Journal of Statistical Research *ISSN 0256 - 422 X*
2012, Vol. 46, No. 2, pp. 113-156

ADAPTIVE MATCHING IN RANDOMIZED TRIALS AND OBSERVATIONAL STUDIES

MARK J. VAN DER LAAN

Division of Biostatistics, University of California, Berkeley
Email: laan@berkeley.edu

LAURA B. BALZER AND MAYA L. PETERSEN

Division of Biostatistics, University of California, Berkeley
Email: lbbalzer@gmail.com, mayaliv@berkeley.edu

SUMMARY

In many randomized and observational studies the allocation of treatment among a sample of n independent and identically distributed units is a function of the covariates of all sampled units. As a result, the treatment labels among the units are possibly dependent, complicating estimation and posing challenges for statistical inference. For example, cluster randomized trials frequently sample communities from some target population, construct matched pairs of communities from those included in the sample based on some metric of similarity in baseline community characteristics, and then randomly allocate a treatment and a control intervention within each matched pair. In this case, the observed data can neither be represented as the realization of n independent random variables, nor, contrary to current practice, as the realization of $n/2$ independent random variables (treating the matched pair as the independent sampling unit). In this paper we study estimation of the average causal effect of a treatment under experimental designs in which treatment allocation potentially depends on the pre-intervention covariates of all units included in the sample. We define efficient targeted minimum loss based estimators for this general design, present a theorem that establishes the desired asymptotic normality of these estimators and allows for asymptotically valid statistical inference, and discuss implementation of these estimators. We further investigate the relative asymptotic efficiency of this design compared with a design in which unit-specific treatment assignment depends only on the units' covariates. Our findings have practical implications for the optimal design and analysis of pair matched cluster randomized trials, as well as for observational studies in which treatment decisions may depend on characteristics of the entire sample.

Keywords and phrases: Cluster randomized trials, matching, asymptotic linearity of an estimator, causal effect, efficient influence curve, empirical process, confounding, dependent treatment allocation, G-computation formula, influence curve, loss function, adaptive randomization, semiparametric statistical model, targeted maximum likelihood estimation, targeted minimum loss based estimation (TMLE).

1 Introduction

In a typical randomized controlled trial, one randomly draws a unit from a target population, measures baseline covariates on the unit, randomly assigns a treatment from among a set of possible treatments according to a known distribution (possibly conditional on the baseline covariates of the unit), and measures the unit's treatment-specific outcome. This experiment is repeated n times resulting in n independent and identically distributed (i.i.d.) copies, providing a firm basis for statistical estimation and inference using the central limit theorem.

In this article we consider an alternative data generating experiment in which one first randomly draws n units from a target population and measures the baseline covariates of each, then assigns n treatments from among some set according to a conditional distribution, *given the n unit-specific baseline covariates*, and finally measures the n treatment-specific outcomes. In such an experiment, the underlying units are independently and identically distributed draws from a common target population, so that the covariates and the underlying treatment-specific outcomes represent an i.i.d sample. However, the treatment assigned to one unit can be a function of the covariates of other units in the sample, creating dependence between the n unit-specific observed data structures. As a result, the data generating design cannot be represented as n repetitions of an experiment, and not even as n independent experiments. The challenge posed to statistical inference by this design is highlighted by the fact that it is unclear how to implement valid bootstrap-based variance estimation. The available data constitute a single repetition of the underlying experiment.

Our study of this problem is motivated, in particular, by a common design-based approach to enforce empirical balance in baseline covariates among the treated and non-treated units in a finite sample. One way to enforce such balance is to partition the random sample of n units into $n/2$ pairs that are maximally similar with respect to covariate values according to some metric, and then to randomly allocate a treatment and a control intervention within each pair. A variation of this design partitions the sample into fewer than $n/2$ pairs, discarding those units for which the poorest matches are obtained. Such pair-matched designs are particularly common in community or cluster randomized controlled trials, motivated both by a desire to improve efficiency, and by the fact that such trials typically enroll far fewer independent units than their individual-level counterparts, and are thus less able to rely on chance alone to achieve the desired covariate balance between treatment groups (see, for example, reviews by Donner and Klar (2000), Hayes and Moulton (2009), Campbell et al. (2007)). Treatment assignment conditional on such a covariate-based partition of the sample preserves randomization while ensuring a degree of covariate balance between treatment and control arms of the trial. However, since the partition (in this case, the construction of the matched pairs) is generated as a function of all n covariate vectors, the treatment assignment of each unit in the sample is now a function of the covariate values of the entire sample.

While the fact that pair matching in randomized trials can introduce dependence between units is well-recognized, the extensive literature on the design and analysis of pair matched trials, including the literature debating the merits and perils of pair-matching, focuses on experimental designs in which the matched pair constitutes the unit of independence (Freedman et al. (1990); Campbell et al. (2007); Hayes and Moulton (2009); Imai et al. (2009); Imai (2008); Donner and Klar (2000); Murray (1998); Donner and Klar (2000); Raudenbush et al. (2007); Klar and Donner (1997); Balzer

et al. (2012)). In one well studied design, two units are sampled from a conditional distribution given a stratum of a baseline covariate, the treatment and control intervention are randomly allocated to the pair, the outcomes for the two units are observed, and the experiment is repeated multiple times at different strata. In such an experiment, the data generating distribution involves independently repeating the stratum-specific experiment of drawing the pair of units from the stratum, assigning the treatments, and measuring the outcomes, across different strata. Therefore, statistical inference can be based on a central limit theorem for sums of independent random variables. If the strata are set by design, then the data for each pair are independent across pairs (with a stratum-specific data distribution for each pair) but are not identically distributed.

A variation of this design is based on randomly sampling a unit and measuring a baseline covariate on the unit, and then sampling a second unit from a conditional distribution, given that the baseline covariate has the same value as the first unit. Treatment and control are allocated within the matched pair, outcomes on each unit in the pair are measured, and the experiment is repeated multiple times. In this case, the data on each pair are not only independent across the pairs but are also identically distributed. van der Laan (2008), Rose and van der Laan (2009), Balzer et al. (2012) discuss formulation of the above two data structures in terms of matched case-control sampling, and present corresponding targeted minimum loss-based estimators.

The focus in the literature on designs in which the matched pair represents the unit of independence may be due in part to the specific studies for which much of the early theory was developed. These include randomized trials in ophthalmology in which the patient's two eyes provide the matched pair, as well as some cluster randomized trials. For example, the Community Intervention Trial for Smoking Cessation (COMMIT) motivated important early work on the use of pair matching in cluster randomized trials (Freedman et al. (1997); Gail et al. (1992); COMMIT Research Group (1991)). This study in fact sampled (albeit not randomly) 11 matched pairs of communities from a larger population of candidate matched pairs.

In contrast, however, many current cluster randomized trials employ a fundamentally different pair matched design. Communities are first sampled, and only then are matched pairs created from among this finite set by applying some algorithm to the baseline characteristics of the communities included in the sample. We refer to this design as "adaptive matching" in order to make explicit its links to the larger literature on adaptive study designs, and specifically adaptive treatment allocation in response to characteristics of the previously observed units: Bai et al. (2002); Andersen et al. (1994); Flournoy and Rosenberger (1995); Hu and Rosenberger (2000); Rosenberger (1996); Rosenberger et al. (1997); Rosenberger and Grill (1997); Rosenberger and Shiram (1997); Tamura et al. (1994); Wei (1979); Wei and Durham (1978); Wei et al. (1990); Zelen (1969); Cheng and Shen (2005); van der Laan (2008); Chambaz and van der Laan (2010); van der Laan and Rose (2012).

Recent cluster randomized trials that have employed adaptive matching include the SPACE study of a school level intervention to improve physical activity in Denmark (Toftager et al., 2011), a cluster randomized trial of routine HIV-1 viral load monitoring in Zambia (Koethe et al., 2010), and the PRISM trial of a community-level intervention to prevent post-partum depression in Australia (Watson et al., 2004). Under adaptive matching, the matched pair no longer represents the independent sampling unit. Instead, such a design corresponds to a special case of the general experimental

design in which the allocation of treatment among a sample of n independent and identically distributed units is a function of the covariates of all sampled units. This raises a number of questions with practical implications for the design and analysis of cluster (as well as individual) randomized trials. When will adaptively pair matched designs result in efficiency gains relative to their non-matched counterparts? What is the optimally efficient approach to estimating the treatment effect in such studies? How should statistical inference be carried out given the dependence between randomized units?

The results developed in this paper also apply to observational studies in which treatment decisions for each participant in a randomly sampled cohort may be influenced by the covariates of all or a subset of the other cohort members, while the participant-specific outcome is only influenced by a participant's own covariates and assigned treatment. Consider, for example, a study that aims to evaluate the impact of enrollment in a weight loss program on participant weight loss. The study protocol might bring subsets of the sampled cohort members together to discuss the program, after which participants are allowed to decide whether or not they wish to enroll. In such a study, enrollment probabilities might differ depending on the characteristics of the subgroup to which enrollees are assigned (for example, the extent to which the subgroup includes charismatic or vocal individuals who have failed similar weight loss programs in the past). As a result, enrollment decisions within a subgroup are no longer independent.

Finally, a special case of the general experiment described in this article is one in which the treatment allocation for each unit in an i.i.d. sample from some target population can be a different function of the sample characteristics of all of the other units in the sample. Conditional on the baseline characteristics of the sample, the treatment assignment of each individual is independent; however, the individual-specific assignment mechanisms are not identical across the individuals. In the example study to evaluate the effect of a weight loss program, the entire sample might be divided up in several subgroups, allowing the subjects within a subgroup to mingle and talk among themselves, before being provided with information about the weight loss program and subsequently deciding whether to enroll. Individual enrollment decisions in such a scenario might depend on the characteristics of other attendees in that subgroup. One might be willing to assume that each individual's enrollment decision is made independently, given what he or she has observed about the characteristics of the other attendees in the subgroup in contrast to the previous example in which decisions within subgroups were dependent. However, an individual's enrollment decision is indexed by the subgroup he or she belongs to, so that treatment allocation is not identical across all individuals.

1.1 Organization of article

In Section 2 we define the statistical estimation problem posed by estimating the additive causal effect of treatment (or average treatment effect) under the general experimental design in which treatment allocation can depend on the characteristics of other units in the sample. Specifically, we define the data generating experiment, the observed data, the likelihood, the statistical model and the target parameter.

In Section 3 we study the fundamental mathematics of the design by determining the tangent

space of the model and the canonical gradient of the pathwise derivative of the target parameter. Section 4 presents a targeted minimum loss based estimator (TMLE) of the additive causal effect of treatment and discusses its implementation. The TMLE presented is double robust. In particular, it remains consistent and asymptotically normally distributed as long as the treatment assignment mechanism for the n treatments is known or well estimated, even if the conditional mean of the outcome given the treatment and covariates is estimated inconsistently. We further present an estimator of the asymptotic variance of this TMLE. Interestingly, it appears that no double robust estimator of this variance is available, so that asymptotic consistent estimation of the variance requires a consistent estimator of the outcome regression. This demonstrates a strong contrast with designs in which treatment is independently assigned. In Section 5 we present a theorem that provides a formal basis for the estimators introduced in Section 4, and in particular establishes the asymptotics of the TMLE and thereby the validity of the statistical inference based on a normal limit distribution.

Section 6 discusses implications of these results for the design and analysis of randomized trials with adaptive pair matching. In particular, we discuss implementation of a TMLE of the average treatment effect and corresponding statistical inference in terms of confidence intervals and testing. While consistent estimation of the variance requires consistent estimation of the outcome regression, for this special case we show that a conservative estimate of the variance is possible. We further contrast the asymptotic variance of the adaptive pair matched design with the asymptotic variance of a design in which the intervention is randomly allocated to each unit independently. This provides insight into the potential benefits of pair matching in cluster randomized trials, beyond that provided by previous literature in which the matched pair constituted the unit of independence. In addition, we contrast the approach to statistical inference presented in this article to the standard approach employed in pair matched randomized trials, in which the average treatment effect is estimated as the sample mean of paired differences, and the variance is estimated as the sample variance treating the pairs as independent. It is shown that this standard approach provides conservative inference, under an explicitly stated assumption which is generally expected to hold.

Section 7 extends these results to the common case in which some units in the initial sample are missing treatment and outcome data. Such a case would occur, for example, in a cluster randomized trial with adaptive pair matching in which treatment were only allocated among the subset of sampled units for which adequate pair matches were identified. We conclude with a summary of the practical implications of our results and identify areas for future work. Proofs of all theorems and an overview of the required empirical process theory are provided in an Appendix.

1.2 Novel contributions of this article

To the best of our knowledge, the estimation problem addressed in this article has not been formally studied. This estimation problem targets the usual average causal effect, but the dependent allocation of treatment allowed by our model makes it different from other estimation problems that the literature has covered.

Even though targeted maximum likelihood estimation is a general method that has been applied to many problems in the literature (see e.g., van der Laan and Rose (2012) for an overview and comprehensive coverage of this method), the actual construction of a targeted maximum likelihood

estimator for a new estimation problem, as defined by the statistical model and target parameter, requires new research: it relies on the construction of a least favorable sub-model for fluctuating an initial estimator and a loss function so that the loss-function specific score of the least favorable sub-model at zero fluctuation spans the efficient influence curve. In particular, this requires determining the efficient influence curve (i.e., canonical gradient of pathwise derivative) for this target parameter in this new model. Indeed, the resulting TMLE as developed in this article is new and not presented anywhere else. In addition, the analysis of this TMLE relies on the state of the art methods in empirical process theory as presented in van der Vaart and Wellner (1996). Finally, the implications of our results for the analysis of cluster randomized trials and observational studies in which treatment allocation depends on the covariates of other units in the sample are new and important. In particular, our theoretical results allow us to formally compare the efficiency of different possible matched pair designs mentioned in the introduction. This work will appear in a future article.

2 Definition of Statistical Estimation Problem

2.1 Observed data

Let $X^n = (X_1, \ldots, X_n)$ be a vector consisting of n i.i.d. observations of $X_i = (W_i, Y_i(0), Y_i(1))$, where W_i denotes the baseline covariates, and $(Y_i(0), Y_i(1))$ denotes the treatment-specific counterfactual outcomes for subject i. (In words, $Y_i(a)$ denotes the outcome that would have been observed for unit i if it had received treatment level $A = a$.) Let $P_{X,0}$ denote the common distribution of X_i. In addition, $g_0^n(A_1, \ldots, A_n \mid X^n)$ is the true conditional distribution of the treatment or intervention $A^n = (A_1, \ldots, A_n)$, conditional on X^n. The observed data are $O_i = (W_i, A_i, Y_i = Y_i(A_i))$, $i = 1, \ldots, n$, so that only one counterfactual outcome, corresponding to the treatment actually received, is observed for each unit. Note that $O^n = (O_1, \ldots, O_n)$ is a many to one function of A^n and X^n, and is thus a missing or censored data structure in which the full data is X^n and the censoring or missingness variable is A^n.

We assume throughout that the conditional distribution of A^n, given X^n, $g_0^n(\cdot \mid X^n)$, is only a function of X^n through $W^n = (W_1, \ldots, W_n)$, which implies the coarsening at random assumption on g_0^n with respect to the full data X^n (Heitjan and Rubin (1991); Jacobsen and Keiding (1995); Gill et al. (1997)). This assumption allows for dependence between $A_1, \ldots, A_n$, as long as it can be explained by covariate vector W^n. One important class of examples covered by such treatment mechanisms g_0^n are studies that first partitions the sample $\{1, \ldots, n\}$ into groups based on the covariate vector W^n and subsequently randomly assign the treatment and control intervention within each group. For example, cluster randomized trials are commonly implemented by first partitioning the sample $\{1, \ldots, n\}$ into $n/2$ pairs based on some metric of similarity in baseline covariates W^n, and then randomly assigning a treatment and a control condition to the two members of each pair. More formally, g_0^n in such a design can be defined as follows: given W^n and thereby the disjoint pairs $C_j(W^n) = \{j_1, j_2\} \subset \{1, \ldots, n\}$ with $C_1(W^n) \cup \ldots \cup C_{n/2}(W^n) = \{1, \ldots, n\}$, within each pair $C_j(W^n)$ assign (1,0) or (0,1) with a flip of a fair coin (i.e. with probability 0.5).

Instead of using the Neyman-Rubin counterfactual formulation above, this observed data gen-

erating distribution can also be described in terms of an non-parametric structural equation model (NPSEM) (Pearl (1995, 2009)) as follows. Let $W_i = f_W(U_{W_i})$, $U_{W,i}$, $i = 1, \ldots, n$, are i.i.d., $A^n = f_{A^n}(W^n, U_{A^n})$, $Y_i = f_Y(W_i, A_i, U_{Y,i})$, with $U_{Y,i}$, $i = 1, \ldots, n$, i.i.d. The analogue of the coarsening at random assumption in terms of this NPSEM is that U_{A^n} is independent of $(U_{Y,i} : i = 1, \ldots, n)$, given W^n. The functions f_Y and f_W are unspecified, but the function f_{A^n} and the distribution of U_{A^n} might be known. For example, the sample might be partitioned into groups according to some known algorithm applied to the baseline characteristics of the sample, and the intervention A randomly assigned within each group.

2.2 Likelihood and statistical model

Under both formulations of the data generating experiment, the observed data is $O_i = (W_i, A_i, Y_i)$, $i = 1, \ldots, n$, and the likelihood of the observed data $O^n = (O_1, \ldots, O_n)$, under distribution P^n, is given by

$$P^n(O_1, \ldots, O_n) = \prod_{i=1}^{n} Q_W(W_i) Q_Y(Y_i \mid W_i, A_i) g^n(A^n \mid W^n),$$

where $Q_W = Q_W(P^n)$ and $Q_Y = Q_Y(P^n)$ denote the common marginal distribution of W and the common conditional distribution of Y, given A, W, respectively. We put no constraints on the sets of possible Q_Y and Q_W, which corresponds with putting no constraints on the common full data distribution $P_{X,0}$ (or no constraints on the NPSEM specified above beyond assumptions on the equation for A^n). Regarding the treatment mechanism g_0^n, we assume that

$$g_0^n(A^n \mid W^n) = \prod_{j=1}^{J} g_{0,j}(A(j) : j \in C_j(W^n) \mid W^n), \tag{2.1}$$

where $C_1(W^n), \ldots, C_J(W^n)$ is a partitioning of the sample $\{1, \ldots, n\}$ into J groups deterministically implied by W^n. Thus, it is assumed that, conditional on W^n, the treatment labels within a group are independently assigned from treatment labels in another group. It is assumed that $\liminf_{n\to\infty} J(n)/n > 0$ so that the asymptotics will still be driven by n. Let g_i^n be the conditional distribution of A_i, given W^n. Although not necessary for deriving the desired asymptotics, we assume that this distribution g_i^n of A_i, given W^n is non-deterministic, $i = 1, \ldots, n$. The set of possible g_0^n will be denoted with $\mathcal{G}^n$. The set of all possible data distributions P^n implied by the nonparametric model on $P_{X,0}$ and the model $\mathcal{G}^n$ for g_0^n represents a statistical model $\mathcal{M}^n$ for the true data distribution P_0^n. This general model will be referred to as $\mathcal{M}^n$. (Generalization of our results to general $J(n)$ with rates of convergence $1/\sqrt{J(n)}$ should be possible as well, but is not pursued here.)

2.2.1 Special models of interest for the treatment mechanism

A special choice for $\mathcal{G}^n$ consists of distributions satisfying $g^n(A^n \mid W^n) = \prod_i g_i(A_i \mid W^n)$. In this particular model it is assumed that, given W^n, $A_1, \ldots, A_n$ are independent with conditional distributions g_i^n of A_i, given W^n, $i = 1, \ldots, n$. This choice, which corresponds to partitions of size

1, allows treatment to be assigned to each unit in the sample according to a distinct unit-specific mechanism that is allowed to depend on the baseline covariates of the entire sample. Such a data generating process might arise in a study, such as the weight loss example presented in the introduction, in which subgroups of individuals are allowed to interact before each assigning themselves independently to the treatment or control condition. Then the baseline covariates of the subgroups influence individual treatment decisions but these decisions are still made independently given the baseline covariates of the cohort. We refer to this choice of $\mathcal{G}^n$ as $\mathcal{G}_1^n$ and refer to corresponding statistical model, implied by the nonparametric model on $P_{X,0}$ and $\mathcal{G}_1^n$ as $\mathcal{M}_1^n$.

A second special choice for $\mathcal{G}_n$ consists of distributions satisfying $g^n(A^n \mid W^n) = \prod_{j=1}^{n/2} g(A(j) : j \in C_j(W^n) \mid W^n)$, where $C_j(W^n) = \{j_1, j_2\}$, $j = 1, \ldots, n/2$, represents a partitioning of the sample $\{1, \ldots, n\}$ into $n/2$ disjoint pairs $C_j(W^n)$. This class of treatment assignment mechanisms describes randomized trials with adaptive pair matching. We refer to this choice of $\mathcal{G}^n$ as $\mathcal{G}_2^n$ and refer to corresponding statistical model, implied by the nonparametric model on $P_{X,0}$ and $\mathcal{G}_2^n$ as $\mathcal{M}_2^n$.

2.3 Target statistical parameter

We focus on the target quantity $\Psi^F(P_X) = E\{Y(1) - Y(0)\}$, a particular parameter of the full-data distribution P_X or, equivalently, a parameter of the distribution of the counterfactuals $(Y(0), Y(1))$ defined by the NPSEM. This quantity is often referred to as the average treatment effect, and corresponds to the causal quantity typically targeted by randomized trials as well as many observational studies. Under coarsening at random, $E_{Q_Y}(Y \mid A = a, W) = E_{P_X}(Y(a)|W)$, while the parameters (Q_Y, Q_W) of P^n are identifiable parameters of P^n. This target quantity is thus identified by the distribution of the data P^n as follows:

$$\begin{aligned} \Psi^F(P_X) &= \Psi(Q) = E_{Q_W}\{E_{Q_Y}(Y \mid A = 1, W) - E_{Q_Y}(Y \mid A = 0, W)\} \\ &= E_{Q_W}\{\bar{Q}(1, W) - \bar{Q}(0, W)\}, \end{aligned}$$

where $Q = (Q_W, \bar{Q})$ denotes the common distribution Q_W of W_i, and common conditional mean $\bar{Q}$ of Y_i, given A_i, W_i. Here $Q = Q(P^n)$ is a parameter of the observed data distribution P^n. This identifiability result defines now a target parameter $\Psi : \mathcal{M}^n \to \mathbb{R}$ of the observed data distribution, defined as $\Psi(P^n) = \Psi(Q)$ (where we abuse notation by using the same Ψ for two different mappings).

The estimation problem is now defined: we want to estimate $\Psi(P^n)$ based on $O^n = (O_1, \ldots, O_n) \sim P^n \in \mathcal{M}^n$, and we also want to provide asymptotic inference in terms of confidence intervals and tests of the null hypotheses.

3 The canonical gradient of the pathwise derivative of the target parameter

In order to construct efficient asymptotically linear estimators, and in particular targeted minimum loss-based estimators (van der Laan and Rubin (2006); van der Laan and Rose (2012)), of $\Psi(P^n)$, we

first determine the tangent space of the model and the canonical gradient of the pathwise derivative of the target parameter.

Let

$$D^*(Q,g)(W,A,Y) = \frac{2A-1}{g(A \mid W)}(Y - \bar{Q}(A,W)) + \bar{Q}(1,W) - \bar{Q}(0,W) - \Psi(Q),$$

which is the efficient influence curve of $\Psi : \mathcal{M} \to \mathbb{R}$ with $\Psi(P) = \Psi(Q)$ under i.i.d. sampling from $P_{Q,g}$, where g is a conditional distribution of A, given W (van der Laan and Robins (2003); van der Laan and Rose (2012)). We will also denote $D^*(Q,g)$ with $D^*(Q,g,\Psi(Q))$ to stress its representation as an estimating function in ψ. We note $D^*(Q,g,\Psi(Q)) = D^*_Y(\bar{Q},g) + D^*_W(Q)$, where

$$\begin{aligned} D^*_Y(\bar{Q},g) &= \frac{2A-1}{g(A \mid W)}(Y - \bar{Q}(A,W)) \\ D^*_W(Q) &= \bar{Q}(1,W) - \bar{Q}(0,W) - \Psi(Q). \end{aligned}$$

The following theorem presents the canonical gradient of the pathwise derivative of the parameter $\Psi : \mathcal{M}^n \to \mathbb{R}$. (For semiparametric efficiency theory, see e.g., Bickel et al. (1997); van der Laan and Robins (2003); van der Vaart (1998).) This canonical gradient is expressed in terms of the above function $D^*(Q,g)$.

Theorem 1. *Let $O_i = (W_i, A_i, Y_i)$, $O^n = (O_1, \ldots, O_n) \sim P^n$, with*

$$P^n(O_1, \ldots, O_n) = \prod_{i=1}^n Q_W(W_i) Q_Y(Y_i \mid W_i, A_i) g^n(A^n \mid W^n),$$

where Q_W is an unspecified marginal distribution, Q_Y is an unspecified conditional distribution of Y, given A, W, and g^n is a conditional distribution of $A^n = (A_1, \ldots, A_n)$, given $W^n = (W_1, \ldots, W_n)$, known to be an element of a set $\mathcal{G}^n$ consisting of distributions satisfying (2.1). Let $\mathcal{M}^n$ be the resulting statistical model for P^n. Let $\mathcal{M}^n(g^n)$ be the model if g^n is known.

Let $\Psi : \mathcal{M}^n \to \mathbb{R}$ be defined by $\Psi(P^n) = E_{Q_W}\{\bar{Q}(1,W) - \bar{Q}(0,W)\}$, where $\bar{Q}(A,W) = E_{Q_Y}(Y \mid A, W)$.

The tangent space at P^n in model $\mathcal{M}^n$ is given by:

$$T(P^n) = \left\{ \sum_{i=1}^n \phi(W_i) : \phi \in T_W \right\} + \left\{ \sum_{i=1}^n \phi(Y_i \mid A_i, W_i) : \phi \in T_Y \right\} + \sum_{j=1}^J T_{C_j}, \qquad (3.1)$$

where $T_W = \{h(W) : Eh(W) = 0\}$,

$$T_Y = \{h(Y \mid A, W) : E_{Q_Y}(h(Y \mid A, W) \mid A, W) = 0\},$$

and

$$T_{C_j} = \{S((A_j : j \in C_j(W^n)) \mid W^n) : E(S \mid W^n) = 0\}.$$

The tangent space at P^n in model $\mathcal{M}^n(g^n)$ is given by

$$T(Q) = \left\{ \sum_{i=1}^{n} \phi(W_i) : \phi \in T_W \right\} + \left\{ \sum_{i=1}^{n} \phi(Y_i \mid A_i, W_i) : \phi \in T_Y \right\}.$$

The statistical parameter Ψ is pathwise differentiable and its canonical gradient at P^n is given by

$$D^{n,*}(P^n) = \frac{1}{n} \sum_{i=1}^{n} D^*(Q, \bar{g}_n)(O_i) = \frac{1}{n} \sum_{i=1}^{n} \{D_W^*(Q)(W_i) + D_Y^*(Q, \bar{g}_n)(O_i)\},$$

where g_i is the conditional distribution of A_i, given W_i, and

$$\bar{g}_n(a \mid W) = \frac{1}{n} \sum_{i=1}^{n} g_i(a \mid W).$$

We note that

$$g_i(1 \mid W_i) = \sum_{(w_j : j \neq i)} g_i(1 \mid (w_j : j \neq i), W_i) \prod_{j \neq i} Q_W(w_j) \tag{3.2}$$

is a function of $g_i(A_i \mid W^n)$ and the common marginal distribution Q_W.

Double robustness of canonical gradient: *We have*

$$E_0 D^{n,*}(\bar{Q}, \bar{g}_n, \psi_0) = 0 \text{ if } \bar{Q} = \bar{Q}_0 \text{ or } \bar{g}_n = \bar{g}_{n,0}, \tag{3.3}$$

assuming that for all i, $0 < g_i(1 \mid W) < 1$ a.e. More generally, if $Q_W = Q_{W,0}$, then for any $\bar{Q}, \bar{g}$, we have

$$E_0 D^{n,*}(\bar{Q}, \bar{g}, \Psi(Q)) = \psi_0 - \Psi(Q) + E_0 \left(\tfrac{\bar{g}_0}{\bar{g}}(1 \mid W) - 1 \right) (\bar{Q}_0 - \bar{Q})(1, W) - E_0 \left(\tfrac{\bar{g}_0}{\bar{g}}(0 \mid W) - 1 \right) (\bar{Q}_0 - \bar{Q})(0, W).$$

The proof is presented in the Appendix.

4 A targeted minimum loss-based estimator (TMLE)

Derivation of the canonical gradient of the pathwise derivative of the target parameter Ψ allows us to construct a targeted minimum loss based estimator (TMLE). In this section we present a TMLE for Ψ for the general statistical model $\mathcal{M}^n$, in which $g_0^n(A^n|W^n) = \prod_{j=1}^{J} g_{0,j}(A(j) : j \in C_j(W^n)|W^n)$ is unknown. This TMLE is thus applicable to studies, such as the example presented in the introduction, in which a cohort of individuals is partitioned into subgroups and individuals within subgroups are allowed to interact in determining their treatment assignment according to some unknown mechanism. It further covers the special cases in which the sample is partitioned into n singletons and in which g_0^n is known (as in a adaptively pair matched trial). Section 6 considers the latter special case in greater detail.

We recall from the literature on TMLE (van der Laan and Rose (2012)) specification of a TMLE of a target parameter $\Psi(Q_0)$ requires specification of a loss function $L(Q)$ and a sub-model $\{Q(\epsilon) : \epsilon\}$ through a Q at $\epsilon = 0$, possibly indexed by nuisance parameter (in our case, $\bar{g}$), so that $\frac{d}{d\epsilon}L(Q(\epsilon))\big|_{\epsilon=0}$ spans the canonical gradient (in our case, $D^{n,*}(Q,\bar{g})$). Since $Q = (Q_W, \bar{Q})$ and $Q_{W,n}$ is already a nonparametric maximum likelihood estimator, the TMLE will only involve fluctuating $\bar{Q}$.

Loss function and initial estimator for $\bar{Q}_0$: Let $Y \in \{0,1\}$ be binary or continuous in $(0,1)$. Let $\bar{Q}_n^0$ be an initial estimator of $\bar{Q}_0$, which can be based on the loss-function

$$-L_i(\bar{Q})(O_i) = \left\{Y_i \log \bar{Q}(W_i, A_i) + (1 - Y_i)\log(1 - \bar{Q}(W_i, A_i))\right\}. \tag{4.1}$$

To understand the validity of this loss, note that

$$-E_0 L_i(\bar{Q})(O_i) = E_{Q_{W,0},g_{0,i}} \bar{Q}_0(A_i, W_i)\log \bar{Q}(A_i, W_i) + (1 - \bar{Q}_0)(A_i, W_i)\log(1 - \bar{Q}(A_i, W_i)),$$

which is indeed minimized at $\bar{Q} = \bar{Q}_0$. This demonstrates that $L_i(\bar{Q})$ is a valid loss function for $\bar{Q}_0$. Specifically, one could fit $\bar{Q}_0$ by minimizing $\sum_{i=1}^n L_i(\bar{Q}_\theta)(O_i)$ over a parametric or semiparametric working model $\{\bar{Q}_\theta : \theta \in \Theta\}$. Furthermore, to select among estimators such as different choices of working models or different algorithms, we can also use this loss to carry out cross-validation based estimator selection. Since conditional on A^n, W^n, the outcomes Y_i, $i = 1, \ldots, n$, are independent, a cross-validation selector that uses (4.1) as loss function and treats $i = 1, \ldots, n$ as the index of the independent units when splitting the sample into training and validation sets will satisfy an oracle type inequality analogue to the ones developed and presented in van der Laan and Dudoit (2003); van der Laan et al. (2006); van der Vaart et al. (2006). Thus an initial estimator of $\bar{Q}_0$ can be based on applying a data-adaptive loss-based learning approach such as super learning, ignoring the dependence between the treatment labels (van der Laan et al. (2007); Polley and van der Laan (2010) and Chapter 3 by Polley, Rose, van der Laan in van der Laan and Rose (2012)).

Least favorable sub-model through initial estimator: Let $\bar{g}_0 = 1/n \sum_i g_{0,i}$, where $g_{0,i}$ is the true conditional distribution of A_i, given W_i. As sub-model for fluctuating $\bar{Q}_n^0$ we use

$$\text{Logit}\bar{Q}_n^0(\epsilon) = \text{Logit}\bar{Q}_n^0 + \epsilon H^*_{\bar{g}_n},$$

where $H^*_{\bar{g}_n}(A, W) = (2A - 1)/\bar{g}_n(A \mid W)$, and $\bar{g}_n$ is an estimator of $\bar{g}_0$. Let $Q_{W,n}$ be the empirical distribution, and $Q_n^0 = (Q_{W,n}, \bar{Q}_n^0)$. We note that

$$\frac{d}{d\epsilon}\sum_i L_i(\bar{Q}_n^0(\epsilon))(O_i)\Big|_{\epsilon=0} = \sum_i D_Y^*(Q_n^0, \bar{g}_n)(O_i)$$

so that this loss function and sub-model indeed generates the crucial component of the canonical gradient of the target parameter, a requirement for the construction of efficient TMLE. The component corresponding with D_W^* is generated by a sub-model $Q_{W,n}(\epsilon)$ through $Q_{W,n}$ at $\epsilon = 0$ with score D_W^*, but since $Q_{W,n}$ is already an NPMLE, the estimated amount of fluctuation according to this sub-model would be zero, so that this sub-model plays no role in the TMLE.

Computing the update of initial estimator: The amount of fluctuation ϵ_n is estimated as

$$\epsilon_n = \arg\min_{\epsilon} \sum_{i=1}^{n} L_i(\bar{Q}_n^0(\epsilon))(O_i).$$

This provides the update $\bar{Q}_n^* = \bar{Q}_n^0(\epsilon_n)$. Let $Q_n^* = (Q_{W,n}, \bar{Q}_n^*)$.

TMLE of target parameter: The TMLE of ψ_0 is the corresponding plug-in estimator

$$\Psi(Q_n^*) = \frac{1}{n}\sum_{i=1}^{n}\{\bar{Q}_n^*(1, W_i) - \bar{Q}_n^*(0, W_i)\}.$$

TMLE solves efficient influence curve equation: By construction, the TMLE solves

$$0 = D^{n,*}(\bar{Q}_n^*, \bar{g}_n, \psi_n^*) = \frac{1}{n}\sum_{i=1}^{n} D^*(\bar{Q}_n^*, \bar{g}_n, \psi_n^*)(O_i).$$

In other words, the TMLE solves the efficient influence curve equation for the model $\mathcal{M}^n$. This equation will form a crucial ingredient for establishing double robustness and asymptotic normality of the TMLE $\Psi(Q_n^*)$.

4.1 Estimation of $\bar{g}_0$

In general $\bar{g}_0$ is not known and must be estimated. In this section we sometimes suppress the n in $\bar{g}_n$ when referring to an estimator of $\bar{g}_0 = 1/n\sum_i g_{0,i}$.

Estimation of $\bar{g}_0$ can be based on the pooled log-likelihood $L(\bar{g})(W^n, A^n) = \sum_i \log \bar{g}(A_i \mid W_i)$, as if we observe a sample of n i.i.d. (W_i, A_i). Let $\bar{g}_n$ be the resulting estimator. The above TMLE is then applied with $\bar{g}_n$ as an estimator of $\bar{g}_0$. Indeed, $L(\bar{g})$ is a valid loss function for $\bar{g}_0$ since

$$E_0 L(\bar{g})(W^n, A^n) = E_0 \sum_{i=1}^{n} \log \bar{g}(a \mid W_i) g_{0,i}(a \mid W_i) = E_{Q_{W,0}, \bar{g}_0} \log \bar{g}(A \mid W),$$

which is minimized at $\bar{g}_0$. Conditional on W^n, the groups $(A_i : i \in C_j(W^n))$ of treatment nodes are independent. Thus, $\bar{g}_0$ can be estimated using loss-based learning and cross-validation, but the cross-validation should, in contrast to estimation of $\bar{Q}_0$, treat the groups indexed by j as the independent units.

In a randomized controlled trial, $g_0(A_i \mid W^n)$ is known by design, while $g_{0,i}(A_i \mid W_i)$ and thus $\bar{g}_0(A_i \mid W_i)$ are not and must thus still be estimated. In such cases, knowledge of the true design g_0^n can be used to get a more accurate estimate of $\bar{g}_0$. Specifically, we have

$$g_i(1 \mid W_i) = \sum_{(w_j : j \neq i)} g_i(1 \mid (w_j : j \neq i), W_i) \prod_{j \neq i} Q_W(w_j).$$

Thus, if g_0^n is known, we can estimate Q_W with the empirical distribution, giving the estimator

$$g_{i,n}(1 \mid W_i) = \sum_{w_j : j \neq i} g_{0,i}(1 \mid (w_j : j \neq i), W_i) \prod_{j \neq i} Q_{W,n}(w_j),$$

and corresponding $\bar{g}_n = 1/n \sum_i g_{i,n}$ of $\bar{g}_0$.

4.2 Statistical inference

In our main Theorem 2 below we assume that the design $g^n = P(A^n \mid W^n)$ is known, but, as mentioned above, this still requires estimation of $\bar{g}_0$ through estimation of $Q_{W,0}$. The asymptotics of Theorem 2 below proves that, under appropriate conditions, the standardized TMLE $\sqrt{n}(\psi_n^* - \psi_0)$ converges to a normal distribution with variance $\sigma_W^2 + \sigma_Y^2$, where σ_Y^2 is consistently approximated with

$$\sigma_{Y,n}^2 = \frac{1}{n} \sum_{j=1}^{J} \{f_{j,n}(\bar{Q}_n^*)(O_i : i \in C_j(W^n))\}^2,$$

with $f_{j,n}(\bar{Q}) = \sum_{i \in C_j(W^n)} f_{i,n}^1(\bar{Q})(O_i)$, and

$$f_{i,n}^1(\bar{Q}) \equiv \tfrac{2A_i - 1}{\bar{g}_n(A_i \mid W_i)}(Y_i - \bar{Q}(A_i, W_i)) - \left\{ \tfrac{g_i(1 \mid W^n)}{\bar{g}_n(1 \mid W_i)}(\bar{Q}_0 - \bar{Q})(1, W_i) - \tfrac{g_i(0 \mid W^n)}{\bar{g}_n(0 \mid W_i)}(\bar{Q}_0 - \bar{Q})(0, W_i) \right\}.$$

Estimation of the asymptotic variance σ_Y^2 appears to rely on a consistent estimator of $\bar{Q}_0$. In addition, σ_W^2 is the variance of a function IC_W of W, which can thus be consistently estimated with $1/n \sum_{i=1}^n IC_{W,n}(W_i)^2$, if $IC_{W,n}$ is a consistent estimate of this unknown function IC_W. Specifically, IC_W is a sum of three functions, one of them being $D_W^*(\bar{Q}^*, Q_{W,0}) = \bar{Q}^*(W) - Q_{W,0}\bar{Q}^*$, where $\bar{Q}(W) = \bar{Q}(1, W) - \bar{Q}(0, W)$, and $Q_{W,0}\bar{Q}^* = \int_w \bar{Q}^*(w) dQ_{W,0}(w)$. This function D_W^* is trivially consistently estimated by plugging in $\bar{Q}_n^*$ and the empirical distribution $Q_{W,n}$. If $\bar{Q}_n^*$ converges to $\bar{Q}_0$, then the other two components of IC_W are equal to zero. If the conditional distribution $g_{0,i}^n$ of A_i, given W^n is equal to the conditional distribution $g_{0,i}$, given W_i, and the latter is constant in i, then these other two components of IC_W are also equal to zero, even if $\bar{Q}_n^*$ is inconsistent.

In general, one of these two components of IC_W is generated by the contribution of $\bar{g}_n$ as an estimator of $\bar{g}_0$, assuming a plug-in estimator is used utilizing that the distribution g_i^n of A_i, given W^n, is known. The influence curve of this contribution can be straightforwardly determined and is presented in Theorem 2. The other component concerns an average of differences $g_{0,i}^n - \bar{g}_n$, indicating that, $g_{0,i}(\cdot \mid W^n)$ has to converge for n going to infinity to a fixed $g_{0,i}(\cdot \mid W_i)$: thus, the dependence on the covariates of the other individuals $l \neq i$ has to be asymptotically negligible. If this convergence occurs fast enough this contribution may be equal to zero, but, in general, we allow for a contribution. This "asymptotic stability of the design" (i.e, g_i^n converging to a fixed g_i) condition is analogue to the condition on the adaptive allocation probabilities in adaptive group sequential designs to establish asymptotic normality of the TMLE, as studied in van der Laan (2008); Chambaz and van der Laan (2010). Either way, consistent estimation of σ_W^2 is possible without relying on a consistent estimator of $\bar{Q}_0$. On the other hand, if $g_{0,i}^n(1 \mid W^n) = g_{0,i}(1 \mid W_i)$, then the

influence curve is easily derived and is specified in Theorem 2, and, if $g_{0,i}$ is constant in i, then the contribution equals zero.

If $g_i^n = g_i$, and $\bar{g}_n = \bar{g}_0$, and $C_j(W^n)$ are singletons, then it can be shown that the asymptotic variance is consistently estimated as

$$\sigma_{I,n}^2 = \frac{1}{n}\sum_{i=1}^{n}\left\{D^*(\bar{Q}_n^*, \bar{g}_n, \psi_n^*)(O_i)\right\}^2, \tag{4.2}$$

which does thus not rely on a consistent estimator of $\bar{Q}_0$ anymore. Note that this latter variance estimator is the estimator one would have used if one treats the sample as n independent observations, and one ignores the adaptivity of the design.

In the special case of adaptive pair matching designs (and thus $\bar{g}_n = \bar{g}_0$ and $IC_W = D_W^*$), we prove below that, under a mild condition, this same estimator (4.2) of the asymptotic variance remains conservative if $\bar{Q}_n^*$ is inconsistent for true $\bar{Q}_0$. It remains to be determined if this result also applies to other group sizes.

Finally, if the design g_0^n is actually unknown and thus also needs to be estimated, and if we assume that this design is consistently estimated, then we conjecture that the asymptotic limit variance described above will be conservative, due to the general result that estimation of an orthogonal factor in the likelihood (i.e., the tangent space of the treatment mechanism is orthogonal to tangent space of relevant Q-factors) generally improves the asymptotic variance (Theorem 2.3 van der Laan and Robins (2003)).

5 Theorem establishing asymptotic normality

We have the following theorem establishing the asymptotic normality of the TMLE presented in Section 4 and thereby in particular the basis for the variance estimator presented in Section 4.2.

Theorem 2. *Let $P_{Q_0,g_0^n} f_i$ represents a conditional expectation of a function, given W^n, which is thus still random through W^n. In this theorem $g_0^n = P_0(A^n \mid W^n)$ is considered known. Let $\mathcal{F}$ be a set of multivariate real valued functions so that $\bar{Q}_n^*$ is an element of $\mathcal{F}$ with probability 1. Define*

$$f_{i,n}^1(\bar{Q}) \equiv \frac{2A_i-1}{\bar{g}_n(A_i|W_i)}(Y_i - \bar{Q}(A_i,W_i)) - \left\{\frac{g_{0,i}(1|W^n)}{\bar{g}_n(1|W_i)}(\bar{Q}_0-\bar{Q})(1,W_i) - \frac{g_{0,i}(0|W^n)}{\bar{g}_n(0|W_i)}(\bar{Q}_0-\bar{Q})(0,W_i)\right\}.$$

Define $X_n(\bar{Q}) = 1/\sqrt{n}\sum_{j=1}^{J}\{\sum_{i\in C_j(W^n)} f_{i,n}^1(\bar{Q})(O_i)\}$, and note that, conditional on W^n, this is a sum of $J = J(n)$ independent mean zero random variables. Let

$$f_{j,n} \equiv \sum_{i\in C_j(W^n)} f_{i,n}^1(O_i).$$

We can represent $X_n(\bar{Q})$ as $X_n(\bar{Q}) \equiv 1/\sqrt{n}\sum_{j=1}^{J} f_{j,n}(O_i : i \in C_j(W^n))$. Let $D_W^*(Q_0) = \bar{Q}_0(W) - \psi_0$.

Uniform bound: *Assume $\max_{i\in\{1,\ldots,n\}} \sup_{\bar{Q}\in\mathcal{F}} \sup_{W^n,O} \mid f_{i,n}(\bar{Q})(W^n,O) \mid < M < \infty$, where the second supremum is over a support of (W^n, A_i, Y_i).*

Asymptotic linearity of function of $\bar{g}_n$: *Assume that for a function $IC_{W,i,\bar{g}}$ of W with mean zero and finite variance (uniformly in i)*

$$\begin{aligned} Z_{W,n,\bar{g}_n} &\equiv \frac{1}{\sqrt{n}} \sum_i P_{Q_0,g_{0,i}} \{D^*(\bar{Q}_n^*, \bar{g}_n, \psi_n^*) - P_{Q_0,g_{0,i}} D^*(\bar{Q}_n^*, \bar{g}_0, \psi_n^*)\} \\ &= \frac{1}{\sqrt{n}} \sum_{i=1}^n IC_{W,i,\bar{g}}(W_i) + o_P(1). \end{aligned}$$

Note that if $g_{0,i}^n = P_0^n(A_i \mid W^n)$ equals $g_{0,i} = P_0(A_i \mid W_i)$ and is thus known, then $\bar{g}_n = \bar{g}_0$ so that $Z_{W,n,\bar{g}_n} = 0$. In general, under the required regularity conditions, we have

$$Z_{W,\bar{g}_n,n} \approx \tfrac{1}{\sqrt{n}} \textstyle\sum_{k=1}^n IC(W_k) - E_0 IC(W_k),$$

where

$$\begin{aligned} IC(W_k) &= \textstyle\int_w \left(\frac{1}{n} \sum_{i=1}^n \sum_{l \neq i}^n g_i(1 \mid W_i = w, W_l = W_k) \right) \frac{\bar{Q}_0 - \bar{Q}}{\bar{g}_0}(1, w) dQ_0(w) \\ &- \textstyle\int_w \left(\frac{1}{n} \sum_{i=1}^n \sum_{l \neq i}^n g_i(0 \mid W_i = w, W_l = W_k) \right) \frac{\bar{Q}_0 - \bar{Q}}{\bar{g}_0}(0, w) dQ_0(w). \end{aligned}$$

and $g_{i,0}(1 \mid W_i, W_l)$ is the conditional distribution of $A_i = 1$, given W_i, W_l.

Asymptotic stability of treatment mechanism as function of covariates: *Let*

$$Z_{W,n,g^n} = Z_{1,g^n} + Z_{W,n},$$

where

$$Z_{1,g^n} = \tfrac{1}{\sqrt{n}} \textstyle\sum_{i=1}^n \frac{g_i(1|W^n) - \bar{g}_n(1|W_i)}{\bar{g}_n(1|W_i)} (\bar{Q}_0 - \bar{Q}_n^*)(1, W_i) - \frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{g_i(0|W^n) - \bar{g}_n(0|W_i)}{\bar{g}_n(0|W_i)} (\bar{Q}_0 - \bar{Q}_n^*)(0, W_i),$$

and

$$Z_{W,n} = \frac{1}{\sqrt{n}} \sum_{i=1}^n \{\bar{Q}_0(W_i) - \psi_0\}.$$

Assume, for a function IC_{W,i,g^n} of W with mean zero and finite variance, $Z_{1,g^n} = \frac{1}{\sqrt{n}} \sum_{i=1}^n IC_{W,i,g^n}(W_i) + o_P(1)$. Note that, if $g_{0,i}^n = g_{0,i}$ and constant in i, then $Z_{1,g^n} = 0$. If $g_{0,i}^n = g_{0,i}$, so that $\bar{g}_n = \bar{g}_0$, then

$$IC_{W,i,g^n}(W_i) = \frac{g_{0,i}(1|W_i) - \bar{g}_0(1|W_i)}{\bar{g}_0(1|W_i)} (\bar{Q}_0 - \bar{Q})(1, W_i) - \frac{g_{0,i}(0|W_i) - \bar{g}_0(0|W_i)}{\bar{g}_0(0|W_i)} (\bar{Q}_0 - \bar{Q})(0, W_i).$$

Convergence of variances: *Assume that for a specified $\{\Sigma_0(\bar{Q}_1, \bar{Q}_2) : \bar{Q}_1, \bar{Q}_2 \in \mathcal{F}\}$, for any $\bar{Q}_1, \bar{Q}_2 \in \mathcal{F}$, $\frac{1}{n} \sum_{j=1}^J P_{Q_0,g_0^n} f_{j,n}(\bar{Q}_1)^2 \to \Sigma_0(\bar{Q}_1, \bar{Q}_1)$ a.s (i.e, for almost every $(W^n, n \geq 1)$), and*

$$\frac{1}{n} \sum_{j=1}^J P_{Q_0,g^n} f_{j,n}(\bar{Q}_1) f_{j,n}(\bar{Q}_2) \to \Sigma_0(\bar{Q}_1, \bar{Q}_2) \text{ a.s.} \tag{5.1}$$

For example, if $C_j(W^n)$ are singletons, the first condition holds if

$$\left[\frac{1}{n}\sum_{i=1}^{n}\frac{g_{0,i}(0 \mid W^n)}{\bar{g}_n^2(0 \mid W_i)}E_0((Y-\bar{Q}(0,W_i))^2 \mid A_i=0,W_i)\right.$$
$$\left.+\frac{1}{n}\sum_{i=1}^{n}\frac{g_{0,i}(1|W^n)}{\bar{g}_n^2(1|W_i)}E_0((Y-\bar{Q}(1,W_i))^2 \mid A_i=1,W_i)\right] \to \Sigma_0(\bar{Q},\bar{Q}) \text{ a.s.}$$

Similarly, for the convergence of covariance. Note that this holds trivially if $g_{0,i}(1 \mid W^n) = g_{0,i}(1 \mid W_i)$.

Convergence of $\bar{Q}_n^*$ to some limit: *For any $\bar{Q}_1, \bar{Q}_2 \in \mathcal{F}$, we define*

$$\sigma_n^2(\bar{Q}_1-\bar{Q}_2) = \frac{1}{n}\sum_{j=1}^{J} P_{Q_0,g_0^n}\{f_{j,n}(\bar{Q}_1)-f_{j,n}(\bar{Q}_2)\}^2,$$

where we note that the right-hand side indeed only depends on $\bar{Q}_1, \bar{Q}_2$ through its difference $\bar{Q}_1 - \bar{Q}_2$.

Assume that for a particular $\bar{Q}^* \in \mathcal{F}$, $\sigma_n^2(\bar{Q}_n^* - \bar{Q}^*) \to 0$ in probability as $n \to \infty$.

Entropy condition: *Let $\mathcal{F}^d = \{f_1 - f_2 : f_1, f_2 \in \mathcal{F}\}$. Let $N(\epsilon, \sigma_n, \mathcal{F}^d)$ be the covering number of the class $\mathcal{F}^d$ w.r.t norm/dissimilarity $\| f \| = \sigma_n(f)$. Assume that the class $\mathcal{F}$ satisfies*

$$\lim_{\delta_n \to 0}\int_0^{\delta_n}\sqrt{\log N(\epsilon,\sigma_n,\mathcal{F}^d)}d\epsilon = 0$$

Asymptotic equicontinuity of process: *Then,*

$$X_n(\bar{Q}_n^*) - X_n(\bar{Q}^*) \text{ converges to zero in probability, as } n \to \infty.$$

First order linear approximation: *As a consequence,*

$$\sqrt{n}(\psi_n^* - \psi_0) = X_{W,n} + X_n(\bar{Q}^*) + o_P(1),$$

where $X_{W,n} = 1/\sqrt{n}\sum_{i=1}^n IC_{W,i}(W_i)$, and $IC_{W,i} = IC_{W,i,\bar{g}} + IC_{W,i,g^n} + D_W^*(Q_0)$.

Asymptotic normality: *In addition, $X_{W,n}$ converges in distribution to $N(0, \sigma_W^2)$, where*

$$\sigma_W^2 = \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n} P_0 IC_{W,i}^2,$$

and $X_n(\bar{Q})$ converges to an independent $N(0, \Sigma_0(\bar{Q}^*, \bar{Q}^*))$, so that

$$\sqrt{n}(\psi_n^* - \psi_0) \text{ converges in distribution to } N(0, \sigma_0^2 \equiv \sigma_W^2 + \Sigma_0(\bar{Q}^*, \bar{Q}^*)).$$

The asymptotic variance $\Sigma_0(\bar{Q}^*, \bar{Q}^*)$ *equals the limit of*

$$\Sigma_n = \frac{1}{n}\sum_{j=1}^{n}\{f_{j,n}(\bar{Q}_n^*)(O_i : i \in C_j(W^n))\}^2.$$

Under certain treatment allocation mechanisms $g_{0,i}^n$, one might have that the contributions captured by $X_{W,n}$ in this theorem require a more general representation $X_{W,n} = 1/\sqrt{n}\sum_{i=1}^n f_i(W)$, where $f_i(W)$ has weak enough dependence on W_j with $j \neq i$, so that such a process still converges weakly to a normal distribution. Depending on applications of interest, we can pursue such a more general representation of this theorem with little extra work.

Note that $f_{j,n}$ in Σ_n still depends on $\bar{Q}_0$, while $\bar{Q}_n^*$ estimates the possibly misspecified limit $\bar{Q}$. Thus Σ_n is not an estimator. In the special case that $C_j(W^n)$ are singletons, it follows that σ_0^2 can be consistently approximated with $1/n\sum_i\{IC_W(Q_n^*)(W_i) + f_{i,n}(\bar{Q}_n^*)(O_i)\}^2$ which equals

$$\sigma_n^2 = \frac{1}{n}\sum_{i=1}^{n}\{D^*(\bar{Q}_n^*, \bar{g}_n, \psi_n^*)(O_i)\}^2,$$

if $g_{0,i}^n = g_{0,i}$ and $\bar{g}_n = \bar{g}_0$ so that $IC_W = D_W^*(\bar{Q}^*, Q_{W,0})$.

Some of the conditions were discussed in the previous section 4 under statistical inference. The entropy condition corresponds with assuming that $\mathcal{F}$ is a Donsker class and is thus a natural condition that puts (minimal) restrictions on the size of the class $\mathcal{F}$. For example, one can define $\mathcal{F}$ as the class of multivariate real valued functions that have a uniform sectional variation norm bounded by a universal $M < \infty$ (van der Laan (1996); Gill et al. (1995)).

In order to demonstrate the condition (5.1), we consider the special case that $C_j(W^n)$ are singletons $j = 1, \ldots, n$, and that $g_i(A_i \mid W^n) = g_i(A_i \mid W_i)$. We wish to show that $\sigma_n^2 = 1/n\sum_{i=1}^n \sigma_i^2 \to \sigma^2$, where

$$\begin{aligned}
\sigma_i^2 &= P_{Q_0,g_i^n}\{(2A-1)/g_i(A_i \mid W_i)(Y_i - \bar{Q}(A_i, W_i))\}^2 - (\bar{Q}(W_i) - \bar{Q}_0(W_i))^2 \\
&= P_{Q_0,g_i^n} 1/g_i^2(A_i \mid W_i)(Y_i - \bar{Q}(A_i, W_i))^2 - (\bar{Q}(W_i) - \bar{Q}_0(W_i))^2 \\
&= \int_{a,y} \frac{1}{g_i^2(a \mid W_i)}(y - \bar{Q}(a, W_i))^2 g_i(a \mid W^n) Q_{Y,0}(y \mid a, W_i) - (\bar{Q}(W_i) - \bar{Q}_0(W_i))^2 \\
&= \int_{a,y} \frac{1}{g_i(a \mid W_i)}(y - \bar{Q}(a, W_i))^2 Q_{Y,0}(y \mid a, W_i) - (\bar{Q}(W_i) - \bar{Q}_0(W_i))^2 \\
&= \frac{1}{g_i(0 \mid W_i)} E_0((Y - \bar{Q}(0, W_i))^2 \mid A_i = 0, W_i) \\
&\quad + \frac{1}{g_i(1 \mid W_i)} E_0((Y - \bar{Q}(1, W_i))^2 \mid A_i = 1, W_i) - (\bar{Q}(W_i) - \bar{Q}_0(W_i))^2.
\end{aligned}$$

So we need that

$$\begin{aligned}
&\tfrac{1}{n}\textstyle\sum_{i=1}^n \frac{1}{g_i(0|W_i)} E_0((Y - \bar{Q}(0, W_i))^2 \mid A_i = 0, W_i) \\
&+\tfrac{1}{n}\textstyle\sum_{i=1}^n \frac{1}{g_i(1|W_i)} E_0((Y - \bar{Q}(1, W_i))^2 \mid A_i = 1, W_i) \\
&-\tfrac{1}{n}\textstyle\sum_{i=1}^n (\bar{Q}(W_i) - \bar{Q}_0(W_i))^2 \to \sigma^2 \text{ a.s.}
\end{aligned}$$

By the law of large numbers, both empirical means converge.

6 Randomized trials with adaptive pair matching

Theorem 2 and the corresponding TMLE of the average treatment effect and variance estimator have important implications for the design and analysis of individual and cluster randomized trials with adaptive pair matching. In particular, previous literature on pair matched trials considered the pair as the unit of independence (Freedman et al. (1990); Campbell et al. (2007); Hayes and Moulton (2009); Imai et al. (2009); Imai (2008); Donner and Klar (2000); Murray (1998); Donner and Klar (2000); Raudenbush et al. (2007); Klar and Donner (1997); Balzer et al. (2012)). This leaves open a number of key questions regarding the design and analysis of trials in which matched pairs are constructed based on applying some algorithm to the baseline characteristics of the entire sample (adaptive pair matching). What is the most efficient estimator of the intervention's effect in such studies? How should the variance of this estimator be estimated, given dependence induced between units? And finally, under what conditions will adaptive pair matching provide a more efficient estimator than that provided by a non-matched design? In this section we consider the implications of Theorem 2 for each of these questions in turn.

6.1 Estimation of the average treatment effect

In a randomized trial with adaptive pair matching, an unadjusted difference of the mean outcome in the treated and untreated units will provide an unbiased estimate of the average treatment effect. However, adjustment for baseline covariates W that predict the outcome Y will result in efficiency gains. This raises the issue of how best to accomplish such adjustment in adaptively pair matched designs, with the dual goals of minimizing the variance of the resulting estimator and ensuring that it remains unbiased. Our Theorem 2 and the corresponding TMLE presented in Section 4 establish such an efficient and unbiased estimator of the average treatment effect.

Specifically, in such trials, $g_{0,i}(A_i \mid W^n) = g_{0,i}(A_i \mid W_i) = \bar{g}_0(A_i \mid W_i)$ is known to be equal to a constant (typically 0.5 for a trial with two arms) and need not be estimated (although estimation of may improve efficiency). The clever covariate in the TMLE thus reduces to a simple $(2A_i - 1)/0.5$. Thus, a TMLE for the average treatment effect in such a trial can be implemented by treating the data as if they were a sample of n i.i.d. units. Implementation of this estimator is described in detail in van der Laan and Rose (2012). Further, because A is randomized, as long as the initial estimator of $\bar{Q}_0$ (the conditional mean of the unit-specific outcome given unit-specific covariates and treatment) is a least squares regression or logistic maximum likelihood regression that includes an intercept and the treatment A as a main term (still allowing for additional interaction terms between A and covariates W), no further update step is needed (the initial estimator is already a TMLE) (Rosenblum and van der Laan (2010); van der Laan and Rose (2012)).

6.2 Statistical inference

Again, we note that in a randomized trial with adaptive pair matching, we have $\bar{g}_n = \bar{g}_0 = g_0$. Our Theorem 2 shows that, under regularity conditions, the standardized TMLE with $\bar{Q}_n^*$ converging to $\bar{Q}^*$ is asymptotically consistent and normally distributed, $\sqrt{n}(\psi_n^* - \psi_0) \Rightarrow_d N(0, \sigma^2)$, where

$\sigma^2 = \sigma_W^2 + \sigma_Y^2$, $\sigma_W^2 = E(\bar{Q}_0(W) - \psi_0)^2$ and σ_Y^2 is the limit of

$$\begin{aligned}
&\tfrac{1}{n}\textstyle\sum_j P_{Q_0,g_0^n}\{f_{j,n}(\bar{Q}^*, \bar{Q}_0)\}^2 \\
&= \tfrac{1}{n}\textstyle\sum_j P_{Q_0,g_0^n}\left(\sum_{i\in C_j(W^n)} H_{\bar{g}_0}(A_i, W_i)(Y_i - \bar{Q}^*)\right)^2 \\
&\qquad -\tfrac{1}{n}\textstyle\sum_j \left(\sum_{i\in C_j(W^n)}(\bar{Q}_0 - \bar{Q}^*)(W_i)\right)^2.
\end{aligned}$$

Given a consistent estimator σ_n^2 of σ^2, $\psi_n^* \pm 1.96\sigma_n/\sqrt{n}$ is an asymptotic 0.95-confidence interval. A test of the null hypothesis $H_0 : \psi_0 = \psi^0$ can be based on the test statistic $T_n = \sqrt{n}(\psi_n^* - \psi^0)/\sigma_n$ which is approximately standard normal under H_0.

In order to implement an estimator of σ_n, σ_W^2 can be estimated as $\frac{1}{n}\sum_{i=1}^n\{\bar{Q}_n(W_i) - \psi_n^*\}^2$. Estimation of σ_Y^2 can be based on Theorem 2, which shows that σ_Y^2 is consistently approximated by $1/n\sum_j\{f_{j,n}(\bar{Q}_n^*, \bar{Q}_0)\}^2$. In implementing a substitution estimator of σ_Y^2, we naturally replace $\bar{Q}^*$ by $\bar{Q}_n^*$ (the updated fit of $\bar{Q}_0$ on which the TMLE substitution estimator of ψ_0 is based). However, $f_{j,n}(\bar{Q}_n^*, \bar{Q}_0)$ still depends on $\bar{Q}_0$. When implementing a consistent estimator of the asymptotic variance σ^2, one may need to estimate $\bar{Q}_0$ with a super learner in order to make it maximally unbiased (van der Laan et al. (2007)). In particular, even if a simple parametric regression based estimator was used as initial estimator of $\bar{Q}_0$ when implementing the TMLE for ψ_0, a more flexible approach to estimating $\bar{Q}_0$ is warranted when estimating σ^2 in order to minimize bias in the resulting variance estimator.

A substitution estimator of this asymptotic variance σ^2 is now given by

$$\begin{aligned}
\sigma_n^2 &= \sigma_{W,n}^2 + \sigma_{Y,n}^2 \\
\sigma_{W,n}^2 &= \frac{1}{n}\sum_{i=1}^n\{\bar{Q}_n(W_i) - \psi_n^*\}^2 \\
\sigma_{Y,n}^2 &= \frac{1}{n}\sum_j\left(\sum_{i\in C_j(W^n)} H_{\bar{g}_0}(A_i, W_i)(Y_i - \bar{Q}_n^*(A_i, W_i)) - \sum_{i\in C_j(W^n)}(\bar{Q}_n - \bar{Q}_n^*)(W_i)\right)^2.
\end{aligned}$$

Interestingly, even though we constructed an estimator of ψ_0 that is guaranteed consistent and asymptotically normally distributed at misspecified $\bar{Q}_n^*$, as long as $g_{0,i}^n$ is known or well estimated, it seems that in general no such robust estimator of the asymptotic variance σ^2 is available. Fortunately, below we will construct conservative variance estimators that do not rely on a consistent estimator of $\bar{Q}_0$.

6.3 Robust conservative estimation of the variance

In the special case of an adaptively pair matched design, we have the following theorem, which establishes a simpler (but) conservative estimate of the asymptotic variance σ^2. The proof is analogue to that of Theorem 4 below.

Theorem 3. *Suppose $g_n^0 \in \mathcal{G}_2^n$, in which case $C_j(W^n)$ are pairs. As above, assume that $g_{0,i}(A_i \mid W^n) = g_0(A_i \mid W_i)$ is known, so that $\bar{g}_0 = g_0$ is known as well. The asymptotic variance σ^2 of the*

TMLE for $\bar{Q}_n^*$ converging to $\bar{Q}^*$, i.e., $\sigma^2 = \sigma_W^2 + \sigma_Y^2$ in Theorem 2, can be represented as follows:

$$\sigma^2 \quad = \quad P_{Q_0,g_0}\{D^*(\bar{Q}^*, g_0, \psi_0)\}^2 - C,$$

where

$$\begin{aligned} C \equiv E_0 \frac{1}{J} \sum_{j=1}^{n/2} (\bar{Q}_0 - \bar{Q}^*)(1, W_{1j})(\bar{Q}_0 - \bar{Q}^*)(0, W_{2j}) \\ + E_0 \frac{1}{J} \sum_{j=1}^{n/2} (\bar{Q}_0 - \bar{Q}^*)(0, W_{1j})(\bar{Q}_0 - \bar{Q}^*)(1, W_{2j}) \\ + E_0 \frac{1}{J} \sum_{j=1}^{n/2} (\bar{Q}_0 - \bar{Q}^*)(1, W_{1j})(\bar{Q}_0 - \bar{Q}^*)(1, W_{2j}) \\ + E_0 \frac{1}{J} \sum_{j=1}^{n/2} (\bar{Q}_0 - \bar{Q}^*)(0, W_{1j})(\bar{Q}_0 - \bar{Q}^*)(0, W_{2j}). \end{aligned} \tag{6.1}$$

Under the assumption that the covariance-term C is positive, a conservative estimate of σ^2 is thus given by:

$$\sigma_{I,n}^2 = \frac{1}{n} \sum_{i=1}^{n} \{D^*(\bar{Q}_n^*, g_0, \psi_n^*)(O_i)\}^2.$$

This theorem, together with Theorem 2, imply that randomized trials with adaptive pair matching can be analyzed ignoring the matching process, both in order to generate an efficient and unbiased point estimator of the treatment effect and for inference on this estimator. Under the assumption that the last covariance-term is positive, as would be expected if units were effectively matched on predictors of the outcome, a variance estimator that treats the data as if they were i.i.d. will be conservative if $\bar{Q}_n^*$ is inconsistent for $\bar{Q}_0$, while it remains asymptotically consistent if $\bar{Q}_n^*$ is consistent.

In general, one can aim to construct a target C_l for which it is known that $C_l \leq C$, and estimate the variance with $\sigma_{I,n}^2 - C_{l,n}$, where $C_{l,n}$ is a consistent estimator of C_l. In this case, one aims to find such a C_l that can be consistently estimated without relying on a consistent estimator of $\bar{Q}_0$. This will be carried out in the next two subsections resulting in a possibly much less conservative variance estimator.

6.4 Comparison of the "naive" variance estimator with the true variance

Let us consider the case that we use the unadjusted regression of Y on A as initial estimator so that $\bar{Q}_n^*(a, W) = \bar{Q}_n^*(a)$, $a \in \{0, 1\}$, does not depend on W, and the TMLE $\psi_n^* = \bar{Q}_n^*(1) - \bar{Q}_n^*(0)$ equals the mean outcome over the $n/2$ pairs $C_j(W^n)$ of $\sum_{i \in C_j(W^n)} A_i Y_i - (1 - A_i) Y_i$, $j = 1, \ldots, n/2$:

$$\psi_n^* = \frac{1}{n/2} \sum_{j=1}^{n/2} \sum_{i \in C_j(W^n)} (A_i Y_i - (1 - A_i) Y_i).$$

Above we presented an expression (6.1) for the true asymptotic variance of the standardized TMLE, $\sqrt{n}(\psi_n^* - \psi_0)$, which can be applied to this simple sample average ψ_n^*. This expression σ^2 was represented as the asymptotic variance σ_I^2 of this TMLE under i.i.d sampling from P_{Q_0,g_0}, i.e. $\sigma_I^2 = P_{Q_0,g_0}D^*(\bar{Q}^*, g_0, \psi_0)^2$, minus a sum of four terms defined as

$$\begin{aligned}
&C \equiv E_0 \tfrac{1}{J} \textstyle\sum_{j=1}^{n/2} (\bar{Q}_0 - \bar{Q}^*)(1, W_{1j})(\bar{Q}_0 - \bar{Q}^*)(0, W_{2j}) \\
&+E_0 \tfrac{1}{J} \textstyle\sum_{j=1}^{n/2} (\bar{Q}_0 - \bar{Q}^*)(0, W_{1j})(\bar{Q}_0 - \bar{Q}^*)(1, W_{2j}) \\
&+E_0 \tfrac{1}{J} \textstyle\sum_{j=1}^{n/2} (\bar{Q}_0 - \bar{Q}^*)(1, W_{1j})(\bar{Q}_0 - \bar{Q}^*)(1, W_{2j}) \\
&+E_0 \tfrac{1}{J} \textstyle\sum_{j=1}^{n/2} (\bar{Q}_0 - \bar{Q}^*)(0, W_{1j})(\bar{Q}_0 - \bar{Q}^*)(0, W_{2j}).
\end{aligned}$$

Let $(W_1, A_1, Y_1), (W_2, A_2, Y_2)$ denote the observations in the two units in a pair, where either $(A_1, A_2) = (1, 0)$ or $(A_1, A_2) = (0, 1)$. For notational convenience, we will denote this term C as

$$\begin{aligned}
&C \equiv E_0(\bar{Q}_0 - \bar{Q}^*)(1, W_1)(\bar{Q}_0 - \bar{Q}^*)(0, W_2) \\
&+E_0(\bar{Q}_0 - \bar{Q}^*)(0, W_1)(\bar{Q}_0 - \bar{Q}^*)(1, W_2) \\
&+E_0(\bar{Q}_0 - \bar{Q}^*)(1, W_1)(\bar{Q}_0 - \bar{Q}^*)(1, W_2) \\
&+E_0(\bar{Q}_0 - \bar{Q}^*)(0, W_1)(\bar{Q}_0 - \bar{Q}^*)(0, W_2).
\end{aligned}$$

Since we use the unadjusted estimator so that $D^*(\bar{Q}^*, g_0, \psi_0)(W, A, Y) = (2A - 1)/g_0(A)(Y - \bar{Q}^*(A))$, and $g_0(A) = 0.5$, we have

$$P_{Q_0,g_0}D^*(\bar{Q}^*, g_0, \psi_0)^2 = 2\left\{\sigma_1^2 + \sigma_0^2\right\},$$

where $\sigma_1^2 = E_0(Y(1) - \psi_0(1))^2$ and $\sigma_0^2 = E_0(Y(0) - \psi_0(0))^2$. We conclude that the true asymptotic variance of $\sqrt{n}(\psi_n^* - \psi_0)$ is given by

$$\sigma^2 = 2\{\sigma_1^2 + \sigma_0^2\} - C.$$

Let us compare this true asymptotic variance σ^2/n of the unadjusted estimator with the variance estimate used in current practice, which we will refer to as the "naive" variance estimator. Current practice assumes that the $n/2$ pairs are i.i.d. and estimates the asymptotic variance of $\sqrt{n/2}(\psi_n^* - \psi_0)$ with the sample variance of the average of the difference across the pairs:

$$0.5\sigma_{n,naive}^2 = \tfrac{1}{n/2} \textstyle\sum_{j=1}^{n/2} (Y_{1j}A_{1j} + Y_{2j}A_{2j} - Y_{1j}(1 - A_{1j}) - Y_{2j}(1 - A_{2j}) - \psi_n^*)^2.$$

This converges for $n \to \infty$ to

$$0.5\sigma_{naive}^2 = \sigma_0^2 + \sigma_1^2 - (\rho_1 + \rho_2),$$

where

$$\begin{aligned}
\rho_1 &= E_0(\bar{Q}_0(1, W_1) - \psi_0(1))(\bar{Q}_0(0, W_2) - \psi_0(0)) \\
\rho_2 &= E_0(\bar{Q}_0(0, W_1) - \psi_0(0))(\bar{Q}_0(1, W_2) - \psi_0(1)).
\end{aligned}$$

The true asymptotic variance and the naive asymptotic variance are given by σ^2/n and $(0.5\sigma^2_{naive})/(n/2) = \sigma^2_{naive}/n$, respectively. As a consequence, the relevant comparison is the comparison of σ^2 with σ^2_{naive}, where

$$\sigma^2 = 2\{\sigma_1^2 + \sigma_0^2\} - C$$
$$\sigma^2_{naive} = 2\{\sigma_1^2 + \sigma_0^2\} - 2(\rho_1 + \rho_2).$$

To show that naive variance estimator represents a conservative variance estimator we would need to show that

$$2(\rho_1 + \rho_2) \leq C.$$

Notice that $C = \rho_1 + \rho_2 + C_1$, where

$$C_1 = E_0(\bar{Q}_0 - \bar{Q}^*)(1, W_1)(\bar{Q}_0 - \bar{Q}^*)(1, W_2) + E_0(\bar{Q}_0 - \bar{Q}^*)(0, W_1)(\bar{Q}_0 - \bar{Q}^*)(0, W_2).$$

Thus, the naive variance estimator would be conservative if $\rho_1 + \rho_2 \leq C_1$. Note that we can also represent this as:

$$\begin{aligned} C_1 - \rho_1 - \rho_2 &= Cov(\tilde{Q}_0(1, W_1), (\tilde{Q}_0(1, W_2)) + Cov(\tilde{Q}_0(0, W_1), \tilde{Q}_0(0, W_2)) \\ &\quad -Cov(\tilde{Q}_0(1, W_1), \tilde{Q}_0(0, W_2)) - Cov(\tilde{Q}_0(0, W_1), \tilde{Q}_0(1, W_2)) \\ &= Cov(\tilde{Q}_0(W_1), \tilde{Q}_0(W_2)), \end{aligned}$$

where $Cov(X, Y) = E(XY)$ denotes the standard covariance between two mean zero random variables X and Y, and we introduced the notation $\tilde{Q}_0(W) = (\tilde{Q}_0(1, W) - \tilde{Q}_0(0, W))$ and $\tilde{Q}_0(a, W) = (\bar{Q}_0 - \bar{Q}^*)(a, W)$. Thus, if the latter covariance-term $Cov(\tilde{Q}_0(W_1), \tilde{Q}_0(W_2))$ is non-negative, then the naive variance estimator is conservative. This is a very reasonable condition certainly expected to hold. Thus, we can conclude that in great generality the naive variance estimator is a conservative estimator. We also note that if in truth there is no treatment effect, conditional on covariates, then this covariance term equals zero, so that the naive variance estimator is unbiased.

6.5 A general conservative estimator of the asymptotic variance of TMLE

Above we presented the naive variance estimator of the unadjusted estimator and showed that it is conservative in great generality. In this subsection we propose a generalization of this estimator to obtain a conservative estimator of the asymptotic variance of the general TMLE (using a general initial estimator).

Recall that $C = (\rho_1 + \rho_2) + C_1$, and note that $\bar{\rho} = \rho_1 + \rho_2$ can be consistently estimated with $\bar{\rho}_n = 2/J \sum_{j=1}^{J} (Y_{1j} - \bar{Q}^*_n(A_{1j}, W_{1j}))(Y_{2j} - \bar{Q}^*_n(A_{2j}, W_{2j}))$. Above we showed that we can obtain a conservative bound for C by replacing C_1 by $\bar{\rho}$. Thus, we can conservatively estimate C by $2\bar{\rho}_n$. Thus, a general conservative estimator of the asymptotic variance σ^2 of $\sqrt{n}(\psi^*_n - \psi_0)$ is given by

$$\sigma^2_n = \sigma^2_{I,n} - 2\bar{\rho}_n,$$

where

$$\sigma^2_{I,n} = \frac{1}{n} \sum_{i=1}^{n} D^*(\bar{Q}^*_n, g_0, \psi^*_n)(O_i)^2.$$

This estimator can be viewed as the generalization of the "naive" variance estimator for the unadjusted estimator of ψ_0, analyzed in the previous subsection.

6.6 A simulation confirming the variance formula for the unadjusted estimator

To confirm our conclusions regarding the asymptotic variance of the unadjusted estimator, consider the following simple simulations. For n units, the baseline covariates $W1$ and $W2$ were independently drawn from $N(0, 0.2^2)$ and $U(-1, 1)$, respectively. Then the following adaptive matching algorithm was employed. First units were classified into a matching category M, representing the 16 quartile combinations of $W1$ and $W2$. Within each stratum of M, units were randomly paired. If there were an odd number of units in a given strata, the remaining unit was set aside. The leftovers were then ordered according to M and pairs created. Next the treatment was randomized within the $n/2$ matched pairs. Finally, the binary outcome Y was drawn independently for each unit with probability

$$p = expit[\beta_0 + \beta_1 A + \beta_2 W1 + \beta_3 W1^* A + \beta_4 W2^2] \tag{6.2}$$

where $expit$ is the inverse logistic function and the coefficients were set as $\beta_0 = -1$, $\beta_1 = -0.5$, $\beta_2 = 3$, $\beta_3 = -2$ and $\beta_4 = 2$. The target causal parameter is the average treatment effect. It had a true value of $\psi_0 = -0.11$ in this data generating experiment ("Scenario 1"). The coefficients were then also varied to examine the asymptotic variance of the unadjusted estimator in different data generating experiments. In Scenario 2 there is no treatment effect: $\beta_1 = \beta_3 = 0$. In Scenario 3 the baseline covariates (used for matching) have no effect on the outcome. Specifically, β_2, β_3 and β_4 were set to zero to yield an average treatment effect of -0.08.

For each scenario, the true finite sample variance $Var(\psi_n^*)$ was estimated as the variance of unadjusted estimator over $R = 10,000$ trials, each of sample size $n = 500$ units and the corresponding asymptotic variance estimate $n\text{Var}(\psi_n^*)$ was reported. Table 1 compares this estimate of the true asymptotic variance with the claimed asymptotic variance $\sigma^2 = \sigma_I^2 - C$ calculated according to Theorem 3, and with the asymptotic naive variance treating the pairs as independent σ_{naive}^2. The asymptotic variances σ^2 and σ_{naive}^2, as well as C and $\bar{\rho}$ were computed with Monte Carlo simulation of 50,000 units. All statistical computing was done in `R` version 2.15.1. In addition, recall our claims that $C - 2\bar{\rho} \geq 0$, implies that $\sigma_{naive}^2 = \sigma_I^2 - 2\bar{\rho}$ is conservative.

In all scenarios, the true asymptotic variance of the TMLE and our claimed true asymptotic variance are in agreement. The simulation for scenario 1 also confirms that $\sigma_{naive}^2 = \sigma_I^2 - 2\bar{\rho}$ is indeed conservative, but close to the true asymptotic variance. In Scenario 2 the correction factors C and $2\bar{\rho}$ are equal when there is no treatment effect: $C = 2\bar{\rho}$, and in Scenario 3 we have $C = 2\bar{\rho} = 0$. Indeed, in both of these scenarios we see perfect agreement between σ_{naive}^2 and the true asymptotic variance σ^2.

	Scenario 1	Scenario 2	Scenario 3
$nVar(\psi_n^*)$	0.8408	0.8708	0.6833
σ^2	0.8523	0.8729	0.6915
σ^2_{naive}	0.8591	0.8729	0.6915
C	0.0712	0.1060	0.0000
$2\bar{\rho}$	0.0643	0.1060	-0.0000

Table 1: Comparing the true finite sample variance of the unadjusted estimator scaled by n $nVar(\psi_n^*)$, the asymptotic variance σ^2 according to Theorem 3 and the naive asymptotic variance treating the pairs as independent σ^2_{naive}. Scenario 1 corresponds to the setting $\beta_0 = -1$, $\beta_1 = -0.5$, $\beta_2 = 3$, $\beta_3 = -2$ and $\beta_4 = 2$ in Eq. 6.2. Scenario 2 corresponds setting $\beta = 1$ and β_3 to zero in order to examine the asymptotic variance if the intervention has no effect on the outcome. Scenario 3 corresponds to setting β_2, β_3 and β_4 to zero in order to examine the asymptotic variance if the baseline covariates (used for matching) have no effect on the outcome. For each scenario, the correction factor C and $2\bar{\rho}$ are also given.

6.7 Efficiency gains due to adaptive pair matching

In this section we compare two design choices regarding g_0^n. In the first, we simply assume that $g_0^n(A^n \mid X^n) = \prod_{i=1}^n g_0(A_i \mid W_i)$ for a common g_0. In this case, (W_i, A_i, Y_i), $i = 1, \ldots, n$, are i.i.d. This design includes classic non-matched randomized trials in which treatment is randomly assigned with some known probability, possibly conditional on unit-specific covariates.

We compare this design to a design employing adaptive pair matching. In other words, in the second design we assume $g_0^n \in \mathcal{G}_2^n$ with $g_{0,i}^n = g_0$, so that $g_0^n(A^n \mid X^n) = \prod_{j=1}^{n/2} g_0(A_i : i \in C_j(W^n) \mid W^n)$ and the marginal $P(A_i = a \mid W^n) = g_0(a \mid W_i)$, $i = 1, \ldots, n$.

We compare the asymptotic variance of the TMLEs under these two designs when $\bar{Q}_n^*$ converges to a possibly misspecified $\bar{Q}^*$. This provides insight into the efficiency gains made possible by adaptive pair matching. We assume that g_0 is known, so that $\bar{g}_n = g_0$, as would be the case in both an non-matched and adaptively matched randomized trial.

Theorem 4. *Under the i.i.d. design, the TMLE is asymptotically linear with influence curve $D^*(\bar{Q}^*, g_0, \psi_0)$, so that its asymptotic variance is given by $\sigma_I^2(\bar{Q}^*) = P_0\{D^*(\bar{Q}^*, g_0, \psi_0)\}^2$. This variance can be represented as*

$$\sigma_I^2(\bar{Q}^*) = E_0\{\bar{Q}_0(W) - \psi_0\}^2$$
$$+E_0E_0\left(H_{g_0}^2(A,W)(Y - \bar{Q}^*(A,W))^2 \mid W\right) - E_0\{\bar{Q}_0(W) - \bar{Q}^*(W)\}^2.$$

For the adaptive paired matching design the asymptotic variance $\sigma^2(\bar{Q}^*)$ of the TMLE is given by

the limit of

$$E_0\{\bar{Q}_0(W)-\psi_0)\}^2+E_0\tfrac{1}{n}\sum_{j=1}^{n/2}P_{Q_0,g^n}\left\{\sum_{i\in C_j(W^n)}H_{g_0}(A_i,W_i)(Y_i-\bar{Q}^*(A_i,W_i))\right\}^2 \\ -E_0\tfrac{1}{n}\sum_{j=1}^{n/2}\left\{\sum_{i\in C_j(W^n)}\{\bar{Q}_0(W_i)-\bar{Q}^*(W_i)\}\right\}^2$$

This can be represented as:

$$\sigma^2(\bar{Q}^*)=E_0\{\bar{Q}_0(W)-\psi_0\}^2 \\ +E_0E_0\left(H^2_{g_0}(A,W)(Y-\bar{Q}^*(A,W))^2\mid W\right)-E_0\{\bar{Q}_0(W)-\bar{Q}^*(W)\}^2-C,$$

where $C=(\rho_1+\rho_2)+C_1$ was defined above as sum of four terms, with

$$C_1=E_0\tfrac{1}{J}\sum_{j=1}^{J}(\bar{Q}_0-\bar{Q}^*)(1,W_{1j})(\bar{Q}_0-\bar{Q}^*)(1,W_{2j}) \\ +E_0\tfrac{1}{J}\sum_{j=1}^{J}(\bar{Q}_0-\bar{Q}^*)(0,W_{1j})(\bar{Q}_0-\bar{Q}^*)(0,W_{2j}).$$

The difference between the two asymptotic variances is thus given by:

$$\sigma_I^2(\bar{Q}^*)-\sigma^2(\bar{Q}^*)=C.$$

If $\bar{Q}^*=\bar{Q}_0$, the two asymptotic variances are equal. If $\bar{Q}^*(A,W)=E_0(Y\mid A)$, then the difference is the sum C of the four covariances.

This theorem teaches us that, while the information bound for the two designs is the same, the TMLE under adaptive pair matching at misspecified $\bar{Q}^*$ will outperform the TMLE under i.i.d. sampling, as long as $C>0$. This theorem further suggests that pair matching will result in efficiency gains over the i.i.d. design to the extent that there are baseline covariates W that are predictive of Y which cannot be adjusted for in the outcome regression. Such a scenario might occur in finite samples due to lack of support in the data. For example, in a cluster randomized trial of an HIV prevention intervention, the sample of communities might include only two communities in proximity to a major trucking route, a community characteristic known to predict higher HIV transmission levels. If by chance in the i.i.d. design both of these communities were assigned to the treatment arm of the trial, lack of data support would preclude adjustment for this community-level covariate and thus pair matching on this covariate would result in efficiency gains.

7 Augmenting the data structure with missingness

Consider the following data generating experiment. Firstly, we sample n i.i.d. $(W_1,Y_1(0),Y_1(1))$, ..., $(W_n,Y_n(0),Y_n(1))$, giving us the vector X^n and vector of baseline covariates W^n. Based on W^n, we run a partitioning algorithm generating pairs $C_j(W^n)$, $j=1,\ldots,J$. However, suppose that the designer does not want to accept pairs that are not similar enough with respect to some metric. Therefore, one applies an algorithm that involves assigning an indicator $\Delta_i(W^n)$, $i=1,\ldots,n$ and applying the partitioning algorithm among the units $\{i:\Delta_i(W^n)=1\}$ resulting in $C_j(W^n)$,

$j = 1, \ldots, J$. Thus $\cup_j C_j(W^n) = \{i : \Delta_i(W^n) = 1\}$. We also note that $\Delta_i(W^n)$ is a deterministic function of W^n. Let n_1 be the number of observations with $\Delta_i(W^n) = 1$. Given W^n, the $\Delta_i(W^n)$ and the pairs $C_j(W^n)$, we draw A^{n_1} from a conditional distribution of

$$g_0^n(A^{n_1} \mid X^n) = g_0^n(A^{n_1} \mid W^n) = \prod_{j=1}^{J} g_0^n((A_i : i \in C_j(W^n)) \mid W^n).$$

We now collect the data $O_i = (W_i, \Delta_i(W^n), \Delta_i(W^n)A_i, \Delta_i(W^n)Y_i(A_i))$, $i = 1, \ldots, n$, giving the observed data $O^n = (O_1, \ldots, O_n)$.

The target quantity of interest remains the average treatment effect $\Psi^F(P_{X,0}) = E_0 Y(1) - E_0 Y(0)$. We have $\psi_0^F = E_W\{\bar{Q}_0(1, W) - \bar{Q}_0(0, W)\}$, where $\bar{Q}_0(a, w) = E(Y(a) \mid W = w) = E_0(Y \mid A = a, W = W)$. We note that Y_i, given W^n, A^{n_1}, is independent across $i = 1, \ldots, n$, and this conditional distribution equals the conditional distribution of Y_i, given W_i, A_i. Therefore,

$$\begin{aligned}
&E(Y_i \mid A_i, W_i, \Delta_i(W^n) = 1) = E(E(Y_i \mid A_i, W_i, W^n) \mid A_i, W_i, \Delta_i(W^n) = 1) \\
&= E(E(Y_i \mid A_i, W_i) \mid A_i, W_i, \Delta_i(W^n) = 1) \\
&= E(Y_i \mid A_i, W_i).
\end{aligned}$$

This proves that $\bar{Q}_0(a, w) = E(Y_i \mid A_i = a, W_i = w, \Delta_i(W^n) = 1)$ and is thus identifiable from the distribution of O^n. This proves the desired identifiability of ψ_0^F:

$$\Psi^F(P_{X,0}) \quad = \quad E_{W,0}\{\bar{Q}_0(1, W) - \bar{Q}_0(0, W)\} = \Psi(P_0^n).$$

The average is with respect to the marginal distribution of W (not conditional on $\Delta_i(W^n) = 1$), so that also the observations with $\Delta_i(W^n) = 0$ are used to identify this target quantity.

This also demonstrates that $E_0 \sum_{i=1}^n I(\Delta_i(W^n) = 1)(Y_i - \bar{Q}(A_i, W_i))^2$ is minimized over $\bar{Q}$ by $\bar{Q}_0$, and thus represents a valid loss function for loss-based learning of $\bar{Q}_0$ based on O^n. Similarly, we can use a log-likelihood loss $\sum_{i=1}^n I(\Delta_i(W^n) = 1)L(\bar{Q})(W_i, A_i, Y_i)$, where $-L(\bar{Q})(W, A, Y) = Y \log \bar{Q}(A, W) + (1 - Y) \log(1 - \bar{Q}(A, W))$.

In order to present a TMLE we first need to derive the canonical gradient, which is presented in the following theorem.

Theorem 5. *Consider the data generating experiment described above.*
Let $O_i = (W_i, \Delta_i(W^n), \Delta_i(W^n)A_i, \Delta_i(W^n)Y_i)$, the observed data is $O^n = (O_1, \ldots, O_n) \sim P^n$ with

$$P^n(O^n) = \prod_{i=1}^{n} Q_W(W_i)\{Q_Y(Y_i \mid W_i, A_i)\}^{\Delta_i(W^n)} g^n((A_i : \Delta_i(W^n) = 1) \mid W^n),$$

where Q_W is an unspecified marginal distribution, Q_Y is an unspecified conditional distribution of Y, given A, W, and g^n is a conditional distribution of $A^{n_1} = (A_i : \Delta_i(W^n) = 1)$, given $W^n = (W_1, \ldots, W_n)$, known to be an element of a set $\mathcal{G}^n$ consisting of distributions satisfying (2.1). Let $\mathcal{M}^n$ be the resulting statistical model for P^n. Let $\mathcal{M}^n(g^n)$ be the model if g^n is known.

Let $\Psi : \mathcal{M}^n \to \mathbb{R}$ *be defined by* $\Psi(P^n) = E_{Q_W}\{\bar{Q}(1,W) - \bar{Q}(0,W)\}$*, where* $\bar{Q}(A,W) = E_{Q_Y}(Y \mid A,W)$.

The tangent space at P^n *in model* $\mathcal{M}^n$ *is given by:*

$$T(P^n) = \left\{\sum_{i=1}^n \phi(W_i) : \phi \in T_W\right\} + \left\{\sum_{i=1}^n \Delta_i(W^n)\phi(Y_i \mid A_i, W_i) : \phi \in T_Y\right\} + \sum_{j=1}^J T_{C_j}, \quad (7.1)$$

where $T_W = \{h(W) : Eh(W) = 0\}$,

$$T_Y = \{h(Y \mid A,W) : E_{Q_Y}(h(Y \mid A,W) \mid A,W) = 0\}, \text{ and}$$
$$T_{C_j} = \{S((A_i : i \in C_j(W^n)) \mid W^n) : E(S \mid W^n) = 0\}.$$

The tangent space at P^n *in model* $\mathcal{M}^n(g^n)$ *is given by*

$$T(Q) = \left\{\sum_{i=1}^n \phi(W_i) : \phi \in T_W\right\} + \left\{\sum_{i=1}^n \phi(Y_i \mid A_i, W_i) : \phi \in T_Y\right\}.$$

Let

$$D^*(\bar{Q}, g, \psi)(W, \Delta, \Delta A, \Delta Y) = D^*_W(\bar{Q}, \psi)(W) + \frac{\Delta(2A-1)}{g(A, 1 \mid W)}(Y - \bar{Q}(A,W)),$$

where g *denotes a distribution of* $g(a, 1 \mid W) = P(A = a, \Delta = 1 \mid W)$*. The statistical parameter* Ψ *is pathwise differentiable and its canonical gradient at* P^n *is given by*

$$D^{n,*}(P^n) = \frac{1}{n}\sum_{i=1}^n D^*(\bar{Q}, \bar{g}_n, \Psi(Q))(O_i),$$

where $g_i(a, 1 \mid W_i) = \Pi_i(1 \mid W_i)g_i(a \mid W_i)$ *is the conditional probability that* $A_i = a$*,* $\Delta_i(W^n) = 1$*, given* W_i*, which can be factored into* $\Pi_i(1 \mid W_i) = P(\Delta_i(W^n) = 1 \mid W_i)$ *and* $g_i(a \mid W_i) = P(A_i = a \mid W_i, \Delta_i(W^n) = 1)$*, and*

$$\bar{g}_n(a, 1 \mid W) = \frac{1}{n}\sum_{i=1}^n g_i(a, 1 \mid W).$$

We note that

$$g_i(a, 1 \mid W_i) = \sum_{(w_j : j \neq i)} \Delta_i((w_j : j \neq i), W_i) g_i(a \mid (w_j : j \neq i), W_i) \prod_{j \neq i} Q_W(w_j) \quad (7.2)$$

is a function of $g_i(A_i \mid W^n)$ *and the common marginal distribution* Q_W*. We have*

$$E_0 D^{n,*}(\bar{Q}, \bar{g}_n, , \psi_0) = 0 \text{ if } \bar{Q} = \bar{Q}_0 \text{ or } \bar{g}_n = \bar{g}_{n,0}, \quad (7.3)$$

assuming that for all i*,* $0 < g_i(1, 1 \mid W_i) < 1$ *a.e.*

The TMLE of $\bar{Q}_0$ is analogue to the TMLE presented in Section 4, with the modification that the clever covariate is now given by $(2A_i - 1)I(\Delta_i(W^n) = 1)/\bar{g}_n(A_i, 1 \mid W_i)$, only the complete observations are used for fitting $\bar{Q}_0$, but the empirical distribution over all $W_1, \ldots, W_n$ is plugged in the target parameter mapping. The same asymptotics can be applied and the formulas for the asymptotic variance are the same as presented earlier, with the only modification that $g_i(a \mid W_i)$ is now replaced by $g_i(a, 1 \mid W_i)$.

8 Summary

This article has investigated efficient estimation and inference for the additive causal effect $E_0\{Y(1)-Y(0)\}$ of treatment on the outcome under a class of designs based on sampling n i.i.d. $(W_i, Y_i(0), Y_i(1)) \sim P_{X,0}$, sampling A^n, given W^n, and collecting (W_i, A_i, Y_i), $i = 1, \ldots, n$. We considered a general class of dependent treatment assignment mechanisms g^n satisfying the assumption that $(A_i : i \in C_j(W^n))$, $j = 1, \ldots, J$, are independent across j, conditionally on W^n, where $C_j(W^n)$, $j = 1, \ldots, J$, is a partitioning of the sample $\{1, \ldots, n\}$ into groups implied by W^n. The number of partitions J was assumed to be proportional to n.

We computed the efficient influence curve of the target parameter for the statistical model implied by this design without making additional assumptions about the common full-data distribution $P_{X,0}$. We defined a corresponding TMLE that is consistent and asymptotically normally distributed under correct specification of g_0^n, and is also efficient if the outcome regression $\bar{Q}_0$ is consistently estimated. This TMLE can be implemented by ignoring the dependency created by the treatment allocation process, with the exception that if cross validation is used to estimate the average $\bar{g}_n$ of $g_{0,i}(A_i \mid W_i)$ across $i = 1, \ldots, n$, the group rather than the unit should be used when partitioning the data into training and validation sets. Thus, construction of training and validation sets for data adaptive estimation of $\bar{Q}_0$ can be based on the sampling unit. We further suggested an alternative plug-in approach to estimating the unit specific treatment mechanism $g_{0,i}$ that makes use of design based knowledge of g_0^n, thus potentially improving estimator robustness and efficiency.

Due to the dependency introduced by the treatment allocation process, no asymptotically consistent bootstrap method appears to be available for the general class of dependent g^n-designs presented in this paper. Further, when groups are size 2 or larger, the asymptotic variance of the TMLE under the dependent sampling relies on a consistent estimator of $\bar{Q}_0$ even when g_0^n is known. In contrast, the asymptotic variance of the TMLE under i.i.d. sampling is fully robust to misspecification of $\bar{Q}_n^*$ in randomized controlled trials.

We further considered adaptively pair matched trials as an important special case of the general dependent treatment allocation design. We formally compared the asymptotic variance of the TMLE under this design with that of the TMLE under i.i.d. sampling. While the information bound for the adaptively pair matched design with $g_i^n = g_i = g_0$ equals the information bound for i.i.d. sampling of (W_i, A_i, Y_i) with $P(A = a \mid W) = g_0(a \mid W)$, we showed that the TMLE under adaptive pair matching and misspecified $\bar{Q}^*$ will outperform the TMLE under i.i.d. sampling as long as the $(\bar{Q}_0 - \bar{Q}^*)(1, \cdot)$ and $(\bar{Q}_0 - \bar{Q}^*)(0, \cdot)$ of the baseline covariates within the groups $C_j(W^n)$ are positively correlated. We also showed that under the paired matching design and the positive correlation condition, an estimate of the variance that treats the n observations as i.i.d. is conservative if $\bar{Q}_n^*$ is inconsistent for $\bar{Q}_0$ and is asymptotically consistent if $\bar{Q}_n^*$ is consistent. We also presented a less conservative variance estimator that relies on an additional reasonable assumption (similar to the above positive correlation assumption). We demonstrated that the estimator of the variance for the unadjusted estimator as currently used by practitioners in the analysis of paired matched trials is valid as well, and our above mentioned less conservative variance estimator is just a generalization of this estimator.

Taken together, these finding teach us that the use of an adaptively pair matched design will

generally result in a more efficient estimator of the treatment effect, while one can still obtain robust conservative variance estimators. However, understanding the complications resulting from the adaptive pair matching requires advanced empirical process theory, and even makes the analysis of the unadjusted estimator a serious challenge.

References

Andersen, J., D. Faries, and R. Tamura (1994). A randomized play-the-winner design for multiarm clinical trials. *Communication in Statistical Theory 23*, 309–323.

Bai, Z., F. Hu, and W. Rosenberger (2002). Asymptotic properties of adaptive designs for clinical trials with delayed response. *Annals of Statistics 30*(1), 122–139.

Balzer, L., M. Petersen, and M. van der Laan (2012). Why match in individually and cluster randomized trials. Technical Report 294, Division of Biostatistics, University of California, Berkeley.

Bickel, P., C. Klaassen, Y. Ritov, and J. Wellner (1997). *Efficient and Adaptive Estimation for Semiparametric Models*. Springer-Verlag.

Campbell, M., A. Donner, and N. Klar (2007). Developments in cluster randomized trials and statistics in medicine. *Statistics in Medicine 26*, 2–19, doi: 10.1002/sim.2731.

Chambaz, A. and M. van der Laan (2010). Targeting the optimal design in randomized clinical trials with binary outcomes and no covariate. Technical Report 258, Division of Biostatistics, University of California, Berkeley.

Cheng, Y. and Y. Shen (2005). Bayesian adaptive designs for clinical trials. *Biometrika 92*(3), 633–646.

Donner, A. and N. Klar (2000). *Design and Analysis of Cluster Randomization Trials in Health Research*. London: Arnold.

Flournoy, N. and W. Rosenberger (1995). *Adaptive Designs*. Hayward, Institute of Mathematical Statistics.

Freedman, L., M. Gail, S. Green, and D. Corle (1997). The efficiency of the matched-pairs design of the community intervention trial for smoking cessation (commit). *Controlled clinical trials 18(2)*, 131–139, doi:10.1016/S0197–2456(96)00115–8.

Freedman, L., S. Green, and D. Byar (1990). Assessing the gain in efficiency due to matching in a community intervention study. *Statistics in Medicine 9*, 943–952.

Gail, M., D. Byar, T. Pechacek, and D. Corle (1992). Aspects of statistical design for the community intervention trial for smoking cessation. *Controlled clinical trials 13*, 6–21.

Gill, R., M. van der Laan, and J. Robins (1997). Coarsening at random: characterizations, conjectures and counter-examples. In D. Lin and T. Fleming (Eds.), *Proceedings of the First Seattle Symposium in Biostatistics*, New York, pp. 255–94. Springer Verlag.

Gill, R., M. van der Laan, and J. Wellner (1995). Inefficient estimators of the bivariate survival function for three models. *Annales de l'Institut Henri Poincaré 31*, 545–597.

COMMIT Research Group. (1991). Summary of design and intervention. *Journal of National Cancer Institute 83(22)*, 1620–1628.

Hayes, R. and L. Moulton (2009). *Cluster Randomized Trials*. Boca Raton: Chapman & Hall/CRC.

Heitjan, D. and D. Rubin (1991, December). Ignorability and coarse data. *Annals of statistics 19*(4), 2244–2253.

Hu, F. and W. Rosenberger (2000). Analysis of time trends in adaptive designs withi appliation to a neurophysiology experiment. *Statistics in Medicine 19*, 2067–2075.

Imai, K. (2008). Variance identification and efficiency analysis in randomized experiments under the matched-pair design. *Statistics in Medicine 27(24)*, 4857–4873.

Imai, K., G. King, and C. Nall (2009). The essential role of pair matching in cluster randomized experiments with applicationto the mexican universal health insurance evaluation. *Statistical Science 24 (1)*, 29–53.

Jacobsen, M. and N. Keiding (1995). Coarsening at random in general sample spaces and random censoring in continuous time. *Annals of Statistics 23*, 774–86.

Klar, N. and A. Donner (1997). The merits of matching in community intervention trials: a cautionary tale. *Statistics in Medicine 16(15)*, 1753–1764.

Koethe, J., A. Westfall, D. Luhanga, G. Clark, J. Goldman, P. Mulenga, R. Cantrell, B. Chi, I. Zulu, M. Saag, and J. S. Stringer (2010). A cluster randomized trial of routine hiv-1 viral load monitoring in zambia. *PlLoS One 5*, 5(3):e9680.

Murray, D. (1998). *Design and Analysis of Community Randomized Trials*. Oxford: Oxford University Press.

Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika 82*, 669–710.

Pearl, J. (2009). *Causality: models, reasoning, and inference* (2nd ed.). New York: Cambridge.

Polley, E. and M. van der Laan (2010). Super learner in prediction. Technical Report 266, Division of Biostatistics, University of California, Berkeley.

Raudenbush, S., A. Martinez, and J. Spybrook (2007). Developments in cluster randomized trials and statistics in medicine. *Educational Evaluation and Policy Analysis 29*, 5–29.

Rose, S. and M. van der Laan (2009). Why match? investigating matched case-control study designs with causal effect estimation. *The International Journal of Biostatistics*, http://www.bepress.com/ijb/vol5/iss1/1/.

Rosenberger, W. (1996). New directions in adaptive designs. *Statistical Science 11*, 137–149.

Rosenberger, W., N. Flournoy, and S. Durham (1997). Asymptotic normality of maximum likelihood estimators from multiparameter response driven designs. *Journal of Statistical Planning and Inference*, 69–76.

Rosenberger, W. and S. Grill (1997). A sequential design for psychophysical experiments: An application to estimating timing of sensory events. *Statistics in Medicine 16*, 2245–2260.

Rosenberger, W. and T. Shiram (1997). Estimation for an adaptive allocation design. *Journal of Statistical Planning and Inference 59*, 309–319.

Rosenblum, M. and M. van der Laan (2010). Targeted maximum likelihood estimation of the parameter of a marginal structural model. *Int J Biostat 6*(2), Article 19.

Tamura, R., D. Faries, J. Andersen, and J. Heiligenstein (1994). A case study of an adaptive clinical trial in the treatment of out-patients with depressive disorder. *Journal of the American Statistical Association 89*, 768–776.

Toftager, M., L. Christiansen, P. Kristensen, and J. Troelsen (2011). Space for physical activity-a multicomponent intervention study: study design and baseline findings from a cluster randomized controlled trial. *BMC Public Health 10*, 711–777.

van der Laan, M. (1996). *Efficient and Inefficient Estimation in Semiparametric Models* (CWI tract 114 ed.). Amsterdam: Centre of Computer Science and Mathematics.

van der Laan, M. (2008). The construction and analysis of adaptive group sequential designs. Technical Report 232, Division of Biostatistics, University of California, Berkeley. http://www.bepress.com/ucbbiostat/paper234.

van der Laan, M. and S. Dudoit (2003). 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 130, Division of Biostatistics, University of California, Berkeley.

van der Laan, M., S. Dudoit, and A. van der Vaart (2006). The cross-validated adaptive epsilon-net estimator. *Statistics and Decisions 24*(3), 373–395.

van der Laan, M., E. Polley, and A. Hubbard (2007). Super learner. *Statistical Applications in Genetics and Molecular Biology 6*(25), 1–21.

van der Laan, M. and J. Robins (2003). *Unified methods for censored longitudinal data and causality*. Berlin Heidelberg New York: Springer.

van der Laan, M. and S. Rose (2012). *Targeted Learning: Causal Inference for Observational and Experimental Data*. New York: Springer.

van der Laan, M. and D. Rubin (2006). Targeted maximum likelihood learning. *The International Journal of Biostatistics 2*(1).

van der Vaart, A. (1998). *Asymptotic Statistics*. Cambridge University Press.

van der Vaart, A., S. Dudoit, and M. van der Laan (2006). Oracle inequalities for multi-fold cross-validation. *Statistics and Decisions 24*(3), 351–371.

van der Vaart, A. and J. Wellner (1996). *Weak Convergence and Empirical Processes*. Springer-Verlag, New York.

Watson, L., R. Small, S. Brown, W. Dawson, and J. Lumley (2004). Mounting a community-randomized trial: sample size, matching, selection, and randomization issues in prism. *Journal of Controlled Clinical Trials 3*, 235–250.

Wei, L. (1979). The generalized polya's urn design for sequential medical trials. *Annals of Statistics 7*, 291–296.

Wei, L. and S. Durham (1978). The randomize play-the-winner rule in medical trials. *Journal of the American Statistical Association 73*, 840–843.

Wei, L., R. Smythe, D. Lin, and T. Park (1990). Statistical inference with data-dependent treatment allocation rules. *Journal of the American Statistical Association 85*, 156–162.

Zelen, M. (1969). Play the winner rule and the controlled clinical trial. *Journal of the American Statistical Association 64*, 131–146.

A Appendix: Proof of Theorem 1

Firstly, we note that

$$\begin{aligned}
&E_0 D^*(\bar{Q}, \bar{g}_n, \psi_0)(O^n) = E_0 \tfrac{1}{n} \textstyle\sum_{i=1}^n \{\bar{Q}(W_i) - \psi_0\} + \\
&E_0 \tfrac{1}{n} \textstyle\sum_{i=1}^n \frac{g_{i,0}(1|W_i)}{\bar{g}_n(1|W_i)} (\bar{Q}_0 - \bar{Q})(1, W_i) - \frac{g_{i,0}(0|W_i)}{\bar{g}_n(0|W_i)} (\bar{Q}_0 - \bar{Q})(0, W_i) \\
&= \Psi(Q) - \psi_0 + \tfrac{1}{n} \textstyle\sum_i \int_w Q_{W,0}(w) \frac{g_{i,0}(1|w)}{\bar{g}_n(1|w)} (\bar{Q}_0 - \bar{Q})(1, w) \\
&\qquad\qquad - \tfrac{1}{n} \textstyle\sum_i \int_w Q_{W,0}(w) \frac{g_{i,0}(0|w)}{\bar{g}_n(0|w)} (\bar{Q}_0 - \bar{Q})(0, w) \\
&= \Psi(Q) - \psi_0 + E_0 \frac{\bar{g}_{0,n}(1|W)}{\bar{g}_n(1|W)} (\bar{Q}_0 - \bar{Q})(1, W) \\
&\qquad\qquad - E_0 \frac{\bar{g}_{n,0}(0|W)}{\bar{g}_n(0|W)} (\bar{Q}_0 - \bar{Q})(0, W).
\end{aligned}$$

Thus, if $\bar{g}_{n,0} = \bar{g}_0$, then this equals $\Psi(Q) - \psi_0 + \psi_0 - \Psi(Q) = 0$. If $Q_0 = Q$, then we also obtain 0. This proves (3.3). We also note that $D^{n,*}(Q_0, \bar{g}_0)$ is an element of the tangent space T_Q.

In addition, for each Q, $D^{n,*}(Q, \bar{g}_0, \psi_0)$ is a gradient in the model $\mathcal{M}(g_0^n)$ with g_0^n known, which shows that $D^{n,*}(Q_0, \bar{g}_0)$ is the canonical gradient of $\Psi : \mathcal{M}^n(g^n) \to \mathbb{R}$ at P_0^n. By factorization of the likelihood, it is also the canonical gradient for any model $\mathcal{M}^n$ that instead assumes that $g_0^n \in \mathcal{G}^n$ for a model $\mathcal{G}^n$. □

B Appendix: Proof of Theorem 2

Recall the notation $Pf = E_P f$. We have

$$P_0^n D^{n,*}(\bar{Q}_n^*, \bar{g}_0, \psi_n^*) \equiv \frac{1}{n}\sum_{i=1}^n P_{Q_0,g_{0,i}} D^*(\bar{Q}_n^*, \bar{g}_0, \psi_n^*) = \psi_0 - \psi_n^*.$$

Here we remind the reader that $\bar{g}_0 = 1/n \sum_i g_{0,i}$ and $g_{0,i}(a \mid w) = P_0(A_i = a \mid W_i = w)$. We also have $D^{n,*}(\bar{Q}_n^*, \bar{g}_n, \psi_n^*) = 0$.

Thus,

$$\begin{aligned}
&\psi_n^* - \psi_0 = \tfrac{1}{n}\textstyle\sum_i \{D^*(\bar{Q}_n^*, \bar{g}_n, \psi_n^*)(O_i) - P_{Q_0,g_{0,i}} D^*(\bar{Q}_n^*, \bar{g}_n, \psi_n^*)\} \\
&+\tfrac{1}{n}\textstyle\sum_i P_{Q_0,g_{0,i}} \{D^*(\bar{Q}_n^*, \bar{g}_n, \psi_n^*) - P_{Q_0,g_{0,i}} D^*(\bar{Q}_n^*, \bar{g}_0, \psi_n^*)\} \\
&\equiv \tfrac{1}{n}\textstyle\sum_i \{D^*(\bar{Q}_n^*, \bar{g}_n, \psi_n^*)(O_i) - P_{0,g_{0,i}} D^*(\bar{Q}_n^*, \bar{g}_n, \psi_n^*)\} + \tfrac{1}{\sqrt{n}} Z_{W,\bar{g}_n,n}.
\end{aligned}$$

We note that, using some straightforward algebra,

$$\begin{aligned}
&Z_{W,\bar{g}_n,n} = \sqrt{n}\textstyle\int_w \frac{\bar{g}_0 - \bar{g}_n}{\bar{g}_n}(1 \mid w)(\bar{Q}_0 - \bar{Q})(1, w) dQ_{W,0}(w) \\
&-\sqrt{n}\textstyle\int_w \frac{\bar{g}_0 - \bar{g}_n}{\bar{g}_n}(0 \mid w)(\bar{Q}_0 - \bar{Q})(0, w) dQ_{W,0}(w) \\
&= \sqrt{n}\textstyle\int_w \frac{\bar{g}_0 - \bar{g}_n}{\bar{g}_0}(1 \mid w)(\bar{Q}_0 - \bar{Q})(1, w) dQ_{W,0}(w) \\
&-\sqrt{n}\textstyle\int_w \frac{\bar{g}_0 - \bar{g}_n}{\bar{g}_0}(0 \mid w)(\bar{Q}_0 - \bar{Q})(0, w) dQ_{W,0}(w) + R(\bar{g}_n, \bar{g}_0)
\end{aligned}$$

where

$$\begin{aligned}
&R(\bar{g}_n, \bar{g}_0) = \sqrt{n}\textstyle\int \frac{(\bar{g}_0 - \bar{g}_n)^2}{\bar{g}_n \bar{g}_0}(1 \mid w)(\bar{Q}_0 - \bar{Q})(1, w) dQ_0(w) \\
&-\sqrt{n}\textstyle\int \frac{(\bar{g}_0 - \bar{g}_n)^2}{\bar{g}_n \bar{g}_0}(1 \mid w)(\bar{Q}_0 - \bar{Q})(1, w) dQ_0(w).
\end{aligned}$$

We assume that the latter is $o_P(1)$. Thus to establish the asymptotic linearity of $Z_{W,\bar{g}_n,n}$ we need to study terms of form $\sqrt{n}\int f(w)(\bar{g}_n - \bar{g}_0)(1|w) dQ(w)$. We now note that

$$\begin{aligned}
&(\bar{g}_n - \bar{g}(1 \mid w) = \tfrac{1}{n}\textstyle\sum_{i=1}^n (g_{i,n} - g_i)(1 \mid w) \\
&\tfrac{1}{n}\textstyle\sum_{i=1}^n \int g_i(1 \mid (W_j : j \neq i), W_i = w) \left(\prod_{j \neq i} Q_{W,n}(w_j) - \prod_{j \neq i} Q_W(w_j) \right) \\
&= \tfrac{1}{n}\textstyle\sum_{i=1}^n \int g_i(1 \mid W_{-i}, W_i = w) \sum_{l=1, l \neq i}^n (Q_{W,n}(w_l) - Q_W(w_l)) \\
&\qquad\qquad \textstyle\prod_{m=1, m \neq i}^{l-1} Q_{W,n}(w_m) \prod_{m=l+1, m \neq i}^n Q_W(w_m) \\
&\approx \tfrac{1}{n}\textstyle\sum_{i=1}^n \sum_{l \neq i} \int g_i(1 \mid W_i = w, W_l = w_l)(Q_{W,n}(w_l) - Q_W(w_l)) \\
&= \tfrac{1}{n}\textstyle\sum_{k=1}^n \tfrac{1}{n}\sum_{i=1}^n \sum_{l \neq i}^n \{g_i(1 \mid W_i = w, W_l = W_k) - g_i(1 \mid W_i = w)\},
\end{aligned}$$

where we suppressed the second order term a formal analysis would have to take into account. Therefore, we can write

$$\begin{aligned}
&\sqrt{n}\int(\bar{g}_n - \bar{g}_0)(1 \mid w)f(w)dQ(w) \\
&= \tfrac{1}{\sqrt{n}}\sum_{k=1}^n \tfrac{1}{n}\sum_{i=1}^n\sum_{l\neq i}^n \int_w \{g_i(1 \mid W_i = w, W_l = W_k) - E_{0,W_k} g_i(1 \mid W_i = w, W_l = W_k)\} f(w)dQ(w) \\
&\equiv \tfrac{1}{\sqrt{n}}\sum_{k=1}^n \{\Phi(W_k) - E_0\Phi(W_k)\},
\end{aligned}$$

where we defined

$$\Phi(W_k) = \frac{1}{n}\sum_{i=1}^n\sum_{l\neq i}^n \int_w g_i(1 \mid W_i = w, W_l = W_k)f(w)dQ(w).$$

Thus such integrals are standardized sums of independent random variables $\Phi_l(W_k) - E_0\Phi_l(W_k)$ with mean zero. Such terms will converge to a normal distribution if the variance of $\Phi(W_k)$ is bounded (uniformly in n, since Φ is really indexed by n as well). This demonstrates that one will need that the $\sum_{l\neq i}$ should essentially only contribute a finite number of terms.

To conclude, under regularity conditions, we might have

$$Z_{W,\bar{g}_n,n} \approx \tfrac{1}{\sqrt{n}}\sum_{k=1}^n IC(W_k) - E_0 IC(W_k),$$

where

$$\begin{aligned}
IC(W_k) &= \int_w \left(\tfrac{1}{n}\sum_{i=1}^n\sum_{l\neq i}^n g_i(1 \mid W_i = w, W_l = W_k)\right)\tfrac{\bar{Q}_0 - \bar{Q}}{\bar{g}_0}(1, w)dQ_0(w) \\
&- \int_w \left(\tfrac{1}{n}\sum_{i=1}^n\sum_{l\neq i}^n g_i(0 \mid W_i = w, W_l = W_k)\right)\tfrac{\bar{Q}_0 - \bar{Q}}{\bar{g}_0}(0, w)dQ_0(w).
\end{aligned}$$

A crucial assumption we made in the theorem is that the variance of $IC(W_k)$ is finite. We will now show that under a reasonable typical assumption we will, in fact, have that $IC(W_k) - E_0 IC(W_k) = 0$. For $i \in C_j(W^n)$, in a typical design one will have that $g_i(a \mid W_i = w_i, W_{-i})$ only depends on $W_i = w_i$. Thus, in that case, for $i \in C_j(W^n)$ we have $g_i(a \mid W^n) = g_i(a \mid W_i)$ for some conditional density $g_i(a \mid w)$. This provides us with the following representation:

$$g_i(a \mid W^n) = \sum_{j=1}^J I(i \in C_j(W^n))g_i(a \mid W_i).$$

This yields the following derivation of $g_i(a \mid W_i, W_l)$:

$$\begin{aligned}
&g_i(a \mid W_i, W_l) = \textstyle\int g_i(a \mid W_i, W_l, W(-i,-l))P(W(-i,-l)) \\
&= \textstyle\int \sum_{j=1}^J I(i \in C_j(W_i, W_l, W(-i,-l))g_i(a \mid W_i)P(W(-i,-l)) \\
&= \textstyle\sum_{j=1}^J g_i(a \mid W_i) \int I(i \in C_j(W_i, W_l, W(-i,-l)))P(W(-i,-l)) \\
&= g_i(a \mid W_i)\textstyle\sum_{j=1}^J P(i \in C_j(W^n) \mid W_i, W_l) \\
&= g_i(a \mid W_i).
\end{aligned}$$

Thus, in this case, we have $IC(W_k)$ is constant in W_k so that $IC(W_k) - E_0 IC(W_k) = 0$.

We now proceed as follows:

$$\begin{aligned}
\psi_n^* - \psi_0 &= \frac{1}{n}\sum_{i=1}^n \{D^*(Q_n^*, \bar{g}_n, \psi_n^*)(O_i) - P_{Q_0,g_i} D^*(Q_n^*, \bar{g}_n, \psi_n^*)\} + \frac{1}{\sqrt{n}} Z_{W,\bar{g}_n,n} \\
&= \frac{1}{n}\sum_{i=1}^n \{D^*(\bar{Q}_n^*, \bar{g}_n, \psi_n^*)(O_i) - P_{Q_0,g_i^n} D^*(\bar{Q}_n^*, \bar{g}_n, \psi_n^*)\} \\
&\quad + \frac{1}{n}\sum_{i=1}^n \{(P_{Q_0,g_i^n} - P_{Q_0,g_i}) D^*(\bar{Q}_n^*, \bar{g}_n, \psi_n^*)\} + \frac{1}{\sqrt{n}} Z_{W,\bar{g}_n,n} \\
&= \frac{1}{n}\sum_{i=1}^n \{D_Y^*(\bar{Q}_n^*, \bar{g}_n)(O_i) - P_{Q_0,g_i^n} D_Y^*(\bar{Q}_n^*, \bar{g}_n)\} \\
&\quad + \frac{1}{n}\sum_{i=1}^n \{(P_{Q_0,g_i^n} - P_{Q_0,g_i}) D^*(\bar{Q}_n^*, \bar{g}_n, \psi_n^*)\} + \frac{1}{\sqrt{n}} Z_{W,\bar{g}_n,n} \\
&\equiv \frac{1}{\sqrt{n}} X_n(\bar{Q}_n^*) + \frac{1}{\sqrt{n}} Z_{W,n,g^n} + \frac{1}{\sqrt{n}} Z_{W,\bar{g}_n,n}.
\end{aligned}$$

Here we used at the third equality that $P_{Q_0,g_{0,i}^n}$ is a conditional expectation, given W^n, so that the empirical process of D_W^* cancels out in the first term. We defined the process only as a function of $\bar{Q}_n^*$, not as a function of $\bar{g}_n$, because $\bar{g}_n$ is only a function of W^n. Note, that

$$\begin{aligned}
&D_Y^*(\bar{Q}, \bar{g}_n)(O_i) - P_{Q_0,g_{0,i}^n} D_Y^*(\bar{Q}, \bar{g}_n) = \tfrac{2A_i - 1}{\bar{g}_n(A_i|W_i)}(Y_i - \bar{Q}(A_i, W_i)) \\
&- \left\{ \tfrac{g_{0,i}^n(1|W^n)}{\bar{g}_n(1|W_i)}(\bar{Q}_0 - \bar{Q})(1, W_i) - \tfrac{g_{0,i}^n(0|W^n)}{\bar{g}_n(0|W_i)}(\bar{Q}_0 - \bar{Q})(0, W_i) \right\} \\
&\equiv f_{i,n}^1(\bar{Q})(O_i).
\end{aligned}$$

Note that $f_{i,n}^1(\bar{Q})$ is a random function of O_i through W^n, while, given W^n, it is a fixed function of O_i. In the special case that $g_{0,i}^n = g_{0,i}$ is constant in i, we have $f_{i,n}^1(\bar{Q})(O_i) = D_Y^*(\bar{Q}, g_{0,i})(O_i) - \{\bar{Q}_0 - \bar{Q}\}(W_i)$. We can represent $X_n(\bar{Q})$ as $X_n(\bar{Q}) = 1/\sqrt{n}\sum_{i=1}^n f_{i,n}^1(\bar{Q})(O_i)$, where $P_{Q_0,g_{0,i}^n} f_{i,n}(\bar{Q}) = 0$.

Let's now determine the form of Z_{W,n,g^n}. We have

$$\begin{aligned}
&1/n\sum_i P_{Q_0,g_{0,i}^n} D_Y^*(\bar{Q}_n^*, \bar{g}_n) \\
&= 1/n\sum_i \tfrac{g_{0,i}(1|W^n)}{\bar{g}_n(1|W_i)}(\bar{Q}_0 - \bar{Q}_n^*)(1, W_i) - \tfrac{g_{0,i}^n(0|W^n)}{\bar{g}_n(0|W_i)}(\bar{Q}_0 - \bar{Q}_n^*)(0, W_i) \\
&= 1/n\sum_i \left(\tfrac{g_{0,i}(1|W^n)}{\bar{g}_n(1|W_i)} - 1\right)(\bar{Q}_0 - \bar{Q}_n^*)(1, W_i) - \left(\tfrac{g_{0,i}^n(0|W^n)}{\bar{g}_n(0|W_i)} - 1\right)(\bar{Q}_0 - \bar{Q}_n^*)(0, W_i) \\
&+ 1/n\sum_i \bar{Q}_0(W_i) - \bar{Q}_n^*(W_i) \\
&1/n\sum_i P_{Q_0,g_{0,i}} D_Y^*(\bar{Q}_n^*, \bar{g}_n) = \textstyle\int_w (\bar{Q}_0 - \bar{Q}_n^*)(w) Q_{W,0}(w) \\
&1/n\sum_i P_{Q_0,g_i^n} - P_{Q_0,g_i} D_W^*(\bar{Q}_n^*, \psi_n^*) = 1/n\sum_i \bar{Q}_n^*(W_i) - P_0 \bar{Q}_n^*
\end{aligned}$$

Thus,

$$
\begin{aligned}
&1/n\sum_i (P_{Q_0,g^n_{0,i}} - P_{Q_0,g_{0,i}})(D^*_Y + D^*_W)(\bar{Q}^*_n, \bar{g}^*_n, \psi^*_n) \\
&= 1/n\sum_i \left(\frac{g_{0,i}(1|W^n)}{\bar{g}_n(1|W_i)} - 1\right)(\bar{Q}_0 - \bar{Q}^*_n)(1, W_i) - \left(\frac{g^n_{0,i}(0|W^n)}{\bar{g}_n(0|W_i)} - 1\right)(\bar{Q}_0 - \bar{Q}^*_n)(0, W_i) \\
&+1/n\sum_i \bar{Q}_0(W_i) - \bar{Q}^*_n(W_i) + \int_w (\bar{Q}_0 - \bar{Q}^*_n)(w) Q_{W,0}(w) \\
&+1/n\sum_i \bar{Q}^*_n(W_i) - P_0\bar{Q}^*_n \\
&= 1/n\sum_i \left(\frac{g_{0,i}(1|W^n)}{\bar{g}_n(1|W_i)} - 1\right)(\bar{Q}_0 - \bar{Q}^*_n)(1, W_i) - \left(\frac{g^n_{0,i}(0|W^n)}{\bar{g}_n(0|W_i)} - 1\right)(\bar{Q}_0 - \bar{Q}^*_n)(0, W_i) \\
&+1/n\sum_i \{\bar{Q}_0(W_i) - \psi_0\}.
\end{aligned}
$$

Thus,

$$
\begin{aligned}
&Z_{W,n,g^n} = \tfrac{1}{\sqrt{n}}\sum_{i=1}^n \left\{\bar{Q}_0(W_i) - \psi_0\right\} \\
&+\tfrac{1}{\sqrt{n}}\sum_{i=1}^n \left\{\frac{g^n_{0,i}(1|W^n) - \bar{g}_n(1|W_i)}{\bar{g}_n(1|W_i)}(\bar{Q}_0 - \bar{Q}^*_n)(1, W_i) - \frac{g^n_{0,i}(0|W^n) - \bar{g}_n(0|W_i)}{\bar{g}_n(0|W_i)}(\bar{Q}_0 - \bar{Q}^*_n)(0, W_i)\right\}.
\end{aligned}
$$

In the special case that $g^n_{0,i} = g_{0,i}$ and constant in i, we have that

$$
Z_{W,n,g^n} = Z_{W,n} \equiv \frac{1}{\sqrt{n}}\sum_{i=1}^n \{\bar{Q}_0(W_i) - \psi_0\}.
$$

In the general case, one can decompose

$$
Z_{W,n,g^n} = Z_{1,g^n} + Z_{W,n},
$$

where

$$
\begin{aligned}
Z_{1,g^n} &= \tfrac{1}{\sqrt{n}}\sum_{i=1}^n \frac{g^n_{0,i}(1|W^n) - \bar{g}_n(1|W_i)}{\bar{g}_n(1|W_i)}(\bar{Q}_0 - \bar{Q}^*_n)(1, W_i) \\
&-\tfrac{1}{\sqrt{n}}\sum_{i=1}^n \frac{g^n_{0,i}(0|W^n) - \bar{g}_n(0|W_i)}{\bar{g}_n(0|W_i)}(\bar{Q}_0 - \bar{Q}^*_n)(0, W_i).
\end{aligned}
$$

Suppose now that $g^n_{0,i} = g_{0,i}$. Then $\bar{g}_n = \bar{g}_0$. Notice that for a function f, we have

$$
\begin{aligned}
\frac{1}{\sqrt{n}}\sum_{i=1}^n E_0\left(\frac{g_{0,i}(1 \mid W_i)}{\bar{g}_0(1 \mid W_i)} - 1\right) f(W_i) &= \frac{1}{\sqrt{n}}\sum_{i=1}^n \int_w \left(\frac{g_{0,i}(1 \mid w)}{\bar{g}_0(1 \mid w)} - 1\right) f(w) dQ_{W,0}(w) \\
&= \frac{1}{\sqrt{n}}\int \left(\frac{\bar{g}_0}{\bar{g}_0}(1 \mid w) - 1\right) f(w) dQ_{W,0}(w) \\
&= 0.
\end{aligned}
$$

This proves that, $Z_{1,g^n}(\bar{Q})$, defined as the process above with $\bar{Q}^*_n$ replaced by $\bar{Q}$, is a standard empirical process $Z_{1,g^n}(\bar{Q}) = 1/\sqrt{n}\sum_i f_i(\bar{Q})(W_i)$ of mean zero and independent random variables

$$
f_i(\bar{Q})(W_i) = \frac{g_{0,i}(1|W_i) - \bar{g}_0(1|W_i)}{\bar{g}_0(1|W_i)}(\bar{Q}_0 - \bar{Q})(1, W_i) - \frac{g_{0,i}(0|W_i) - \bar{g}_0(0|W_i)}{\bar{g}_0(0|W_i)}(\bar{Q}_0 - \bar{Q})(0, W_i).
$$

Such a process can be analyzed with methods we use below, showing that $Z_{1,g^n}(\bar{Q}_n^*) = Z_{1,g^n}(\bar{Q}^*) + o_P(1)$, and $Z_{1,g^n}(\bar{Q}^*) = 1/\sqrt{n}\sum_i IC_{1,g^n,i}(W_i) + o_P(1)$, where $IC_{1,g^n,i} = f_i(\bar{Q}^*)$. We conclude that

$$\sqrt{n}(\psi_n^* - \psi_0) = X_n(\bar{Q}_n^*) + Z_{W,n} + Z_{1,g^n} + Z_{W,\bar{g}_n,n},$$

where our assumptions guarantee that $Z_{W,n} + Z_{1,g^n} + Z_{W,\bar{g}^n,n} = 1/\sqrt{n}\sum_i IC_{W,i}(W_i) + o_P(1)$. So we showed that $\sqrt{n}(\psi_n^* - \psi_0) = X_{W,n} + X_n(\bar{Q}_n^*)$, where $X_{W,n} = 1/\sqrt{n}\sum_i IC_{W,i}(W_i) + o_p(1)$ for some influence curve $IC_{W,i}$. Thus $X_{W,n}$ is understood and converges to a normal distribution with mean zero and variance $\sigma_W^2 = \lim_n \frac{1}{n}\sum_{i=1}^n P_0 IC_{W,i}^2$, if the variance of $IC_{W,i}$ is bounded uniformly in i.

Below we establish that, conditional on W^n, $X_n(\bar{Q}_n^*)$ converges in distribution to a Gaussian random variable. The separate weak convergence of $X_{W,n}$ and $X_n(\bar{Q}_n^*)$ implies the desired weak convergence of $X_{W,n}$ and $X_n(\bar{Q}_n^*)$ jointly as follows. For notational convenience, let X_n denote $X_n(\bar{Q}_n^*)$ and X denotes its limit in distribution. Let $W^\infty = (W^n : n = 1, \ldots,)$. Note that $P(X_{W,n} \in A, X_n \in B) = E_{W^\infty} I(X_{W,n} \in A) P(X_n \in B \mid W^\infty)$. Since $P(X_n \in B \mid W^\infty)$ converges to $P(X \in B)$ for almost every W^∞, we obtain

$$P(X \in B) E_{W^\infty} I(X_{W,n} \in A) \to P(X \in B) P(X_W \in A)$$

plus a term $E_{W^\infty} I(X_{W,n} \in A)(P(X_n \in B) \mid W^\infty) - P(X \in B))$. The latter term converges to zero by the dominated convergence theorem. The joint convergence implies the weak convergence of the sum $X_{W,n} + X_n(\bar{Q}_n^*)$ to $X_W + X$.

So it remains to study $X_n(\bar{Q}_n^*)$. By application of a CLT for sums of independent random variables, under the stated conditions, one can show that, conditional on W^n, $(X_n(\bar{Q}_j) : j)$ for fixed $\bar{Q}_j \in \mathcal{F}$ converges to a multivariate normal distribution with covariance matrix defined by $(\bar{Q}_1, \bar{Q}_2) \to \Sigma_0(\bar{Q}_1, \bar{Q}_2)$. Weak convergence of $X_n(\bar{Q})$ for a fixed $\bar{Q}$ or finite collection of $\bar{Q}$'s is not enough for establishing the desired asymptotic linearity. In order to understand terms such as $X_n(\bar{Q}_n^*) - X_n(\bar{Q})$ (and that our proposed variance estimator is consistent) we need to understand the process $(X_n(\bar{Q}) : \bar{Q} \in \mathcal{F})$ with respect to supremum norm over a set $\mathcal{F}$ that contains $\bar{Q}_n^*$ with probability tending to 1. Again, we will study this process conditional on $(W^n : n \geq 1)$.

Let $d_n^2(\bar{Q}_1, \bar{Q}_2) = 1/n \sum_j P_{Q_0,g^n}\{f_{j,n}(\bar{Q}_1) - f_{j,n}(\bar{Q}_2)\}^2$. We note that $X_n(\bar{Q}_1) - X_n(\bar{Q}_2) = X_n'(\bar{Q}_1 - \bar{Q}_2)$ for a slightly different process X_n'. Thus, $d_n^2(\bar{Q}_1, \bar{Q}_2) = 1/n \sum_j P_{Q_0,g^n}\{f_{j,n}'(\bar{Q}_1 - \bar{Q}_2)\}^2$ for a specified $f_{j,n}'(\bar{Q}_1 - \bar{Q}_2) = \sum_{i \in C_j(W^n)}\{f_{i,n}(\bar{Q}_1) - f_{i,n}(\bar{Q}_2)\}(O_i)$. Note that $d_n^2(\bar{Q}_1, \bar{Q}_2)$ is the conditional variance of $X_n(\bar{Q}_1) - X_n(\bar{Q}_2)$, conditional on W^n, or equivalently, it is the conditional variance of $X_n'(\bar{Q}_1 - \bar{Q}_2)$. We will denote this conditional variance also with $\sigma_n^2(\bar{Q}_1 - \bar{Q}_2) = d_n^2(\bar{Q}_1, \bar{Q}_2)$.

Recall that $\mathcal{F}^d = \{f_1 - f_2 : f_1, f_2 \in \mathcal{F}\}$. Given the entropy condition on $\mathcal{F}$, we will prove asymptotic equicontinuity of $(X_n(\bar{Q}) : \bar{Q} \in \mathcal{F}^d)$ with respect to this semi-metric d_n: for each $\epsilon > 0$ and sequence $\delta_n \to 0$,

$$P\left(\sup_{d_n(f,g) \leq \delta_n} \mid X_n(f) - X_n(g) \mid > \epsilon\right) \to 0 \text{ as } n \to \infty.$$

This is equivalent with establishing the following asymptotic equicontinuity of $(X_n'(f) : f \in \mathcal{F}^d)$ w.r.t semi-metric σ_n: for each $\epsilon > 0$ and sequence $\delta_n \to 0$,

$$P\left(\sup_{\sigma_n(f)\leq\delta_n} \mid X_n'(f) \mid > \epsilon\right) \to 0 \text{ as } n \to \infty.$$

If $d_n(\bar{Q}_n^*, \bar{Q}) \to 0$ in probability, and $\bar{Q}_n^* - \bar{Q} \in \mathcal{F}^d$ with probability tending to 1, then this asymptotic equicontinuity proves that $X_n(\bar{Q}_n^*) - X_n\bar{Q}) = X_n'(\bar{Q}_n^* - \bar{Q})$ converges to zero in probability, as $n \to \infty$.

To establish the asymptotic equicontinuity result, we use a number of fundamental building blocks. Note that $X_n'(f)/\sigma_n(f)$ is a sum of J independent mean zero bounded random variables and the variance of this sum equals 1. Bernstein's inequality states that $P(\mid \sum_j Y_j \mid > x) \leq 2\exp\left(-\frac{1}{2}\frac{x^2}{v+Mx/3}\right)$, where $v \geq \text{VAR}\sum_j Y_j$. Thus, by Bernstein's inequality, conditional on W^n, we have

$$P\left(\frac{\mid X_n'(f) \mid}{\sigma_n(f)} > x\right) \leq 2\exp\left(-\frac{1}{2}\frac{x^2}{1+Mx/3}\right) \leq K\exp(-Cx^2),$$

for a universal K and C.

As stated in our review section, this implies $\parallel X_n'(f)/\sigma_n(f) \parallel_{\psi_2} \leq (1 + K/C)^{0.5}$, where for a given convex function ψ with $\psi(0) = 0$, $\parallel X \parallel_\psi \equiv \inf\{C > 0 : E\psi(\mid X \mid /C) \leq 1\}$ is the so called Orlics norm, and $\psi_2(x) = \exp(x^2) - 1$. Thus $\parallel X_n'(f) \parallel_{\psi_2} \leq C_1\sigma_n(f)$ for $f \in \mathcal{F}^d$. This result allows us to apply Theorem 2.2.4 in van der Vaart and Wellner (1996) (this theorem is copied below in the appendix): for each $\delta > 0$ and $\eta > 0$, we now have

$$\parallel \sup_{\sigma_n(f)\leq\delta} \mid X_n'(f) \mid \parallel_{\psi_2} \leq K\left\{\int_0^\eta \psi_2^{-1}(N(\epsilon, \sigma_n, \mathcal{F}^d)d\epsilon + \delta\psi_2^{-1}(N^2(\eta, \sigma_n, \mathcal{F}^d))\right\}, \qquad \text{(B.1)}$$

Convergence to zero with respect to ψ_2-orlics norm implies convergence in expectation to zero and thereby convergence to zero in probability. Let δ_n be a sequence converging to zero, and let η_n also converge to zero but slowly enough so that the term $\delta_n\psi_2^{-1}(N^2(\eta_n, \sigma_n, \mathcal{F}^d))$ converges to zero as $n \to \infty$. By assumption, $\int_0^{\delta_n} \psi_2^{-1}(N(\epsilon, \sigma_n, \mathcal{F}^d)d\epsilon$ converges to zero. Thus,

$$\lim_{\delta_n\to 0}\left\{\int_0^{\delta_n} \psi_2^{-1}(N(\epsilon, \sigma_n, \mathcal{F}^d)d\epsilon + \delta_n\psi_2^{-1}(N^2(\eta_n, \sigma_n, \mathcal{F}^d))\right\} = 0.$$

This proves that

$$E\left(\sup_{\sigma_n(f)\leq\delta_n} \mid X_n'(f) \mid\right) \to 0,$$

and thereby the asymptotic equicontinuity of X_n'.

We now prove the convergence to the limit variance: If $\sigma_n(\bar{Q}_n^* - \bar{Q}) \to 0$ in probability, then

$$\frac{1}{n}\sum_{j=1}^J \{f_{j,n}(\bar{Q}_n^*)(O_i)\}^2 - \frac{1}{n}\sum_{j=1}^J P_{Q_0,g^n}\{f_{j,n}(\bar{Q})\}^2 \to 0 \text{ in probability.}$$

We can write this difference as a sum of the following two differences:

$$\begin{aligned}
&\tfrac{1}{n}\textstyle\sum_{j=1}^{J}\{f_{j,n}(\bar{Q}_n^*)(\bar{O}_j)\}^2 - \tfrac{1}{n}\textstyle\sum_{j=1}^{J} P_{Q_0,g^n} f_{j,n}(\bar{Q}_n^*)^2 \\
&\tfrac{1}{n}\textstyle\sum_{j=1}^{J} P_{Q_0,g^n} f_{j,n}(\bar{Q}_n^*)^2 - \tfrac{1}{n}\textstyle\sum_{j=1}^{n} P_{Q_0,g^n} f_{j,n}(\bar{Q})^2 \\
&= \tfrac{1}{n}\textstyle\sum_{j=1}^{J} P_{Q_0,g^n}\left\{f_{j,n}(\bar{Q}_n^*)^2 - f_{j,n}(\bar{Q})^2\right\} \\
&= \tfrac{1}{n}\textstyle\sum_{j=1}^{J} P_{Q_0,g^n}\left\{f'_{j,n}(\bar{Q}_n^* - \bar{Q})\right\}\left\{f_{j,n}(\bar{Q}_n^*) + f_{j,n}(\bar{Q})\right\} \\
&\leq \left(\tfrac{1}{n}\textstyle\sum_{j=1}^{J} P_{Q_0,g^n}\left\{f'_{j,n}(\bar{Q}_n^* - \bar{Q})\right\}^2\right)^{0.5}\left(\tfrac{1}{n}\textstyle\sum_{j=1}^{J} P_{Q_0,g^n}\left\{f_{j,n}(\bar{Q}_n^*) + f_{j,n}(\bar{Q})\right\}^2\right)^{0.5},
\end{aligned}$$

where we used Cauchy-Schwarz inequality at the last inequality. The last term can thus be bounded by $Md_n(\bar{Q}_n^*, \bar{Q})$, so that it converges to zero in probability, since $d_n(\bar{Q}_n^*), \bar{Q})$ converges to zero in probability.

We now consider the first term, which can be represented as

$$\frac{1}{n}\sum_{j=1}^{J} h_{j,n}(\bar{Q}_n^*),$$

where

$$h_{j,n}(\bar{Q}) \equiv f_{j,n}^2(\bar{Q})(O_i) - P_{Q_0,g^n} f_{j,n}(\bar{Q})^2.$$

Define the process $Y_n(\bar{Q}) = 1/n\sum_j h_{j,n}(\bar{Q})$. Note that $h_{j,n}(\bar{Q})$ has conditional mean zero given W^n. Thus, conditional on W^n, $Y_n(\bar{Q})$ is a sum of independent mean zero random variables. The process $\sqrt{n}Y_n(\bar{Q})$ has exactly same structure as process $X_n(\bar{Q})$ we analyzed above. Therefore, under our conditions, we have $\sup_{\bar{Q}\in\mathcal{F}} \mid Y_n(\bar{Q}) = O_P(1/\sqrt{n})$. This implies, in particular, that the first term converges to zero in probability. This proves the convergence to the desired limit.

C Appendix: Proof of Theorem 4

We can decompose $D^*(\bar{Q}^*, g_0, \psi_0)$ orthogonally in a function of W and a function of Y, A, W, which has conditional mean zero, given W, as follows:

$$D^*(\bar{Q}^*, g_0, \psi_0) = \bar{Q}_0(W) - \psi_0 + H_{g_0}(A, W)(Y - \bar{Q}^*(A, W)) - \{\bar{Q}_0(W) - \bar{Q}^*(W)\}.$$

Thus, the variance is given by:

$$\begin{aligned}
P_0\{D^*(\bar{Q}^*, g_0, \psi_0)\}^2 = E_0\{\bar{Q}_0(W) - \psi_0\}^2 + E_0E_0\left(H_{g_0}^2(A, W)(Y - \bar{Q}^*(A, W))^2 \mid W\right) \\
- E_0\{\bar{Q}_0(W) - \bar{Q}^*(W)\}^2.
\end{aligned}$$

Note that

$$\begin{aligned}
E_0E_0(H_{g_0}^2(A, W)(Y - \bar{Q}^*(A, W))^2 \mid W) &= E_0E_0\left(\frac{1}{g_0^2(A \mid W)}E_0((Y - \bar{Q}^*(A, W))^2 \mid A, W) \mid W\right) \\
&= E_0\sum_a \frac{1}{g_0(a \mid W)}\sigma_0^2(\bar{Q}^*)(a, W),
\end{aligned}$$

where $\sigma_0^2(\bar{Q}^*)(a,W) \equiv E_0((Y-\bar{Q}^*(A,W))^2 \mid A=a,W)$. Thus, we have obtained the following expression:

$$\sigma_I^2(\bar{Q}^*) = E_0\{\bar{Q}_0(W)-\psi_0\}^2 + E_0 \sum_a \frac{1}{g_0(a \mid W)}\sigma^2(\bar{Q}^*)(a,W) - E_0\{\bar{Q}_0(W)-\bar{Q}^*(W)\}^2.$$

For the paired matching design the asymptotic variance σ^2 of the TMLE is given by the limit of

$$\begin{aligned} E_0\{\bar{Q}_0(W)-\psi_0)\}^2 + E_0 \tfrac{1}{n}\textstyle\sum_{j=1}^{n/2} P_{Q_0,g^n}\left\{\sum_{i\in C_j(W^n)} H_{g_0}(A_i,W_i)(Y_i-\bar{Q}^*(A_i,W_i))\right\}^2 \\ -E_0 \tfrac{1}{n}\textstyle\sum_{j=1}^{n/2}\left\{\sum_{i\in C_j(W^n)}\{\bar{Q}_0(W_i)-\bar{Q}^*(W_i)\}\right\}^2 \end{aligned}$$

Each $\sum_{i\in C_j(W^n)}$ is a sum over two terms. We use that $(a+b)^2 = a^2+b^2+2ab$. The contribution a^2+b^2 from the square terms yields:

$$\begin{aligned} &E_0 \tfrac{1}{n}\textstyle\sum_{i=1}^{n}\left\{P_{Q_0,g^n}\left\{H_{g_0}(A_i,W_i)(Y_i-\bar{Q}^*(A_i,W_i))\right\}^2 - \{\bar{Q}_0(W_i)-\bar{Q}^*(W_i)\}^2\right\} \\ &= E_0 \textstyle\sum_a \frac{1}{g_0(a|W)}\sigma_0^2(\bar{Q}^*)(a,W) - E_0\{\bar{Q}_0(W)-\bar{Q}^*(W)\}^2. \end{aligned}$$

This equals the corresponding expression we have for $\sigma_I^2(\bar{Q}^*)$. The contribution $2ab$ from the cross-terms yields:

$$\begin{aligned} &2E_0 \tfrac{1}{n}\textstyle\sum_{j=1}^{n/2} P_{Q_0,g^n} H_{g_0,1j}(Y_{1j}-\bar{Q}^*(A_{1j},W_{1j}))H_{g_0,2j}(Y_{2j}-\bar{Q}^*(A_{2j},W_{2j})) \\ &-2E_0 \tfrac{1}{n}\textstyle\sum_{j=1}^{n/2}\{\bar{Q}_0(W_{1j})-\bar{Q}^*(W_{1j})\}\{\bar{Q}_0(W_{2j})-\bar{Q}^*(W_{2j})\} \\ &= -4E_0 \tfrac{1}{n}\textstyle\sum_{j=1}^{n/2}\left\{(Y_{1j}(1)-\bar{Q}^*(1,W_{1j}))(Y_{2j}(0)-\bar{Q}^*(0,W_{2j}))\right\} \\ &-4E_0 \tfrac{1}{n}\textstyle\sum_{j=1}^{n/2}\left\{(Y_{1j}(0)-\bar{Q}^*(0,W_{1j}))(Y_{2j}(1)-\bar{Q}^*(1,W_{2j}))\right\} \\ &-2E \tfrac{1}{n}\textstyle\sum_{j=1}^{n/2}\{\bar{Q}_0(W_{1j})-\bar{Q}^*(W_{1j})\}\{\bar{Q}_0(W_{2j})-\bar{Q}^*(W_{2j})\}. \end{aligned}$$

To conclude, the asymptotic variance under the paired matching design is given by:

$$\begin{aligned} \sigma^2(\bar{Q}^*) = E_0\{\bar{Q}_0(W)-\psi_0)\}^2 + E_0 \textstyle\sum_a \frac{1}{g_0(a|W)}\sigma^2(\bar{Q}^*)(a,W) \\ -E_0\{\bar{Q}_0(W)-\bar{Q}^*(W)\}^2 \\ -4E_0 \tfrac{1}{n}\textstyle\sum_{j=1}^{n/2}\left\{(Y_{1j}(1)-\bar{Q}^*(1,W_{1j}))(Y_{2j}(0)-\bar{Q}^*(0,W_{2j}))\right\} \\ -4E_0 \tfrac{1}{n}\textstyle\sum_{j=1}^{n/2}\left\{(Y_{1j}(0)-\bar{Q}^*(0,W_{1j}))(Y_{2j}(1)-\bar{Q}^*(1,W_{2j}))\right\} \\ -2E \tfrac{1}{n}\textstyle\sum_{j=1}^{n/2}\{\bar{Q}_0(W_{1j})-\bar{Q}^*(W_{1j})\}\{\bar{Q}_0(W_{2j})-\bar{Q}^*(W_{2j})\} \end{aligned}$$

Thus, the difference between the two asymptotic variances is given by:

$$\begin{aligned}
&\sigma_I^2 - \sigma^2 = 4E_0\tfrac{1}{n}\textstyle\sum_{j=1}^{n/2}\left\{(Y_{1j}(1) - \bar{Q}^*(1,W_{1j}))(Y_{2j}(0) - \bar{Q}^*(0,W_{2j}))\right\}\\
&+4E_0\tfrac{1}{n}\textstyle\sum_{j=1}^{n/2}\left\{(Y_{1j}(0) - \bar{Q}^*(0,W_{1j}))(Y_{2j}(1) - \bar{Q}^*(1,W_{2j}))\right\}\\
&+2E\tfrac{1}{n}\textstyle\sum_{j=1}^{n/2}\{\bar{Q}_0(W_{1j}) - \bar{Q}^*(W_{1j})\}\{\bar{Q}_0(W_{2j}) - \bar{Q}^*(W_{2j})\}\\
&= 2E_0\tfrac{1}{J}\textstyle\sum_{j=1}^{n/2}\left\{(\bar{Q}_0(1,W_{1j}) - \bar{Q}^*(1,W_{1j}))(\bar{Q}_0(0,W_{2j}) - \bar{Q}^*(0,W_{2j}))\right\}\\
&+2E_0\tfrac{1}{J}\textstyle\sum_{j=1}^{n/2}\left\{(\bar{Q}_0(0,W_{1j}) - \bar{Q}^*(0,W_{1j}))(\bar{Q}_0(1,W_{2j}) - \bar{Q}^*(1,W_{2j}))\right\}\\
&+E\tfrac{1}{J}\textstyle\sum_{j=1}^{n/2}\{\bar{Q}_0(W_{1j}) - \bar{Q}^*(W_{1j})\}\{\bar{Q}_0(W_{2j}) - \bar{Q}^*(W_{2j})\}\\
&= E_0\tfrac{1}{J}\textstyle\sum_{j=1}^{n/2}(\bar{Q}_0 - \bar{Q}^*)(1,W_{1j})(\bar{Q}_0 - \bar{Q}^*)(0,W_{2j})\\
&+E_0\tfrac{1}{J}\textstyle\sum_{j=1}^{n/2}(\bar{Q}_0 - \bar{Q}^*)(0,W_{1j})(\bar{Q}_0 - \bar{Q}^*)(1,W_{2j})\\
&+E_0\tfrac{1}{J}\textstyle\sum_{j=1}^{n/2}(\bar{Q}_0 - \bar{Q}^*)(1,W_{1j})(\bar{Q}_0 - \bar{Q}^*)(1,W_{2j})\\
&+E_0\tfrac{1}{J}\textstyle\sum_{j=1}^{n/2}(\bar{Q}_0 - \bar{Q}^*)(0,W_{1j})(\bar{Q}_0 - \bar{Q}^*)(0,W_{2j})\\
&\equiv C,
\end{aligned}$$

and $\sigma^2 = \sigma_I^2 - C$. □

D Appendix: Proof of Theorem 5

The proof is analogue to the proof of Theorem 1. Therefore, we suffice with proving (7.3). Firstly, we note that

$$\begin{aligned}
&E_0 D^{n,*}(\bar{Q},\bar{g},\bar{\Pi},\psi_0)(O^n) = E_0\tfrac{1}{n}\textstyle\sum_{i=1}^{n}\{\bar{Q}(W_i) - \psi_0\} +\\
&E_0\tfrac{1}{n}\textstyle\sum_{i=1}^{n}\frac{g_{i,0}(1,1|W_i)}{\bar{g}_n(1,1|W_i)}(\bar{Q}_0 - \bar{Q})(1,W_i) - \frac{g_{i,0}(0,1|W_i)}{\bar{g}_n(0,1|W_i)}(\bar{Q}_0 - \bar{Q})(0,W_i)\\
&= \Psi(Q) - \psi_0 + \tfrac{1}{n}\textstyle\sum_i \int_w Q_{W,0}(w)\frac{g_{i,0}(1,1|w)}{\bar{g}_n(1,1|w)}(\bar{Q}_0 - \bar{Q})(1,w)\\
&\qquad\qquad -\tfrac{1}{n}\textstyle\sum_i \int_w Q_{W,0}(w)\frac{g_{i,0}(0,1|w)}{\bar{g}_n(0,1|w)}(\bar{Q}_0 - \bar{Q})(0,w)\\
&= \Psi(Q) - \psi_0 + E_0\frac{\bar{g}_{0,n}(1,1|W)}{\bar{g}_n(1,1|W)}(\bar{Q}_0 - \bar{Q})(1,W)\\
&\qquad\qquad -E_0\frac{\bar{g}_{n,0}(0,1|W)}{\bar{g}_n(0,1|W)}(\bar{Q}_0 - \bar{Q})(0,W).
\end{aligned}$$

Thus, if $\bar{g}_{n,0} = \bar{g}_0$, then this equals $\Psi(Q) - \psi_0 + \psi_0 - \Psi(Q) = 0$. If $Q_0 = Q$, then we also obtain 0. This proves (3.3). □

E Appendix: Review of relevant empirical process/weak convergence theory

We refer to van der Vaart and Wellner (1996), Section 2.2. on maximal inequalities and covering numbers. For a real valued random variable X and convex function ψ with $\psi(0) = 0$, the Orlics

norm is defined as $\| X \|_\psi \equiv \inf\{C > 0 : E\psi(| X | /C) \leq 1\}$. Setting $\psi(x) = x^p$ gives the L_p-norms $\| X \|_p = E(| X |^p)^{1/p}$, $p \geq 1$. Another important choice for empirical processes is $\psi_p(x) = \exp(x^p) - 1$. Sums of independent bounded random variables and Gaussian random variables have bounded ψ_2-norm. There is an important relation between the orlics norm and a bound on the tail probability of the random variable. In particular, we have (page 96 in van der Vaart and Wellner (1996))

$$P(| X | > x) \leq \frac{1}{\psi(x/ \| X \|_\psi)}.$$

For $\psi_p(x)$ this leads to tail estimates $\exp(-Cx^p)$ for any random variable with a finite ψ_p norm. Conversely, an exponential tail bound of this type shows that $\| X \|_{\psi_p}$ is finite: Lemma 2.2.1 states that if $P(| X | > x) \leq K \exp(-Cx^p)$ for every x, for constants K and C, and for $p \geq 1$, then its orlics norm satisfies $\| X \|_{\psi_p} \leq ((1 + K)/C)^{1/p}$. So if we have an exponential tail probability for $X_n(f)$, then we can translate this into a bound on the ψ_p-orlics norm.

Given a sequence of random variables X_i, we have (page 96)

$$\| \max_{i \leq m} X_i \|_\psi \leq K\psi^{-1}(m) \max_i \| X_i \|_\psi .$$

Thus, if we can bound the orlics norm of $X_n(f)$ in terms of a norm on f, then this result allows us to bound the orlics norm of a maximum over m functions. This bound combined with chaining gives the typical entropy type bounds. As we will see one of the main things we will need is a bound on $\| X_n(f) \|_\psi$ in terms of $d(f, f)$ for a semi-metric d on $\mathcal{F}$.

Bounding orlics norm: Let (T, d) be an arbitrary semi-metric space. The covering number $N(\epsilon, d)$ is the minimal number of balls of radius ϵ needed to cover T. Call a collection of points ϵ-separated if the distance between each pair of points is strictly larger than ϵ. The packing number $D(\epsilon, d)$ is the maximum number of ϵ-separated points in T. Entropy numbers are the logarithms of the covering or packing number. Since $N(\epsilon, d) \leq D(\epsilon, d) \leq N(0.5\epsilon, d)$, bounds in packing number map into a bound in covering number and vice versa.

For our purpose, we will need Theorem 2.2.4 in van der Vaart and Wellner (1996), which is stated here for completeness.

Theorem 6. *(Theorem 2.2.4, van der Vaart and Wellner, 96) Let ψ be a convex non-decreasing non zero function with $\psi(0) = 0$ and $\limsup_{x,y\to\infty} \psi(x)\psi(y)/\psi(cxy) < \infty$ for some constant c. Let $(X_t : t \in T)$ be a separable stochastic process (that is, $\sup_{d(s,t)<\delta} | X_s - X_t |$ remains almost surely the same if the index set T is replaced by a suitable countable subset) with*

$$\| X_s - X_t \|_\psi \leq Cd(s, t) \text{ for every } s, t,$$

for some semimetric d on T and a constant C. Then, for any $\eta, \delta > 0$,

$$\| \sup_{d(s,t)\leq\delta} | X_s - X_t | \|_\psi \leq K \left\{ \int_0^\eta \psi^{-1}(D(\epsilon, d))d\epsilon + \delta\psi^{-1}(D^2(\eta, d)) \right\}$$

for a constant K depending on ψ and C only. In particular, the constant K can be chosen so that

$$\| \sup_{s,t} \mid X_s - X_t \mid \|_\psi \leq K \int_0^{diamT} \psi^{-1}(D(\epsilon, d))d\epsilon,$$

where $diam(T)$ is the diameter of T. This result also gives

$$\| \sup_{t} \mid X_t \mid \|_\psi \leq \| X_{t_0} \|_\psi + \int_0^{diam(T)} \psi^{-1}(D(\epsilon, d))d\epsilon.$$

The bound shows that the sample paths of X are uniformly continuous in ψ-norm, whenever the covering integral $\int_0^\eta \psi^{-1}(D(\epsilon, d))d\epsilon$ is finite/exists for some $\eta > 0$. In order to have that this integral is bounded for classes T with covering numbers that behave as ϵ^{-p}, one will need to use an Orlics norm with $\psi(x) = x^p$, and if one wants the integral to be bounded for any p, then one needs $\psi(x) = \exp(x^q) - 1$ for some q.

If one can prove that $\| X_n(s) - X_n(t) \|_\psi \leq Cd(s, t)$ for a constant C independent of n, and each X_n is a separable stochastic process, then this theorem teaches us that for any sequence δ_n, and $\eta_n > 0$, we have that there exists a constant K depending on ψ, C only (not dependent on n!) so that

$$\| \sup_{d(s,t)\leq\delta_n} \mid X_n(s) - X_n(t) \mid \|_\psi \leq K \left\{ \int_0^\eta \psi^{-1}(D(\epsilon, d))d\epsilon + \delta_n \psi^{-1}(D^2(\eta, d)) \right\}.$$

We can now apply this inequality for a sequence $\delta_n \to 0$ for $n \to \infty$. Since η can be chosen arbitrary small, it follows that, if $\int_0^\eta \psi^{-1}(D(\epsilon, d))d\epsilon < \infty$ for some $\eta > 0$, then

$$\| \sup_{d(s,t)\leq\delta_n} \mid X_n(s) - X_n(t) \mid \|_\psi \to 0 \text{ as } n \to \infty.$$

So we can state the following useful corollary:

Corollary E.1. Suppose there exists a $\eta > 0$ so that $\int_0^\eta \psi^{-1}(D(\epsilon, d))d\epsilon < \infty$. In addition, assume

$$\| X_n(s) - X_n(t) \|_\psi \leq Cd(s, t)$$

for a constant C independent of n, and each X_n is a separable stochastic process with respect to d. Then for any sequence $\delta_n \to 0$, we have

$$\| \sup_{d(s,t)\leq\delta_n} \mid X_n(s) - X_n(t) \mid \|_\psi \to 0 \text{ as } n \to \infty.$$

This corollary provides us with conditions under which X_n is asymptotically uniformly d-equicontinuous in probability. Theorem 1.5.7. in van der Vaart and Wellner (1996) now states that X_n is asymptotically tight in $\ell^\infty(T)$ if $X_n(t)$ is asymptotically tight for every t, (T, d) is totally bounded, and X_n is asymptotically d-equicontinuous in probability. In addition, Theorem 1.5.4 states that if X_n is asymptotically tight and its marginals converge weakly to the marginals $X(t_1), \ldots, X(t_k)$ of a stochastic process X, then there is a version of X with uniformly bounded sample paths and X_n converges weakly to X. Thus, we can state the following result:

Lemma E.1. *Let ψ be one of the following functions: $\psi(x) = x^p$ for some p, or $\psi(x) = \exp(x^1)-1$, $\psi(x) = \exp(x^2) - 1$. Let d be a semi-metric on T so that $(\ell^\infty(T), d)$ is totally bounded, and there exists a $\eta > 0$ so that $\int_0^\eta \psi^{-1}(D(\epsilon, d))d\epsilon < \infty$. In addition, assume*

$$\| X_n(s) - X_n(t) \|_\psi \leq Cd(s,t)$$

for a constant C independent of n, and each X_n is a separable stochastic process with respect to d. Then for any sequence $\delta_n \to 0$, we have for each $x > 0$

$$Pr\left(\sup_{d(s,t)\leq\delta_n} | X_n(s) - X_n(t) | > x \right) \to 0 \text{ as } n \to \infty, \tag{E.1}$$

and X_n is asymptotically tight.

If $X_n(t_1), \ldots, X_n(t_k)$ converges weakly to $(X(t_1), \ldots, X(t_k)$, then there exists a version X with uniformly bounded sample paths and $X_n \Rightarrow_d X$.

If X is Gaussian process X in $\ell^\infty(T)$, and $d(s,t) = \rho_p(s,t) \equiv \| (X(f) - X(g)) \|_p$, then there exists a version of X which is tight Borel measurable map into $\ell^\infty(T)$.

Actually (page 41), if X is Gaussian, then X_n converges weakly to X in $\ell^\infty(T)$ if and only if for some p (and then for all p) (i) the marginals of X_n converge to the corresponding marginals of X, (ii) X_n is asymptotically equicontinuous in probability with respect to

$$d(s,t) = \rho_p(s,t) \equiv \| X(s) - X(t) \|_p,$$

as defined in (E.1), and (iii) T is totally bounded for $d = \rho_p$.