use in the Appendix based on weak-convergence theory in van der Vaart and Wellner [41]. As a consequence, one could pose a parametric model for $\bar{Q}_{0,t}$, say $\{Q_{t,\theta} : \theta\}$, and use standard MLE

$$\theta_N = \arg\min_{\theta} L_t(Q_{t,\theta}),$$

as if the observations $(L_i(t), C_{i,t}^L)$, $i = 1, \ldots, N$, are independent and identically distributed and we are targeting this common conditional density of $L_i(t)$ given $C_{i,t}^L$. More importantly, we can use loss-based cross-validation and super-learning to fit this function $\bar{Q}_{0,t}$ of $(l(t), c_t^L)$, thereby allowing for adaptive estimation of $\bar{Q}_{0,t}$. Specifically, consider a collection of candidate estimators $\hat{\bar{Q}}_{t,k}$ that maps a data set $\{(L_i(t), C_{t,i}^L) : i\}$ into an estimate, $k = 1, \ldots, K$, and let $P_N^t$ denote the empirical distribution that puts mass $1/N$ onto each $L_i(t), C_{t,i}^L$. Given a random split vector $B_N \in \{0,1\}^N$, define $P_{N,B_N}^{t,1}$ and $P_{N,B_N}^{t,0}$ as the empirical distributions of the validation sample $\{i : B_n(i) = 1\}$ and training sample $\{i : B_N(i) = 0\}$, respectively. We can now define the cross-validation selector $k_n$ of $k$ as

$$k_n = \arg\min_k E_{B_N} P^{t,1}_{N,B_N} L_t(\hat{\bar{Q}}_{t,k}(P^{t,0}_{N,B_N}))$$
$$= \arg\min_k E_{B_N} \sum_{i:B_N(i)=1} \log \hat{\bar{Q}}_{t,k}(P^{t,0}_{N,B_N})(L_i(t)|C^L_{t,i}).$$

If $L_i(t)$ is continuous, one could code $L_i(t)$ in terms of binary variables $I(L_i(t) = l)$ across the different levels $l$ of $L_i(t)$, and model the conditional distribution/hazard of $I(L_i(t) = l)$, given $L_i(t) \geq l$ and $\bar{A}(t-1), \bar{L}(t-1)$, as a function of $C^L_{t,i}$ and $l$, as in van der Laan [60, 61]. One could now construct candidate estimators of this conditional hazard, possibly smoothing in the level $l$, by utilizing estimators of predictors of binary variables in the machine learning literature, including standard logistic regression software for fitting parametric models. Similarly, this can be extended to multivariate $L_i(t)$ by first factorizing the conditional distribution of $L_i(t)$ in univariate conditional distributions. In this manner, one obtains then candidate estimators of $\bar{Q}_{0,L(t)}$ based on a large variety of algorithms from the literature.

We could fit each $\bar{Q}_{0,t}$ separately for $t = 1, \ldots, \tau + 1$, but it is also possible to pool across $t$ by constructing estimators and using cross-validation based on the sum loss function

$$L(Q) = \sum_t L_t(\bar{Q}_t).$$

Similarly, we can use the log-likelihood loss function for $\bar{g}_t$:

$$L_t(\bar{g}_t) = -\sum_{i=1}^N \log \bar{g}_t(A_i(t)|C^A_{t,i}),$$

and use loss-based cross-validation and super-learning to fit $\bar{g}_t$, possibly pooling across time based on the sum loss function

$$L(g) = \sum_t L_t(\bar{g}_t).$$

Given the resulting estimator $Q_N$ of $Q_0$, one can evaluate $\Psi(Q_N)$ as estimator of $\psi_0 = \Psi(Q_0)$, according to the iterative conditional expectation mapping presented earlier. Since $Q_N$ is optimized to fit $Q_0$ (i.e. involving trading off bias and variance w.r.t. $Q_0$, not $\psi_0$), such a data-adaptive plug-in estimator, although it inherits the (e.g. minimax adaptive) rate of convergence at which $Q_N$ converges to $Q_0$, it is overly biased for $\Psi(Q_0)$, so that $\Psi(Q_N)$ will generally not converge to $\Psi(Q_0)$ at rate $1/\sqrt{N}$.

**TMLE:** TMLE will involve modifying an initial estimator $\bar{Q}_{t,N}$ into a targeted version $\bar{Q}^*_{t,N}$, $t = 1, \ldots, \tau + 1$, through utilization of an estimator $g_N$ of $g_0$, a least-favorable submodel (w.r.t. target parameter $\psi_0$) $\{\bar{Q}^k_{t,N}(\epsilon, g_N) : \epsilon\}$ through a current fit $\bar{Q}^k_{t,N}$ at $\epsilon = 0$, fitting $\epsilon$ for each $t$ and each step $k$ with standard MLE $\varepsilon_{N,t,k}$, iterative updating $\bar{Q}^{k+1}_{t,N} = \bar{Q}^k_{t,N}(\varepsilon_{N,t,k})$, $t = 1, \ldots, \tau + 1$, till convergence in $k = 1, 2, \ldots$. The resulting TMLE of $\psi_0$ is defined accordingly as the substitution estimator $\Psi(Q^*_N)$. Thus, a TMLE will also involve estimation of the intervention mechanism $g_0 = (\bar{g}_{0,t} : t = 0, \ldots, \tau)$. To define such a TMLE, we need to determine the efficient influence curve of the statistical target parameter, which will imply these least-favorable submodels. We refer to our technical report van der Laan [62] for a derivation of the efficient influence curve, a study of its robustness, and a detailed presentation of this general TMLE. (In the next section, we also showcase the formula for this efficient influence curve.) Instead, in this article, we will focus on the single time-point longitudinal data structure with $O_i = (W_i = L_i(0), A_i = A_i(0), Y_i)$ and present a complete self-contained analysis of the TMLE.

# 5 Characterizing the optimal asymptotic variance of the MLE in terms of efficient influence curve

Due to our sequential conditional independence assumption, the log-likelihood of $O$, i.e. the log of the data-density (eq. 4) of $O$, can be represented as a double sum over time-points $t$ and units $i$, and for each $t$, the

sum over $i$ consists of independent random variables, conditional on the past. As a consequence, under regularity conditions, one can show that the log-likelihood is asymptotically normally distributed. Therefore, we conjecture that we can establish so-called local asymptotic normality of our statistical model, which involves establishing asymptotic normality of log-likelihood under sampling from fluctuations/submodels $P_{\varepsilon=1/\sqrt{N}} \subset \mathcal{M}$ of a fixed data distribution $P$ across all possible fluctuations. As shown in van der Vaart [35], for models satisfying the local asymptotic normality condition, the normal limit distribution of an MLE is an optimal limit distribution based on the convolution theorem [34]. In this section, we informally demonstrate the importance of the efficient influence curve as the random variable whose variance characterizes the normal limit distribution of an MLE of the target parameter for our semiparametric model for $N \to \infty$, and thereby characterizes the normal limit distribution of optimal estimators. As part of this we use a template for establishing the normal limit distribution of the MLE, which can be equally well applied to the TMLE.

Even though it is well known that a regular estimator based on sample of $n$ i.i.d. observations is efficient if and only if it is asymptotically linear with influence curve equal to the efficient influence curve, here we are not interested in asymptotics when we observe $n$ of our data structures that are indexed by this parameter $N$ (like observing an i.i.d. sample $O_1, \ldots, O_n$, where each $O_i$ describes the data on $N$ causally connected units), but we are interested in the asymptotics in $N$ based on a single draw of $O$. Therefore, we think it is important to point out the asymptotic behavior of the MLE based on such a single $O^N$ when $N \to \infty$, showing that the asymptotic variance of the MLE is still characterized by the efficient influence curve. Our lesson is that our goal should still be to construct an estimator that is asymptotically normally distributed with variance equal to the variance of the efficient influence curve, appropriately normalized, and our proposed TMLE achieves this goal by using least-favorable submodels whose score span the efficient influence curve.

Specifically, we show that, under appropriate regularity conditions required for an MLE to be valid (i.e. all observables are discrete, so that MLE is well defined asymptotically), the asymptotic variance of a standardized MLE $\sqrt{N}(\psi_N - \psi_0)$ of the target parameter equals the limit in $N$ of $NP_0\{D^*(Q_0, g_0)\}^2$, where $P_0\{D^*(Q_0, g_0)\}^2$ is the variance of the efficient influence curve $D^*(Q_0, g_0)$. The formal analysis of an MLE requires understanding of an empirical process $(Z_N(Q) : Q)$ (specified below) uniformly in $Q$, which is challenging due to the fact that, contrary to $Z_N(Q_0)$, at misspecified $Q$, the time-specific components of $Z_N(Q)$ cannot be represented as sums of independent random variables, conditional on the history at that time. Since the TMLE is tailored to deal with the curse of dimensionality (and MLE is a special case of TMLE by defining the initial estimator for the TMLE as the MLE, assuming this MLE is a well-defined estimator), while a regularized MLE will *not* be asymptotically normally distributed when the observables are continuous valued, the analysis of a TMLE is more important. Such a formal analysis is presented for the point-treatment $K = 0$ case in a later section and much can be learned from that analysis for the purpose of analyzing the TMLE or MLE for general $K$. Nonetheless, the template below can be used to establish the asymptotic normality for both the MLE and also for the TMLE under the assumption that initial estimator $Q_N, g_N$ is consistent for $Q_0, g_0$.

Let $Q_N$ be an MLE, assuming it is well defined for $N$ large enough (i.e. all covariates are discrete). We wish to analyze the plug-in MLE $\Psi(Q_N)$ of $\psi_0$. We can represent the efficient influence curve as $D^*(Q, g) = D^*(Q, h(Q, g), \Psi(Q))$ for some parameter $h(Q, g)$, as shown in our technical report. In our accompanying technical report we show that $P_0 D^*(Q_N, h(g_0, Q_N), \Psi(Q_N)) = (\psi_0 - \Psi(Q_N)) + R_N$, where $R_N$ is a second-order term defined as sum of two terms $R_N(Q_N, Q_0)$ and $R_N((h_N, h_0), (Q_N, Q_0))$. The first involves square differences of $Q_N, Q_0$, while the second involves the product of differences $h(g_0, Q_N) - h(g_0, Q_0)$ and $Q_N - Q_0$. We will assume that $R_N = o_P(1/\sqrt{N})$, which basically corresponds with assuming that relevant parts of $Q_0$ are estimated by $Q_N$ at a rate faster than $N^{-1/4}$. Since $Q_N$ is an MLE, and $D^*(Q_N, h(g_0, Q_N), \Psi(Q_N))$ is a score at $P_{Q_N, g_0}$, we have that the MLE solves the efficient influence curve equation

$$D^*(Q_N, h(g_0, Q_N), \Psi(Q_N)) = 0.$$

We also have $P_0D^*(Q_0,h,\psi_0)=0$ for any $h$, as explicitly shown in our technical report. This allows us to establish a first-order expansion of the standardized MLE:

$$\begin{aligned}(\Psi(Q_N)-\psi_0) &= -P_0D^*(Q_N,h(g_0,Q_N),\Psi(Q_N))+R_N\\ &= D^*(Q_N,h(g_0,Q_N),\Psi(Q_N))-P_0D^*(Q_N,h(g_0,Q_N),\Psi(Q_N))+R_N.\end{aligned}$$

Thus under the assumption that $R_N=o_P(1/\sqrt{N})$, it follows that the asymptotic distribution of $\sqrt{N}(\Psi(Q_N)-\Psi(Q_0))$ equals the limit distribution of

$$Z_N(Q_N)\equiv\sqrt{N}(D^*(Q_N,h(g_0,Q_N),\Psi(Q_N))-P_0D^*(Q_N,h(g_0,Q_N),\Psi(Q_N))).$$

A non-trivial analysis as carried out for the case $\tau=0$, and using appropriate conditions, can be used to establish that $Z_N(Q_N)-Z_N(Q_0)=o_P(1)$, so that $Z_N(Q_N)$ behaves as $Z_N(Q_0)=\sqrt{N}(D^*(Q_0,h_0,\psi_0)-P_0D^*(Q_0,h_0,\psi_0))$. Under these assumptions, it then remains to investigate weak convergence of $Z_N(Q_0)$ as $N$ converges to infinity.

In our technical report, we establish the following representation of the efficient influence curve:

$$D^*(Q_0,g_0)=\tfrac{1}{N}\sum_i D^*_{L_i(0)}(Q_0)(L_i(0))+\tfrac{1}{N}\sum_t\sum_j\tfrac{1}{N}\sum_m\tfrac{h^*_{t,m}}{\bar h_t}(C^L_{t,j})\sum_l D_{l,t,m}(L_j(t),C^L_{t,j}),$$

where

$$D_{l,t,m}=E(Y(l)|L_m(t)=L_j(t),C^L_{t,m}=C^L_{t,j})-E(Y(l)|C^L_{t,m}=C^L_{t,j}),$$

$h^*_{t,m}(c)=P_{Q_0,g^*}(C^L_{t,m}=c)$, $h_{t,m}(c)=P_{Q_0,g_0}(C^L_{t,m}=c)$, and $\bar h_t=\frac{1}{N}\sum_m h_{t,m}$. Here, we assumed that $L_i(0)$, $i=1,\ldots,N$, are independent. Thus, we can represent the efficient influence curve as

$$D^*(Q_0,g_0)=\frac{1}{N}\sum_i D^*_{L_i(0)}(Q_0)(L_i(0))+\frac{1}{N}\sum_t\sum_j D^*_t(Q_0,g_0)(L_j(t),C^L_{t,j}),$$

where we defined

$$D^*_t(Q_0,g_0)(L_j(t),C^L_{t,j})=\frac{1}{N}\sum_m\frac{h^*_{t,m}}{\bar h_t}(C^L_{t,j})\sum_l D_{l,t,m}(L_j(t),C^L_{t,j}).$$

Note that $D^*_t(Q_0,g_0)$ has conditional mean zero, given $C^L_{t,j}$. In order to claim that $D^*_t(Q_0,g_0)$ has finite variance one needs that the summation over $l$ reduces essentially to a finite sum due to $L_m(t)$ being conditionally independent of $Y(l)$, given $C^L_{t,m}$, for most $m$.

This yields the following representation (suppressing the dependence of $D^*$ on $P_0$):

$$Z_N(Q_0)=\frac{1}{\sqrt{N}}\sum_i D_{L_i(0)}(L_i(0))+\frac{1}{\sqrt{N}}\sum_{t,i}D^*_t(L_i(t),C^L_{t,i})),$$

where $D^*_t$ is a function of $L_i(t)$ and $C^L_{t,i}$ with conditional mean zero, given $C^L_{t,i}$. Due to factorization of the likelihood in terms of $\prod_{t,i}\bar Q_t(L_i(t)|C^L_{t,i})$ and that $D^*_t$ is a score of $\bar Q_t$, it follows that $Z_N(Q_0)$ is an orthogonal sum over $t,i$ in $L^2_0(P_0)$, so that the variance of $Z_N(Q_0)$ is given by

$$\mathrm{VARZ_N}(Q_0)=\frac{1}{N}\sum_i P_0D_{L_i(0)}(L_i(0))^2+\frac{1}{N}\sum_{t,i}P_0\{D^*_t(L_i(t),C^L_{t,i})\}^2.$$

We have

$$P_0\{D^*_t(L_i(t),C^L_{t,i})\}^2=\int_{l(t),c(t)}D^*_t(l(t),c(t))^2\bar Q_{0,t}(l(t)|c(t))h_{t,i}(c(t)).$$

Thus, the asymptotic variance of $Z_N(Q_0)$ is given by limit of

$$\sigma_0^2\equiv\sigma^2_{L(0)}+\lim_{N\to\infty}\sum_t\int_{l(t),c(t)}\{D^*_t(l(t),c(t))\}^2\bar Q_{0,t}(l(t)|c(t))\bar h_t(c(t)),$$

where $\bar{h}_t = \frac{1}{N}\sum_i h_{t,i}$, and it can be expected that $\bar{h}_t$ converges to a fixed function as $N \to \infty$. Here we defined

$$\sigma^2_{L(0)} = \lim_{N\to\infty} \frac{1}{N}\sum_i P_0 D_{L_i(0)}(L_i(0))^2.$$

Note that we also have that $\sigma_0^2 = \lim_{N\to\infty} NP_0\{D^*\}^2$ equals $N$ times the variance of the efficient influence curve $D^* = D^*(P_0)$ for the target parameter $\Psi$ at $P_0$. This demonstrates that the asymptotic variance of $Z_N(Q_0)$ (and thus the asymptotic variance of the standardized MLE $\sqrt{N}(\Psi(Q_N) - \psi_0)$ is given by $\sigma_0^2 = \lim_{N\to\infty} NP_0\{D^*(P_0)\}^2$.

This does not demonstrate the asymptotic normality of the MLE yet. For that purpose, we note that $Z_N(Q_0) = Z_{L(0),N} + \sum_t Z_{N,t}(Q_0)$, where $Z_{N,t}(Q_0) = 1/\sqrt{N}\sum_i D^*_{t,i}(Q_0)$, with $D^*_{t,i} = D^*_t(L_i(t), C^L_{t,i})$, is a sum of independent random variables $L_i(t)$, conditional on $\bar{A}(t-1), \bar{L}(t-1)$. As a consequence of the latter, it follows that $EZ_{N,t_1}(Q_0)Z_{N,t_2}(Q_0) = 0$ for $t_1 < t_2$ (i.e. just condition on $\bar{A}(t_2-1), \bar{L}(t_2-1)$, making $Z_{N,t_1}$-fixed, and use that $E(Z_{N,t_2}(Q_0)|\bar{A}(t_2-1), \bar{L}(t_2-1)) = 0$). Using CLTs, we can therefore establish that for each $t = 0, \ldots, \tau+1$, $Z_{N,t}(Q_0)$ converges weakly to a normal distribution $Z_t(Q_0)$. Under weak regularity conditions, this also implies that $E(Z_{t_1}(Q_0)Z_{t_2}(Q_0)) = 0$ for $t_1 < t_2$ and thus that these $t$-specific limit normally distributed random variables $Z_t(Q_0)$ are pairwise independent. As a consequence, the sum across $t$ converges to a normal distribution with variance equal to the sum of the $t$-specific variances, and thus $\sigma_0^2$ as defined above. To conclude, under appropriate regularity conditions, we will have that $\sqrt{N}(\Psi(Q_N) - \psi_0) \approx Z_N(Q_0)$ converges weakly to $N(0, \sigma_0^2)$

This demonstrates that the efficient influence curve characterizes the limit distribution of the maximum likelihood estimator, and thus indeed characterizes an asymptotically optimal mean zero normal limit distribution with variance equal to the asymptotic variance of the "efficient influence curve empirical process" $Z_N(Q_0)$.

# 6 The TMLE of causal effect of single time-point intervention

We will present the TMLE for the point-treatment intervention case (i.e. $\tau = 0$). This case is of great interest itself, extends estimation of a causal effect of a single time-point intervention to dependent data of the form studied in this article, and thereby covers important applications. In the next section, we will formally analyze this TMLE. The tools of the proof will be generalizable to the general $\tau$ case. In addition, the single time-point case allows for a TMLE that is actually double robust in the sense that it remains consistent if either $Q_0$ or $a\ h_0(Q_0, g_0)$ is consistently estimated, while the efficient influence curve for the general case with $\tau > 0$ appears to not satisfy such a double robustness result as is evident from the efficient influence curve representation provided in our technical report van der Laan [62].

## 6.1 Structural equation model

Using notation $W_i$ for the baseline covariate $L_i(0)$, and $A_i$ for $A_i(0)$, the structural equation model for the $\tau = 0$ case reduces now to

$$\begin{aligned} W_i &= L_i(0) = f_{W_i}(U_{W_i}) \\ A_i &= A_i(0) = f_A(c_i^A(W), U_{A_i}) \\ Y_i &= L_i(1) = f_Y(c_i^Y(W,A), U_{Y_i}) \\ i &= 1, \ldots, N, \end{aligned}$$

where the fixed-dimensional summary measures $c_i^A(W)$ and $c_i^Y(W,A)$ are determined by $W = (W_1, \ldots, W_N)$ and $(W,A)$ with $A = (A_1, \ldots, A_N)$, respectively. We assume throughout that $A$ is discrete valued, so that conditional densities of $A$, given $W$, are just conditional probability distributions: this is by no means a necessary condition, but simplifies presentation. The "friends" $F_i$ of subject $i$ may be included in

$W_i$: $F_i \subset W_i$. The function $c_i^A(W)$ includes $W_i$, beyond summary measures of $(W_j : j \in F_i)$ and might be defined as $c_i^A(W) = (W_i, (W_j : j \in F_i))$, assuming that $|F_i| \leq K < \infty$ for some fixed $K$, so that $c_i^A(W)$ can indeed be defined as a fixed multivariate dimensional function not depending on $N$. Similarly, the function $c_i^Y(W, A)$ includes $(W_i, A_i)$ beyond summary measures of $((W_j, A_j) : j \in F_i)$ and might be defined as $c_i^Y(W, A) = (W_i, A_i, ((W_j, A_j) : j \in F_i))$. We also use the short-hand notation $C_i^Y = c_i^Y(A, W)$ and $C_i^A = c_i^A(W)$. The above structural equation model assumes that $A_i$ and $Y_i$ are the same function of this dimension reduction $(W_i, (W_j : j \in F_i))$ and $(W_i, A_i, (W_j, A_j : j \in F_i))$, respectively, for each $i$, so that two units with the same number of friends who have the same individual covariate and treatment values, and also have the same values for the covariates and treatments of their friends, will be subjected to the same conditional distribution for drawing their treatment and outcome. In our asymptotics theorem in the next section, we treat $F_i$, $i = 1, \ldots, N$, as fixed, so that also the probability distribution of $O$ and the target parameter $\psi_0$ are indexed by the fixed value of $(F_i : i = 1, \ldots, N)$.

In addition, we assume that conditional on $W$, (1) $(U_{A_i}, U_{Y_i})$, $i = 1, \ldots, N$, are i.i.d. and (2) for each $i$, $U_{A_i}$ is independent of $U_{Y_i}$. In one model, we assume that $U_{W_i}$, $i = 1, \ldots, N$, are i.i.d.: note that (since $f_{W_i}$ is allowed to be different for each $i$) this corresponds with assuming that $W_1, \ldots, W_N$ are independent, but not necessarily identically distributed. We will highlight the case that this latter assumption is considerably weakened, which will be made explicit in our theorem. These independence assumptions on the $U_i$'s imply that (1) $W_1, \ldots, W_N$ are independent (or more generally, their dependence is weak enough), (2) conditional on $W = (W_1, \ldots, W_N)$, $A_1, \ldots, A_N$ are independent, and (3) conditional on $(W, A)$, $Y_1, \ldots, Y_N$ are independent. Thus, all the dependence between units is explained by the observed pasts of the units themselves and of their friends.

**Causal quantity:** Let $g^*$ be a user-supplied conditional distribution of $A$, given $W$, and let us denote the random variable with this distribution with $A_* = (A_{1*}, \ldots, A_{N*})$. For simplicity, let us assume that under this $g^* A_{i*}$ are conditionally independent, given $W$, and that $g_i^*(A_{i,*}|W) = \bar{g}^*(A_{i,*}|C_i^{A,*})$ for a common conditional density $\bar{g}^*$ and summary measure $C_i^{A,*} = c_i^{A,*}(W)$. Our goal is to estimate the mean of the counterfactual outcome of $\bar{Y} = 1/N \sum_{i=1}^N Y_i$ under the stochastic intervention $g^*$. Let $Y_{g^*} = (Y_{g^*,i} : i = 1, \ldots, N)$ be the counterfactual indexed by a stochastic intervention $g^*$ on $A$ and $\bar{Y}_{g^*} = 1/N \sum_{i=1}^N Y_{g^*,i}$. The causal quantity of interest is defined as $\Psi^F(P_{U,W,A,Y}) = E_{P_{U,W,A,Y}} \bar{Y}_{g^*}$, which is a parameter of the distribution of $(U, W, A, Y)$ modeled by the above structural equation model. In this expectation defining $\Psi^F$, we actually condition on the vector $F = (F_1, \ldots, F_N)$ of sets of friends.

**Identifiability from observed data distribution:** We observe $O = (O_1, \ldots, O_N)$, where $O_i = (W_i, A_i, Y_i)$. Due to the above assumptions, the probability distribution of $O$ is given by:

$$P(O) = Q_W(W) \prod_{i=1}^N \bar{g}(A_i|C_i^A) \bar{Q}_Y(Y_i|C_i^Y), \tag{7}$$

where $\bar{g}(\cdot|c^A)$ is a common (in $i$) density for $A_i$ for each value $c^A$, and $\bar{Q}_Y(\cdot|c^Y)$ is a common density for $Y_i$ for each value $c^Y$. Our model also implies a model $\mathcal{Q}_W$ on the distribution $Q_W$ of $W$, such as the model that assumes that all $W_i$ are independent.

Since our assumptions imply the randomization assumption stating that $A = (A_1, \ldots, A_N)$ is independent of $U_Y = (U_{Y_i} : i = 1, \ldots, N)$, given $W = (W_1, \ldots, W_N)$, the post-intervention probability distribution $P_{g^*}$ of $(W, Y_{g^*}) = (W_i, Y_{i,g^*} : i = 1, \ldots, N)$ is identified by the following $G$-computation formula applied to the probability distribution $P$ of $O$:

$$\begin{aligned} P_{g^*}(W, A_*, Y) &= Q_W(W) \prod_{i=1}^N \bar{Q}_Y(Y_i|C_i^{Y,*}) \bar{g}^*(A_{i,*}|C_i^{A,*}) \\ &\equiv P^{g^*}(W, A_*, Y), \end{aligned} \tag{8}$$

where $\bar{Q}_Y(\cdot|C_i^{Y,*})$ is defined as the conditional distribution of $Y_i$, given $c_i^Y(A_*, W)$ with $A$ in the parents $c_i^Y(A, W)$ replaced by $A_*$. We denoted the probability distribution of on the right-hand side with $P^{g^*}$, which is thus always defined as a parameter of the data distribution $P$ of $O$ for a $P$ in the statistical model for

$P = P(P_{U,O})$ implied by our causal model for the underlying distribution $P_{U,O}$. The random variable with distribution $P^{g^*}$ is denoted with $(W, A_*, Y^{g^*})$.

**Statistical model, statistical target parameter, and statistical estimation problem:** Let $\mathcal{M}$ be the statistical model for the data distribution $P$ of $O$ defined by (eq. 7) in which $Q_W \in \mathcal{Q}_W$ for a specified model $\mathcal{Q}_W$, and the common $\bar{g} \in \mathcal{G}$ for some model $\mathcal{G}$, while $\bar{Q}_Y$ is unspecified. Thus, the density of $O$ factorizes in three factors:

$$P(O) = Q_W(W) \prod_{i=1}^{N} \bar{Q}_Y(Y_i|C_i^Y) \prod_{i=1}^{N} \bar{g}(A_i|C_i^A),$$

where $Q_W \in \mathcal{Q}_W$, $\bar{Q}_Y$ is unspecified, and $\bar{g} \in \mathcal{G}$. This defines the statistical model $\mathcal{M}$.

Let the statistical target parameter mapping $\Psi : \mathcal{M} \to \mathbb{R}$ be defined as $\Psi(P) = E_{P^{g^*}} \bar{Y}^{g^*}$. Under the stated causal model and identifiability assumptions under which $P = P_{P_{U,W,A,Y}}$, we have $\Psi(P) = \Psi^F(P_{U,W,A,Y})$, so that in that case $\Psi(P)$ can be interpreted as the desired causal quantity. Our goal is to construct an estimator of $\psi_0 = \Psi(P_0)$ based on $O = (O_1, \ldots, O_N) \sim P_0 \in \mathcal{M}$, which defines the statistical estimation problem.

Let $\bar{Q}(C^Y) = \int y \bar{Q}_Y(y|C^Y) d\mu(y)$ be the conditional mean under $\bar{Q}_Y$. Note that $E(Y_i|A, W) = \bar{Q}(C_i^Y)$. The target parameter $\Psi(P)$ only depends on $P$ through $Q_W$, and $\bar{Q}$:

$$\begin{aligned} \psi_0 &= E_0 \bar{Y}^{g^*} \\ &= \Psi(\bar{Q}_0, Q_{W,0}) \\ &\equiv \frac{1}{N} \sum_{j=1}^{N} \int_{a,w} \bar{Q}_0(c_j^Y(a,w)) g^*(a \mid w) Q_{W,0}(dw), \end{aligned} \tag{9}$$

where $Q_{W,0}(dw) = Q_{W,0}(w) d\mu_W(w)$ denotes integration w.r.t. measure implied by density $Q_{W,0}$ w.r.t. some dominating measure $\mu_W$. If we want to emphasize that $\Psi(P)$ only depends on $P$ through $Q(P) = (Q_W, \bar{Q})$, then we will also use (and abuse) the notation $\Psi(Q)$ to indicate the mapping from $Q$ into the desired estimand.

## 6.2 Efficient influence curve

In our technical report van der Laan [62] we established a general representation of the efficient influence curve of $E\bar{Y}_{g^*}$ for the longitudinal data structure and the model $\mathcal{Q}_W$ that assumes that the baseline covariates $L_1(0), \ldots, L_N(0)$ are independent, and it is given by:

$$\begin{aligned} D^*(Q,g) = &\sum_{j=1}^{N} \left\{ E_{Q,g^*}(\bar{Y} \mid L_j(0)) - E_{Q,g^*} \bar{Y} \right\} \\ &+ \frac{1}{N} \sum_{t=1, j,m} \frac{h_{t,m}^*}{\bar{h}_t}(C_{t,j}^L) \left\{ E_{Q,g^*}(\bar{Y}|L_m(t) = L_j(t), C_{t,m}^L = C_{t,j}^L) - E_{Q,g^*}(\bar{Y}|C_{t,m}^L = C_{t,j}^L) \right\}. \end{aligned}$$

For a different model for the covariate distribution of $L(0)$, only the first component would be different. In our case, we have $\tau = 0$, giving the following two terms:

$$\begin{aligned} D^*(Q,g) = &\sum_{j=1}^{N} \{ E_{Q,g^*}(\bar{Y}|L_j(0)) - E_{Q,g^*} \bar{Y} \} \\ &+ \sum_{j=1}^{N} \frac{1}{N} \sum_{m=1}^{N} \frac{h_m^*}{\bar{h}}(C_j^Y) \left\{ E_{Q,g^*}(\bar{Y}|Y_m = Y_j, C_m^Y = C_j^Y) - E_{Q,g^*}(\bar{Y}|C_m^Y = C_j^Y) \right\}, \end{aligned}$$

where $h_m^*(c) = P_{Q,g^*}(C_m^Y = c)$, $h_m(c) = P_{Q,g}(C_m^Y = c)$, and $\bar{h} = \frac{1}{N} \sum_{m=1}^{n} h_m(c)$ are densities w.r.t. some appropriate dominating measure $\mu$. We have, using short-hand notation $E_*$ for $E_{Q,g^*}$,

$$
\begin{aligned}
E_{Q,g^*}(\bar{Y}|Y_m, C_m^Y) &= \frac{1}{N}\sum_{j\neq m} E_*(Y_j|Y_m, C_m^Y) + 1/NY_m \\
&= \frac{1}{N}\sum_{j\neq m} E_*(E_*(Y_j|Y_m, W, A)|Y_m, C_m^Y) + 1/NY_m \\
&= \frac{1}{N}\sum_{j\neq m} E_*E_*(Y_j|W, A)|Y_m, C_m^Y) + 1/NY_m \\
&= \frac{1}{N}\sum_{j\neq m} E_*(\bar{Q}(C_j^Y(W, A))|Y_m, C_m^Y) + 1/NY_m,
\end{aligned}
$$

and

$$
\begin{aligned}
P(W, A|Y_m, C_m^Y) &= I(c_m^Y(W, A) = C_m^Y)\frac{P(W, A, Y_m)}{P(Y_m, C_m^Y)} \\
&= I(c_m^Y(W, A) = C_m^Y)\frac{P(Y_m|W, A)P(W, A))}{P(Y_m|C_m^Y)P(C_m^Y)} \\
&= I(c_m^Y(W, A) = C_m^Y)\frac{P(Y_m|C_m^Y)P(W, A))}{P(Y_m|C_m^Y)P(C_m^Y)} \\
&= P(W, A|C_m^Y),
\end{aligned}
$$

and thereby

$$
E^*(\bar{Y}|Y_m, C_m^Y) = \frac{1}{N}\sum_{j\neq m} E^*(\bar{Q}(c_j^Y(W, A))|C_m^Y) + 1/NY_m.
$$

Thus,

$$
E^*(\bar{Y}|C_m^Y) = \frac{1}{N}\sum_{j\neq m} E^*(\bar{Q}(c_j^Y(W, A))|C_m^Y) + 1/N\bar{Q}(C_m^Y).
$$

Therefore,

$$
\left\{E_{Q,g^*}(\bar{Y}|Y(m) = Y(j), C_m^Y = C_j^Y) - E_{Q,g^*}(\bar{Y}|C_m^Y = C_j^Y)\right\} = \tfrac{1}{N}\{Y_j - \bar{Q}(C_j^Y)\},
$$

which does thus not depend on $m$.

This proves the following representation of the efficient influence curve in the case $\tau = 0$:

$$
\begin{aligned}
D^*(Q, g) = &\sum\nolimits_{j=1}^N \{E_{Q,g^*}(\bar{Y}|L_j(0)) - E_{Q,g^*}\bar{Y}\} \\
&+ \frac{1}{N}\sum\nolimits_{j=1}^N \frac{\bar{h}^*}{\bar{h}}(C_j^Y)\left\{Y_j - \bar{Q}(C_j^Y)\right\}.
\end{aligned}
$$

We will state this result and the double robustness of the efficient influence curve in the following theorem.

**Theorem 1** *Consider the model $\mathcal{M}$ in which $W_1, \ldots, W_N$ are assumed to be independent. The efficient influence curve $D^*(P)$ at $P \in \mathcal{M}$ of target parameter $\Psi : \mathcal{M} \to \mathbb{R}$ is given by*

$$
\begin{aligned}
D^*(P) &= \sum\nolimits_{i=1}^N D_{W_i}^*(Q_W, \bar{Q})(W_i) + \sum\nolimits_{i=1}^N \frac{1}{N}\frac{\bar{h}(g^*, Q_W)(C_i^Y)}{\bar{h}(g, Q_W)(C_i^Y)}(Y_i - \bar{Q}(C_i^Y)), \\
&\equiv D_W^*(P) + D_Y^*(P)
\end{aligned}
$$

*where*

$$
\begin{aligned}
&D_{W_i}^*(Q_W, \bar{Q})(W_i) = E_{Q,g^*}(\bar{Y}|W_i) - E_{Q,g^*}(\bar{Y}) \\
&= \frac{1}{N}\sum_{j=1}^N \int_{a,w_{-i}} g^*(a|w_{-i}, W_i)\bar{Q}(c_j^Y(a, w_{-i}, W_i))\prod_{l\neq i} Q_{W_l}(w_l) - \psi \\
&= \frac{1}{N}\sum_{j=1}^N \{E(Y_j^{g^*}|W_i) - E_{W_i}E(Y_j^{g^*}|W_i)\},
\end{aligned}
$$

$$h_i(g,Q_W)(c) \equiv \int_{a,w,c_i^Y(a,w)=c} g(a|w) \prod_{l=1}^{N} Q_W(dw_l) = E_W g_i(c|W),$$

*and* $g_i(c|W=w) = P_0(c_i^Y(A,W)=c|W=w)$ *is the conditional probability that* $c_i^Y(A,W)$ *equals c, given* $W=w$*, which is a probability determined by* $g(A|W)$*. In addition,* $\bar{h} = \frac{1}{N}\sum_i h_i$ *and* $\bar{h}^* = \frac{1}{N}\sum_i h_i^*$ *with* $h_i^* = h_i(g^*, Q_W)$ *are densities defined w.r.t. a dominating measure* $\mu$ *and it is assumed that* $\bar{h}^*/\bar{h}$ *is uniformly bounded on a set that contains with probability 1* $C_i^Y$ *for all i.*

**Double robustness of efficient influence curve:** *Represent the efficient influence curve as* $D^*(\bar{Q}, Q_W, g) = D_W^*(Q_W, \bar{Q}) + D_Y^*(Q_W, \bar{Q}, g)$*. We have*

$$P_0 D_W^*(Q_{W,0}, \bar{Q}) = 0,$$

$$P_0 D_Y^*(\bar{Q}, Q_{W,0}, g_0) = \psi_0 - \Psi(\bar{Q}, Q_{W,0}),$$

*so that*

$$P_0 D^*(\bar{Q}, Q_{W,0}, g_0) = \psi_0 - \Psi(\bar{Q}, Q_{W,0}).$$

*Since the efficient influence curve at* $P_0$ *depends on* $g_0$ *only through* $\bar{h}(g_0, Q_{W,0})$*, we have that if* $\bar{h}(g, Q_{W,0}) = \bar{h}(g_0, Q_{W,0})$*, then*

$$P_0 D_Y^*(\bar{Q}, Q_{W,0}, g_0) = \psi_0 - \Psi(\bar{Q}, Q_{W,0}),$$

*and thus*

$$P_0 D^*(\bar{Q}, Q_{W,0}, g) = \psi_0 - \Psi(\bar{Q}, Q_{W,0}).$$

*Let* $P_0^W$ *denote the conditional distribution of O, given W, and let* $Q_{W,N}$ *be the degenerate distribution of W that puts mass 1 on W. We also note that*

$$P_0^W D_Y^*(\bar{Q}, Q_{W,N}, g_0) = \Psi(\bar{Q}_0, Q_{W,N}) - \Psi(\bar{Q}, Q_{W,N}). \tag{10}$$

*We also have that for all g,*

$$P_0 D_Y^*(\bar{Q}_0, Q_{W,0}, g) = 0.$$

**Explicit proof of double robustness:** Even though our general theorem in the technical report can be applied to this single time-point case and this double robustness result follows by noting that the second-order term $R(Q, Q_0)$ in that theorem equals 0, here we provide an explicit proof of the stated double robustness for this single time-point case. Firstly, we have

$$\begin{aligned} E_0 D_{W_i}^*(Q_{W,0}, \bar{Q})(W_i) &= \frac{1}{N}\sum_{j=1}^{N} \int_{a,w} g^*(a|w)\bar{Q}(c_j^Y(a,w)) Q_{W,0}(dw) - \Psi(\bar{Q}, Q_{W,0}) \\ &= 0. \end{aligned}$$

We also have

$$\begin{aligned} E_0 \sum_i D_{Y_i}^*(\bar{Q}, Q_{W,0}, g_0) &= \frac{1}{N}\sum_i E_0 \frac{\bar{h}_0^*(C_i^Y)}{\bar{h}_0(C_i^Y)}(Y_i - \bar{Q}(C_i^Y)) \\ &= \frac{1}{N}\sum_i \int \frac{\bar{h}_0^*(c)}{\bar{h}_0(c)}(\bar{Q}_0 - \bar{Q})(c) h_{i,0}(c) d\mu(c) \\ &= \int \bar{h}_0^*(c)(\bar{Q}_0 - \bar{Q})(c) d\mu(c) \\ &= \psi_0 - \Psi(\bar{Q}, Q_{W,0}). \end{aligned}$$

This derivation with $P_0$ replaced by $P_0^W$ also establishes (eq. 10).

This proves that with $D_i^* = D_{W_i}^* + D_{Y_i}^*$, $D^* = \sum_i D_i^*$, we have

$$E_0 \sum_i D_i^*(Q_{W,0}, \bar{Q}, g_0) = 0 + \psi_0 - \Psi(\bar{Q}, Q_{W,0}).$$

This proves the robustness w.r.t. misspecification of $\bar{Q}$. In addition, it follows trivially that $E_0 D^*(\bar{Q}_0, Q_{W,0}, g) = 0$ for any choice $g$.

## 6.3 Double robustness for an inefficient influence curve

In the following lemma, we present an inefficient influence curve and establish its double robustness. This could be used to construct an inefficient TMLE analogue to the efficient TMLE presented below.

**Lemma 1** *Suppose $C_i^Y(A, W)$ only depends on $A, W$ through $(A_i, W_i), ((A_j, W_j) : j \in F_i)$. For notational convenience, in this lemma let $F_i$ include $i$ itself: $i \in F_i$. Define the conditional probability densities $g_i^*(A_j : j \in F_i | W_j : j \in F_i)$ and $g_{i,0}(A_j : j \in F_i | W_j : j \in F_i)$, and define*

$$D_{Y_i,1}^*(\bar{Q}, Q_{W,0}, g_0) = \frac{1}{N} \frac{g_{i,0}^*(C_i^Y)}{g_{i,0}(C_i^Y)} (Y_i - \bar{Q}(C_i^Y)).$$

*Let $D_{i,1}^* = D_{W_i}^* + D_{Y_i,1}^*$ and $D_1^* = \sum_i D_{i,1}^*$. We have*

$$E_0 D_1^*(Q_{W,0}, \bar{Q}, g_0) = \psi_0 - \Psi(\bar{Q}, Q_{W,0}).$$

*We also have $E_0 D_1^*(Q_{W,0}, \bar{Q}_0, g) = 0$ for all $g$.*

**Proof:** We have

$$\begin{aligned}
E_0 \sum_i D_{Y_i}^*(\bar{Q}, Q_{W,0}, g_0) &= \frac{1}{N} \sum_i E_0 \frac{g_{i,0}^*(C_i^Y)}{g_{i,0}(C_i^Y)} (\bar{Q}_0(C_i^Y) - \bar{Q}(C_i^Y)) \\
&= \frac{1}{N} \sum_i E_0 \int_{a_j : j \in F_i} \frac{g_{i,0}^*(a_j : j \in F_i | W_j : j \in F_i)}{g_{i,0}(a_j : j \in F_i | W_j : j \in F_i)} (\bar{Q}_0 - \bar{Q})(a_j, W_j : j \in F_i) \\
&\quad g_{i,0}(a_j : j \in F_i | W_j : j \in F_i) \\
&= \frac{1}{N} \sum_i E_0 \int_{a_j : j \in F_i} g_{i,0}^*(a_j : j \in F_i | W_j : j \in F_i) (\bar{Q}_0 - \bar{Q})(a_j, W_j : j \in F_i) \\
&= \psi_0 - \Psi(\bar{Q}, Q_{W,0}). \ \square
\end{aligned}$$

## 6.4 Estimating equation approach

Consider the efficient influence curve and let us represent it as an estimating function in $\psi$:

$$D^*(Q, g, \psi) = \sum_{i=1}^N (D_{W_i}^*(Q) - \psi) + D_{Y_i}^*(Q, g_0),$$

where now $D_{W_i}^* = E_Q(\bar{Y}^{g^*} | W_i)$. We will represent it as $D^*(Q, \bar{h}, \psi)$ to stress that it only relies on $(Q, g)$ through $(Q, \bar{h}(Q, g))$. We have $E_0 D^*(Q, \bar{h}_0, \psi) = \psi_0 - \psi$, so that $D^*$ is a targeted estimating function for fitting $\psi_0$. Given an estimator $\bar{h}_N$ and $Q_N$ of $\bar{h}_0$ and $Q_0$, respectively, based on the data $O$, we can estimate $\psi$ with the solution of

$$0 = D^*(Q_N, \bar{h}_N, \psi)(O).$$

Since $D^*(Q, \bar{h}, \psi) = D^*(Q, \bar{h}) - N\psi$, this solution is given by

$$\psi_N = \frac{1}{N}\sum_{i=1}^{N}\{D^*_{W_i}(Q_N) + D^*_{Y_i}(Q_N, \bar{h}_N)\}.$$

This estimator, as the TMLE presented below, is double robust w.r.t. misspecification of $(\bar{h}_0, \bar{Q}_0)$ and is asymptotically efficient if both are estimated consistently, assuming the required regularity conditions hold (as presented in our theorem below). Since it is not a substitution estimator, it will be more sensitive to practical violations of the positivity assumptions due to $\bar{h}^*_N/\bar{h}_N$ being large.

**Remark regarding the balance of the two contributions in the efficient influence curve:** The factor $1/N$ in $D^*_{Y_i}$ might come as a surprise in relation to $D^*_{W_i}$. Let us consider the case that $g^*(a|w) = \prod_{i=1}^n g^*_i(a_i|w_i)$. To intuitively understand that this efficient influence curve does indeed represent a balance between these two contributions, we note the following:

$$\begin{aligned}\sum_i D^*_{W_i} &= \sum_i \{E_Q(\bar{Y}^{g^*}|W_i) - \Psi(Q)\} \\ &= \sum_{i=1}^N \left\{\frac{1}{N}\sum_{j=1}^N E_Q(Y^*_j|W_i) - \Psi(Q)\right\} \\ &= \frac{1}{N}\sum_{i=1}^N\sum_{j=1}^N \Big\{I(i \in F_j)(E_Q(Y^*_j|W_i) - \Psi_j(Q))\Big\},\end{aligned}$$

where $\Psi(Q) = 1/N\sum_j \Psi_j(Q)$, and $\Psi_j(Q) = E_Q Y^*_j$. Thus, indeed the contribution $\sum_i D^*_{W_i}$ is of the same size as function of $N$ as $\sum_i D^*_{Y_i}$, under the assumption that $|F_i| \leq K < \infty$ for some $K < \infty$, which is indeed an assumption we made to establish $\sqrt{N}$-asymptotics.

## 6.5 TMLE

Recall the target parameter representation $\Psi(Q_W, \bar{Q})$ defined by (eq. 9).

Let $\bar{Q}_N$ be an estimator of $\bar{Q}_0$, where $\bar{Q}_0(c) = E_0(Y_i|C^Y_i = c)$. Suppose $Y_i \in \{0,1\}$ or that $Y_i$ is continuous with values in $(0,1)$. This estimator $\bar{Q}_N$ could be based on the log-likelihood loss function

$$-L(\bar{Q})(O) = \sum_{i=1}^N \log\Big\{\bar{Q}(C^Y_i)^{Y_i}(1 - \bar{Q}(C^Y_i))^{1-Y_i}\Big\}.$$

For example, suppose that we assume a logistic regression model $\bar{Q}_\theta(c) = \frac{1}{1+\exp(-m_\theta(c))}$. Then we can estimate $\theta$ with the standard maximum likelihood based logistic regression estimator:

$$\theta_N = \arg\max_\theta \sum_{i=1}^N \log\Big\{\bar{Q}_\theta(C^Y_i)^{Y_i}(1 - \bar{Q}_\theta(C^Y_i))^{1-Y_i}\Big\}.$$

More generally, one can also use cross-validation based on this loss function and thereby estimate $\bar{Q}_0$ with an $L(\bar{Q})$-based super-learner. The super-learner takes as input a library of candidate logistic regression estimators (including machine learning algorithms) and uses cross-validation to select the optimal weighted-combination of this library of estimators, where the weight is obtained by minimizing the cross-validated risk based on this loss function. Thus, if one uses $V$-fold cross-validation, then one divides up the sample $(Y_i, C^Y_i)$, $i = 1, \ldots, N$, in $V$-subgroups, one defines one of the subgroups as validation sample, and the remainder as training sample. One then trains the $j$th algorithm on the $v$th training sample, and one evaluates the $v$-specific cross-validated risk $-\sum_{i \in Val(v)} \log \bar{Q}_{N,Tr(v),j}(C^Y_i)^{Y_i}(1 - \bar{Q}_{N,Tr(v),j}(C^Y_i))^{1-Y_i}$ for this $j$th algorithm. This is done for each choice of sample split $v \in \{1, \ldots, V\}$, and the $V$ cross-validated risks are averaged, giving a single cross-validated risk for the $j$th algorithm. One could now select the best choice $j_N$ by selecting the algorithm that has the smallest cross-validated risk. The estimator $\bar{Q}_{N,J_N}$ is referred to as the

discrete super-learner. Similarly, one can define a candidate algorithm $\bar{Q}_{N,\alpha} = \sum_j \alpha_j \bar{Q}_{N,j}$ for a vector of weights $\alpha$ and select the optimal choice $\alpha_N$ that minimizes the cross-validated risk of $\bar{Q}_{N,\alpha}$ over the choice $\alpha$. This estimator $\bar{Q}_{N,\alpha_N}$ is referred to as the super-learner. The estimator could also be based on a squared error loss function

$$L_2(\bar{Q})(O) = \sum_{i=1}^{N} (Y_i - \bar{Q}(C_i^Y))^2.$$

Let $\bar{Q}_{W,N}$ be a nonparametric maximum likelihood estimator of $Q_W \in \mathcal{Q}_W$, thus respecting the model $\mathcal{Q}_W$ for the joint distribution of $W_1, \ldots, W_N$. For example, if $W_i$ are i.i.d., then we would estimate this marginal distribution of $W_i$ with the empirical distribution of $(W_1, \ldots, W_N)$. If $W_1, \ldots, W_N$ are only known to be independent, then we would estimate each marginal distribution of $W_i$ with the discrete distribution that puts mass 1 on the singleton $W_i$, $i = 1, \ldots, N$: note that this empirical distribution is equivalent with the joint distribution that puts mass 1 on $(W_1, \ldots, W_N)$. If the model $\mathcal{Q}_W$ is larger than the independence model, then we would still estimate $Q_W$ with this degenerate distribution $Q_{W,N}$.

Given the estimator $\bar{Q}_N$ and $Q_{W,N}$ of $\bar{Q}_0$ and $Q_{W,0}$, one could now define a corresponding plug-in estimator $\Psi(\bar{Q}_N, Q_{W,N})$. However, the TMLE differs from this estimator using a targeted version $\bar{Q}_N^*$ of $\bar{Q}_N$ instead.

Let $\bar{g}_N$ be an estimator of $\bar{g}_0$, and let $g_N$ be the corresponding estimator of the conditional distribution $g_0$ of $A$, given $W$. Given the model assumption $g(A|W) = \prod_i \bar{g}(A_i|c_i^A(W))$ for a common conditional density $\bar{g}$, this estimator can be based on the log-likelihood loss:

$$L(\bar{g})(O) = -\sum_{i=1}^{N} \log \bar{g}(A_i|C_i^A).$$

As explained above, this could be a simple logistic regression estimator or a super-learner based on this loss function based on the sample $(A_i, C_i^A = c_i^A(W))$, $i = 1, \ldots, N$.

Given $\bar{g}_N$, $Q_{W,N}$, and $\bar{Q}_N$, let $\{\bar{Q}_N(\epsilon) : \epsilon\}$ be a target-parameter-specific submodel through $\bar{Q}_N$ defined by

$$\text{Logit}\bar{Q}_N(\varepsilon) = \text{Logit}\bar{Q}_N + \epsilon \frac{\bar{h}(g^*, Q_{W,N})}{\bar{h}(g_N, Q_{W,N})},$$

where $\bar{h}_N(c) = \frac{1}{N}\sum_{i=1}^N h_{i,N}(c)$, with $h_{i,N}(c) = P_{Q_{W,N},g_N}(c_i^Y(A,W) = c)$, and, similarly, $\bar{h}_N^*(c) = \frac{1}{N}\sum_{i=1}^N h_{i,N}^*(c)$, with $h_{i,N}^*(c) = P_{Q_{W,N},g^*}(c_i^Y(A_*,W) = c)$, all defined as densities w.r.t. a dominating measure $\mu$.

Let

$$\epsilon^N = \arg\min_{\epsilon} L(\bar{Q}_N(\epsilon))(O)$$

be the maximum likelihood estimator, which simply involves running univariate logistic regression on a pooled data set with outcomes $Y_i$ and covariate $\frac{\bar{h}(g^*,Q_{W,N})}{\bar{h}(g_N,Q_{W,N})}(C_i^Y)$, using as off-set $Logit\,\bar{Q}_N$. This defines now an update $\bar{Q}_N^* = \bar{Q}_N(\epsilon^N)$.

The TMLE of $\psi_0$ is defined as the corresponding plug-in estimator

$$\psi_N^* = \Psi(\bar{Q}_N^*, Q_{W,N}).$$

We note that this TMLE solves the efficient influence curve equation

$$D^*(\bar{Q}_N^*, Q_{W,N}, g_N, \psi_N^*)(O) = 0,$$

which is a key ingredient in our proof of asymptotic normality of $\psi_N^*$. Or, using the notation $\bar{h}_N = \bar{h}(Q_{W,N}, g_N)$, and $D^*(\bar{Q}, Q_W, \bar{h}, \psi)$, we can write this as

$$D^*(\bar{Q}_N^*, Q_{W,N}, \bar{h}_N, \psi_N^*) = 0.$$

Specifically, being a substitution estimator $\Psi(Q_N^*)$ and using an NPMLE of $Q_{W,0}$, we have $\sum_i D^*_{W_i}(Q_N^*) = 0$, while the targeted update $\bar{Q}_N^*$ of $\bar{Q}_N$ guarantees that

$$\sum_i D^*_{Y_i}(\bar{Q}_N^*, \bar{h}(Q_{W,N}, g_N)) = 0.$$

## 6.6 The clever covariate

Computation of the above TMLE requires the construction of an estimator $\bar{h}_N^*/\bar{h}_N$ of the clever covariate $\bar{h}_0^*/\bar{h}_0$ (density ratio). This estimator needs to be evaluated at $C_i^Y$ for each $i = 1, \ldots, N$, in order to compute the TMLE update $\bar{Q}_N^*$. In addition, since $\Psi(Q_{W,N}, \bar{Q}_N^*)$ involves integration of $\bar{Q}_N^*(C_i^Y(a^*, w))$ over any point in support of $W, A_*$ w.r.t. product measure $Q_{W,N} \times g^*$, we also need to evaluate $\bar{h}_N^*/\bar{h}_N$ at any such point. One possible estimator is a plug-in estimator

$$\frac{\bar{h}_N^*}{\bar{h}_N}(c) = \frac{\sum_i P_{Q_{W,N},g^*}(c_i^Y(A,W) = c)}{\sum_i P_{Q_{W,N},g_N}(c_i^Y(A_*,W) = c)},$$

obtained by plugging in our empirical counterpart $Q_{W,N}$ for $Q_{W,0}$, and an estimator $g_N$ of $g_0$. Let us consider the case that $Q_{W,N}$ puts mass 1 on $W$. In that case, this simplifies to

$$\frac{\sum_i \int_{a^*} I(c_i^Y(a^*,W) = c) g^*(a|W)}{\sum_i \int_{a^*} I(c_i^Y(a^*,W) = c) g_N(a|W)}.$$

In addition, one can use that $c_i^Y(a,w) = (a_j, w_j : j \in F_i)$ so that for each $i$, the integral only integrates over $(a_j : j \in F_i)$, where we used the convention that $i \in F_i$. Nonetheless, this type of implementation can easily be quite computationally overwhelming.

Therefore, we use this subsection to formulate insights about the clever covariate that will allow a much easier implementation of an estimator of this clever covariate. The basic idea is that we will directly estimate $\bar{h}_0$ instead of indirectly through plugging in estimators of $Q_W$ and $g_0$. These insights are formulated in the following lemma, where we consider the case that $C_i^Y = (W_j, A_j : j \in F_i)$ with $i \in F_i$.

**Lemma 2** *We note that $\bar{h}_0 = 1/N \sum_i h_{i,0}$ is a mixture of densities $h_{i,0}$ of $C_i^Y$ (living in a single space $C_Y$ common in i) and thus represents a density of a random variable which we will denote with $C^Y \in C_Y$. Suppose $C^Y = (W_j^c, A_j^c : j = 1, \ldots, k)$ for some k, representing covariates and treatment values of the subject and its friends.*

- *We have*

$$\bar{h}_0 = \arg\max_{\bar{h}} E_0 \sum_{i=1}^N \log \bar{h}(C_i^Y), \tag{11}$$

*where we maximize over a set of densities of $C^Y$ that contains the true $\bar{h}_0$.*

- *The density $\bar{h}_0$ can be factorized as*

$$\bar{h}_0(W_j^c, A_j^c : j = 1, \ldots, k) = \bar{g}_0^c(A_j^c : j | W_j^c : j) \bar{Q}_W^c(W_j^c : j),$$

*where $\bar{g}_0^c$ is the conditional density of $(A_j^c : j)$, given $(W_j^c : j)$, and $\bar{Q}_W^c$ is the marginal density of $(W_j^c : j)$, under the joint density $\bar{h}_0$.*

- *We also have*

$$\bar{g}_0^c = \arg\max_{\bar{g}^c} E_0 \sum_{i=1}^N \log \bar{g}^c(A_j : j \in F_i | W_j : j \in F_i), \tag{12}$$

*where we maximize over a set of conditional densities of $(A_j^c : j)$, given $(W_j^c : j)$, that contains the true $\bar{g}_0^c$, and,*

$$\bar{Q}^c_{W,0} = \arg\max_{\bar{Q}^c_W} E_0 \sum_{i=1}^{N} \log \bar{Q}^c_W(W_j : j \in F_i). \tag{13}$$

• *By the same arguments,* $\bar{h}^*_0 = \bar{g}^{*c}_0 \bar{Q}^c_{W,0}$, *where* $\bar{h}^*_0$ *is a density of random variable* $C^{Y,*} = (W^{c,*}_j, A^{c,*}_j : j)$,

$$\bar{h}^*_0 = \arg\max_{\bar{h}} E_{P_{Q_0,g^*}} \sum_{i=1}^{N} \log \bar{h}(C^{Y,*}_i), \tag{14}$$

$\bar{g}^{*c}_0$ *is the conditional density of* $(A^{c,*}_j : j)$, *given* $(W^{c,*}_j : j)$, *and* $\bar{Q}^c_{W,0}$ *is the marginal of* $(W^{c,*}_j : j)$, *under the joint density* $\bar{h}^*_0$. *The latter equals the* $\bar{Q}^c_{W,0}$ *defined above as the marginal density under* $\bar{h}_0$.

• *As a consequence, we can conclude that*

$$\frac{\bar{h}^*_0}{\bar{h}_0}(C^Y) = \frac{\bar{g}^{*c}_0}{\bar{g}^c_0}(C^Y). \tag{15}$$

Thus, the take home point of this lemma is (eq. 15) teaching us that we only need to estimate $\bar{g}^{*c}_0$ and $\bar{g}^c_0$, where $\bar{g}^c_0$ can be fitted as if we are estimating a conditional density of $(A^c_j : j)$, given $(W^c_j : j)$, based on data $(A_i, (A_j : j \in F_i))$, $(W_i, (W_j : j \in F_i))$, $i = 1, \ldots, N$, as if these $N$ observations are i.i.d. That is, an important practical implementation is to fit $\bar{g}^c_0$ with maximum likelihood based estimation, treating $C^Y_i$ as i.i.d., as if we are fitting the common conditional distribution of $(A_j : j \in F_i)$, given $(W_j : j \in F_i)$. For example, if $A_j$ is binary, $|F_i| = k$, then such a conditional distribution could be factorized in terms of a product of $k$ binary conditional distributions. Each of these binary conditional distributions can be fitted with logistic regression, possibly incorporating adaptive estimation. The asymptotic consistency of such a maximum likelihood based estimator, and the validity of cross-validation ignoring the dependence, would rely on $C^Y_i$ only being dependent on $C^Y_j$ for a finite (universal in $N$) number of $j \neq i$. Such an estimator yields an actual fitted function $\bar{g}^c_N$ that is easily evaluated at any required value.

Suppose now that $g_0$ is known, as in an RCT. The above-mentioned approach would ignore the knowledge on $g_0$ and is thus not necessarily appropriate. If $g_0$ is very simple, as if often the case in an RCT, then one might simply be able to show that $\bar{g}^c_0$ is known (e.g. if the randomization probability for $A_i$ does not depend on covariates) in which case there is no need to estimate $\bar{g}^c_0$. In such cases, one could also use a simple marginal empirical distribution for this conditional density $\bar{g}^c_0$ in the estimation procedure outlined in previous paragraph. Consider now the case that $g_0$ is known, but that it is a quite complex function. In that case, one could decide to simulate a very large number of $(W, A)$ from $(Q_{W,N}, g_0)$ and use an adaptive maximum likelihood based estimator of $\bar{g}^c_0$ based on this large sample using the method presented in previous paragraph. This maximum likelihood based estimator would obviously utilize that it is known that $g_0$ only depends on certain covariates, so that the estimator can be simplified as much as possible. That is, we use the above-described estimation procedure for estimation of $\bar{g}^c_0$, but now applied to a very large data set simulated from the distribution of $(W, A)$ under $Q_{W,N} \times g_0$. In this manner, one can still obtain excellent approximation of the true $\bar{g}^c_0$ that fully utilizes that we know the true $g_0$.

Let us now discuss estimation of $\bar{g}^{*c}_0$. Given that we know $g^*$, as above for the case that $g_0$ is known, one might either be able to determine $\bar{g}^{*c}_0$ (e.g. if the randomization probabilities of $A_{i,*}$ do not depend on covariates), and for complex $g^*$, we can simulate a very large number $(W, A_*)$ from $(Q_{W,N}, g^*)$, and use an adaptive maximum likelihood based estimator of $\bar{g}^{*c}_0$ based on this large sample using the above-described estimation procedure.

In this manner, we obtain a functional form that approximates $\bar{g}^c_0$ and $\bar{g}^{*c}_0$ well (by utilization of $g_0, g^*$ being known), and that one can evaluate for any $C^Y$. The TMLE $\bar{Q}^*_N$ can now be computed, and the target parameter evaluation $\Psi(\bar{Q}^*_N, Q_{W,N})$ as well.

Suppose now that $C^Y_i = (A_i, \bar{A}^c_i, (W_j : j \in F_i))$, where $\bar{A}^c$ is a summary measure of the treatment nodes $(A_j : j \in F_i)$ of the friends of subject $i$. In this case, by a simple generalization of the lemma above, it follows that $\bar{g}^c_0$ only involves fitting the conditional density of $(A_i, \bar{A}^c_i)$, given $(W_j : j \in F_i)$, treating these $i$-specific data points as i.i.d., as above. Thus, a reduction of the dependence of $C^Y_i$ on the treatment nodes (i.e. a

model assumption in our model) would result in a significantly less variable estimated clever covariate $\bar{h}_0^*/\bar{h}_0$, and the above method can still be applied. For example, one might feel that it is a reasonable assumption to assume that the mean outcome for unit $i$ depends on $A, W$ only through the treatment node for subject $i$ and the proportion of treated among the friends of $i$, beyond dependence on all the covariates.

# 7 Asymptotic normality of TMLE of counterfactual mean of single time-point stochastic intervention

In this section, we state a theorem establishing the asymptotics of the TMLE of $\psi_0$ under conditions. Subsequently, we discuss the implications of this theorem regarding statistical inference in terms of confidence intervals. The proof is deferred to the Appendix. In the Appendix of our technical report, we demonstrate that our proof is generalizable to the general longitudinal data structures. In this section, we define $F_i$ to include $i$ itself: i.e. $i \in F_i$.

**Theorem 2** *Consider the statistical formulation of data $O = (O_1, \ldots, O_N) \sim P_0 \in \mathcal{M}$, $O_i = (W_i, A_i, Y_i)$, statistical model $\mathcal{M}$, and statistical target parameter $\Psi : M \to \mathbb{R}$, all defined conditionally on the network-profile $F = (F_1, \ldots, F_N)$. Recall that this network-profile $F$ implies* that *$Y_i$ only depends on $(W, A)$ through $(W_j, A_j : j \in F_i)$ and that $A_i$ depends on $W$ through $(W_j : j \in F_i)$. Suppose $\bar{g}_0 \in \mathcal{G}$, and that $Q_{W,0} \in \mathcal{Q}_W$ satisfies an independence assumption specified below, and $\bar{Q}_{Y,0}$ is unspecified. A probability distribution of $O$ is thus parameterized* by *$Q_W, \bar{g}, \bar{Q}_Y$ as follows:*

$$P(O) = Q_W(W) \prod_{i=1}^{N} \bar{g}(A_i \mid C_i^A) \bar{Q}(Y_i \mid C_i^Y), \tag{16}$$

*where $C_i^Y = c_i^Y(A, W) \in \mathcal{C}^Y \subset \mathbb{R}^{d_1}$, $C_i^A = c_i^A(W) \in \mathcal{C}^A \subset \mathbb{R}^{d_2}$, $\bar{Q}_Y(\cdot|c)$ is a density for $Y$ for each possible $c \in \mathcal{C}^Y$, but is otherwise unspecified, $\bar{g}(\cdot|c)$ is a density for $A$ for each possible $c \in \mathcal{C}^A$, and $Q_W \in \mathcal{Q}_W$. This defines the statistical model $\mathcal{M}$ for the probability distribution of $O$.*

*For a specified stochastic intervention $g^*$, the target parameter $\Psi : \mathcal{M} \to \mathbb{R}$ is defined by*

$$\begin{aligned}\Psi(P_0) &= E_{P_0} \bar{Y}^{g^*} = \Psi(\bar{Q}_0, Q_{W,0}) \\ &= \frac{1}{N} \sum_{j=1}^{N} \int_{a,w} \bar{Q}_0(c_j^Y(a, w)) g^*(a|w) Q_{W,0}(dw),\end{aligned}$$

*where $Q_{W,0}(w) = P_0(W = w)$ (defined as density w.r.t. some dominating measure), $Q_{W,0}(dw) = Q_{W,0}(w) d\mu_W(w)$ denotes integration w.r.t. the measure implied by $Q_{W,0}$, $\bar{Q}_0(c_j^Y(A, W)) = E_0(Y_j|A, W)$, and $\bar{Q}_0(c) = \int_y y \bar{Q}_Y(dy|c)$ is the mean under density $\bar{Q}_Y(\cdot|c)$.*

*Let $D^*(\bar{Q}, Q_W, g_0)(O)$ be the efficient influence curve of $\Psi$ as defined in Theorem 1:*

$$D^*(\bar{Q}, Q_W, g) = \sum_{i=1}^{N} \{D_{W_i}^*(Q_W, \bar{Q})(W_i) + D_{Y_i}^*(\bar{Q}, Q_W, g)\},$$

*where*

$$D_{Y_i}^*(\bar{Q}, Q_W, g) = \frac{1}{N} \frac{\bar{h}(g^*, Q_W)(C_i^Y)}{\bar{h}(g, Q_W)(C_i^Y)} (Y_i - \bar{Q}(C_i^Y)),$$

*and $D_{W_i}^* = E(\bar{Y}^{g^*}|W_i) - \Psi(P)$. We will also denote these functions with $D^*(\bar{Q}, Q_W, \bar{h})$ and $D_{Y_i}^*(\bar{Q}, Q_W, \bar{h})$ to emphasize that they only depend on $g$ through $\bar{h}$. We use the definitions of $\bar{h}_0(c) = \frac{1}{N} \sum_{i=1}^{N} h_{0,i}(c)$, $\bar{h}_0^* = \frac{1}{N} \sum_{i=1}^{N} h_{0,i}^*$, $h_{0,i}(c) = P_{g_0, Q_{W,0}}(c_i^Y(A, W) = c)$, $h_{0,i}^*(c) = P_{g^*, Q_{W,0}}(c_i^Y(A, W) = c)$, defined as densities w.r.t. a dominating measure $\mu$, and let $h_0 = \bar{h}_0^*/\bar{h}_0$.*

*Let $Q_{W,N}$ be the distribution that puts mass 1 on $(W_1, \ldots, W_N)$. Consider the TMLE $\psi_N^* = \Psi(Q_N^*) = \Psi(\bar{Q}_N^*, Q_{W,N})$ defined above using $\bar{g}_N$ in $\bar{h}_N = \bar{h}(g_N, Q_{W,N})$. As shown above, this TMLE solves*

$$D^*(\bar{Q}_N^*, Q_{W,N}, \bar{h}_N)(O) = 0.$$

*Note that $\tilde{h}_N$ is a plug-in estimator of $\tilde{h}_0$ implied by $\bar{g}_N \in \mathcal{G}$ and $Q_{W,N}$.*

*We make the following assumptions:*

**Entropy condition:** *Consider a class $\mathcal{F}_Y$ of functions $c^Y \to \bar{Q}(c^Y)$ on a set in $\mathcal{C}^Y \subset \mathbb{R}^d$ that contains $c^Y(A, W)$ with probability 1. Assume that $\bar{Q}_N^* \in \mathcal{F}_Y$ with probability 1. Consider a class $\mathcal{F}_h$ of functions $c^Y \to \bar{h}(c^Y)$ on $\mathcal{C}^Y \subset \mathbb{R}^d$. Assume that $\tilde{h}_N^* \in \mathcal{F}_h$ with probability 1. Define the dissimilarity measure on the Cartesian product of $\mathcal{F} = \mathcal{F}_Y \times \mathcal{F}_h \times \mathcal{G}$:*

$$d((\tilde{h}_1, \bar{Q}_1, \bar{g}_1), (\tilde{h}, \bar{Q}, \bar{g})) = \max\left(\sup_{c \in \mathcal{C}^Y} |\tilde{h}_1 - \tilde{h}|, \sup_{c \in \mathcal{C}^Y} |\bar{Q}_1 - \bar{Q}|, \sup_{c \in \mathcal{C}^A} |\bar{g}_1 - \bar{g}|\right).$$

*Assume that there exists some $\eta > 0$, so that $\int_0^\eta \sqrt{\log(N(\epsilon, \mathcal{F}, d))} d\epsilon < \infty$, where $N(\epsilon, \mathcal{F}, d)$ is the number of balls of size $\epsilon$ w.r.t. metric $d$ needed to cover $\mathcal{F}$.*

*In particular, this assumption holds if $\sup_{\theta \in \mathcal{F}_Y} \| \theta \|_v^* < \infty$, $\sup_{\theta \in \mathcal{F}_h} \| \theta \|_v^* < \infty$, $\sup_{\bar{g} \in \mathcal{G}} \| \bar{g} \|_v^* < \infty$, where $\| \theta \|_v^*$ is the uniform sectional variation norm as defined in Gill et al. [41] and van der Laan [63].*

**Universal bound:** *Assume $\sup_{\theta \in F, O} |f|(O) < \infty$, where the supremum of $O$ is over a set that contains $O$ with probability 1. This assumption will typically be a consequence of the entropy condition, such as it is a consequence of the uniform sectional variation norm condition above.*

**Uniform consistency and rate condition:** *Assume $d(\tilde{h}_N, \bar{Q}_N^*, \bar{g}_N), (\tilde{h}_0, \bar{Q}^*, \bar{g}_0)) \to 0$ in probability as $N \to \infty$,*

$$R_{N,1} \equiv -\int_c \left(\frac{\bar{h}_N^*}{\bar{h}_N} - \frac{\bar{h}_0^*}{\bar{h}_0}\right)(\bar{Q}_N^* - \bar{Q}^*)(c)\bar{h}_0(c)d\mu(c) = o_P(1/\sqrt{N})$$

*and*

$$R_{N,4} = \int_c \left\{\frac{\bar{h}_N^*}{\bar{h}_N} - \frac{\bar{h}_0^*}{\bar{h}}\right\} \frac{1}{\bar{h}_0}(\bar{h}_N - \bar{h}_0)(\bar{Q}_0 - \bar{Q}^*)(c)\bar{h}_0 d\mu(c) = o_P\left(\frac{1}{\sqrt{N}}\right).$$

**Asymbiotic linearity condition on $\bar{g}_N$:**

$$\begin{aligned}&\int_c \frac{\bar{h}_0^*}{\bar{h}_0^2} \frac{\bar{h}(g_N - g_0, Q_{W,0})}{N}(\bar{Q}^* - \bar{Q}_0)(c)\bar{h}_0(c)d\mu(c)\\ &= \frac{1}{N}\sum_{i=1}^N f_{A,i}^1(O) + o_P(1/\sqrt{N}),\end{aligned}$$

*where $f_{A,i}^1(O)$ only depends on $O$ through $(A_i, (W_j : j \in F_i))$, and $E_0(f_{A,i}^1(O)|W) = 0$.*

**Positivity condition:** *Assume*

$$\sup_{c \in C^Y} \frac{\bar{h}^*(g^*, Q_{W,0})}{\bar{h}(g_0, Q_{W,0})}(c) < \infty.$$

**Universal bound on connectivity between units:** *Assume that there exists a $K < \infty$ so that $\sup_i |F_i| < K$ for all $i = 1, \ldots,$ a.s.*

**Universal bound on dependence of $W$-distribution, and stochastic intervention:** *Assume that there exists a $K < \infty$, so that $g^*(A_j : j \in F_i|W)$ only depends on $(W_j : j \in R_i)$ with $\max_i |R_i| < K$, and, for each $i$, $W_i$ is independent of $(W_j : j \in S_i^c)$ with $\max_i |S_i| \leq K$, where $S_i^c = \{j : j \notin S_i\}$, and $K$ does not depend on $N$.*

**First-order approximation:** *Then*,

$$\psi_N^* - \psi_0 = \frac{1}{N}\sum_{i=1}^{N}\{f_i(O) - P_0 f_i\} + o_P(1/\sqrt{N}),$$

*where*

$$f_i = D_{Y,i}^*(\bar{Q}^*, Q_{W,0}, g_0) + f_{A,i}^1 + f_{W,i}^1 + f_{W,i}^2$$

$$f_{W,i}^1(W) = \int_a \bar{Q}^*(c_i^Y(a,W))g^*(a|W)$$

$$f_{W,i}^2(W) = \int_c \left\{\frac{h_{i,N}^*}{\bar{h}_0} - \frac{\bar{h}_0^*}{\bar{h}_0^2}h_i(g_0, Q_{W,N})\right\}(c)(\bar{Q}_0 - \bar{Q}^*)(c)\bar{h}_0(c)d\mu(c)$$

$$h_{i,N}^*(c) = \int_a I(c_i^Y(a,W) = c)g^*(a|W) = g_i^*(c|W)$$

$$h_i(g_0, Q_{W,N})(c) = \int_a I(c_i^Y(a,W) = c)g_0(a|W) = g_{0,i}(c|W).$$

**Weak convergence of first-order approximation:** We can orthogonally decompose

$$f_i(O) - P_0 f_i = f_{Y,i}(O) + f_{A,i}(O) + f_{W,i}(O),$$

*where*

$$f_{Y,i} = D_{Y,i}^* - E_0(D_{Y,i}^*|A,W)$$

$$= \frac{\bar{h}_0^*}{\bar{h}_0}(C_i^Y)(Y_i - \bar{Q}_0(C_i^Y))$$

$$f_{A,i} = E_0(D_{Y,i}^*|A,W) - E_0(D_{Y,i}^*|W) + f_{A,i}^1$$

$$= \frac{\bar{h}_0^*}{\bar{h}_0}(C_i^Y)(\bar{Q}_0 - \bar{Q}^*)(C_i^Y)$$

$$- \int_c \frac{\bar{h}_0^*}{\bar{h}_0}(c)(\bar{Q}_0 - \bar{Q}^*)(c)g_{0,i}(c|W) + f_{A,i}^1$$

$$E_0(D_{Y,i}^*|W) = \int_c \frac{\bar{h}_0^*}{\bar{h}_0}(\bar{Q}_0 - \bar{Q}^*)(c)g_{0,i}(c|W)$$

$$f_{W,i} = f_{W,i}^1 + f_{W,i}^2 + E_0(D_{Y,i}^*|W) - P_0\{f_{W,i}^1 + f_{W,i}^2 + E_0(D_{Y,i}^*|W)\}$$

$$= \int_a \bar{Q}_0(c_i^Y(a,W))g^*(a|W) - \int_{a,w} \bar{Q}_0(c_i^Y(a,W))g^*(a|w)Q_{W,0}(dw)$$

$$= \int_c \bar{Q}_0(c)g_i^*(c|W) - \int_{c,w} \bar{Q}_0(c)g_i^*(c|w)Q_{W,0}(dw).$$

*For $(i,j) \in \{1,\ldots,N\}^2$, let $R_W(i,j)$ be the indicator that $f_{W,i}$ and $f_{W,j}$ are dependent, $R_A(i,j) = I(F_i \cap F_j \neq \emptyset)$, and $R_2(i,j) = I(R_A(i,j) = 1 \text{ or } R_W(i,j) = 1)$. For example, if $W_1,\ldots,W_N$ are independent, then $R_W(i,j) = I(R_i \cap R_j \neq \emptyset)$. We have*

$$\frac{1}{\sqrt{N}} \sum_i \{f_i(O) - P_0 f_i\} \Rightarrow_d N(0,\sigma^2), \text{ where } \sigma^2 = \sigma_Y^2 + \sigma_A^2 + \sigma_W^2,$$

*and*

$$\sigma_Y^2 = \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^N P_0 f_{Y,i}^2$$

$$\sigma_A^2 = \lim_{N\to\infty} \frac{1}{N} \sum_{i_1,i_2} R_A(i_1,i_2) P_0 f_{A,i_1} f_{A,i_2}$$

$$\sigma_W^2 = \lim_{N\to\infty} \frac{1}{N} \sum_{i_1,i_2} R_W(i_1,i_2) P_0 f_{W,i_1} f_{W,i_2},$$

*assuming these limits exist, and $P_0 f$ denotes the marginal expectation of $f(O)$, given $F$. As a consequence, $\sqrt{N}(\psi_N^* - \psi_0) \Rightarrow_d N(0,\sigma^2)$.*

**Alternative expression of asymptotic variance:** *One can also represent $\sigma^2$ as*

$$\sigma^2 = \lim_{N\to\infty} \frac{1}{N} \sum_{i_1,i_2} R_2(i_1,i_2) P_0 f_{i_1} f_{i_2}.$$

To provide the reader with a general understanding of the asymptotic normality of the TMLE we note the following. In the Appendix, we provide general conditions under which a process $Z_N = (Z_N(\theta) : \theta \in \mathcal{F})$, where $Z_N(\theta) = 1/\sqrt{N} \sum_i f_i(\theta)(O)$, converges weakly to a Gaussian process $Z = (Z(\theta) : \theta \in \mathcal{F})$ as random functionals in the Banach space $\ell^\infty(\mathcal{F})$ of real valued functionals on a family $\mathcal{F}$ of functions, endowed with the supremum norm [41], where the dependence between the $f_i(\theta)$s is restricted by assuming that $f_i(\theta)(O)$ can only depend on a set of maximally $K$ $f_j(\theta)(O)$s, where the integer bound $K$ does not depend on $N$. For completeness, we provide here the general theorem that is a corollary from the results established in the Appendix and provides the key building block for the probabilistic component of our proofs:

**Theorem 3** *Consider a process $Z_N = (Z_N(\theta) : \theta \in \mathcal{F})$, with $Z_N(\theta) = 1/\sqrt{N} \sum_{i=1}^N f_i(\theta)(O)$, where $E_0 f_i(\theta)(O) = 0$, for each $i$, $f_i(\theta)$ is independent of $\{f_j(\theta) : j \in S_i^c\}$ for a set $S_i \subset \{1,\ldots,N\}$ with $\max_i |S_i| < K$ for a universal $K$, where $S_i^c = \{j : j \notin S_i\}$, and $\mathcal{F}$ is a set of multivariate uniformly bounded real valued functions $\theta : \mathbb{R}^d \to \mathbb{R}$. Let $R(i,j)$ be the indicator that $f_i(\theta)$ and $f_j(\theta)$ are dependent. We make the following additional assumptions:*

- *For all integers $p > 0$, $\{E_0 f_i(\theta)(O)^p\}^{1/p} \leq C \parallel \theta \parallel_\infty$ for supremum norm $\parallel \theta \parallel$ on $F$, and universal $C < \infty$.*
- *There exists an $\eta > 0$ so that the entropy integral $\int_0^\eta \sqrt{\log N(\epsilon, \mathcal{F}, \parallel \cdot \parallel_\infty)} d\epsilon < \infty$ for $\mathcal{F}$ w.r.t. norm $\parallel \cdot \parallel_\infty$ is finite.*
- *The marginal distributions $Z_N(\theta)$ converge to a normal distribution $Z(\theta)$ for all $\theta \in \mathcal{F}$.*

*Then $Z_N$ converges weakly to a Gaussian process $Z$ identified by the covariance operator $\Sigma(\theta_1,\theta_2)$ defined by*

$$\Sigma(\theta_1,\theta_2) = \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^N \sum_{j=1}^N R(i,j) E_0 f_i(\theta_1) f_j(\theta_2).$$

*In particular, $Z_N$ is asymptotically equicontinuous in the sense that if $\theta_{1N} - \theta_{2N} \to 0$ w.r.t. supremum norm, where $\theta_{1N}, \theta_{2N} \in F$, then $Z_N(\theta_{1N}) - Z_N(\theta_{2N})$ converges to zero in probability.*

## 7.1 Statistical inference

One can estimate $\sigma^2$ by plugging in estimators $Q_{W,N}, \bar{Q}_N^*, \bar{h}_N$ in the expressions for $f_{Y,i}, f_{W,i}, f_{A,i}$. Given an estimator of $\sigma_N^2$, one can then construct a confidence interval $\psi_N^* \pm 1.96\sigma_N/\sqrt{N}$. If $\sigma_N$ is consistent for $\sigma$, then this will be an asymptotically valid 0.95-confidence interval. The expression for $\sigma^2$ suggests that a consistent estimator of $\sigma^2$ relies on consistent estimation of $\bar{Q}_0$, even though the consistency of $\psi_N^*$ only relies on a consistent estimator of $\bar{h}_0$ and thus the relevant part of $g_0$ (since the expectation w.r.t. $W$ is consistently estimated). Even if $\psi_N^*$ relied on a less nonparametric estimator of $\bar{Q}_0$, this suggests using a super-learner using flexible machine learning algorithms when estimating this asymptotic variance $\sigma^2$. However, below, we provide alternative estimators of the asymptotic variance that appear to avoid having to estimate $\bar{Q}_0$.

**Ignoring contribution of $\bar{g}_N$:** We claim that if $g_0$ is unknown, and one uses an MLE $g_N$ according to some model, then ignoring the contribution $f_{A,i}^1$ in $f_{A,i}$ due to estimation of $g_0$ will result in an upper bound $\sigma_{N,u}^2$ for the actual asymptotic variance $\sigma^2$ of the TMLE, based on a generalization of the result in van der Laan and Robins [3]. This result relies on the fact that $g_0$ is an orthogonal nuisance parameter w.r.t. $\psi_0$. Such a result would then allow us to use this simplified plug-in estimator $\sigma_{N,u}^2$ (using $g_N$ for $g_0$) in the statistical model $\mathcal{M}$ in which $g_0$ is not known but a correctly specified model for $g_0$ (i.e. $\bar{g}_0$) is available. Again, such a result will need to be formally established in future research.

In the sequel of this subsection, we suggest the following practical proposals for variance estimation.

**Assuming a consistent $\bar{Q}_N^*$:** Suppose that one is willing to assume that $\bar{Q}_N^*$ is consistent for $\bar{Q}_0$. In that case, ignoring the $f_{A,i}^1$ contribution by the argument above, it follows that $f_{A,i} = 0$, so that we can estimate $\sigma^2$ with

$$\sigma_N^2 = \frac{1}{N}\sum_{i=1}^N \left\{ \frac{\bar{h}_N^*}{\bar{h}_N}(C_i^Y)(Y_i - \bar{Q}_N^*(C_i^Y)) \right\}^2 \tag{17}$$

$$+ \frac{1}{N}\sum_{i,j} f_{W,i}(\bar{Q}_N^*, Q_{W,N}) f_{W,j}(\bar{Q}_N^*, Q_{W,N}),$$

where

$$f_{W,i}(\bar{Q}, Q_W) = \int_a \bar{Q}(c_i^Y(A,W))g^*(a|W) - \int_{a,w} \bar{Q}(c_i^Y(a,W))g^*(a|w)Q_W(dw).$$

**Assuming rare outcome:** suppose now that one is not willing to assume that $\bar{Q}_N^*$ is consistent but it is known that $\bar{Q}_0$ is close to zero (e.g. rare outcome). In addition, assume that $\bar{Q}^* \approx 0$ as well, which can be guaranteed by incorporating such a constraint in the logistic regressions submodel of the TMLE as in Balzer and van der Laan [64]. In that case it follows that the contributions to the variance of $f_{A,i}$ and $f_{W,i}$ are second-order relative to the contributions of $f_{Y,i}$ w.r.t. $\bar{Q}_0 \approx 0$. As a consequence, in that case, it would be appropriate to still use this estimate $\sigma_N^2$ (eq. 17), and the inconsistency of $\bar{Q}_N^*$ will only make the estimate of $\sigma_Y^2$ conservative. In fact, by this argument one could even drop the $\sigma_W^2$ contribution, but for the sake of being conservative, we would recommend including this term.

**A generally appropriate variance estimator:** We now proceed with deriving a more general variance estimator under reasonable assumptions. Firstly, we will ignore the contribution $f_{A,i}^1$ due to estimation of $g_0$, and as mentioned above we conjecture (based on i.i.d. theory) that this will only make the variance estimator conservative. Secondly, we note that (recall $f_{W,i} = f_{W,i}(\bar{Q}_0, Q_{W,0})$)

$$\begin{aligned} f_i &= f_{Y,i} + f_{A,i} + f_{W,i} \\ &= \tilde{h}_0(C_i^Y)(Y_i - \bar{Q}^*(C_i^Y)) + f_{W,i} \\ &= \tilde{h}_0(C_i^Y)(Y_i - \bar{Q}^*(C_i^Y)) + f_{W,i}(\bar{Q}^*, Q_{W,0}) + \{f_{W,i}(\bar{Q}_0^*, Q_{W,0}) - f_{W,i}(\bar{Q}^*, Q_{W,0})\}, \end{aligned}$$

where $f_{W,i}(\bar{Q}^*, Q_{W,0}) = \int_c \bar{Q}^*(c) g_i^*(c|W) - \int_{c,w} \bar{Q}^*(c) g_i^*(c|w) Q_{W,0}(dw)$. We now note that

$$\begin{aligned}\frac{1}{N}\sum_{i=1}^N \{f_{W,i}(\bar{Q}_0^*, Q_{W,0}) - f_{W,i}(\bar{Q}^*, Q_{W,0})\} &= \frac{1}{N}\sum_{i=1}^N \int_c (\bar{Q}_0 - \bar{Q}^*)(c) g_i^*(c|W) \\ &\quad - E_{Q_{W,0}} \frac{1}{N}\sum_{i=1}^N \int_c (\bar{Q}_0 - \bar{Q}^*)(c) g_i^*(c|W) \\ &= \Psi(\bar{Q}_0, Q_{W,N}) - \Psi(\bar{Q}^*, Q_{W,N}) - E_W\{\Psi(\bar{Q}_0, Q_{W,N}) - \Psi(\bar{Q}^*, Q_{W,N}),\end{aligned}$$

where in this last expression $Q_{W,N}$ denotes the empirical distribution that puts mass 1 on $W$. As shown in the next section, it follows that $\sqrt{N}(\Psi(\bar{Q}_N^*, Q_{W,N}) - \Psi(\bar{Q}_0, Q_{W,N})$ converges to a normal distribution, and therefore one expects that the conditional bias $\sqrt{N}(\Psi(\bar{Q}^*, Q_{W,N}) - \Psi(\bar{Q}_0, Q_{W,N})) = o_P(1)$. We will assume that indeed $\sqrt{N}(\Psi(\bar{Q}^*, Q_{W,N}) - \Psi(\bar{Q}_0, Q_{W,N}) = o_P(1)$. Under this assumption, we have that $\frac{1}{N}\sum_{i=1}^N \{f_{W,i}(\bar{Q}_0^*, Q_{W,0}) - f_{W,i}(\bar{Q}^*, Q_{W,0})\} = o_P(1/\sqrt{N})$. As a consequence,

$$\begin{aligned}\frac{1}{\sqrt{N}}\sum_{i=1}^N f_i &\approx \frac{1}{\sqrt{N}}\sum_{i=1}^N \tilde{h}_0(C_i^Y)(Y_i - \bar{Q}^*(C_i^Y)) \\ &\quad + \frac{1}{\sqrt{N}}\sum_{i=1}^N \left\{\int_c \bar{Q}^*(c) g_i^*(c|W) - \int_{c,w} \bar{Q}^*(c) g_i^*(c|w) Q_{W,0}(dw)\right\}.\end{aligned}$$

In addition, the first sum on the right-hand side already has conditional mean zero, given $W$, so that the asymptotic variance of the left-hand side equals the variance of the first sum plus the variance of the second sum. The second variance can be consistently estimated with

$$\sigma_{W,N}^2 = \frac{1}{N}\sum_{i_1,i_2} R_W(i_1, i_2) f_{W,i_1}(\bar{Q}_N^*, Q_{W,N}) f_{W,i_2}(\bar{Q}_N^*, Q_{W,N}).$$

The variance of the first sum can be represented as:

$$\begin{aligned}&P_0 \frac{1}{N}\sum_{i,j} R_A(i,j) \tilde{h}_0(C_i^Y)(Y_i - \bar{Q}^*(C_i^Y)) \tilde{h}_0(C_j^Y)(Y_j - \bar{Q}^*(C_j^Y)) \\ &- \frac{1}{N}\sum_{i,j} R_A(i,j) P_0 \tilde{h}_0(C_i^Y)(Y_i - \bar{Q}^*(C_i^Y)) P_0 \tilde{h}_0(C_j^Y)(Y_j - \bar{Q}^*(C_j^Y)).\end{aligned}$$

If one is willing to assume that

$$\lim_{N\to\infty} \frac{1}{N}\sum_{i,j} R_A(i,j) P_0 \tilde{h}_0(C_i^Y)(Y_i - \bar{Q}^*(C_i^Y)) P_0 \tilde{h}_0(C_j^Y)(Y_j - \bar{Q}^*(C_j^Y)) \geq 0,$$

then a conservative estimate of the first variance is defined as:

$$\sigma_{Y,N}^2 = \frac{1}{N}\sum_{i,j} R_A(i,j) D_{Y,i}^*(\bar{Q}_N^*, Q_{W,N}, \bar{h}_0) D_{Y,j}^*(\bar{Q}_N^*, Q_{W,N}, \bar{h}_0).$$

Till what degree this is a reasonable assumption will need to be further studied. Under this assumption, the proposed estimator of the asymptotic variance $\sigma^2$ is given by

$$\sigma_{N,1}^2 = \sigma_{W,N}^2 + \sigma_{Y,N}^2.$$

## 8 TMLE of intervention-specific mean, conditional on *W*

In our target parameter, we conditioned on the network information $F = (F_1, \ldots, F_N)$, but marginalized over $W$, given $F$. As a consequence, in order to establish asymptotic normality of the TMLE we had to rely on an

independence assumption on the joint distribution of $W$ (given $F$), such as that all $W_1, \ldots, W_N$ are independent, or only that each $W_i$ only depends on maximally $K$ $W_j$'s. In this section, we define the target parameter conditional on all of $W$, which happens to equal $\Psi(\bar{Q}_0, Q_{W,N})$, where $Q_{W,N}$ is the empirical distribution that puts mass 1 on $W$. Our target parameter is now a parameter of the conditional distribution $P_0^W$ of $O$, given $W$, modeled in same way as above (but without need to model a distribution of $W$). Its efficient influence curve is now just the $D_Y^*(\bar{Q}_0, Q_{W,N}, \bar{h}(g_0, Q_{W,N}))$-component, where for the sake of notational convenience we will still denote $\bar{h}(g_0, Q_{W,N})$ with $\bar{h}_0$ (just in this section and in proof of next theorem in the Appendix). We will use the same TMLE as presented in the previous sections. In the Appendix, we show how our template for analyzing the TMLE can be modified to analyze the TMLE with respect to this conditional $W$-specific target parameter, and that essentially the terms due to estimation of $Q_{W,0}$ now drop while the other terms are essentially the same. As a consequence, there is no need to redo all the technical proofs. Our proof now relies on the identity $P_0^W D_Y^*(\bar{Q}, Q_{W,N}, \bar{h}_0) = \Psi(\bar{Q}_0, Q_{W,N}) - \Psi(\bar{Q}, Q_{W,N})$, as established by (eq. 10). This results in the following Theorem 4. This theorem differs from Theorem 2 in that it dropped the independence assumption on the distribution of $W$ and that the asymptotic variance of the TMLE (w.r.t. $\Psi(\bar{Q}_0, Q_{W,N})$ instead of $\Psi(\bar{Q}_0, Q_{W,0})$) does not include the $\sigma_W^2$-term anymore. Thus, by changing our target parameter to this conditional version, we removed a restrictive assumption and we reduced the asymptotic variance of the TMLE w.r.t. this conditional target parameter.

**Theorem 4** *The conditional probability distribution of O, given W, is parameterized* by $\bar{g}, \bar{Q}_Y$ *as follows:*

$$P^W(O) = \prod_{i=1}^N \bar{g}(A_i|C_i^A)\bar{Q}_Y(Y_i|C_i^Y), \tag{18}$$

*where* $C_i^Y = c_i^Y(A, W) \in \mathcal{C}^Y \subset \mathbb{R}^{d_1}$, $C_i^A = c_i^A(W) \in \mathcal{C}^A \subset \mathbb{R}^{d_2}$, $\bar{Q}_Y(\cdot|c)$ *is a density for Y for each possible* $c \in \mathcal{C}^Y$, *but is otherwise unspecified,* $\bar{g}(\cdot|c)$ *is a density for A for each possible* $c \in \mathcal{C}^A$, *and* $\bar{g} \in \mathcal{G}$. *This defines the statistical model* $\mathcal{M}^W$ *for the conditional probability distribution* $P_0^W$ *of O, given W. Let* $Q_{W,N}$ *denote the probability distribution of W that puts mass 1 on the observed* $W = (W_1, \ldots, W_N)$.

*For a specified stochastic intervention* $g^*$, *the target parameter* $\Psi : \mathcal{M}^W \to \mathbb{R}$ *is defined by*

$$\begin{aligned}\Psi^W(P^W) &= \Psi(\bar{Q}, Q_{W,N}) \\ &= \frac{1}{N}\sum_{j=1}^N \int_{a,w} \bar{Q}(c_j^Y(a,w))g^*(a|w)Q_{W,N}(dw) \\ &= \frac{1}{N}\sum_{j=1}^N \int_a \bar{Q}(c_j^Y(a,W))g^*(a|W),\end{aligned}$$

*where* $\bar{Q}(c_j^Y(A,W)) = E_P(Y_j|A,W)$. *Since* $\Psi^W(P^W)$ *only depends on* $P^W$ *through* $\bar{Q}$, *we will also denote this parameter with* $\Psi^W(\bar{Q})$.

*The efficient influence curve of* $\Psi^W$ *at* $P^W$ *is given by:*

$$D_Y^*(\bar{Q}, Q_{W,N}, \bar{h}) = \textstyle\sum_{i=1}^N D_{Y_i}^*(\bar{Q}, Q_{W,N}, \bar{h}),$$

*where*

$$D_{Y,i}^*(\bar{Q}, Q_{W,N}, \bar{h}) = \frac{1}{N}\frac{\bar{h}(g^*, Q_{W,N})(C_i^Y)}{\bar{h}(g, Q_{W,N}))(C_i^Y)}(Y_i - \bar{Q}(C_i^Y)).$$

*Consider the TMLE* $\psi_N^* = \Psi(Q_N^*) = \Psi(\bar{Q}_N^*, Q_{W,N})$ *defined above using* $\bar{g}_N$ *in* $\bar{h}_N = \bar{h}(g_N, Q_{W,N})$. *As shown above, this TMLE solves*

$$D_Y^*(\bar{Q}_N^*, Q_{W,N}, \bar{h}_N)(O) = 0.$$

*We use the definitions* $\bar{h}(c) = \frac{1}{N}\sum_{i=1}^N h_i(c)$, $\bar{h}_N^* = \frac{1}{N}\sum_{i=1}^N h_{i,N}^*$, $h_i(c) = P_{g,Q_{W,N}}(c_i^Y(A,W) = c)$, $h_{i,N}^*(c) = P_{g^*,Q_{W,N}}(c_i^Y(A,W) = c)$, *defined as densities w.r.t. a dominating measure* $\mu$, *and let* $\tilde{h} = \bar{h}_N^*/\bar{h}$, $\tilde{h}_0 = \bar{h}_N^*/\bar{h}_0$, *where* $\bar{h}_0 = \bar{h}(g_0, Q_{W,N})$. *Note that* $\tilde{h}_N = \bar{h}_N^*/\bar{h}_N$ *is a plug-in estimator of* $\tilde{h}_0 = \bar{h}_N^*/\bar{h}_0$ *implied by* $\bar{g}_N \in \mathcal{G}$ *and* $Q_{W,N}$.

*We make the following assumptions:*

**Entropy condition:** Consider a class $\mathcal{F}_Y$ of functions $c^Y \to \bar{Q}(c^Y)$ on a set in $\mathcal{C}^Y \subset \mathbb{R}^d$ that contains $c^Y(A,W)$ with probability 1. Assume that $\bar{Q}_N^* \in \mathcal{F}_Y$ with probability 1. Consider a class $\mathcal{F}_h$ of functions $c^Y \to \tilde{h}(c^Y)$ on $\mathcal{C}^Y \subset \mathbb{R}^d$. Assume that $\tilde{h}_N \in \mathcal{F}_h$ with probability 1. Define the dissimilarity measure on the Cartesian product of $\mathcal{F} = \mathcal{F}_Y \times \mathcal{F}_h \times \mathcal{G}$:

$$d((\tilde{h}_1, \bar{Q}_1, \bar{g}_1), (\tilde{h}, \bar{Q}, \bar{g})) = \max\left(\sup_{c \in \mathcal{C}^Y} |\tilde{h}_1 - \tilde{h}|, \sup_{c \in \mathcal{C}^Y} |\bar{Q}_1 - \bar{Q}|, \sup_{c \in \mathcal{C}^A} |\bar{g}_1 - \bar{g}|\right).$$

*Assume that there exists some $\eta > 0$, so that $\int_0^\eta \sqrt{\log(N(\epsilon, \mathcal{F}, d))}\, d\epsilon < \infty$, where $N(\epsilon, \mathcal{F}, d)$ is the number of balls of size $\epsilon$ w.r.t. metric $d$ needed to cover $\mathcal{F}$.*

*In particular, this assumption holds if $\sup_{\theta \in \mathcal{F}_Y} \max(\| \tilde{h} \|_v^*, \| \bar{Q} \|_v^*, \| \bar{g} \|_v^*) < \infty$, where $\| \cdot \|_v^*$ is the uniform sectional variation norm as defined in Gill et al. [65] and van der Laan [63].*

**Universal bound:** *Assume $\sup_{\theta \in \mathcal{F}, O} |f|(O) < \infty$, where the supremum of $O$ is over a set that contains $O$ with probability 1. This assumption will typically be a consequence of the entropy condition, such as it is a consequence of the uniform sectional variation norm condition above.*

**Uniform consistency and rate condition:** *Assume $d(\tilde{h}_N, \bar{Q}_N^*, \bar{g}_N), (\tilde{h}_0, \bar{Q}^*, \bar{g}_0)) \to 0$ in probability as $N \to \infty$,*

$$R_{N,1} \equiv -\int_c \left(\frac{\bar{h}_N^*}{\bar{h}_N} - \frac{\bar{h}_N^*}{\bar{h}_0}\right)(\bar{Q}_N^* - \bar{Q}^*)\bar{h}_0(c)\, d\mu(c) = o_P(1/\sqrt{N})$$

*and*

$$R_{N,4} = \int_c \left\{\frac{\bar{h}_N^*}{\bar{h}_N} - \frac{\bar{h}_N^*}{\bar{h}_0}\right\} \frac{1}{\bar{h}_0}(\bar{h}_N - \bar{h}_0)(\bar{Q}_0 - \bar{Q}^*)(c)\bar{h}_0 d\mu(c) = o_P\left(\frac{1}{\sqrt{N}}\right).$$

**Asymptotic linearity condition on $\bar{g}_N$:**

$$\int_c \frac{\bar{h}_N^*}{\bar{h}_0^2} \frac{\bar{h}(g_N - g_0, Q_{W,N})}{N}(\bar{Q}^* - \bar{Q}_0)(c)\bar{h}_0(c)\, d\mu(c)$$
$$= \frac{1}{N}\sum_{i=1}^N f_{A,i}^1(O) + o_P(1/\sqrt{N}),$$

*where $f_{A,i}^1(O)$ only depends on $O$ through $(A_i, (W_j : j \in F_i))$, and $E_0(f_{A,i}^1(O)|W) = 0$.*

**Positivity condition:**

$$\sup_{c \in \mathcal{C}^Y} \frac{\bar{h}^*(g^*, Q_{W,N})}{\bar{h}(g_0, Q_{W,N})}(c) < \infty.$$

**Universal bound on connectivity:** *Assume that there exists a $K < \infty$ so that $\sup_i |F_i| < K$ for all $i = 1, \ldots,$ a.s.*

**Restriction on stochastic intervention:** *Assume $g^*(A_j : j \in F_i|W)$ only depends on $W$ through $(W_j : j \in R_i)$ with $\max_i |R_i| < K$ for some universal $K < \infty$.*

**First-order approximation:** *Then,*

$$\psi_N^* - \psi_0^W = \frac{1}{N}\sum_{i=1}^N \{f_i^W(O) - P_0^W f_i^W\} + o_P(1/\sqrt{N}),$$

*where*

$$f_i^W = D_{Y,i}^*(\bar{Q}^*, Q_{W,N}, \bar{h}_0) + f_{A,i}^1.$$

**Weak convergence of first-order approximation:** *We can orthogonally decompose*

$$f_i^W(O) - P_0^W f_i = f_{Y,i}(O) + f_{A,i}(O),$$

*where*

$$\begin{aligned} f_{Y,i} &= D^*_{Y,i} - E_0(D^*_{Y,i}|A,W) \\ &= \frac{\bar{h}^*_N}{\bar{h}_0}(C_i^Y)(Y_i - \bar{Q}_0(C_i^Y)), \\ f_{A,i} &= E_0(D^*_{Y,i}|A,W) - E_0(D^*_{Y,i}|W) + f^1_{A,i} \\ &= \frac{\bar{h}^*_N}{\bar{h}_0}(C_i^Y)(\bar{Q}_0 - \bar{Q}^*)(C_i^Y) \\ &\quad - \int_c \frac{\bar{h}^*_N}{\bar{h}_0}(c)(\bar{Q}_0 - \bar{Q}^*)(c) g_{0,i}(c|W) + f^1_{A,i}. \end{aligned}$$

*For* $(i,j) \in \{1,\ldots,N\}^2$, *let* $R_A(i,j) = I(F_i \cap F_j \neq \emptyset)$. *We have*

$$\frac{1}{\sqrt{N}} \sum_i \{f_i^W(O) - P_0^W f_i^W\} \Rightarrow_d N(0, \sigma^{2,W}), \text{ where } \sigma^{2,W} = \sigma_Y^2 + \sigma_A^2,$$

*and*

$$\begin{aligned} \sigma_Y^2 &= \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^N P_0^W f_{Y,i}^2 \\ \sigma_A^2 &= \lim_{N\to\infty} \frac{1}{N} \sum_{i_1,i_2} R_A(i_1,i_2) P_0^W f_{A,i_1} f_{A,i_2} \end{aligned}$$

*assuming these limits exist, and* $P_0^W f$ *denotes the conditional expectation of* $f(O)$, *given* $W$. *As a consequence,* $\sqrt{N}(\psi^*_N - \psi_0^W) \Rightarrow_d N(0, \sigma^{2,W})$.

**Alternative expression of asymptotic variance:** One can also represent $\sigma^{2,W}$ as

$$\sigma^{2,W} = \lim_{N\to\infty} \frac{1}{N} \sum_{i_1,i_2} R_A(i_1,i_2) P_0^W f_{i_1} f_{i_2}.$$

## 8.1 Variance estimation

**Known $g_0$ and consistent $\bar{Q}^*_N$:** Let us consider an RCT so that $g_N = g_0$ and the term $f^1_{A,i} = 0$. If one is willing to assume that $\bar{Q}^*_N$ is consistent for $\bar{Q}_0$, then $f_{A,i} = f^1_{A,i} = 0$. Therefore, in this case, the asymptotic variance can be estimated as

$$\sigma_N^{2,W} = \frac{1}{N} \sum_{i=1}^N \left\{ \frac{\bar{h}^*_N}{\bar{h}_N}(C_i^Y)(Y_i - \bar{Q}^*_N(C_i^Y)) \right\}^2. \tag{19}$$

**Known $g_0$, rare outcome:** Suppose now that we still have an RCT, but we are not willing to assume $\bar{Q}^*_N$ is consistent, but $\bar{Q}_0$ is close to zero (e.g. rare outcome). In addition, assume $\bar{Q}^* \approx 0$: i.e. one might incorporate this constraint on $\bar{Q}_0$ in the submodel of the TMLE, as in Balzer and van der Laan [64]. It follows that a first-order (w.r.t. $\bar{Q}_0$ approximating zero) approximation of the asymptotic variance $\sigma^{2,W}$ can still ignore the $f_{A,i}$-contribution. As a consequence, in that case an appropriate approximation of the asymptotic variance is given by $\sigma_Y^2 = \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^N P_0^W f_{Y,i}^2$. That is, this asymptotic variance $\sigma_Y^2$ is approximated by

$$\frac{1}{N} \sum_{i=1}^N \left\{ \frac{\bar{h}^*_N}{\bar{h}_N}(C_i^Y)(Y_i - \bar{Q}_0(C_i^Y)) \right\}^2.$$

However, the latter is conservatively estimated by using a possibly inconsistent $\bar{Q}_N^*$, showing that we can still use (eq. 19) as the estimator of the asymptotic variance.

**Ignoring contribution of $g_N$ is conservative:** Even when $g_0$ is estimated with $g_N$, as argued before, we suggest that the contribution $f_{A,i}^1$ only reduces the asymptotic variance, so that ignoring this contribution will be fine for the sake of reliable statistical inference. Thus, our overall conclusion is that (eq. 19) is an appropriate (possibly conservative) estimator for the asymptotic variance when either $\bar{Q}_N^*$ is consistent or if $\bar{Q}_0 \approx 0$.

**A general variance estimator:** Assume that

$$\Psi(\bar{Q}_0, Q_{W,N}) - \Psi(\bar{Q}^*, Q_{W,N}) = o_P(1/\sqrt{N}).$$

Since this represents the bias term of the TMLE $\Psi(\bar{Q}_N^*, Q_{W,N})$, and we have asymptotic normality of $\sqrt{N}(\Psi(\bar{Q}_N^*, Q_{W,N}) - \Psi(\bar{Q}_0, Q_{W,N})$, and $\bar{Q}_N^*$ is consistent for $\bar{Q}^*$, this should be true under the assumptions of the previous theorem. However, under this assumption we have that, ignoring $f_{A,i}^1$,

$$f_{A,i} = \frac{\bar{h}_N^*}{\bar{h}_0}(C_i^Y)(\bar{Q}_0 - \bar{Q}^*)(C_i^Y) + o_P(1/\sqrt{N}),$$

and, as a consequence,

$$f_i^W = \frac{\bar{h}_N^*}{\bar{h}_0}(C_i^Y)(Y_i - \bar{Q}^*(C_i^Y)) = D_{Y,i}^*(\bar{h}_0, \bar{Q}^*).$$

Note that indeed

$$\begin{aligned} P_0^W \frac{1}{N}\sum_{i=1}^N \tilde{h}(g_0, Q_{W,N})(C_i^Y)(Y_i - \bar{Q}^*(C_i^Y)) &= \Psi(\bar{Q}_0, Q_{W,N}) - \Psi(\bar{Q}^*, Q_{W,N}) \\ &= o_P(1/\sqrt{N}). \end{aligned}$$

Thus, under the assumptions of the Theorem, and ignoring the $f_{A,i}^1$ contribution from $g_N$, we have

$$\sqrt{N}(\psi_N^* - \psi_0) = \frac{1}{\sqrt{N}}\sum_{i=1}^N \tilde{h}(g_0, Q_{W,N})(Y_i - \bar{Q}^*(C_i^Y)) + o_P(1),$$

where the linear term has conditional mean zero w.r.t. $P_0^W$. The conditional variance of the linear term on the right-hand side is thus given by the following expression:

$$\sigma^{2W} = \frac{1}{N}\sum_{i,j} R_A(i,j)\left\{P_0^W f_i^W f_j^W - P_0^W f_i^W P_0^W f_j^W\right\}.$$

Suppose that

$$\lim_{N\to\infty} \frac{1}{N}\sum_{i,j} R_A(i,j) P_0^W D_{Y,i}^*(\bar{Q}^*, Q_{W,N}, \bar{h}_0) P_0^W D_{Y,j}^*(\bar{Q}^*, Q_{W,N}, \bar{h}_0) \geq 0.$$

Then a conservative estimate of the last expression $\sigma^{2W}$ is defined as:

$$\sigma_N^{2W} = \frac{1}{N}\sum_{i,j} R_A(i,j) D_{Y,i}^*(\bar{Q}_N^*, Q_{W,N}, \bar{h}_0) D_{Y,j}^*(\bar{Q}_N^*, Q_{W,N}, \bar{h}_0).$$

Note that if $\bar{Q}_N^*$ is consistent for $\bar{Q}_0$, then this estimator is asymptotically equivalent with (eq. 19), but we expect the latter to be significantly larger for finite samples when $\bar{Q}_N^*$ is not a good approximation of $\bar{Q}_0$. If $g_0$ is not known, then $\bar{h}_0$ is replaced by its estimator $\bar{h}_N$.

# 9 Summary and concluding remarks

We formulated a general causal model for the longitudinal data structure generated by a finite population of causally connected units. This allows us to define counterfactuals indexed by interventions on the treatment nodes of the units, and corresponding causal contrasts. We established identifiability of the causal quantities from the data observed on the units when observing all units or a random sample of the units, assuming that the size of the population converges to infinity, under appropriate assumptions. Our causal assumptions implied conditional independence across units at time $t$, conditional on the past of all units, resulting in a factorized likelihood of the observed data (even though the observed data is generated by a single experiment, not by a repetition of independent experiments). To deal with the curse of dimensionality we assumed that a unit's dependence on the past of other units can be summarized by a finite dimensional measure and that this dependence is described by a common function across the units. This describes now the statistical model for the data distribution and the statistical target parameter, and thereby the statistical estimation problem. We demonstrated that we can use cross-validation and super-learning to estimate the different factors of the likelihood. Given the statistical model and statistical target parameter that identifies the counterfactual mean under an intervention, we derived the efficient influence curve of the target parameter. We showed that this efficient influence curve characterizes the normal limit distribution of a maximum likelihood estimator and thus still represents an optimal asymptotic variance among estimators of the target parameter. However, due to the curse of dimensionality, maximum likelihood estimators will be ill-defined for finite samples, and smoothing will be needed.

Such smoothed/regularized maximum likelihood estimators are not targeted and will thereby be overly biased w.r.t. the target parameter, and, as a consequence, generally not result in asymptotically normally distributed estimators of the statistical target parameter. Therefore, we formulated targeted maximum likelihood estimators of this estimand and showed that the robustness of the efficient influence curve implies that the bias of the TMLE will be a second-order term involving squared differences $\bar{h}_n - \bar{h}_0$ and $Q_n - Q_0$ for two nuisance parameters $\bar{h}_0 = \bar{h}(g_0, Q_0)$ and the relevant factor of likelihood $Q_0$. Subsequently, as showcased in this article, we focussed on defining and analyzing the TMLE of causal effects of an intervention on a single treatment node on a future outcome. In this special case, we showed that the efficient influence curve is double robust w.r.t. these two nuisance parameters $\bar{h}_0, \bar{Q}_0$, where $\bar{h}_0$ depends on the intervention mechanism and the distribution of the covariates, and $\bar{Q}_0$ is a common conditional mean function for the outcome. We established two formal asymptotic normality theorems for the TMLE under the assumption that each unit is only connected to fewer than $K$ other units for a universal $K$.

In future work, it will be of interest to extend our asymptotics theorem to the case that a unit can depend on a fixed (in $N$)-dimensional summary measure that can depend on a number of units that can converge to infinity with sample size. We can also be less-restrictive and allow that these summary measures have a dimension $K$ that increases with $N$, and then establishes rates of convergence that are slower than $1/\sqrt{N}$ and establishes corresponding (e.g. normal) limit distributions. In addition, in future work, the finite sample behavior of these estimators and confidence intervals will need to be evaluated through simulation studies. We will also generalize our TMLE to the TMLE of parameters defined by marginal structural working models for the causal dose–response curve for a collection of stochastic interventions. We also plan to investigate if there are other causal models for causally connected units that might allow the formulation of TMLE for the general longitudinal data structure in terms of sequential regressions, as in the double robust estimating equation based estimators for i.i.d. data presented in Bang and Robins [58] and subsequent analogue TMLE in van der Laan and Gruber [59] and Petersen et al. [66].

Overall, we believe that the statistical study of these causal models for dynamic networks of units provides a fascinating and important area of future research, relying on deep advances in empirical process and statistical estimation theory, while raising new challenges. In the mean time, these advances will be needed to move forward statistical practice.

**Acknowledgments:** This research was supported by NIH grant R01 AI074345-05. The author owes thanks to Elizabeth Ogburn, Maya Petersen, and Oleg Sofrygin for helpful discussions. The author also thanks Tyler vanderWeele for suggesting to weaken the independence assumptions on $W$ resulting in our Theorem 4.

# Appendix

## Introduction to Appendix

We start out with presenting a general template of our proof of Theorem 2 which establishes the asymptotics of the TMLE for the case $\tau = 0$. In this template, we define the remaining ingredients (eq. A1), (eq. A2), and (eq. A3) that will need to be established in the remainder of the proof. Each of these three ingredients is carried out in a separate section. These sections are themselves organized by special tasks that need to be carried out. We conclude with a similar template of the proof of Theorem 4, demonstrating that the technical components are the same as needed for Theorem 2. At the end of the Appendix, we provide a notation index that will be helpful to read through the article as well as through the Appendix.

## General template of proof of Theorem 2

Recall that $D^* = 1/N \sum_{j=1}^N D_j^*(O)$ is a sum over the units $j$. We will use the notation $P_N D^* = D^*(O) = 1/N \sum_{j=1}^N D_j^*(O)$, while $P_0 D^* = 1/N \sum_j E_{P_0} D_j^*(O)$ is its expectation w.r.t. distribution $P_0$. Due to Theorem 1, we have $D^* = D_W^* + D_Y^*$, $P_0 D_W^*(\bar{Q}_N^*, Q_{W,0}) = 0$, $P_0 D_Y^*(\bar{Q}_N^*, Q_{W,0}, \bar{h}_0) = \psi_0 - \Psi(\bar{Q}_N^*, Q_{W,0})$, and $P_N D_W^*(\bar{Q}_N^*, Q_{W,N}) = P_N D_Y^*(\bar{Q}_N^*, Q_{W,N}, \bar{h}_N) = 0$. In particular, this yields

$$P_0 D^*(\bar{Q}_N^*, Q_{W,0}, \bar{h}_0) = \psi_0 - \Psi(\bar{Q}_N^*, Q_{W,0}).$$

We now proceed as follows:

$$\begin{aligned}
\Psi(\bar{Q}_N^*, Q_{W,N}) - \psi_0 &= \Psi(\bar{Q}_N^*, Q_{W,N}) - \Psi(\bar{Q}_N^*, Q_{W,0}) + \Psi(\bar{Q}_N^*, Q_{W,0}) - \psi_0 \\
&= \Psi(\bar{Q}_N^*, Q_{W,N}) - \Psi(\bar{Q}_N^*, Q_{W,0}) - P_0 D_Y^*(\bar{Q}_N^*, Q_{W,0}, \bar{h}_0) \\
&= \Psi(\bar{Q}_N^*, Q_{W,N}) - \Psi(\bar{Q}_N^*, Q_{W,0}) + (P_N - P_0) D_Y^*(\bar{Q}_N^*, Q_{W,0}, \bar{h}_0) \\
&\quad + P_N \{D_Y^*(\bar{Q}_N^*, Q_{W,N}, \bar{h}_N) - D_Y^*(\bar{Q}_N^*, Q_{W,0}, \bar{h}_0)\} \\
&= \Psi(\bar{Q}_N^*, Q_{W,N}) - \Psi(\bar{Q}_N^*, Q_{W,0}) + (P_N - P_0) D_Y^*(\bar{Q}_N^*, Q_{W,0}, \bar{h}_0) \\
&\quad + (P_N - P_0) \{D_Y^*(\bar{Q}_N^*, Q_{W,N}, \bar{h}_N) - D_Y^*(\bar{Q}_N^*, Q_{W,0}, \bar{h}_0)\} \\
&\quad + P_0 \{D_Y^*(\bar{Q}_N^*, Q_{W,N}, \bar{h}_N) - D_Y^*(\bar{Q}_N^*, Q_{W,0}, \bar{h}_0)\}.
\end{aligned}$$

We note that

$$\begin{aligned}
&\{D_Y^*(\bar{Q}_N^*, Q_{W,N}, \bar{h}_N) - D_Y^*(\bar{Q}_N^*, Q_{W,0}, \bar{h}_0)\} \\
&= \frac{1}{N} \sum_{i=1}^N \left( \frac{\bar{h}_N^*}{\bar{h}_N} - \frac{\bar{h}_0^*}{\bar{h}_0} \right) (Y_i - \bar{Q}_N^*(C_i^Y)),
\end{aligned}$$

where $\bar{h}_N^* = \bar{h}(g^*, Q_{W,N})$, $\bar{h}_0^* = \bar{h}(g^*, Q_{W,0})$, and $\bar{h}_0 = \bar{h}(g_0, Q_{W,0})$. From this, it follows that

$$\begin{aligned}
&P_0\{D_Y^*(\bar{Q}_N^*, Q_{W,N}, \bar{h}_N) - D_Y^*(\bar{Q}_N^*, Q_{W,0}, \bar{h}_0)\} \\
&= P_0 \frac{1}{N}\sum_{i=1}^N \left(\frac{\bar{h}_N^*}{\bar{h}_N} - \frac{\bar{h}_0^*}{\bar{h}_0}\right)(\bar{Q}_0 - \bar{Q}^*)(C_i^Y) \\
&\quad - P_0 \frac{1}{N}\sum_{i=1}^N \left(\frac{\bar{h}_N^*}{\bar{h}_N} - \frac{\bar{h}_0^*}{\bar{h}_0}\right)(\bar{Q}_N^* - \bar{Q}^*)(C_i^Y) \\
&\equiv \int_c \left(\frac{\bar{h}_N^*}{\bar{h}_N} - \frac{\bar{h}_0^*}{\bar{h}_0}\right)(\bar{Q}_0 - \bar{Q}^*)\bar{h}_0(c)d\mu(c) + R_{N,1},
\end{aligned}$$

where we used that for a given function $f$ $P_0 \frac{1}{N}\sum_{i=1}^n f(C_i^Y) = \int_c f(c)\bar{h}_0(c)d\mu(c)$. We assumed that the second-order term $R_{N,1} = o_P(1/\sqrt{N})$. In addition, we define

$$\begin{aligned}
R_{N,2} &\equiv (P_N - P_0)\{D_Y^*(\bar{Q}_N^*, Q_{W,N}, \bar{h}_N) - D_Y^*(\bar{Q}_N^*, Q_{W,0}, \bar{h}_0)\} \\
&= \frac{1}{N}\sum_{i=1}^N \left\{ \left(\frac{\bar{h}_N^*}{\bar{h}_N} - \frac{\bar{h}_0^*}{\bar{h}_0}\right)(Y_i - \bar{Q}_N^*(C_i^Y)) - P_0\left(\frac{\bar{h}_N^*}{\bar{h}_N} - \frac{\bar{h}_0^*}{\bar{h}_0}\right)(Y_i - \bar{Q}_N^*(C_i^Y)) \right\}.
\end{aligned}$$

We also note that

$$\begin{aligned}
&\Psi(\bar{Q}_N^*, Q_{W,N}) - \Psi(\bar{Q}_N^*, Q_{W,0}) \\
&= \frac{1}{N}\sum_{i=1}^N \left\{ \int_a \bar{Q}_N^*(c_i^Y(a, W))g^*(a|W) - \int \bar{Q}_N^*(c_i^Y(a, w))g^*(a|w)Q_{W,0}(dw) \right\} \\
&\equiv \frac{1}{N}\sum_{i=1}^N \{f_{W,i}^1(W) - P_0 f_{W,i}^1\} + R_{N,0},
\end{aligned}$$

where

$$f_{W,i}^1 = \int_a \bar{Q}^*(c_i^Y(a, W))g^*(a|W) = \int \bar{Q}(c)g_i^*(c|W),$$

and

$$R_{N,0} = \frac{1}{N}\sum_{i=1}^N \left\{ \int_c (\bar{Q}_N^* - \bar{Q}^*)(c)g_i^*(c|W) - \int (\bar{Q}_N^* - \bar{Q}^*)(c)g_i^*(c|w)Q_{W,0}(dw) \right\}.$$

We used here that $\int_a \bar{Q}(c_i^Y(a, W))g^*(a|W) = \int_c \bar{Q}(c)g_i^*(c|W)$. Define the process $Z_{W,N}^1(\bar{Q}) = \frac{1}{\sqrt{N}}\sum_{i=1}^N \{f_{W,i}^1(\bar{Q}) - P_0 f_{W,i}^1(\bar{Q})\}$ indexed by $\bar{Q}$. Note that $\sqrt{N}R_{N,0} = Z_{W,N}^1(\bar{Q}_N^* - \bar{Q}^*)$. As a consequence, showing that $R_{N,0} = o_P(1/\sqrt{N})$ corresponds with proving that $Z_{W,N}^1(\epsilon_N) = o_P(1)$ for a sequence $\epsilon_N$ that converges to zero w. r.t. supremum norm. Therefore, our proof will involve studying this empirical process $Z_{W,N}^1$ and establishing the required asymptotic equicontinuity. In this manner, we will establish

$$R_{N,0} = o_P(1/\sqrt{N}) \tag{A2}$$

Thus, we have obtained the following expansion:

$$\psi_N^* - \psi_0 = \frac{1}{N}\sum_{i=1}^{N}\{f_{W,i}^1(W) - P_0 f_{W,i}\} + (P_N - P_0)D_Y^*(\bar{Q}_N^*, Q_{W,0}, \bar{h}_0) + \int_c \left(\frac{\bar{h}_N^*}{\bar{h}_N} - \frac{\bar{h}_0^*}{\bar{h}_0}\right)(\bar{Q}_0 - \bar{Q}^*)(c)\bar{h}_0(c)d\mu(c) + R_{N,1} + R_{N,2} + o_P(1/\sqrt{N}),$$

We have

$$\begin{aligned}(P_N - P_0)D_Y^*(\bar{Q}_N^*, Q_{W,0}, \bar{h}_0) &= (P_N - P_0)D_Y^*(\bar{Q}^*, Q_{W,0}, \bar{h}_0) \\ &\quad + (P_N - P_0)\{D_Y^*(\bar{Q}_N^*, Q_{W,0}, \bar{h}_0) - D_Y^*(\bar{Q}^*, Q_{W,0}, \bar{h}_0)\} \\ &\equiv (P_N - P_0)D_Y^*(\bar{Q}^*, Q_{W,0}, \bar{h}_0) + R_{N,3}.\end{aligned}$$

We will show that

$$R_{N,2} = o_P(1/\sqrt{N}) \text{ and } R_{N,3} = o_P(1/\sqrt{N}) \tag{A3}$$

To understand these last two terms, define the process

$$Z_N(\tilde{h}, \bar{Q}) = \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\left\{\tilde{h}(C_i^Y)(Y_i - \bar{Q}(C_i^Y)) - P_0\tilde{h}(C_i^Y)(Y_i - \bar{Q}(C_i^Y))\right\},$$

which is a sum of the form $Z_N(\tilde{h}, \bar{Q}) = \frac{1}{\sqrt{N}}\sum_i f_i(\bar{Q}, h)(O_i)$ indexed by $(\tilde{h}, \bar{Q})$, where $\tilde{h}$ plays role of $\bar{h}^*/\bar{h}$. Note that $R_{N,2} = Z_N(\tilde{h}_N - \tilde{h}_0, \bar{Q}_N^*)$, while $R_{N,3} = Z_N(\tilde{h}_0, \bar{Q}_N^* - \bar{Q}^*)$. Thus, showing that $R_{N,2} = o_P(1/\sqrt{N})$ and $R_{N,3} = o_P(1/\sqrt{N})$ comes down to showing that $Z_N(\varepsilon_N) = o_P(1/\sqrt{N})$ for $\varepsilon_N$ converging to zero w.r.t. supremum norm. Therefore, our proof will involve studying this process $Z_N()$ and establishing the required asymptotic equicontinuity. Specifically, we will decompose this process in three orthogonal processes that can be represented as sums over functions of conditionally independent random variables identified by the sets $F_i$ (analogue to orthogonal decomposition below of the first-order approximation) and establish this asymptotic equicontinuity for each of the three orthogonal processes.

Consider now the term

$$\int_c \left(\frac{\bar{h}_N^*}{\bar{h}_N} - \frac{\bar{h}_0^*}{\bar{h}_0}\right)(\bar{Q}_0 - \bar{Q}^*)\bar{h}_0(c)d\mu(c). \tag{20}$$

This term equals

$$\begin{aligned}&\int_c \left\{\frac{\bar{h}_N^*(c)}{\bar{h}_N(c)} - \frac{\bar{h}_0^*(c)}{\bar{h}_0(c)}d\mu(c)\right\}(\bar{Q}_0 - \bar{Q}^*)(c)\bar{h}_0(c)d\mu(c) \\ &= \int_c \left\{\frac{\bar{h}_N^* - \bar{h}_0^*}{\bar{h}_0}(c) - \frac{\bar{h}_0^*}{\bar{h}_0^2}(\bar{h}_N - \bar{h}_0)(c)\right\}(\bar{Q}_0 - \bar{Q}^*)(c)\bar{h}_0(c)d\mu(c) + R_{N,4},\end{aligned}$$

where

$$R_{N,4} = \int_c \left\{\frac{\bar{h}_N^*}{\bar{h}_N} - \frac{\bar{h}_0^*}{\bar{h}_0}\right\}\frac{1}{\bar{h}_0}(\bar{h}_N - \bar{h}_0)(\bar{Q}_0 - \bar{Q}^*)(c)\bar{h}_0(c)d\mu(c).$$

We assumed that $R_{N,4} = o_P(1/\sqrt{N})$. Using that $\bar{h}_0 = \frac{1}{N}\sum_i h_i(g_0, Q_{W,0})$, $\bar{h}_N = \frac{1}{N}\sum_{i=1}^N h_i(g_N, Q_{W,N})$, and $\bar{h}_N^* = \frac{1}{N}\sum_i h_i(g^*, Q_{W,N})$, it follows that (eq. 20) reduces to

$$\begin{aligned}
&\frac{1}{N}\sum_{i=1}^N \int_c \left\{ \frac{h_i(g^*, Q_{W,N})}{\bar{h}_0}(c) - \frac{\bar{h}_0^*}{\bar{h}_0^2} h_i(g_N, Q_{W,N}) \right\} (\bar{Q}_0 - \bar{Q}^*)(c)\bar{h}_0(c)d\mu(c) + o_P(1/\sqrt{N}) \\
&= \frac{1}{N}\sum_{i=1}^N \int_c \left\{ \frac{h_i(g^*, Q_{W,N})}{\bar{h}_0}(c) - \frac{\bar{h}_0^*}{\bar{h}_0^2} h_i(g_0, Q_{W,N}) \right\} (\bar{Q}_0 - \bar{Q}^*)(c)\bar{h}_0(c)d\mu(c) \\
&\quad - \frac{1}{N}\sum_{i=1}^N \int_c \frac{\bar{h}_0^*}{\bar{h}_0^2} h_i(g_N - g_0, Q_{W,N})(\bar{Q}_0 - \bar{Q}^*)(c)\bar{h}_0(c)d\mu(c) + o_P(1/\sqrt{N}) \\
&\equiv \frac{1}{N}\sum_{i=1}^N f_{W,i}^2(W) \\
&\quad - \frac{1}{N}\sum_{i=1}^N \int_c \frac{\bar{h}_0^*}{\bar{h}_0^2} h_i(g_N - g_0, Q_{W,N})(\bar{Q}_0 - \bar{Q}^*)(c)\bar{h}_0(c)d\mu(c) + o_P(1/\sqrt{N}) \\
&\equiv Z_{W,N}^2/\sqrt{N} - \frac{1}{N}\sum_{i=1}^N \int_c \frac{\bar{h}_0^*}{\bar{h}_0^2} h_i(g_N - g_0, Q_{W,0})(\bar{Q}_0 - \bar{Q}^*)(c)\bar{h}_0(c)d\mu(c) \\
&\quad - \frac{1}{N}\sum_{i=1}^N \int_c \frac{\bar{h}_0^*}{\bar{h}_0^2} h_i(g_N - g_0, Q_{W,N} - Q_{W,0})(\bar{Q}_0 - \bar{Q}^*)(c)\bar{h}_0(c)d\mu(c) + o_P(1/\sqrt{N}) \\
&\equiv Z_{W,N}^2/\sqrt{N} + \int_c \frac{\bar{h}_0^*}{\bar{h}_0^2} \frac{\bar{h}(g_N - g_0, Q_{W,0})}{N} (\bar{Q}^* - \bar{Q}_0)(c)\bar{h}_0(c)d\mu(c) - R_{N,5} + o_P(1/\sqrt{N}),
\end{aligned}$$

where we note that $P_0 f_{W,i}^2 = 0$, and we defined

$$\begin{aligned}
R_{N,5} &= \frac{1}{N}\sum_{i=1}^N \int_c \frac{\bar{h}_0^*}{\bar{h}_0^2} h_i(g_N - g_0, Q_{W,N} - Q_{W,0})(\bar{Q}_0 - \bar{Q}^*)(c)\bar{h}_0(c)d\mu(c) \\
&= \frac{1}{N}\sum_{i=1}^N \int_c \frac{\bar{h}_0^*}{\bar{h}_0^2} (\bar{Q}_0 - \bar{Q}^*)(c)\bar{h}_0(c)d\mu(c) \left\{ \int_a I(c_i^Y(a, W) = c)(g_N - g_0)(a|W) \right. \\
&\quad \left. - \int_w \int_a I(c_i^Y(a, w) = c)(g_N - g_0)(a|w)Q_{W,0}(dw) \right\} \\
&= \frac{1}{N}\sum_{i=1}^N \int_a \frac{\bar{h}_0^*}{\bar{h}_0^2} (\bar{Q}_0 - \bar{Q}^*)(c_i^Y(a, W))(g_N - g_0)(a|W) \\
&\quad - \frac{1}{N}\sum_{i=1}^N \int_w \int_a \frac{\bar{h}_0^*}{\bar{h}_0^2} (\bar{Q}_0 - \bar{Q}^*)(c_i^Y(a, w))(g_N - g_0)(a|w)Q_{W,0}(dw) \\
&\equiv \frac{1}{N}\sum_{i=1}^N \left\{ f_{W,i}^3(\bar{g}_N) - f_{W,i}^3(\bar{g}_0) \right\}(W) \\
&\equiv Z_{W,N}^3(\bar{g}_N)/\sqrt{N} - Z_{W,N}^3(\bar{g}_0)/\sqrt{N},
\end{aligned}$$

where we defined the process $Z_{W,N}^3(\bar{g}) = \frac{1}{\sqrt{N}}\sum_{i=1}^N \{f_{W,i}^3(\bar{g})(W) - P_0 f_{W,i}^3(\bar{g})\}$ with

$$f_{W,i}^3(\bar{g})(W) = \int_a \frac{\bar{h}_0^*}{\bar{h}_0^2} (\bar{Q}_0 - \bar{Q}^*)(c_i^Y(a, W))g(a|W).$$

The term $Z^2_{W,N} = 1/\sqrt{N}\sum_i\{f^2_{W,i}(W) - P_0 f^2_{W,i}\}$ is included in the first-order expansion and thus partly characterizes the normal limit distribution of $\psi^*_N$, so that its analysis will be part of the analysis of the first-order approximation. Since $h_i(g, Q_{W,N})$ only depends on $(W_1, \ldots, W_N)$ through $(W_j : j \in F_i)$, where we condition on $F_1, \ldots, F_N$, we will indeed be able to show that a term $Z^2_{W,N}$ is nicely behaved empirical process (converging to a normal distribution), even though each $i$-specific term is correlated with the $j$-specific terms when $F_i \bigcap F_j \neq \emptyset$.

Showing that $Z^3_{W,N}(\bar{g}_N) - Z^3_{W,N}(\bar{g}_0) = o_P(1)$ comes down to showing that $(Z^3_{W,N}(\bar{g}) : \bar{g} \in G)$ is an asymptotic equicontinuous process w.r.t. supremum norm, and that $\bar{g}_N - \bar{g}_0$ converges to zero w.r.t. the supremum norm. In this manner, we show that

$$R_{N,5} = o_P(1/\sqrt{N}) \tag{A4}$$

We assumed that

$$\int_c \frac{\bar{h}^*_0}{\bar{h}^2_0}\frac{\bar{h}(g_N - g_0, Q_{W,0})}{N}(\bar{Q}^* - \bar{Q}_0)(c)\bar{h}_0(c)d\mu(c)$$

$$= \frac{1}{N}\sum_{i=1}^N f^1_{A,i}(O_i) + o_P(1/\sqrt{N}),$$

where $f^1_{A,i}(O)$ only depends on $O$ through $(A_i, (W_j : j \in F_i))$, and $E_0(f^1_{A,i}(O)|W) = 0$.

Thus, if we prove (eq. A2), (eq. A3), and (eq. A4), then we have obtained the following first-order expansion:

$$\begin{aligned}\psi^*_N - \psi_0 = &\frac{1}{N}\sum_{i=1}^N\{f^1_{W,i}(W) - P_0 f^1_{W,i}\} + \frac{1}{N}\sum_i\{f^2_{W,i}(W) - P_0 f^2_{W,i}\}\\ &+ (P_N - P_0)D^*_Y(\bar{Q}^*, Q_{W,0}, \bar{h}_0)\\ &+ \frac{1}{N}\sum_{i=1}^N f^1_{A,i}(O) + o_P(1/\sqrt{N}).\end{aligned}$$

**Analysis of first-order approximation:** Let $\bar{f}_{W,i} = f^1_{W,i} + f^2_{W,i}$. The first-order approximation equals

$$\begin{aligned}&1/N\sum_i\{D^*_{Y,i}(\bar{Q}^*, Q_{W,0}, \bar{h}_0)(O_i) + \bar{f}_{W,i}(W) + f^1_{A,i}(O) - P_0\{D^*_{Y,i} + \bar{f}_{W,i}\}\}\\ &\equiv 1/N\sum_i f_i(O).\end{aligned}$$

It remains to prove that this first-order expansion converges to a normal limit distribution. This proof has its own outline. Firstly, we decompose $1/N\sum_i f_i(O)$ by $f_i = f_{W,i} + f_{A,i} + f_{Y,i}$, where $f_{W,i} = E_0(f_i|W) - E_0 f_i$, $f_{A,i} = E_0(f_i|A, W) - E_0(f_i|W)$, and $f_{Y,i} = f_i - E_0(f_i|A, W)$. We can represent $\frac{1}{N}\sum_i f_i(O)$ as $Z_{NY}/\sqrt{N} + Z_{NA}/\sqrt{N} + Z_{NW}/\sqrt{N}$, where $Z_{NW} = 1/\sqrt{N}\sum_i f_{W,i}$, $Z_{NA} = 1/\sqrt{N}\sum_i f_{A,i}$, and $Z_{NY} = 1/\sqrt{N}\sum_i f_{Y,i}$. It follows that $f_{W,i}$ simplifies to:

$$\begin{aligned}f_{W,i}(W) &= \int_a \bar{Q}_0(c^Y_i(a, W))g^*(a|W)\\ &\quad - \int_{a,w}\bar{Q}_0(c^Y_i(a, W))g^*(a, W)Q_{W,0}(dw)\\ &= \int_c \bar{Q}_0(c)g^*_i(c|W) - \int_{c,w}\bar{Q}_0(c)g^*_i(c|W)Q_{W,0}(dw).\end{aligned}$$

In addition,

$$f_{Y,i} = D^*_{Y,i} - E_0(D^*_{Y,i}|A, W) = \frac{\bar{h}^*_0}{\bar{h}_0}(C^Y_i)(Y_i - \bar{Q}_0(C^Y_i)),$$

and $f_{A,i} = E_0(D^*_{Y,i}|A, W) - E_0(D^*_{Y,i}|W) + f^1_{A,i}$. We also note that, conditional on $W, A$, $Z_{NY}$ is a sum of independent mean zero random variables $f_{Y,i}(Y_i)$ (functions of $Y_i$); conditional on $W$, $Z_{NA}$ is a sum of

$f_{A,i}(A_j : j \in F_i)$ with conditional mean zero, given $W$; $A_i$, $i = 1, \ldots, N$ are (conditionally) independent, given $W$; and, finally, $Z_{NW} = 1/\sqrt{N} \sum_i \{f_{W,i}(W_j : j \in R_i) - P_0 f_{W,i}\}$, with $W$ satisfying our independence assumption (e.g. $W_1, \ldots, W_N$ are independent). Recall that the sets $R_i$ are defined such that $g^*(A_j : j \in F_i | W)$ only depends on $W$ through $(W_j : j \in R_i)$.

Exploiting these independence structures, we will show that

$$\begin{aligned} Z_{NY} &\Rightarrow_d N(0, \sigma_Y^2) \\ Z_{NA} &\Rightarrow_d N(0, \sigma_A^2) \\ Z_{NW} &\Rightarrow_d N(0, \sigma_W^2) \end{aligned} \tag{A1}$$

with the expressions for $\sigma_Y^2$, $\sigma_A^2$, and $\sigma_W^2$ as specified in the theorem. Here (eq. A1) represents all three convergence statements. Due to the orthogonality of the three empirical processes, using moment generating functions, our results also imply $Z_{NY} + Z_{NA} + Z_{NW} \Rightarrow_d N(0, \sigma^2 = \sigma_Y^2 + \sigma_A^2 + \sigma_W^2)$. For example, we can analyze $E(Z_{NY} + Z_{NA})^p = \sum_t c(p,k) E\{Z_{NY}\}^k E\{Z_{NA}\}^{p-k}$ and use convergence of moments of each process separately to establish convergence to $E(Z_Y + Z_A)^p$. Once we have convergence of all moments, and we can bound $E(Z_{NY} + Z_{NA})^p \leq C^p$ for some $C < \infty$, which follows from results established in our separate analysis, then we obtain convergence in moment generating function, and thereby weak convergence of the sum $Z_{NY} + Z_{NA}$. In this manner, the desired weak convergence of the sum $Z_{NY} + Z_{NA} + Z_{NW}$ is shown.

This finishes the outline of the proof. It remains to establish (eq. A1), (eq. A2), (eq. A3), and (eq. A4).

## (A3)

### (A3): Outline of proof

Let $\tilde{h} = \bar{h}^*/\bar{h}$ and we will denote $D_Y^*$ with $D_Y^*(\tilde{h}, \bar{Q})$. Our goal is to prove that $\sqrt{N}(P_N - P_0)\{D_Y^*(\tilde{h}_0, \bar{Q}_N^*) - D^*(\tilde{h}_0, \bar{Q}^*)\} = o_P(1)$ and $\sqrt{N}(P_N - P_0)\{D_Y^*(\tilde{h}_N, \bar{Q}_N^*) - D_Y^*(\tilde{h}_0, \bar{Q}_N^*)\} = o_P(1)$. Let $P_{0,Y|A,W} f$, $P_{0,A|W} f$, and $P_{0,W} f$, denote the expectation operators w.r.t. their respective conditional distributions. We have

$$\begin{aligned} Z_N(\tilde{h}, \bar{Q}) &= \sqrt{N}(P_N - P_0) D_Y^*(\tilde{h}, \bar{Q}) = \sqrt{N}(P_N - P_{0,Y|A,W}) D_Y^*(\tilde{h}, \bar{Q}) \\ &\quad + \sqrt{N}(P_N - P_{0,A|W}) P_{0,Y|A,W} D_Y^*(\tilde{h}, \bar{Q}) \\ &\quad + \sqrt{N}(P_N - P_{0,W}) P_{0,A|W} P_{0,Y|A,W} D_Y^*(\tilde{h}, \bar{Q}) \\ &= \frac{1}{\sqrt{N}} \sum_{i=1}^N \tilde{h}(C_i^Y)(Y_i - \bar{Q}_0(C_i^Y)) \\ &\quad + \frac{1}{\sqrt{N}} \sum_{i=1}^N \{\tilde{h}(C_i^Y)(\bar{Q}_0 - \bar{Q})(C_i^Y) - \int_c \tilde{h}(c)(\bar{Q}_0 - \bar{Q})(c) g_{0,i}(c|W)\} \\ &\quad + \frac{1}{\sqrt{N}} \sum_{i=1}^N \{\int_c \tilde{h}(\bar{Q}_0 - \bar{Q})(c) g_{0,i}(c|W) - P_0 D_Y^*(\tilde{h}, \bar{Q})\} \\ &\equiv Z_{NY}(\tilde{h}) + Z_{NA}(\tilde{h}, \bar{Q}) + Z_{NW}(\tilde{h}, \bar{Q}). \end{aligned}$$

We now note that, for a fixed $(\tilde{h}, \bar{Q})$, conditional on $(W, A)$, $Z_{NY}$ is a sum of independent mean zero random variables $f_{Y,i}(Y_i)$ (functions of $Y_i$). We also note that for a fixed $(\tilde{h}, \bar{Q})$, conditional on $W$, $Z_{NA}$ is a sum of mean zero $f_{A,i}((A_j : j \in F_i))$, where $A_i$, $i = 1, \ldots, N$ are (conditionally) independent. Finally, for a fixed $(\tilde{h}, \bar{Q})$, $Z_{NW} = 1/\sqrt{N} \sum_i f_{W,i}(W_j : j \in F_i) - P_0 f_{W,i}$, and, by assumption on $Q_{W,0}$, for each $i$, $f_{W,i}$ is only dependent on maximally $K$ $f_{W,j}$.

Let $\bar{Q}^*$ be the limit of $\bar{Q}_N$, and let $\tilde{h}_0 = \bar{h}_0^*/\bar{h}_0$ be the limit of $\tilde{h}_N$. By exploiting these independence structures, we will use empirical process theory to establish that

$$Z_{NY}(\tilde{h}_N) = Z_N^Y(\tilde{h}_0) + o_P(1)$$

$$Z_{NA}(\tilde{h}_N, \bar{Q}_N^*) = Z_{NA}(\tilde{h}_0, \bar{Q}^*) + o_P(1)$$

$$Z_{NW}(\tilde{h}_N, \bar{Q}_N^*) = Z_{NW}(\tilde{h}_0, \bar{Q}^*) + o_P(1).$$

This then establishes $R_{N,2} = o_P(1/\sqrt{N})$ and $R_{N,3} = o_P(1/\sqrt{N})$.

### (A3): Outline of establishing asymptotic equicontinuity of a process

For that purpose, we will apply Lemma 5 in van der Vaart and Wellner [41], which concerns establishing weak convergence of a process $(Z_N(\theta) : \theta \in \mathcal{F})$, indexed by a $\theta = (\tilde{h}, \bar{Q}) \in \mathcal{F}$. Given that $\mathcal{F}$ is a subset of some metric space of functions with metric $d$, one defines $N(\epsilon, F, d)$ as the minimal number of balls of size $\epsilon$ needed to cover $\mathcal{F}$. In addition, for a given strictly monotone function $\lambda : \mathbb{R}_{\geq 0} \to \mathbb{R}$, let $\| Z_N(\theta) \|_\lambda = \inf\{c_0 : \lambda(|Z_N(\theta)|/c_0) \leq 1\}$ be the so-called orlics norm of the random variable $Z_N(\theta)$.

For example, one can select the $L^p$-norm $\| Z_N(\theta) \|_p = \left\{\int E\{Z_N(\theta)\}^p\right\}^{1/p}$ of $Z_N(\theta)$ for arbitrary large $p$ which correspond with the choice of orlics norm defined by $\lambda_p(x) = x^p$. The orlics norm implied by $\lambda_{2,e}(x) = \exp(x^2) - 1$ is the typical orlics norm pursued in the case of sums of independent random variables, and this is the one we will also use.

This Lemma 5 states that, if (1) $\| Z_N(\theta_1) - Z_N(\theta_2) \|_\lambda$ is bounded by $cd(\theta_1, \theta_2)$ for some universal constant $c$ and metric $d(\cdot, \cdot)$, (2) $\mathcal{F}$ is totally bounded w.r.t. this metric $d$, (3) for some $\eta > 0$, $\int_0^\eta \lambda^{-1}(N(\epsilon, \mathcal{F}, d))d\epsilon < \infty$, (4) the marginal distributions $Z_N(\theta)$ converge to a normal distribution $Z(\theta)$, then $Z_N$ converges weakly to a Gaussian process $Z$ in $\ell^\infty(\mathcal{F})$, where $\ell^\infty(\mathcal{F})$ is the metric space of functions $G : \mathcal{F} \to \mathbb{R}$ endowed with supremum norm $\| G \|_F = \sup_{\theta \in \mathcal{F}} |G(\theta)|$. We assumed that our parameter space $\mathcal{F}$ for $(\tilde{h}_0, \bar{Q}_0)$ consists of uniformly bounded functions on a set $\mathcal{C}^Y$ that contains $C_i^Y(A, W)$ with probability 1, and we defined the metric $d$ as the supremum norm. Thus, (2) holds. We posed (3) as an entropy condition on the parameter space $\mathcal{F}$, which will thus hold by assumption. For example, $\mathcal{F}$ could be the class of functions on $\mathcal{C}^Y \subset \mathbb{R}^d$ that have uniform sectional variation norm bounded by a $M < \infty$, in which case this entropy condition holds. Under conditions 1–3 we have that the process $Z_N$ is asymptotically tight, and, for any sequence $\delta_n \to 0$, we have for each $x > 0$,

$$P\left(\sup_{d(\theta_1, \theta_2) < \delta_n} |Z_N(\theta_1) - Z_N(\theta_2)| > x\right) \to 0 \text{ as } N \to \infty.$$

So once we have established the orlics-norm condition (1), then this tightness can be used to establish that terms $Z_N(\theta_N) - Z_N(\theta) = o_P(1)$ for random $\theta_N \in \mathcal{F}$ converging to $\theta \in \mathcal{F}$ w.r.t. metric $d$ in probability, assuming $\mathcal{F}$ satisfies the entropy condition and is totally bounded w.r.t. this metric $d$.

### Bounding the orlics norm of our empirical processes

The orlics norm $\| \cdot \|_\lambda$ indexed by function $\lambda_2(x) = \exp(x^2) - 1$ is defined as

$$\| X \|_\lambda = \inf\left\{c > 0 : E \exp(|X|^2/C) - 1 \leq 1\right\}.$$

We consider a stochastic process $X_N(\theta)$ indexed by $\theta \in \mathcal{F}$ for a class of functions $\mathcal{F}$. In our application, we have that, for example, $\theta = (\bar{Q}, \tilde{h}) \in \mathcal{F}$ represents two real valued functions $\bar{Q}$ and $\tilde{h}$ defined on a set $\mathcal{C}^Y \subset \mathbb{R}^d$ that contains $\{c_i^Y(A, W) : i\}$ with probability 1. In addition, our processes can be represented as

$X_N(\theta) = 1/\sqrt{N}\sum_{i=1}^N f_i(\theta)(O)$, where, for example, for each $i$ there is an associated set $F_i \subset \{1,\ldots,N\}$, and, if $F_i \cap F_j = \emptyset$, then $f_i(\theta)(O)$ and $f_j(\theta)(O)$ are independent, and, in general, it is known that for each $i$, $f_i(\theta)$ is independent of $\{f_j(\theta) : j \in S_i^c\}$ for sets $S_i$ with $\max_i |S_i| \leq K$. For some of our processes, these independencies are conditional on a random variable (e.g. conditional on infinite sequence $(W_1,\ldots,),(A_1,\ldots)$). In that case, we will apply our general proof below conditional on this random variable and obtain a bound on the orlics norm that holds for almost every value of the conditioning random variable. For example, one establishes a universal bound $C$ in $\| X_N(\theta_1) - X_N(\theta_2) \|_\lambda < C \| f_1 - f_2 \|$ with the $P$ in the orlics norm being a conditional distribution, given a value of the conditioning random variable) where $C$ does not depend on the value of the random variable one conditions upon. Finally, we really need to bound $\| X_N(\theta_1) - X_N(\theta_2) \|_\lambda$, so that we will apply the lemmas below to $X_N(\theta_1) - X_N(\theta_2)$ instead of $X_N(\theta)$.

So our goal is to bound $\| X_N(\theta) \|_\lambda \leq C \| \theta \|$ for some universal (in $N$ and $\theta \in \mathcal{F}$) . As outlined in previous subsection, the choice of orlics norm and norm $\| \theta \|$ for $\theta \in \mathcal{F}$ is important, since the corresponding entropy requirement on $\mathcal{F}$ is that $\int_0^\eta \lambda^{-1}(N(\epsilon, \| \cdot \|, \mathcal{F}))d\epsilon < \infty$. We will establish our results for the strongest orlics norm which corresponds with $\lambda_2(x)$, while we select the supremum norm $\| \theta \| = \max(\| \bar{Q} \|_\infty, \| \tilde{h} \|_\infty)$ for the functions $\theta \in \mathcal{F}$.

**Lemma 3** Let $\| X \|_\lambda$ *be the orlics norm defined above w.r.t.* $\lambda(x) = \exp(x^2) - 1$. *Suppose that for each* $p$

$$E|X_N(\theta)|^p \leq C(N,p) \| \theta \|^p .$$

*Let* $D(N)$ *be a number so that*

$$\sum_{p=1}^\infty C(N,2p)D(N)^{2p}/p! \leq 1.$$

*Then*,

$$\| X_N(\theta) \|_\lambda \leq \frac{1}{D(N)} \| \theta \| .$$

*In particular, if* $C(N,p)$ *can be bounded from above by* $C(p)$ *constant in N, and one finds a D (constant in N) so that* $\sum_{p=1}^\infty C(2p)D^{2p}/p! \leq 1$, *then it follows that* $\| X_N(\theta) \|_\lambda \leq \frac{1}{D} \| \theta \|$.

**Proof.** We first note

$$E\exp\{(X_N(\theta)/C)\}^2 - 1 = \sum_{p=1}^\infty \frac{E(|X_N(\theta)|/C)^{2p}}{p!} = \sum_{p=1}^\infty \frac{E|X_N(\theta)|^{2p}}{C^{2p}p!}.$$

Suppose that for each even $p$ $E|X_N(\theta)|^p \leq C(N,p) \| \theta \|^p$ . Then, we have

$$E\exp\{X_N(\theta)/C\}^2 - 1 \leq \sum_{p=1}^\infty \frac{C(N,2p) \| \theta \|^{2p}}{C^{2p}p!}.$$

So $\| X_N(\theta) \|_\lambda$ is bounded by a $C$ chosen so that

$$\sum_{p=1}^\infty \frac{C(N,2p)}{p!}\left(\frac{\| \theta \|}{C}\right)^{2p} \leq 1.$$

Let $D(N)$ be a number so that

$$\sum_{p=1}^\infty C(N,2p)D(N)^{2p}/p! \leq 1.$$

Then, $C$ can be selected so that $\| f \| /C \leq D(N)$, or equivalently, $C \geq \| \theta \| /D(N)$. Thus, we have shown that $\| X_N(\theta) \|_\lambda \leq \frac{1}{D(N)} \| \theta \|$ . The last statement is straightforwardly shown. □

Thus, apparently, it suffices to establish a bound of the type $E|X_N(\theta)|^p \leq C(N,p) \parallel \theta \parallel^p$ for some $C(N,p)$ that is somewhat well behaved as function in $p$ for $p \to \infty$ so that the previous lemma applies.

We use the following lemma to bound the $p$th moment of $X_N(\theta)$.

**Lemma 4** *Assume that, for each $i = 1, \ldots, N$, and each integer p, we have a universal constant C so that*

$$\parallel f_i(\theta) \parallel_p \equiv (E(f_i(\theta)(O))^p)^{1/p} \leq C \parallel \theta \parallel . \tag{21}$$

*Then, we have*

$$E\prod_{j=1}^{p} f_j \leq \prod_{j=1}^{p} \parallel f_j \parallel_{2j} \leq C^p \parallel \theta \parallel^p .$$

The bounding (eq. 21) is a straightforward consequence of our conditions stated in the theorem, where we use the supremum norm on $\theta \in F$, thereby allowing us to apply this lemma.

**Proof.** By repeatedly applying Cauchy–Schwarz inequality, it follows that

$$E\prod_{j} f_j(\theta) \leq \prod_{j=1}^{p} \left(Ef_j(\theta)^{2j}\right)^{1/(2j)} = \prod_{j=1}^{p} \parallel f_j(\theta) \parallel_{2j} .$$

By assumption, $\parallel f_j(\theta) \parallel_{2j} \leq C \parallel \theta \parallel$, so that the latter is bounded by $C^p \parallel \theta \parallel^p$ . □

The following lemma provides us with an upper bound for $C(N,p)$ so that $E|X_N(\theta)|^p \leq C(N,p) \parallel \theta \parallel$ .

**Lemma 5** *Assume that, for each i, and each integer p, we have a universal constant C so that*

$$\parallel f_i(\theta) \parallel_p = (E(f_i(\theta)(O))^p)^{1/p} \leq C \parallel \theta \parallel .$$

*Let $R(i_1, \ldots, i_p)$ be an indicator, identified by indices $\vec{i} = (i_1, \ldots, i_p) \in \{1, \ldots, N\}^p$, which equals 1 if there exist a set $F(i_l)$ among the sets $F(i_1), \ldots, F(i_p)$ that is disjoint from the other sets. More generally, we can define $R(i_1, \ldots, i_p)$ equals 1 if there exists an element $j \in \{i_1, \ldots, i_p\}$ so that $f_j(\theta)$ is independent of $f_k(f)$ for all $k \in \{i_1, \ldots, i_p\}$ with $k \neq j$.*

*Let*

$$C(N,p) \equiv N^{-p/2} \sum_{\vec{i}} (1 - R(\vec{i})).$$

*Then*

$$\parallel X_N(\theta) \parallel_p^p \leq C(N,p) C^p \parallel \theta \parallel^p .$$

**Proof.** We have

$$E\left(1/\sqrt{N}\sum_i f_i\right)^p = N^{-p/2} \sum_{i_1,\ldots,i_p} E\prod_{j=1}^{p} f_{i_j}$$

$$= N^{-p/2} \sum_{i_1,\ldots,i_p} (1 - R(i_1, \ldots, i_p)) E\prod_{j=1}^{p} f_{i_j}.$$

By the previous lemma, we have $E\prod_{j=1}^{p} f_j \leq C^p \parallel \theta \parallel^p$ for a $C < \infty$, so that we obtain

$$E\left(1/\sqrt{N}\sum_i f_i\right)^p \leq N^{-p/2} \sum_{i_1,\ldots,i_p} (1 - R(i_1, \ldots, i_p)) C^p \parallel \theta \parallel^p . \square$$

By putting a bound on $|F_i|$, we can obtain a nice bound on $\sum_{i_1,\ldots,i_p} (1 - R(i_1, \ldots, i_p))$, so that the previous lemma combined with Lemma 3 results in the following lemma providing the desired universal bound on the orlics norm.

**Lemma 6** *Assume that, for each i, and each p, we have a universal constant C so that*

$$\| f_i(\theta) \|_p = (E(f_i(\theta)(O))^p)^{1/p} \leq C \| \theta \| .$$

*Assume that $f_i$ is independent of $\{f_j : j \in S_i^c\}$ for a set $S_i \subset \{1,\ldots,N\}$ and $\max_i |S_i| \leq K$. For p an integer, we have $E|X_N(\theta)|^p \leq C(N,p)C^p \| \theta \|^p$ , where*

$$\begin{aligned} C(N,p) &\equiv N^{-p/2} \sum_{i_1,\ldots,i_p} (1 - R(i_1,\ldots,i_p)) \\ &\leq 2^p (Kp)^{p/2} (N - Kp)^{p/2} N^{-p/2}. \end{aligned}$$

*For $\lambda(x) = e^{x^2} - 1$, we have $\| X_N(\theta) \|_\lambda \leq C_1(K) \| \theta \|$ for some $C_1(K) \leq C\sqrt{K}$ for some universal $C < \infty$.*

**Proof.** We first need to show that $C(N,p) \leq 2^p (Kp)^{p/2} (N - Kp)^{p/2} N^{-p/2}$. Selection of one particular $\vec{i}$ corresponds with $p$ times in a row selecting an element in $\{1,\ldots,N\}$. Without restrictions on this sequence of $p$ draws, one has $N$ options at each of the subsequent $p$ steps resulting in $N^p$ vectors $\vec{i}$. Suppose we have arrived at the $l$th draw, so that we have a sequence $(i_1,\ldots,i_{l-1})$ with corresponding sets $S(i_1),\ldots,S(i_{l-1})$. For a next $i_l$ we define a binary $B(i_l) = 1$ if $S(i_l) \cap \cup_{s=1}^{l-1} S(i_s) = \emptyset$. Suppose $B(i_l) = 1$. $\{i_1,\ldots,i_l\}$, $i_l$ is an island, *and* one cannot find a single element $i_m$ in $\{1,\ldots,N\}/\{i_1,\ldots,i_l\}$ for which $i_m$ is an element of both (1) $S_{i_l}$ and (2) $\cup_{s<l-1} S_{i_s}$, since we arranged that $S(i_l) \cap \cup_{s=1}^{l-1} S(i_s) = \emptyset$. As a consequence, an element with $B(i_l) = 1$ will need at least one future $s > l$ selection with $B(i_s) = 0$ in order to connect $i_l$ with $i_s$, and such a future selection $s$ cannot simultaneously connect with another $i_j$ with $j < l$. As a consequence, if the sequence of $p$ elements $(B(i_1),\ldots,B(i_p))$ has more than $p/2$ 1's, then there will be at least one island $\{i_l\}$ among $\{i_1,\ldots,i_p\}$ of size 1 with $B(i_l) = 1$. Thus, in that case $1 - R(\vec{i}) = 0$. Thus, we only need to count the vectors $\vec{i}$ for which $B(i_1),\ldots,B(i_p)$ has at most $p/2$ 1's.

For a choice with $B(i_l) = 1$, we have at most $N - Kp$ possible choices since we cannot select any of the elements in $S(i_1),\ldots,S(i_{l-1})$. For a choice with $B(i_l) = 0$, we have maximally $Kp$ choices. The total number of sequences $B(i_1),\ldots,B(i_p)$ for which there are at most $p/2$ 1's is upper-bounded by $2^p$. The total number of sequences $\vec{i}$ present in one such sequence is given by $(Kp)^{p/2}(N - Kp)^{p/2}$. To conclude, we have the following upper bound

$$C(N,p) \leq 2^p (Kp)^{p/2} (N - Kp)^{p/2} N^{-p/2},$$

which proves our first result.

Thus, we have $E|X_N(\theta)|^p \leq C^p C(N,p) \| \theta \|^p$ with $C(N,p)$ bounded by this upper bound. We now want to bound the orlics norm $\| X_N(\theta) \|_{\lambda_2}$ . Let us first do this for the orlics norm $\lambda_1(x) = \exp(x) - 1$. Using that $p! \geq (p/2)!(p/2)^{p/2}$, $(N - Kp)/N \leq 1$, we have

$$\begin{aligned} \| X_N(\theta) \|_{\lambda_1} &= \inf\left\{ c_0 : \sum_{p=1}^{\infty} \frac{C^p C(p,N)}{p!} \frac{\| \theta \|^p}{c_0^p} \leq 1 \right\} \\ &= \inf\left\{ c_0 : \sum_{p=1}^{\infty} \frac{C^p 2^p}{p!} (Kp)^{p/2} \frac{(N - Kp)^{p/2}}{N^{p/2}} \frac{\| \theta \|^p}{c_0^p} \leq 1 \right\} \\ &\leq \inf\left\{ c_0 : \sum_{p=1}^{\infty} \left( \frac{2C\sqrt{K} \| \theta \|}{c_0} \right)^p \frac{p^{p/2}}{(p/2)!(p/2)^{p/2}} \leq 1 \right\} \\ &= \inf\left\{ c_0 : \sum_{p=1}^{\infty} \left( \frac{2\sqrt{2}C\sqrt{K} \| \theta \|}{c_0} \right)^p \frac{1}{(p/2)!} \leq 1 \right\}. \end{aligned}$$

Thus there exists *a* $c_0 = c_0(K,C) \| \theta \|$ so that the term on the left of the inequality is smaller or equal than 1, so that we have shown $\| X_N(\theta) \|_{\lambda_1} \leq c_0(K,C) \| \theta \|$ . It also follows that $c_0(K,C)$ can be bounded by a universal constant times $\sqrt{K}$. This completes the proof for this orlics norm identified by $\lambda_1$.

Let us now do the proof for the orlics norm identified by $\lambda_2$. Note $E\exp\{(X_N(\theta)/c_0)^2\}-1=\sum_{p=1}^{\infty}\frac{EX_N(\theta)^{2p}}{c_0^{2p}p!}$. Thus, we have

$$\begin{aligned}\| X_N(\theta)\|_{\lambda_2} &= \inf\left\{c_0 : \sum_{p=1}^{\infty}\frac{C^{2p}C(2p,N)}{p!}\frac{\| \theta\|^{2p}}{c_0^{2p}}\le 1\right\}\\ &= \inf\left\{c_0 : \sum_{p=1}^{\infty}\frac{C^{2p}2^{2p}}{p!}(K2p)^p\frac{(N-K2p)^p}{N^p}\frac{\| \theta\|^{2p}}{c_0^{2p}}\le 1\right\}\\ &\le \inf\left\{c_0 : \sum_{p=1}^{\infty}\left(\frac{2\sqrt{2}C\sqrt{K}\| \theta\|}{c_0}\right)^{2p}\frac{p^p}{p!}\le 1\right\}.\end{aligned}$$

The term within $()^{2p}$ can be made smaller than an arbitrary number $\delta>0$ by just selecting $c_0$ large enough. Therefore, we need to show that $\sum_{p=1}^{\infty}\delta^p p^p/p!$ is bounded for some small enough $\delta>0$. The proof then proceeds as above for the $\lambda_1$-orlics norm. Now, we note that, using $1-x\approx\exp(-x)$ for $x\approx 0$, and $\sum_{j=1}^{p}(j-1)=p(p-1)/2$,

$$\begin{aligned}\frac{p!}{p^p}&=\prod_{j=1}^{p}\{1-(j-1)/p\}\approx\exp(-\textstyle\sum_{j=1}^{p}(j-1)/p)\\ &=\exp(-\textstyle\sum_{j,=1}^{p}(j-1)/p)=\exp(-1/ppp(p-1)/2)=\exp(-(p-1)/2).\end{aligned}$$

Thus, $\sum_{p=1}^{\infty}\delta^p p^p/p!$ behaves as $\sum_{p=1}^{\infty}\delta^p\exp((p-1)/2)$. Since $\exp((p-1)/2)\le\exp(p)$, by selecting $\delta$ small enough with $\delta*\exp(1)<1$, this sum can be made arbitrarily small. As before it follows that $c_0$ can be bounded by universal constant times $\sqrt{K}$. □

### (A3): Asymptotic equicontinuity of $Z_{NY}(\tilde{h})$

The process $Z_{NY}=1/\sqrt{N}\sum_i f_{Y,i}$ is a sum of independent random variables conditional on $(W,A)$, so that its analysis is a simple imitation of the general analysis presented in previous subsection, conditional on $W,A$. The proof that the $\|\cdot\|_p$ -norm (conditional on $W,A$) of $f_{Y,i}(\tilde{h}_1)-f_i^Y(\tilde{h}_2)$ is bounded by a universal constant times the supremum norm of $\tilde{h}_1-\tilde{h}_2$ is as follows:

$$\begin{aligned}E_{|A,W}\{(\tilde{h}_1-\tilde{h}_2)(C_i^Y)(Y_i-\bar{Q}_0(C_i^Y))\}^p &\le \| \tilde{h}_1-\tilde{h}_2\|_\infty^p E_{|A,W}(Y_i-\bar{Q}_0(C_i^Y))^p\\ &\equiv C^p\| \tilde{h}_1-\tilde{h}_2\|_\infty^p,\end{aligned}$$

where, because $Y_i\le 1$ and $\bar{Q}_0\le 1$, we have that $C\le 1$.

### (A3): Asymptotic equicontinuity of $Z_{NA}(\tilde{h},\bar{Q})$

Conditional on $W$, for a fixed $\theta=(\tilde{h},\bar{Q})$, we can represent this process as $1/\sqrt{N}\sum_i\{f_{A,i}(O)-P_0^W f_{A,i}\}$, where $f_{A,i}$ depends on $A$ through $(A_j : j\in F_i))$, while all $A_i$, $i=1,\ldots,N$, are independent. As a consequence, for each $i$, conditional on $W$, $f_{A,i}$ is independent of $(f_{A,j} : j, F_j\cap F_i=\emptyset)$. Again, the above general analysis can be applied, and the proof that the $\|\cdot\|_p$ -norm of $f_{A,i}(\theta_1)-f_{A,i}(\theta_2)$ is bounded by a universal constant times the supremum norm of $\theta_1-\theta_2$ is as follows. Firstly,

$$\begin{aligned}f_{A,i}(\tilde{h}_1,\bar{Q}_1)-f_{A,i}(\tilde{h}_2,\bar{Q}_2) &= \tilde{h}_1(\bar{Q}_0-\bar{Q}_1)-\tilde{h}_2(\bar{Q}_0-\bar{Q}_2)\\ &\quad-(\int\tilde{h}_1(\bar{Q}_0-\bar{Q}_1)g_{0,i}-\int\tilde{h}_2(\bar{Q}_0-\bar{Q}_2)g_{0,i}\\ &=(\tilde{h}_1-\tilde{h}_2)(\bar{Q}_0-\bar{Q}_2)+\tilde{h}_1(\bar{Q}_2-\bar{Q}_1)\\ &\quad-(\int(\tilde{h}_1-\tilde{h}_2)(\bar{Q}_0-\bar{Q}_2)g_{0,i}+\int\tilde{h}_1(\bar{Q}_2-\bar{Q}_1)g_{0,i}.\end{aligned}$$

We have $\| (\tilde{h}_1 - \tilde{h}_2)(\bar{Q}_0 - \bar{Q}_2) \|_p \leq \| \tilde{h}_1 - \tilde{h}_2 \|_\infty$ and $\| \tilde{h}_1(\bar{Q}_2 - \bar{Q}_1) \|_p \leq \| \tilde{h}_1 \|_\infty \| \bar{Q}_2 - \bar{Q}_1 \|_\infty$. By our uniform bound on the class of functions $\mathcal{F}$ we have that $\| \tilde{h}_1 \|_\infty < M < \infty$ for some $M < \infty$. We also have

$$\| \int (\tilde{h}_1 - \tilde{h}_2)(\bar{Q}_0 - \bar{Q}_2) g_{0,i} \|_p \leq \| \int_c (\bar{Q}_0 - \bar{Q}_2)(c) g_{0,i}(c|W) \|_\infty \| \tilde{h}_1 - \tilde{h}_2 \|_\infty,$$

where $\| \int_c (\bar{Q}_0 - \bar{Q}_2)(c) g_{0,i}(c|W) \|_\infty \leq \sup_W \int_c g_{0,i}(c|W) = 1$. The same bounding applies to $\| \tilde{h}_1(\bar{Q}_2 - \bar{Q}_1) g_{0,i}$. This proves that indeed $\| f_{A,i}(\tilde{h}_1, \bar{Q}_1) - f_{A,i}(\tilde{h}_2 - \bar{Q}_2) \|_p$ is bounded by $C$ times $\max(\| \tilde{h}_1 - \tilde{h}_2 \|_\infty, \| \bar{Q}_1 - \bar{Q}_2 \|_\infty)$, which completes the proof.

#### (A3): Asymptotic equicontinuity of $Z_{NW}(\tilde{h}, \bar{Q})$

Conditional on $F_1, \ldots, F_N$, we can represent $Z_{NW}(\tilde{h}, \bar{Q})$ as $1/N \sum_i \{ f_{W,i}(\tilde{h}, \bar{Q})(W_j : j \in R_i) - P_0 f_{W,i} \}$. Specifically, $f_{W,i}(\tilde{h}, \bar{Q}) = \int \tilde{h}(\bar{Q} - \bar{Q}_0) g_{0,i}(c|W)$. Under our independence assumption, we know that for each $i$, $f_{W,i}$ only depends on maximally $K$ $f_{W,i}$. Thus, we can apply our general proof above to establish the bound of its orlics norm. As above, we can show that the $\| \cdot \|_p$ norm of $f_i^W(\theta)$ is bounded by a constant $C$ times the supremum norm of $\theta$.

## Proof of (A2)

Define the process $Z^1_{W,N}(\bar{Q}) = \frac{1}{\sqrt{N}} \sum_{i=1}^N \{ f^1_{W,i}(\bar{Q}) - P_0 f^1_{W,i}(\bar{Q}) \}$ indexed by $\bar{Q}$, where $f^1_{W,i}(\bar{Q}) = \int \bar{Q}(c) g_i^*(c|W)$. We need to prove that $R_{N,0} = Z^1_{W,N}(\bar{Q}^*_N - \bar{Q}^*) = o_P(1)$. This proof is completely analogue to our proof above for establishing asymptotic equicontinuity of the other $Z_{W,N}(\tilde{h}, \bar{Q})$ process analyzed above, but now with respect to the supremum norm for $\bar{Q}$.

## Proof of (A4)

Recall the definition of the process

$$Z^3_{W,N}(\bar{g}) = \frac{1}{\sqrt{N}} \sum_{i=1}^N f^3_{W,i}(\bar{g}),$$

where

$$f^3_{W,i}(\bar{g}) = \int_a \frac{\bar{h}^*_0}{\bar{h}^2_0}(\bar{Q}_0 - \bar{Q}^*)(c_i^Y(a, W)) g(a|W) - \int_w \int_a \frac{\bar{h}^*_0}{\bar{h}^2_0}(\bar{Q}_0 - \bar{Q}^*)(c_i^Y(a, w)) g(a|w) Q_{W,0}(w),$$

and $g(a|w)$ is the conditional distribution of $A = (A_1, \ldots, A_N)$, given $W$, implied by $\bar{g}$. We need to prove that $Z^3_{W,N}(\bar{g}_N) - Z^3_{W,N}(\bar{g}_0) = o_P(1)$. This proof is completely analogue to our proof above for establishing asymptotic equicontinuity of the other $Z_{W,N}(\tilde{h}, \bar{Q})$ process analyzed above, but now with respect to the supremum norm for $\bar{g}$.

## (A1): Establishing weak convergence of first-order approximation of standardized estimator

### Outline of proof

Recall

$$\sqrt{N}(\psi_N^* - \psi_0) \approx \frac{1}{\sqrt{N}} \sum_i f_i(O) = \frac{1}{\sqrt{N}} \sum_i \{f_{Y,i} + f_{A,i} + f_{W,i}\} = Z_{NY} + Z_{NA} + Z_{NW},$$

where

$$f_{Y,i} = \frac{\bar{h}(g^*)}{\bar{h}(g_0)}(C_i^Y)(Y_i - \bar{Q}_0(C_i^Y)),$$

$$f_{A,i} = \frac{\bar{h}(g^*)}{\bar{h}(g_0)}(C_i^Y)(\bar{Q}_0 - \bar{Q}^*)(C_i^Y) - \int_c \frac{\bar{h}(g^*)}{\bar{h}(g_0)}(c)(\bar{Q}_0 - \bar{Q}^*)(c) g_{0,i}(c|W) + f_{A,i}^1$$

$$f_{W,i} = \int_c \bar{Q}_0(c) g_i^*(c|W) - \int_{c,w} \bar{Q}_0(c) g_i^*(c|W) Q_{W,0}(dw).$$

We will establish weak convergence of each of the three terms separately.

The proof of weak convergence of $Z_{NY}$ can be based on standard CLT since, conditional on $(A, W)$, $Z_{NY}$ is a sum of mean zero independent random variables.

**Lemma 7** *$Z_{NY} = 1/\sqrt{N} \sum_{i=1}^N f_{Y,i}$ converges weakly to a normal distribution with mean zero and variance*

$$\sigma_Y^2 = \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^N P_0 f_{Y,i}^2 = \lim_{N\to\infty} \int \tilde{h}_0(c) \sigma_Y^2(c) \bar{h}_0^*(c) d\mu(c),$$

*assuming this limit exists, where*

$$\sigma_Y^2(C_i^Y) = E_0(\{Y_i - \bar{Q}_0(C_i^Y)\}^2 | A, W) = E_0(\{Y_i - \bar{Q}_0(C_i^Y)\}^2 | c_i^Y(A, W)).$$

*For example, if $Y_i$ is binary, then the latter expression equals*

$$\sigma_Y^2(C_i^Y) = \bar{Q}_0(1 - \bar{Q}_0)(C_i^Y).$$

*Recall that $\bar{h}_0^* = \frac{1}{N} \sum_i h_{0,i}^*$.*

We establish weak convergence of $Z_{NA}$ by establishing convergence of its $p$th moment. Specifically, we establish that $E(Z_{NA})^p \to \bar{\rho}^{p/2} \frac{p!}{(p/2)! 2^{p/2}}$ for $p$ even, and $E(Z_{NA})^p \to 0$ for $p$ odd, as $N \to \infty$, where $\bar{\rho}$ represents the limit of the second moment $E(Z_{NA})^2$. This convergence in moments implies that $Z_N$ converges weakly to a normal distribution $N(0, \sigma^2 = \bar{\rho})$, where we utilize the following two lemmas.

**Lemma 8** *A random variable $Z$ with $EZ^p = \bar{\rho}^{p/2} \frac{p!}{2^{p/2}(p/2)!}$ for $p$ even, and $EZ^p = 0$ for $p$ odd has probability distribution equal to $N(0, \sigma^2 = \bar{\rho})$, the normal distribution with mean zero and variance $\bar{\rho}$.*

**Proof.** We have

$$\begin{aligned} E\exp(tZ) &= \sum_{p=0}^{\infty} \frac{t^p}{p!} EZ^p = \sum_{p=0}^{\infty} \frac{t^{2p}}{(2p)!} EZ^{2p} \\ &= \sum_{p=0}^{\infty} \frac{t^{2p}}{(2p)!} \bar{\rho}^p \frac{(2p)!}{p!2^p} \\ &= \sum_{p=0}^{\infty} \frac{(0.5t^2\bar{\rho})^p}{p!} \\ &= \exp(0.5t^2\bar{\rho}), \end{aligned}$$

which is the moment generating function of $N(0, \sigma^2 = \bar{\rho})$, i.e. a normal distribution with mean zero and variance equal to $\bar{\rho}$. □

**Lemma 9** Suppose $EZ_N^p \leq C^p$ *for a universal* $C < \infty$. *Suppose that* $EZ_N^p \to \bar{\rho}^{p/2} p!/(p/2)!2^{p/2}$ *for* $p$ *even, and* $EZ_N^p \to 0$ *for* $p$ *odd, as* $N \to \infty$. *Then* $Z_N$ *converges in distribution* to $Z = N(0, \sigma^2 = 2d)$, *as* $N \to \infty$.

**Proof.** Consider the moment generating function $E\exp(tZ_N)$ when $EZ_N^p \to \bar{\rho}^{p/2} \frac{p!}{(p/2)!2^{p/2}}$. By Fubini's theorem,

$$E \sum_{p=0}^{\infty} \tfrac{t^p}{(p)!} Z_N^p = \sum_{p=0}^{\infty} \tfrac{t^p}{p!} EZ_N^p.$$

Because $EZ_N^p \leq C^p$, we have

$$\sum_{p=M}^{\infty} \frac{t^p}{p!} EZ_N^p \leq \sum_{p=M}^{\infty} \frac{t^p}{p!} C^p,$$

which converges to zero in $M \to \infty$. Therefore, we can truncate the summation defining the moment generating function of $Z_N$ and focus on establishing convergence of $E\sum_{p=0}^{M} \frac{t^p}{p!} Z_N^p$, but the latter follows from $EZ_N^p \to EZ^p$ as $N \to \infty$. This proves that

$$E\exp(tZ_N) \to E\exp(tZ).$$

This proves that $Z_N(Q)$ converges in distribution to $Z(Q) = N(0, \sigma^2 = \bar{\rho})$ as $N \to \infty$. □

### (A1): Establishing convergence of *p*th moment for $Z_{NA}$

We consider the case that $W_1, \ldots, W_N$ are independent, given $F$. The proof can be generalized to handle our weaker independence assumption on the distribution of $W$.

**Lemma 10** *Consider the empirical mean* $Z_{NA} = \frac{1}{\sqrt{N}} \sum_i f_{A,i}$, *Let*

$$\rho(j_1, j_2|W) = E_0(f_{A,j_1} f_{A,j_2}|W).$$

*For example, if* $g_N = g_0$, *we have*

$$\begin{aligned} \rho(j_1, j_2|W) = &\int \frac{\bar{h}_0^*(c_1)}{h_0(c_1)} (\bar{Q}_0 - \bar{Q})(c_1) \frac{\bar{h}_0^*}{h_0}(c_2)(\bar{Q}_0 - \bar{Q})(c_2) g_{0,j_1,j_2}(c_1, c_2|W) \\ &- \int \frac{\bar{h}_0^*}{h_0}(c)(\bar{Q}_0 - \bar{Q})(c) g_{0,j_1}(c|W) \int \frac{\bar{h}_0^*}{h_0}(c)(\bar{Q}_0 - \bar{Q})(c) g_{0,j_2}(c|W), \end{aligned}$$

where $g_{0,i,j}$ is the conditional distribution of $(C_i(A, W), C_j(A, W))$, given $W$, which only depends on $A$ through $(A_l : l \in F_i \cup F_j)$.

Let $\rho_A(j_1,j_2)=E(\rho(j_1,j_2|W)|F)$. For two integers $(i_1,i_2)$, define $R_2(i_1,i_2)$ as the indicator that the intersection of $F_{i_1}$ and $F_{i_2}$ is non-empty. Assume that for a constant $\bar{\rho}_A$, we have

$$\frac{1}{N}\sum_{i_1,i_2,R_2(i_1,i_2)=1}\rho_A(i_1,i_2)\to_{N\to\infty}\bar{\rho}_A.$$

*We have for p even,*

$$E\left(\frac{1}{\sqrt{N}}\sum_i f_{A,i}\right)^p\to\frac{p!}{(p/2)!2^{p/2}}\bar{\rho}_A^{p/2}\ as\ N\to\infty.$$

*For p odd, this pth moment converges to zero.*

**Proof.** Given an index $\vec{i}=(i_1,\ldots,i_p)\in\{1,\ldots,N\}^p$ (one among $N^p$), we can draw a graph by drawing a line between two elements $i_{l_1},i_{l_2}$ in $\{i_1,\ldots,i_p\}$ whenever the two corresponding sets $F(i_{l_1})$ and $F(i_{l_2})$ have a non-empty intersection. Classify an element $(i_1,\ldots,i_p)$ by the sizes of the connected sets that make up the graph of $(i_1,\ldots,i_p)$. One category of indices is that each connected set is of size 2, assuming $p$ is even, and let $R_2(\vec{i})$ be the indicator of falling in this category. For each of the other categories with all connected sets of size *larger or equal than 2, but at least one larger than 2*, we can show that its number $X$ of elements is of smaller order than $N^{-p/2}$: $N^{-p/2}X\to 0$ as $N\to\infty$, using that $|F_i|<K$. The latter shows, in particular, that the moment for $p$ odd converges to zero. In addition, for $\vec{i}$ with $R_2(\vec{i})=1$, let $j=1,\ldots,p/2$ index the $p/2$ pairs that are connected, and let $j_1(\vec{i}),j_2(\vec{i})$ denote the two indices in $\{i_1,\ldots,i_p\}$ corresponding with each $j$th pair. We also note that $(f_{A,j_1},f_{A,j_2})$ are independent across the pairs $j$, conditional on $W$. We have

$$\begin{aligned}E_{|W}\left(\frac{1}{\sqrt{N}}\sum_i f_{A,i}\right)^p&=N^{-p/2}\sum_{i_1,\ldots,i_p}R_2(i_1,\ldots,i_p)\prod_{j=1}^{p/2}Ef_{A,j_1(\vec{i})}f_{A,j_2(\vec{i})}+o(1)\\&=N^{-p/2}\sum_{i_1,\ldots,i_p}R_2(i_1,\ldots,i_p)\prod_{j=1}^{p/2}\rho(j_1,j_2|W)+o(1).\end{aligned}$$

Let $(F_i,W_i)$ represent the $i$-specific baseline covariates, so that $F_i$ is separate from $W_i$. We now want to take a conditional expectation, given $F_1,\ldots,F_N$, of the last expression in order to obtain an expression for the $p$th moment only conditioning on $F$. Conditional on $F_1,\ldots,F_N$, the indicators $R_2(\vec{i})$ are fixed. Since $\rho(j_1,j_2|W)$ only depends on $W$ through $(W_i:i\in F_{j_1}\cup F_{j_2})$, the sets $F_{j_1}\cup F_{j_2}$ in the product over $j$ are disjoint across $j$, and $W_1,\ldots,W_N$ are independent, it follows that, conditional on $F=(F_1,\ldots,F_N)$,

$$E\left(\frac{1}{\sqrt{N}}\sum_i f_{A,i}\right)^p\approx N^{-p/2}\sum_{i_1,\ldots,i_p}R_2(i_1,\ldots,i_p)\prod_{j=1}^{p/2}E(\rho(j_1,j_2|W)|F).$$

Let $\rho_A(j_1,j_2)=E(\rho(j_1,j_2|W)|F)$. For two integers $(i_1,i_2)$, define $R_2(i_1,i_2)$ as the indicator that the intersection of $F_{i_1}$ and $F_{i_2}$ is non-empty. Let $R_2=\{(i_1,i_2)\in\{1,\ldots,N\}^2:R_2(i_1,i_2)=1\}$, and $R_2^{p/2}$ is the Cartesian product of this set. Let $R=\{(i_1,\ldots,i_p):R_2(\vec{i})=1\}$, where we are reminded that $R_2(\vec{i})$ is the indicator of all connected sets among $\{i_1,\ldots,i_p\}$ being of size 2. We have the following lemmas.

**Lemma 11** *We have*

$$\begin{aligned}&N^{-p/2}\sum_{((j_1,j_2):j=1,\ldots,p/2)\in R_2^{p/2}}R_2(j_1,j_2:j=1,\ldots,p/2)\prod_{j=1}^{p/2}\rho_A(j_1,j_2)\\&\quad=N^{-p/2}\sum_{((j_1,j_2):j=1,\ldots,p/2)\in R_2^{p/2}}\prod_{j=1}^{p/2}\rho_A(j_1,j_2)+o(1).\end{aligned}$$

**Proof of Lemma 11.** Note that the right-hand side sums over vectors in $\mathcal{R}_2^{p/2}$ while the left-hand side sums over vectors that are both in $\mathcal{R}_2^{p/2}$ *and* satisfy that the corresponding $p$-dimensional vector is an element of $\mathcal{R}$. Since a vector made up of $p/2$-connected pairs can correspond with connected sets of larger size than 2, we have that $\mathcal{R} \subset \mathcal{R}_2^{p/2}$, i.e. the right-hand side sums over more elements. However, the number of these extra vectors $\vec{i} \in \mathcal{R}_2^{p/2}/\mathcal{R}$ that should not have been counted is of smaller order than $N^{p/2}$, so that the contribution is negligible. □

**Lemma 12** *We have*

$$N^{-p/2} \sum_{i_1,\ldots,i_p} R_2(i_1,\ldots,i_p) \prod_{j=1}^{p/2} \rho_A(j_1,j_2)$$
$$= \frac{p!}{(p/2)!2^{p/2}} N^{-p/2} \sum_{((j_1,j_2):j)\in R_2^{p/2}} R_2(j_1,j_2 : j = 1,\ldots,p/2) \prod_{j=1}^{p/2} \rho_A(j_1,j_2).$$

**Proof of Lemma 12:** Consider a vector of three connected pairs $(1,1),(2,2),(3,3)$ (i.e. $p=6$). These three connected pairs appear 3! (i.e. $(p/2)!$) times on right-hand side. However, on the left-hand side, any vector of length 6 with two 1s, two 2s, and two 3s is counted, and there are $6!/2^3$ (i.e. $p!/2^{p/2}$) of such vectors: the number of ordered vectors of length 6 is 6!, but flipping the two 1's or two 2's or two 3's does not yield a different vector. □

Finally, we state the following trivial result

**Lemma 13** *We have*

$$\sum_{((j_1,j_2):j=1,\ldots,p/2)\in R_2^{p/2}} \prod_{j=1}^{p/2} \rho_A(j_1,j_2) = \left( \sum_{\{(i_1,i_2):R_2(i_1,i_2)=1\}} \rho_A(i_1,i_2) \right)^{p/2}.$$

This proves that

$$N^{-p/2} \sum_{i_1,\ldots,i_p} R_2(i_1,\ldots,i_p) \prod_{j=1}^{p/2} \rho_A(j_1,j_2)$$
$$= \frac{p!}{(p/2)!2^{p/2}} N^{-p/2} \sum_{((j_1,j_2):j)\in R_2^{p/2}} R_2(j_1,j_2 : j = 1,\ldots,p/2) \prod_{j=1}^{p/2} \rho_A(j_1,j_2)$$
$$\approx \frac{p!}{(p/2)!2^{p/2}} N^{-p/2} \sum_{((j_1,j_2):j)\in R_2^{p/2}} \prod_{j=1}^{p/2} \rho_A(j_1,j_2)$$
$$= \frac{p!}{(p/2)!2^{p/2}} \left( 1/N \sum_{\{(i_1,i_2):R_2(i_1,i_2)=1\}} \rho_A(i_1,i_2) \right)^{p/2}.$$

Finally, we assumed that the latter summation within the power converges to $\bar{\rho}$. Thus, for $p$ even, we have

$$E\left( \frac{1}{\sqrt{N}} \sum_i f_{A,i} \right)^p \to \frac{p!}{(p/2)!2^{p/2}} \bar{\rho}^{p/2}. \square$$

### (A1): Convergence of $p$th moment of $Z_{NW}$.

The same proof can be applied to establish the convergence of the $p$th moment of $Z_{NW}$ resulting in the following lemma.

**Lemma 14** *Let $Z_{NW} = \sum_i (f_{W,i}(W) - P_0 f_{W,i})$, and $f_{W,i}(W) = f_{W,i}(W_j : j \in R_i)$ for set $R_i$ defined by $F$ with $|R_i| < K$ for some fixed $K < \infty$, where we condition on $F$. Let*

$$\rho_W(j_1,j_2) = E_0(f_{W,j_1}(W) f_{W,j_2}(W)|F) - E_0(f_{W,j_1}(W)|F) E_0(f_{W,j_2}(W)|F).$$

*Specifically, for $Z_{NW}$ we have*

$$f_{W,i}(W) = \int_a \bar{Q}_0(c_i^Y(a,W))g^*(a|W) = \int_c \bar{Q}_0(c)g^*_{0,i}(c|W).$$

*We assumed that $g^*((A_j : j \in F_i)|W)$ only depends on $(W_j : j \in R_i)$ for sets $R_i$ implied by F. Thus, in this case*

$$\rho_W(j_1,j_2) = E_W \int_{c_1,c_2} \bar{Q}_0(c_1)\bar{Q}_0(c_2)g_{0,j_1}(c_1|W)g_{0,j_2}(c_2|W) - E_W \int_c \bar{Q}_0(c)g_{0,j_1}(c|W)E_W \int_c \bar{Q}_0(c)g_{0,j_2}(c|W).$$

*For two integers $(i_1,i_2)$, define $R_2(i_1,i_2)$ as the indicator that the $f_{W,i_1}$ and $f_{W,i_2}$ are dependent (conditional on F). Assume*

$$\frac{1}{N}\sum_{i_1,i_2,R_2(i_1,i_2)=1} \rho_W(i_1,i_2) \to_{N\to\infty} \bar{\rho}.$$

*We have for p even,*

$$E\left(\frac{1}{\sqrt{N}}\sum_i f_{W,i}\right)^p \to_{N\to\infty} \frac{p!}{(p/2)!2^{p/2}}\bar{\rho}^{p/2}.$$

*For p odd, this pth moment converges to zero.*

## General template of proof of Theorem 4

We have $P_0^W D_Y^*(\bar{Q}_N^*, Q_{W,N}, \bar{h}(Q_{W,N}, g_0)) = \Psi(\bar{Q}_0, Q_{W,N}) - \Psi(\bar{Q}_N^*, Q_{W,N})$. We now proceed as follows:

$$\begin{aligned}
\Psi(\bar{Q}_N^*, Q_{W,N}) - \Psi(\bar{Q}_0, Q_{W,N}) &= -P_0^W D_Y^*(\bar{Q}_N^*, Q_{W,N}, \bar{h}(Q_{W,N}, g_0)) \\
&= (P_N - P_0^W)D_Y^*(\bar{Q}_N^*, Q_{W,N}, \bar{h}(Q_{W,N}, g_0)) \\
&\quad + P_N\{D_Y^*(\bar{Q}_N^*, Q_{W,N}, \bar{h}_N) - D_Y^*(\bar{Q}_N^*, Q_{W,N}, \bar{h}(Q_{W,N}, g_0))\} \\
&= (P_N - P_0^W)D_Y^*(\bar{Q}_N^*, Q_{W,N}, \bar{h}(Q_{W,N}, g_0)) \\
&\quad + (P_N - P_0^W)\{D_Y^*(\bar{Q}_N^*, Q_{W,N}, \bar{h}_N) - D_Y^*(\bar{Q}_N^*, Q_{W,N}, \bar{h}(Q_{W,N}, g_0))\} \\
&\quad + P_0^W\{D_Y^*(\bar{Q}_N^*, Q_{W,N}, \bar{h}_N) - D_Y^*(\bar{Q}_N^*, Q_{W,N}, \bar{h}(Q_{W,N}, g_0))\}.
\end{aligned}$$

The second term we denote with $R_{N,2}$. We note that $\{D_Y^*(\bar{Q}_N^*, Q_{W,N}, \bar{h}_N) - D_Y^*(\bar{Q}_N^*, Q_{W,N}, \bar{h}(Q_{W,N}, g_0))\}$ equals

$$\frac{1}{N}\sum_{i=1}^N \left(\frac{\bar{h}_N^*}{\bar{h}_N} - \frac{\bar{h}_N^*}{\bar{h}(Q_{W,N}, g_0)}\right)(Y_i - \bar{Q}_N^*(C_i^Y)).$$

Thus, we have obtained the following expansion:

$$\begin{aligned}
\psi_N^* - \psi_0 &= (P_N - P_0^W)D_Y^*(\bar{Q}_N^*, Q_{W,N}, \bar{h}(Q_{W,N}, g_0)) \\
&\quad + P_0^W \frac{1}{N}\sum_{i=1}^N \left(\frac{\bar{h}_N^*}{\bar{h}_N} - \frac{\bar{h}_N^*}{\bar{h}(Q_{W,N}, g_0)}\right)(\bar{Q}_0 - \bar{Q}^*)(C_i^Y) + R_N,
\end{aligned}$$

where

$$\begin{aligned}
R_N &= P_0^W \frac{1}{N}\sum_{i=1}^N \left(\frac{\bar{h}_N^*}{\bar{h}_N} - \frac{\bar{h}_N^*}{\bar{h}(Q_{W,N}, g_0)}\right)(\bar{Q}^* - \bar{Q}_N^*)(C_i^Y) \\
&\quad + (P_N - P_0^W)\frac{1}{N}\sum_{i=1}^N \left(\frac{\bar{h}_N^*}{\bar{h}_N} - \frac{\bar{h}_N^*}{\bar{h}(Q_{W,N}, g_0)}\right)(Y_i - \bar{Q}_N^*(C_i^Y)) \\
&\equiv R_{N,1} + R_{N,2}.
\end{aligned}$$

By assumption, we have $R_{N,1} = o_P(1/\sqrt{N})$.

We have

$$\begin{aligned}(P_N - P_0^W)D_Y^*(\bar{Q}_N^*, Q_{W,N}, \bar{h}(Q_{W,N}, g_0)) &= (P_N - P_0^W)D_Y^*(\bar{Q}^*, Q_{W,N}, \bar{h}(Q_{W,N}, g_0)) \\ &\quad + (P_N - P_0^W)\{D_Y^*(\bar{Q}_N^*, Q_{W,N}, \bar{h}(Q_{W,N}, g_0)) - D_Y^*(\bar{Q}^*, Q_{W,N}, \bar{h}(Q_{W,N}, g_0))\} \\ &\equiv (P_N - P_0^W)D_Y^*(\bar{Q}^*, Q_{W,N}, \bar{h}(Q_{W,N}, g_0)) + R_{N,3}.\end{aligned}$$

Analogue to our proof for Theorem 2, we can show that

$$R_{N,2} = o_P(1/\sqrt{N}) \text{ and } R_{N,3} = o_P(1/\sqrt{N}) \tag{A3}$$

Consider now the term

$$P_0^W \frac{1}{N}\sum_{i=1}^N \left(\frac{\bar{h}_N^*}{\bar{h}_N} - \frac{\bar{h}_N^*}{\bar{h}(Q_{W,N}, g_0)}\right)(\bar{Q}_0 - \bar{Q}^*)(C_i^Y).$$

We have

$$\begin{aligned}&P_0^W \frac{1}{N}\sum_{i=1}^N \left\{\frac{\bar{h}_N^*(C_i^Y)}{\bar{h}_N)(C_i^Y)} - \frac{\bar{h}_N^*(C_i^Y)}{\bar{h}(Q_{W,N}, g_0)(C_i^Y)}\right\}(\bar{Q}_0 - \bar{Q}^*)(C_i^Y) \\ &= \int_c \left\{\frac{\bar{h}_N^*(c)}{\bar{h}_N(c)} - \frac{\bar{h}_N^*(c)}{\bar{h}(Q_{W,N}, g_0)(c)}\right\}(\bar{Q}_0 - \bar{Q}^*)(c)\bar{h}(Q_{W,N}, g_0)(c) \\ &= -\int_c \frac{\bar{h}_N^*}{\bar{h}(Q_{W,N}, g_0)^2}(\bar{h}(Q_{W,N}, g_N) - \bar{h}(Q_{W,N}, g_0))(c)(\bar{Q}_0 - \bar{Q}^*)(c)\bar{h}(Q_{W,N}, g_0)(c) \\ &\quad + R_{N,4},\end{aligned}$$

where

$$R_{N,4} = \int_c \left\{\frac{\bar{h}_N^*}{\bar{h}_N} - \frac{\bar{h}_N^*}{\bar{h}(Q_{W,N}, g_0)}\right\}\frac{1}{\bar{h}_0}(\bar{h}_N - \bar{h}(Q_{W,N}, g_0))(\bar{Q}_0 - \bar{Q}^*)(c)\bar{h}(Q_{W,N}, g_0)(c).$$

We assumed that $R_{N,4} = o_P(1/\sqrt{N})$. We also assumed

$$\begin{aligned}&-\int_c \frac{\bar{h}_N^*}{\bar{h}(Q_{W,N}, g_0)^2}(\bar{h}(Q_{W,N}, g_N) - \bar{h}(Q_{W,N}, g_0))(c)(\bar{Q}_0 - \bar{Q}^*)(c)\bar{h}(Q_{W,N}, g_0)(c) \\ &= \frac{1}{N}\sum_{i=1}^N f_{A,i}^1(O) + o_P(1/\sqrt{N}),\end{aligned}$$

where $f_{A,i}^1(O)$ only depends on $O$ through $(A_i, (W_j : j \in F_i))$, and $E_0(f_{A,i}^1(O)|W) = 0$.

Thus, we have obtained the following first-order expansion:

$$\begin{aligned}\psi_N^* - \Psi(Q_{W,N}, \bar{Q}_0) &= (P_N - P_0^W)D_Y^*(\bar{Q}^*, Q_{W,N}, \bar{h}(Q_{W,N}, g_0)) \\ &\quad + \frac{1}{N}\sum_{i=1}^N \{f_{A,i}^1 - P_0^W f_{A,i}^1\} + o_P(1/\sqrt{N}).\end{aligned}$$

**Analysis of first-order approximation:** Let

$$f_i = D_{Y,i}^*(\bar{Q}^*, Q_{W,N}, \bar{h}(Q_{W,N}, g_0)) + f_{A,i}^1.$$

Then, the first-order approximation is given by $1/N\sum_i\{f_i(O) - P_0^W f_i\}$, where $P_0^W f_i = 0$. It remains to prove that this first-order expansion converges to a normal limit distribution. This proof has its own outline. Firstly, we decompose $1/N\sum_i f_i(O)$ using $f_i = f_{A,i} + f_{Y,i}$, where $f_{A,i} = E_0(f_i|A, W) - E_0(f_i|W)$, and $f_{Y,i} = f_i - E_0(f_i|A, W)$. Denote the two corresponding terms with $Z_{NY}/\sqrt{N} + Z_{NA}/\sqrt{N}$.

Note that

$$f_{Y,i} = D^*_{Y,i} - E_0(D^*_{Y,i}|A,W) = \frac{\bar{h}^*_0}{\bar{h}_0}(C^Y_i)(Y_i - \bar{Q}_0(C^Y_i)),$$

and $f_{A,i} = E_0(D^*_{Y,i}|A,W) - E_0(D^*_{Y,i}|W) + f^1_{A,i}$. We also note that, conditional on $(W,A)$, $Z_{NY}$ is a sum of independent mean zero random variables $f_{Y,i}$ (functions of $Y_i$), and, conditional on $W$, $Z_{NA} = 1/\sqrt{N}\sum_i f_{A,i}$ for some $f_{A,i}$ which depends on $A$ through $(A_j : j \in F_i)$, while $A_i$, $i = 1,\ldots,N$ are pairwise (conditionally) independent.

Analogue to our proof of Theorem 2, we can show that

$$Z_{NY} \Rightarrow_d N(0,\sigma^2_Y)$$
$$Z_{NA} \Rightarrow_d N(0,\sigma^2_A)$$

with the expressions for $\sigma^2_Y$, $\sigma^2_A$ as specified in the Theorem. Due to the orthogonality of the two empirical processes, using moment generating functions, it also follows that $Z_{NY} + Z_{NA} \Rightarrow_d N(0,\sigma^2 = \sigma^2_Y + \sigma^2_A)$. □

## Notation index

TMLE: Targeted Minimum Loss-Based Estimation/Estimator

$O_i$: Data observed on unit $i$. In general, $O_i = (L_i(0), A_i(0), \ldots, L_i(K), A_i(\tau), L_i(\tau+1) = Y_i)$, and the special case $\tau = 0$ is denoted with $O_i = (W_i, A_i, Y_i)$

$A_i(t)$: Intervention node for unit $i$ at time $t$

$L_i(t)$: Measurements/covariates for unit $i$ at time $t$ in between intervention nodes $A_i(t-1)$ and $A_i(t)$

$Y_i$: Final outcome for unit $i$

$\bar{L}_i(t)$: $\bar{L}_i(t) = (L_i(0),\ldots,L_i(t))$. Similarly, we define $\bar{A}_i(t)$

$L_i$: $L_i = (L_i(0),\ldots,L_i(\tau+1) = Y_i)$

$F_i(t)$: Friends of unit $i$ at time $t$ indicating that $L_i(t)$ and $A_i(t)$ causally only depends on the history of all subjects through the history of unit $i$ itself and the history of its friends $j \in F_i(t)$. This defines exclusion restrictions in the structural equation model for the equations for $A_i(t)$ and $L_i(t)$. For the $\tau = 0$-data structure, we denote this set with $F_i$

$O$: $O = (O_1,\ldots,O_N)$ is the collection of all data on the $N$ units

$P_0$: $P$ is possible probability distribution of data $O$ under our model assumptions, and $P_0$ is the true probability distribution of $O$

$L$: $L = (L_1,\ldots,L_N)$

$A$: $A = (A_1,\ldots,A_N)$

$\bar{L}(t)$: $\bar{L}(t) = (\bar{L}_i(t) : i = 1,\ldots,N)$ the history of $L$ for all $N$ units

$\bar{A}(t)$: $\bar{A}(t) = (\bar{A}_i(t) : i = 1,\ldots,N)$ the history of treatment/intervention process $A$ on all $N$ subjects $Y$: $Y = (Y_1,\ldots,Y_N)$ the outcomes on all $N$ subjects

$\bar{Y}$: $\bar{Y} = \frac{1}{N}\sum_{i=1}^N Y_i$ the average outcome for the combined $N$ units

$Pa(A(t))$: $Pa(A(t)) = (\bar{L}(t), \bar{A}(t-1))$, parent nodes of $A(t)$ according to the following time-ordering only: $O = (L(0), A(0),\ldots,L(\tau),A(\tau),Y)$

$Pa(L(t))$: $Pa(L(t)) = (\bar{L}(t-1), \bar{A}(t-1))$ parent nodes of $L(t)$ according to time-ordering only

$F$: $F = (F_1,\ldots,F_N)$ the friend-process/network-process for all $N$ units. In all probability distributions, we always condition on $F(0) = (F_1(0),\ldots,F_N(0))$

$U$: $U = (U_{L(0)}, U_{A(0)},\ldots,U_{L(\tau)},U_{A(\tau)},U_Y)$ the exogenous errors in the structural equation model for $O$ defined as $L(0) = f_{L(0)}(U_{L(0)}), A(0) = f_{A(0)}(L(0),U_{A(0)}),\ldots, L(\tau) = f_{L(\tau)}(Pa(L(\tau)),U_{L(\tau)})$, $A(\tau) = f_{A(\tau)}(Pa(A(\tau)),U_{A(\tau)})$, $Y = f_Y(Pa(Y),U_Y)$, where $f_\cdot$ are functions of the parent nodes and exogenous errors, modeled as in article

$P^F$: A possible probability distribution of $(O,U)$ as modeled by the structural equation model

$P_0^F$: The true distribution of $(O,U)$

$\mathcal{M}^F$: The set of possible probability distributions of $(O,U)$ as specified by the structural equation model formulated in the article. We also refer to this as the full-data model

$\mathcal{M}$: The set of possible probability distributions of $O$, implied by $\mathcal{M}^F$, or defined without reference to the underlying model $\mathcal{M}^F$. $\mathcal{M}$ is called the statistical model for the data distribution $P_0$

$g^*$: $g^* = \prod_t g_t^*$, $g_t^*$ is a conditional distribution of $a$ $A^*(t)$, given $Pa(A^*(t))$, $t = 0,\ldots,\tau$. The distribution $g_{A(t)}^*$ is modeled through a common $\bar{g}_t^*$: $g_t^*(A(t)|Pa(A(t))) = \prod_{i=1}^N \bar{g}_t^*(A_i^*(t)|Pa^*(A^*(t)))$. $\bar{g}^* = (\bar{g}_t^* : t = 0,\ldots,\tau)$. $g^*$ represents a stochastic intervention on the intervention nodes $A$ representing the intervention that replaces the true conditional distribution of $A(t)$, given $Pa(A(t))$, by this user-supplied choice $\bar{g}_t^*$, for all $t = 0,\ldots,\tau$. One can also denote $A_* = A_{g^*}$

$g$: $g = \prod_t g_t$. $g_t$ is a possible conditional distribution of $A(t)$, given $Pa(A(t))$, $t = 0,\ldots,\tau$. The distribution $g_{A(t)}$ is modeled through a common $\bar{g}_t$: $g_t(A(t)|Pa(A(t))) = \prod_{i=1}^N \bar{g}_t(A_i(t)|Pa(A(t)))$. $\bar{g} = (\bar{g}_t : t = 0,\ldots,\tau)$. $g_0$ is the true conditional distribution parametrized in terms of the true $\bar{g}_0$

$L_{g^*}$: The post-intervention random version of $L$ obtained by replacing the structural equations for $A$ by the stochastic intervention $g^*$. It is also called a intervention-specific counterfactual

$Y_{g^*}$: The post-intervention random version of $Y$. Note $Y_{g^*}$ is a component of $L_{g^*}$

$\bar{Y}_{g^*}$: $\bar{Y}_{g^*} = \frac{1}{N}\sum_{i=1}^N Y_{g^*,i}$, the average outcome under intervention $g^*$

$P_{g^*}$: A possible probability distribution of the counterfactual $L_{g^*}$. $P_{0,g^*}$ the true probability distribution of $L_{g^*}$ implied by the true distribution $P_0^F$ of $(O,U)$.

$P^{g^*}$: The G-computation formula expression for $P_{g^*}$, purely defined as a function of $P$. Under the posed causal model $\mathcal{M}^F$, we would have $P_{g^*,0} = P^{g^*,0}$

$L^{g^*}$: A random variable with probability distribution $P_0^{g^*}$. Similarly, we define $Y^{g^*}$ and $\bar{Y}^{g^*}$

$\Psi^F$: $\Psi^F : \mathcal{M}^F \to \mathbb{R}$ represents the parameter mapping that maps a distribution of the underlying $(O,U)$ into the desired quantity of interest: $\Psi^F(P_0^F)$ represents the true causal quantity value. In this article, we defined $\Psi^F(P^F) = E_{P_{g^*}}\bar{Y}_{g^*}$ [Ψ:]$\Psi : \mathcal{M} \to \mathbb{R}$ represents the parameter mapping that maps a distribution of $P$ of $O$ into a parameter value of interest. $\Psi(P_0)$ represents the true statistical parameter value/estimand. In this article, $\Psi(P_0) = E_{P_0^{g^*}}\bar{Y}^{g^*}$, i.e. the expectation of $\bar{Y}$ under the $G$-computation distribution $P_0^{g^*}$. Under the causal model $\mathcal{M}^F$, we have $\Psi^F(P_0^F) = \Psi(P_0)$

Statistical estimation problem: Estimation of $\psi_0 = \Psi(P_0)$ based on $O \sim P_0$, i.e. defined separately from the underlying causal model, but the causal model allows a causal interpretation $E_{P_0^F}\bar{Y}_{g^*}$

$C_{t,i}^L, C_{t,i}^A$: $C_{t,i}^L = c_{t,i}^L(\bar{L}(t-1),\bar{A}(t-1))$, $C_{t,i}^A = c_{t,i}^A(\bar{L}(t),\bar{A}(t-1))$ are $i$-specific summary measures of the past that $L_i(t)$ and $A_i(t)$ depend upon, respectively

$Q,\bar{g}$: $P(O) = P_{Q,\bar{g}}(O) \equiv Q_{L(0)}(L(0))\prod_{t=1}^{\tau+1}\prod_{i=1}^N \bar{Q}_{L(t)}(L_i(t)|C_{t,i}^L)\prod_{t=0}^{\tau}\prod_{i=1}^n \bar{g}_t(A_i(t)|C_{t,i}^A)$. $Q = (Q_{L(0)}, \bar{Q}_{L(t)} : t = 1,\ldots,\tau+1)$. The statistical model $M = \{P_{Q,\bar{g}} : Q, \bar{g} \in G\}$, where $Q$ is left-unspecified, and $G$ is some model for $\bar{g}$ We denote $\bar{Q} = (\bar{Q}_{L(t)} : t = 1,\ldots,\tau+1)$ and $\bar{g} = (\bar{g}_{A(t)} : t = 0,\ldots,\tau)$. Note $\bar{Q}_{L(t)}$ and $\bar{g}_{A(t)}$ denote common (in $i$) conditional distributions of $L_i(t)$ and $A_i(t)$, respectively. We also use short-hand $\bar{Q}_t = \bar{Q}_{L(t)}$

$C_{t,i}^{L,*}$: $c_{t,i}^L(\bar{L}(t-1),\bar{A}^*(t-1))$, i.e. same summary measure as $c_{t,i}^L$ but with $A$ replaced by $A_*$.

$\Psi(Q)$: Same as $\Psi(P)$, but stressing that $\Psi$ only depends on $P$ through $Q$

$D*(Q,g)$: The canonical gradient/efficient influence curve of $\Psi : \mathcal{M} \to \mathbb{R}$ at $P = P_{Q,\bar{g}}$. Also denoted with $D^*(Q,h,\Psi(Q))$ to stress that it only depends on $g$ through a specified $h(Q,g)$ and can be viewed as estimating function in $\psi$

$L(Q)$: A loss function $(O,Q) \to L(Q)(O)$ for $Q$ satisfying $Q_0 = \arg\min_Q P_0 L(Q)$. In our case, we define a loss for $\bar{Q}_{L(t)}$ for each $t$ and define $L(\bar{Q})$ as the sum-loss: $L(\bar{Q})(O) = \sum_{t=0}^{\tau+1} L_t(\bar{Q}_{L(t)})(O)$. For example, one can use the log-likelihood loss. We use a separate loss function for $Q_{L(0)}$

$L(\bar{g})$: A loss function $(O,\bar{g}) \to L(\bar{g})(O)$ for $\bar{g}$. See $L(\bar{Q})$ for sum-loss representation

Cross-validation: For example, suppose we want to estimate $\bar{Q}_{0,t} = \bar{Q}_{0,L(t)}$, the common conditional distribution of $L_i(t)$, given $C^L_{i,t}$. Create a data set $(L_i(t), C^L_{i,t})$, $i = 1,\ldots,N$. Consider a $V$-fold sample split of these $N$ observations in a so-called validation sample $Val(v) \subset \{1,\ldots,N\}$ and its complement, the so-called training sample, $Tr(v)$, $v = 1,\ldots,V$. Let $Q_{t,N,Tr(v)}$ be an estimator applied to the training sample $\{(L_i(t), C^L_{i,t}) : i \in Tr(v)\}$. The cross-validated risk w.r.t. loss $L_t()$ of this estimate is defined as $\sum_v \sum_{i \in Val(v)} L_t(Q_{t,N,Tr(v)})(L_i(t), C^L_{i,t})$. A cross-validation selector among a set of candidate estimators is defined as the one that minimizes this cross-validated risk across the candidate estimators. Similarly, we can define cross-validation for estimation of $\bar{g}_t$

$h$: $h_{i,t}(Q,g)(c_t) = P_{Q,g}(C^L_{i,t} = c_t)$, $i = 1,\ldots,N$, $t = 1,\ldots,\tau+1$. Similarly, we define $h_{i,t}(Q,g^*) = P_{Q,g^*}(C^L_{i,t} = c_t)$. Short-hand notations are $h_{i,t}$ and $h^*_{i,t}$. In addition, $\bar{h}_t = \frac{1}{N}\sum_{i=1}^N h_{i,t}$ and $\bar{h}^*_t = \frac{1}{N}\sum_{i=1}^N h^*_{i,t}$

Analogue point-treatment notation: $P_W$, $P_{A|W}$, $P_{Y|A,W}$, $C^A_i = c^i_A(W)$, $C^Y_i = c^i_Y(A,W)$, $h_i(Q,g)(c) = P_{Q,g}(C^Y_i = c)$, $\bar{h}(Q,g)(c) = \frac{1}{N}\sum_{i=1}^N h_i(Q,g)(c)$, $\bar{h}^* = \bar{h}(Q,g^*)$, $D^*(P)$, $D^*(Q,g)$, $D^*(Q,\bar{h}(Q,g),\Psi(Q))$, $Q = (Q_W,\bar{Q})$, $\bar{Q}_Y$ common conditional density of $Y_i$, given $C^Y_i$, $Q_W$ is the probability density of $W$ and $\bar{Q}_0(C^Y_i) = E_{\bar{Q}_Y}(Y_i|A,W) = E_{P_0}(Y_i|C^Y_i)$, $\bar{g}(A_i|C^A_i)$ common density of $A_i$, given $C^A_i$

$D^*_Y(P), D^*_W(P)$: $D^*(P) = D^*_W(P) + D^*_Y(P)$, orthogonal decomposition in function of $W$ and function of $O = (W,A,Y)$ with conditional mean zero, given $A,W$, both are elements of the tangent space at $P$ of the statistical model $\mathcal{M}$

$P^W$: $P^W$ is conditional distribution of $O$, given $W$. $P^W f = E_P(f(O)|W)$

$\Psi^W$: $\Psi^W(P) = \Psi(\bar{Q}, Q_{W,N})$, where $Q_{W,N}$ is probability distribution of $W$ that puts mass 1 on the observed $W$.

$Pf$: $f$ always represents a function of $O$: $O \to f(O)$. $Pf = E_P f(O)$

$P_N f$: $P_N f = f(O)$ since $P_N$ represents probability distribution that puts mass 1 on observed $O$

$Z_N(\theta)$: $Z_N(\theta) = \frac{1}{\sqrt{N}}\sum_{i=1}^N \{f_i(\theta)(O) - P_0 f_i(\theta)\}$ for specified $f_i(\theta)$. $(Z_N(\theta) : \theta \in \mathcal{F})$ represents a process indexed by class of functions $\mathcal{F}$, which we aim to analyze. In our processes $\theta$ plays role of $(\bar{Q}, \bar{g}, \bar{h})$

$f_i$, $f_{Y,i}$, $f_{A,i}$, $f_{W,i}$: Given $a$ $f_i(O)$, we orthogonally decompose $f_i = f_{Y,i} + f_{A,i} + f_{W,i}$ with $f_{Y,i} = f_i - P_0^{A,W} f_i$, $f_{A,i} = P_0^{A,W} f_i - P_0^W f_i$, $f_{W,i} = P_0^W f_i - P_0 f_i$, where $P_0^X$ represents the conditional distribution of $O$, given $X$.

$f^1_{W,i}, f^2_{W,i}, f^1_{A,i}$, etc.: $f_{W,i}$ indicates that it only depends on $O$ through $W$ and will be centered marginally, $f_{A,i}$ indicates that it only depends on $O$ through $(A,W)$ and will be centered conditional on $W$, and $f_{Y,i}$ indicates that it is centered to have mean zero conditionally, given $A,W$. In addition, we use superscripts to have notation for multiple of such functions if part of a single proof: e.g. $f^1_{W,i}, f^2_{W,i}$. In different separate parts of proofs, we often use same notation so that $f_{W,i}$ can denote one thing in one proof and another in another proof.

$Z_{NA}$, $Z_{NY}$, $Z_{NW}$: Given a mean zero centered process $(Z_N(\theta) = \frac{1}{\sqrt{N}}\sum_i f_i(\theta) : \theta \in F)$, we define a corresponding orthogonal decomposition $Z_N = Z_{NY} + Z_{NA} + Z_{NW}$ with $Z^Y_N(\theta) = \frac{1}{\sqrt{N}}\sum_{i=1}^N f_{Y,i}(\theta)$, $Z^A_N(\theta) = \frac{1}{\sqrt{N}}\sum_{i=1}^N f_{A,i}(\theta)$, and $Z^W_N(\theta) = \frac{1}{\sqrt{N}}\sum_{i=1}^N f_{W,i}(\theta)$

$N(\epsilon, \mathcal{F}, d)$: Number of balls of size $\epsilon$ needed to cover $\mathcal{F}$ w.r.t. metric $d$

$\mathcal{C}_Y$, $\mathcal{C}_A$: $C_Y$ is set that contains $C^Y_i$ for all $i$ with probability 1. It is a subset of $\mathbb{R}^k$ for some $k$ (constant in $N$). Similarly, $C_A$ is set that contains $C^A_i$ with probability 1

$\|.\|_\lambda$: The orlics norm of a random variable implied by a strictly monotone function $\lambda : \mathbb{R}_{\geq 0} \to \mathbb{R}$. We are concerned with bounding the orlics norm of the random variable $Z_N(\theta)$ uniformly in $\theta$ and $N$.

# References

1. Rubin DB. Matched sampling for causal effects. Cambridge, MA: Cambridge University Press, 2006.
2. Pearl J. Causality: models, reasoning, and inference, 2nd edn. New York: Cambridge University Press, 2009.

3. van der Laan MJ, Robins JM. Unified methods for censored longitudinal data and causality. New York: Springer, 2003.
4. Tsiatis AA. Semiparametric theory and missing data. New York: Springer, 2006.
5. Hernán MA, Robins JM. Causal inference. New York: Chapman & Hall/CRC, 2012. Unpublished.
6. van der Laan MJ, Rose S. Targeted learning: causal inference for observational and experimental data. New York: Springer, 2012.
7. Halloran ME, Struchiner CJ. Causal inference in infectious diseases. Epidemiology 1995;6:142–51.
8. Hudgens MG, Halloran ME. Toward causal inference with interference. J Am Stat Assoc 2008;1030:832–42. PMCID: PMC2600548.
9. VanderWeele TJ, Vandenbrouke JP, Tchetgen Tchetgen EJ, Robins JM. A mapping between interactions and interference: implications for vaccine trials. Epidemiology 2012;230:285–92. PMID: 22317812.
10. Tchetgen Tchetgen EJ, VanderWeele TJ. On causal inference in the presence of interference. Stat Meth Med Res 2012;210:55–75. PMID: 21068053.
11. Sobel M. What do randomized studies of housing mobility demonstrate? Causal inference in the face of interference. J Am Stat Assoc 2006;101:1398–407.
12. Donner A, Klar N. Design and analysis of cluster randomization trials in health research. London: Arnold, 2000.
13. Hayes RJ, Moulton LH. Cluster randomized trials. Boca Raton, FL: Chapman & Hall/CRC, 2009.
14. Campbell MJ, Donner A, Klar N. Developments in cluster randomized trials and statistics in medicine. Stat Med 2007;26:2–19. DOI:10.1002/sim.2731.
15. Petersen ML, van der Laan MJ. A general roadmap for the estimation of causal effects. Unpublished, Division of Biostatistics, University of California, Berkeley, CA, 2012.
16. Robins JM, Rotnitzky A, Scharfstein DO. Sensitivity analysis for selection bias and unmeasured confounding in missing data and causal inference models. In: Halloran ME, Berry D, editors. Statistical models in epidemiology, the environment and clinical trials. IMA volume 116. New York: Springer-Verlag, 1999:1–92.
17. Rotnitzky A, Scharfstein D, Su T-Li, Robins J. Methods for conducting sensitivity analysis of trials with potentially nonignorable competing causes of censoring. Biometrics 2001;570:103–13. ISSN 1541-0420.
18. Diaz I, van der Laan MJ. Sensitivity analysis for causal inference under unmeasured confounding and measurement error problems. Technical Report 303, Division of Biostatistics, University of California, Berkeley, CA, 2012. Submitted to IJB, also technical report. Available at: http://www.bepress.com/ucbbiostat/paper303
19. van der Laan MJ. Estimation based on case-control designs with known prevalence probability. Int J Biostat 2008. Available at: http://www.bepress.com/ijb/vol4/iss1/17/
20. van der Laan MJ, Rubin DB. Targeted maximum likelihood learning. Int J Biostat 2006;2:Article 11.
21. Chambaz A, van der Laan MJ. Targeting the optimal design in randomized clinical trials with binary outcomes and no covariate, theoretical study. Int J Biostat 2011;70:1–32. Working paper 258. Available at: www.bepress.com/ucbbiostat
22. Chambaz A, van der Laan MJ. Targeting the optimal design in randomized clinical trials with binary outcomes and no covariate, simulation study. Int J Biostat 2011;70:33–. Working paper 258. Available at: www.bepress.com/ucbbiostat
23. van der Laan MJ, Balzer LB, Petersen ML. Adaptive matching in randomized trials and observational studies. J Stat Res 2013;46:113–56.
24. Hu F, Rosenberger WF. The theory of response adaptive randomization in clinical trials. New York: Wiley, 2006.
25. Carrington PJ, Scott J, Wasserman S. Models and methods in social network analysis (structural analysis in the social sciences). New York: Cambridge University Press, 2005.
26. Bakshy E, Eckles D, Yan R, Rosenn I. Social influence in social advertising: evidence from field experiments. EC 2012: Proceedings of the ACM Conference on Electronic Commerce, ACM, 2012. Available at: http://arxiv.org/abs/1206.4327
27. Airoldi EM, Toulis P, Kao E, Rubin DB. Estimation of causal peer influence effects. Proceedings of the 30th International Conference on Machine Learning, Atlanta, GA, JMLR: W&CP volume 28, 2013.
28. Aronow PM, Samii C. Estimating average causal effects under general interference. Technical report, Yale University and New York University, 2013. Unpublished manuscript.
29. van der Laan MJ, Dudoit S. Unified cross-validation methodology for selection among estimators and a general cross-validated adaptive epsilon-net estimator: finite sample oracle inequalities and examples. Technical report, Division of Biostatistics, University of California, Berkeley, CA, November 2003.
30. van der Laan MJ, Dudoit S, van der Vaart AW. The cross-validated adaptive epsilon-net estimator. Stat Decisions 2006;240:373–95.
31. van der Vaart AW, Dudoit S, van der Laan MJ. Oracle inequalities for multi-fold cross-validation. Stat Decisions 2006;240:351–71.
32. Polley EC, Rose S, van der Laan MJ. Super learning. In: van der Laan MJ and Rose S, editors. Targeted learning: causal inference for observational and experimental data. New York: Springer, 2012.
33. van der Laan MJ, Polley E, Hubbard A. Super learner. Stat Appl Genet Mol Biol 2007;60. ISSN 1.
34. Bickel PJ, Klaassen CA, Ritov Y, Wellner J. Efficient and adaptive estimation for semiparametric models. New York: Springer-Verlag, 1997.
35. van der Vaart AW. Asymptotic statistics. New York: Cambridge University Press, 1998.

36. Scharfstein DO, Rotnitzky A, Robins JM. Adjusting for nonignorable drop-out using semiparametric non-response models (with discussion). J Am Stat Assoc 1999;94:1096–146.
37. Scharfstein DO, Rotnitzky A, Robins JM. Adjusting for nonignorable drop-out using semiparametric nonresponse models, (with discussion and rejoinder). J Am Stat Assoc 1999;94:1096–120 (1121–46).
38. Gruber S, van der Laan MJ. A targeted maximum likelihood estimator of a causal effect on a bounded continuous outcome. Int J Biostat 2010;6:Article 26. Available at: www.bepress.com/ijb/vol6/iss1/26
39. Rosenblum M, van der Laan MJ. Targeted maximum likelihood estimation of the parameter of a marginal structural model. Int J Biostat 2010;60:Article 19.
40. Robins JM, Rotnitzky A. Recovery of information and adjustment for dependent censoring using surrogate markers. In: Jewell N, Dietz K, Farewell V, editors. Aids epidemiology. Methodological issues. Basel: Birkhäuser, 1992:296–31.
41. van der Vaart AW, Wellner JA. Weak convergence and empirical processes. New York: Springer-Verlag, 1996.
42. Neyman J. On the application of probability theory to agricultural experiments. Stat Sci 1990;5:465–80.
43. Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. J Educ Psychol 1974;64:688–701.
44. Holland PW. Statistics and causal inference. J Am Stat Assoc 1986;810:945–60.
45. Robins JM. Addendum to: "A new approach to causal inference in mortality studies with a sustained exposure period – application to control of the healthy worker survivor effect" [math. Modelling **7** (1986), no. 9–12, 1393–1512; MR 87m:92078]. Comput Math Appl 1987;140:923–45. ISSN 0097-4943.
46. Robins JM. A graphical approach to the identification and estimation of causal parameters in mortality studies with sustained exposure periods. J Chron Dis (40, Suppl) 1987;2:139s–61s.
47. Heitjan DF, Rubin DB. Ignorability and coarse data. Ann Stat 1991;190:2244–53.
48. Jacobsen M, Keiding N. Coarsening at random in general sample spaces and random censoring in continuous time. Ann Stat 1995;23:774–86.
49. Gill RD, van der Laan MJ, Robins JM. Coarsening at random: characterizations, conjectures and counter-examples. In: Lin DY and Fleming TR, editors. Proceedings of the first Seattle symposium in biostatistics. New York: Springer Verlag, 1997:255–94.
50. Robins JM. Causal inference from complex longitudinal data. In: Berkane M, editor. Latent variable modeling and applications to causality. New York: Springer Verlag, 1997:69–117.
51. Robins JM. [Choice as an alternative to control in observational studies]: comment. Stat Sci. 1999;140:281–93.
52. Gill R, Robins JM. Causal inference in complex longitudinal studies: continuous case. Ann Stat 2001;290.
53. Yu Z, van der Laan MJ. Measuring treatment effects using semiparametric models. Technical report, Division of Biostatistics, University of California, Berkeley, CA, 2003.
54. Dawid A, Didelez V. Identifying the consequences of dynamic treatment strategies: a decision theoretic overview. Stat Surv 2010;4:184–231.
55. Didelez V, Dawid AP, Geneletti S. Direct and indirect effects of sequential treatments. Proceedings of the 22nd Annual Conference on Uncertainty in Artificial Intelligence, Cambridge, MA, 2006:138–46.
56. Diaz I, van der Laan MJ. Population intervention causal effects based on stochastic interventions. Biometrics 2012;68:541–9
57. Zheng W, van der Laan MJ. Cross-validated targeted minimum loss based estimation. In: van der Laan MJ and Rose S, editors. Targeted learning: causal inference for observational and experimental studies. New York: Springer, 2011:459–74.
58. Bang H, Robins JM. Doubly robust estimation in missing data and causal inference models. Biometrics 2005;61:962–72.
59. van der Laan MJ, Gruber S. Targeted minimum loss based estimation of causal effects of multiple time point interventions. Int J Biostat 2012;88. DOI:10.1515/1557-4679.1370. PMID: 22611591.
60. van der Laan MJ. Targeted maximum likelihood based causal inference: part I. Int J Biostat 2010;60:Article 2.
61. van der Laan MJ. Targeted maximum likelihood based causal inference: part II. Int J Biostat 2010;60:Article 3.
62. van der Laan MJ. Causal inference for networks. Technical Report 300, Division of Biostatistics, University of California, Berkeley, CA. Submitted to JCI, also technical report. Available at: http://www.bepress.com/ucbbiostat/paper300, 2012.
63. van der Laan MJ. Efficient and inefficient estimation in semiparametric models. Centre of Computer Science and Mathematics, Amsterdam, CWI tract 114 edition, 1996.
64. Balzer LB, van der Laan MJ. Estimating effects on rare outcomes: knowledge is power. Technical Report 310, Division of Biostatistics, University of California, Berkeley, CA, 2013.
65. Gill RD, van der Laan MJ, Wellner JA. Inefficient estimators of the bivariate survival function for three models. Ann De l'Inst Henri Poincaré 1995;31:545–97.
66. Petersen M, Schwab J, Gruber S, Blaser N, Schomaker M, van der Laan MJ. Targeted minimum loss based estimation of marginal structural working models. J Causal Inference 2013;Submitted, technical report. Available at: http://biostats.bepress.com/ucbbiostat/paper312/