On the Efficiency of Finely Stratified Experiments *

Yuehao Bai
Department of Economics
University of Southern California
yuehao.bai@usc.edu

Jizhou Liu
Booth School of Business
University of Chicago
jliu32@chicagobooth.edu

Azeem M. Shaikh
Department of Economics
University of Chicago
amshaikh@uchicago.edu

Max Tabord-Meehan
Department of Economics
University of Chicago
maxtm@uchicago.edu

July 31, 2023

Abstract

This paper studies the efficient estimation of a large class of treatment effect parameters that arise in the analysis of experiments. Here, efficiency is understood to be with respect to a broad class of treatment assignment schemes for which the marginal probability that any unit is assigned to treatment equals a pre-specified value, e.g., one half. Importantly, we do not require that treatment status is assigned in an i.i.d. fashion, thereby accommodating complicated treatment assignment schemes that are used in practice, such as stratified block randomization and matched pairs. The class of parameters considered are those that can be expressed as the solution to a restriction on the expectation of a known function of the observed data, including possibly the pre-specified value for the marginal probability of treatment assignment. We show that this class of parameters includes, among other things, average treatment effects, quantile treatment effects, local average treatment effects as well as the counterparts to these quantities in experiments in which the unit is itself a cluster. In this setting, we establish two results. First, we derive a lower bound on the asymptotic variance of estimators of the parameter of interest in the form of a convolution theorem. Second, we show that the naïve method of moments estimator achieves this bound on the asymptotic variance quite generally if treatment is assigned using a "finely stratified" design. By a "finely stratified" design, we mean experiments in which units are divided into groups of a fixed size and a proportion within each group is assigned to treatment uniformly at random so that it respects the restriction on the marginal probability of treatment assignment. In this sense, "finely stratified" experiments lead to efficient estimators of treatment effect parameters "by design" rather than through *ex post* covariate adjustment.

KEYWORDS: Convolution Theorem, Efficiency, Experiment, Experimental design, Finely stratified experiment, Matched pairs, Randomized controlled trial

JEL classification codes: C12, C14

*The fourth author acknowledges support from NSF grant SES-2149408.

1 Introduction

This paper studies the efficient estimation of a large class of treatment effect parameters that arise in the analysis of experiments. Our analysis includes a broad class of treatment assignment schemes for which the marginal probability that any unit is assigned to treatment equals a pre-specified value, e.g., one half. In particular, we do not require that treatment status is assigned in an i.i.d. fashion. In this way, our framework accommodates complicated treatment assignment schemes that are used routinely throughout the sciences, such as stratified block randomization and matched pairs. For a discussion of such treatment assignment schemes focused on clinical trials, see Rosenberger and Lachin (2015); for reviews focused on development economics, see Duflo et al. (2007) and Bruhn and McKenzie (2009). The class of parameters we consider are those that can be characterized as the solution to a restriction on the expectation of a known function of the observed data, including possibly the pre-specified value for the marginal probability of treatment assignment. We show in Section 2 below that this class of parameters includes many treatment effect parameters of interest: average treatment effects, quantile treatment effects, and local average treatment effects as well as the counterparts to these quantities in experiments in which the unit is itself a cluster.

In the setting described above, we establish two results. First, we derive a lower bound on the asymptotic variance of "regular" estimators of the parameter of interest in the form of a convolution theorem. The richness of the possible treatment assignment schemes complicates the derivation of such a result in our setting. As explained further in the discussion following Theorem 3.1 in Section 3, this feature precludes the use of standard arguments which could be used in establishing versions of these results if assignment were i.i.d.; see, for example, van der Vaart (1998). Second, we show that the naïve method of moments estimator achieves this bound on the asymptotic variance quite generally if treatment status is assigned using a "finely stratified" design. By a "finely stratified" design, we mean experiments in which units are divided into groups of a fixed size and a proportion within each group is assigned to treatment uniformly at random so that it respects the restriction on the marginal probability of treatment assignment. When the fixed size equals two and the pre-specified value for the marginal probability of treatment assignment equals one half, such a design is simply a matched pairs design. An attractive feature of this result is that, in contrast to other estimators that achieve the same efficiency bound when treatment is possibly assigned in other ways, it does not require any *ex post* covariate adjustment. Such adjustments frequently involve the nonparametric estimation of conditional expectations or similar quantities; see, for example, Zhang et al. (2008), Tsiatis et al. (2008), Jiang et al. (2022a), Jiang et al. (2022b) and Rafi (2023).[1] Moreover, with the exception of Zhang et al. (2008), these adjustments are generally developed for specific parameters of interest. Our results show that "finely stratified" experiments remarkably lead to efficient estimators for a large class of treatment effect parameters "by design" rather than through any such *ex post* covariate adjustment and, in this way, circumvent nonparametric estimation. In this sense, our results generalize similar observations made in Bai et al. (2022), Bai (2022), and Cytrynbaum (2023b) in the special case of estimating the average treatment effect.

Our paper builds upon two strands of literature. The first strand of literature concerns bounds on the

[1] For related results in the context of observational data see Newey (1994), Hahn (1998), Heckman et al. (1998), Frolich (2007), Firpo (2007), Farrell (2015), and Chernozhukov et al. (2017).

efficiency with which treatment effect parameters can be estimated in experiments. Hahn (1998) establishes a lower bound on the asymptotic variance of "regular" estimators of the average treatment effect when treatment status is assigned in an i.i.d. fashion. He finds, in particular, that the variance-minimizing treatment assignment scheme is the i.i.d. treatment assignment scheme that assigns units according to a conditional Neyman allocation in which units are assigned to treatment (control) with conditional probability proportional to the conditional variance of the potential outcome under treatment (control). Armstrong (2022) shows that this treatment assignment scheme remains variance-minimizing among a much larger class of treatment assignment schemes. As explained further in Remark 3.4, it may, however, be desirable to consider narrower classes of treatment assignment schemes that may not include the conditional Neyman allocation. Motivated by such concerns, Rafi (2023) derives a lower bound on the asymptotic variance of "regular" estimators of the average treatment effect for the class of *finitely*-stratified treatment assignment schemes considered in Bugni et al. (2019) with pre-specified values for proportions of treatment within each of the finitely many strata. As a result, neither treatment assignment schemes that implement the conditional Neyman allocation nor "finely stratified" designs are permitted as possible treatment assignment schemes. We note that results analogous to those in Hahn (1998) for the local average treatment effect and quantile treatment effect have been derived in Frolich (2007) and Firpo (2007), respectively; results analogous to those in Rafi (2023) for the the local average treatment effect have been derived in Jiang et al. (2022a). Our efficiency bound differs from these prior results in two ways: first, they apply to a general class of treatment effect parameters, including, but not limited to, those mentioned above; second, our analysis, like Rafi (2023), rules out treatment assignment schemes that implement the conditional Neyman allocation, but permits a richer class of treatment assignment schemes, including "finely stratified" designs that appear, according to our results, to be useful in achieving efficiency.

The second strand of literature concerns the analysis of "finely stratified" experiments. Within this literature, our analysis is most closely related to Bai et al. (2022), who derived the asymptotic behavior of the difference-in-means estimator of the average treatment effect when treatment is assigned according to a matched pairs design. Cytrynbaum (2023b) extends these results to permit the proportion of units assigned to treatment to vary with the baseline covariates; Bai et al. (2023d) extends these results to permit multiple treatments; finally, Bai et al. (2023c) extends these results for the analysis of different cluster-level average treatment effects; and Jiang et al. (2021) develop results analogous to those in Bai et al. (2022) for suitable estimators of the quantile treatment effect. As in our analysis in Section 4, the main requirement underlying the results in these papers is that units are paired so that they are suitably "close" in terms of the observed, baseline covariates. Finally, we note that some finite-sample optimality properties of matched pairs-designs for estimation of the average treatment effect are developed in Bai (2022).

The remainder of this paper is organized as follows. In Section 2, we describe our setup and notation. We emphasize in particular the way in which our framework can accommodate various treatment effect parameters of interest. In Section 3, we develop our lower bound on the asymptotic variance of "regular" estimators of these parameters. Section 4 derives the asymptotic behavior of the naïve method of moments estimator of our parameter of interest when treatment is assigned using a "finely stratified" design and shows, in particular, that its asymptotic variance achieves the bound established in the preceding section. In Section

5, we illustrate the practical relevance of our theoretical results by comparing the mean-squared errors of the naïve method of moments estimators of the average treatment effect and local average treatment effect when treatment status is assigned according to a matched pairs design versus that of an estimator using *ex post* covariate adjustment when treatment status is assigned in an i.i.d. fashion. Finally, we conclude in Section 6 with some recommendations for empirical practice guided by both these simulations and our theoretical results. Proofs of all results can be found in the Appendix.

2 Setup and Motivation

Let $A_i \in \{0,1\}$ denote the treatment status of the ith unit, and let $X_i \in \mathbf{R}^{d_x}$ denote their observed, baseline covariates. For $a \in \{0,1\}$, let $R_i(a) \in \mathbf{R}^{d_r}$ denote a *vector* of potential responses. As we illustrate below, considering a vector of responses allows us to accommodate certain parameters of interest. Let $R_i \in \mathbf{R}^{d_r}$ denote the vector of observed responses obtained from $R_i(a)$ once treatment is assigned. As usual, the observed responses and potential responses are related to treatment status by the relationship

$$R_i = R_i(1)A_i + R_i(0)(1 - A_i) \ . \tag{1}$$

We assume throughout that our sample consists of n units. For any random vector indexed by i, for example A_i, we define $A^{(n)} = (A_1, \ldots, A_n)$. Let P_n denote the distribution of the observed data $(R^{(n)}, A^{(n)}, X^{(n)})$, and Q_n the distribution of $(R^{(n)}(1), R^{(n)}(0), X^{(n)})$. We assume $Q_n = Q^n$, where Q is the marginal distribution of $(R_i(1), R_i(0), X_i)$. Given Q_n, P_n is then determined by (1) and the mechanism for determining treatment assignment. We assume that treatment assignment is performed such that a standard unconfoundedness assumptions holds and such that the probability of assignment given X is some known constant for every $1 \leq i \leq n$:

Assumption 2.1. Treatment status is assigned so that

$$(R^{(n)}(1), R^{(n)}(0)) \perp\!\!\!\perp A^{(n)} | X^{(n)} \ , \tag{2}$$

and such that $P\{A_i = 1 | X_i = x\} = \eta$, for some $\eta \in (0,1)$ for all $1 \leq i \leq n$.

Note that given Assumption 2.1, (X_i, A_i, R_i) are identically distributed for $1 \leq i \leq n$, and their marginal distribution does not change with n (see Lemma A.5 in the Appendix). As a consequence we denote the marginal distribution of (X_i, A_i, R_i) by P. Next we define our parameters of interest, denoted generically by $\theta_0 \in \Theta \subset \mathbf{R}^{d_\theta}$. We consider parameters θ_0 which can be defined as the solution to a set of moment equalities. In particular, if $m : \mathbf{R}^{d_x} \times \{0,1\} \times \mathbf{R}^{d_r} \to \mathbf{R}^{d_\theta}$ is a known measurable function, then we consider parameters θ_0 which uniquely solve the moment equality

$$E_P[m(X_i, A_i, R_i, \theta_0)] = 0 \ , \tag{3}$$

where we emphasize that $m(\cdot)$ is not a function of any unknown nuisance parameters, but may depend on

the known value of η in Assumption 2.1. We present five examples of well-known parameters which can be described as (functions of) solutions to a set of moment conditions as in (3).

Example 2.1 (Average Treatment Effect)**.** Let $Y_i(a) = R_i(a)$ denote a scalar potential outcome for the ith unit under treatment $a \in \{0,1\}$, and let $Y_i = R_i$ denote the observed outcome. Let $\theta_0 = E_Q[Y_i(1) - Y_i(0)]$ denote the average treatment effect (ATE). Under Assumption 2.1, θ_0 solves the moment condition in (3) with $m(\cdot)$ given by

$$m(X_i, A_i, R_i, \theta) = \frac{Y_i A_i}{\eta} - \frac{Y_i(1-A_i)}{1-\eta} - \theta \ . \tag{4}$$

For a list of papers which consider estimators based on (4), see Hirano et al. (2003). ■

Example 2.2 (Quantile Treatment Effect)**.** Let $Y_i(a) = R_i(a)$ denote a scalar potential outcome for the ith unit under treatment $a \in \{0,1\}$, and let $Y_i = R_i$ denote the observed outcome. Let $\tau \in (0,1)$ and $\theta_0 = (\theta_0(1), \theta_0(0))' = (q_{Y(1)}(\tau), q_{Y(0)}(\tau))'$, where

$$q_{Y(a)}(\tau) = \inf\{\lambda \in \mathbf{R} : Q\{Y_i(a) \leq \lambda\} \geq \tau\} \ .$$

In other words, θ_0 is defined to be the vector of τth quantiles of the marginal distributions of $Y_i(1)$ and $Y_i(0)$. If we assume $q_{Y(a)}(\tau)$ is unique for $a \in \{0,1\}$ in the sense that $Q\{Y(a) \leq q_{Y(a)}(\tau)+\epsilon\} > Q\{Y(a) \leq q_{Y(a)}(\tau)\}$ for all $\epsilon > 0$, then it follows from Assumption 2.1 and Lemma 1 in Firpo (2007) that θ_0 solves the moment condition in (3) with $m(\cdot)$ given by

$$m(X_i, A_i, R_i, \theta) = \begin{pmatrix} \dfrac{A_i(\tau - I\{Y_i \leq \theta(1)\})}{\eta} \\ \dfrac{(1-A_i)(\tau - I\{Y_i \leq \theta(0)\})}{1-\eta} \end{pmatrix} \ ,$$

for $\theta = (\theta^{(1)}, \theta^{(0)})'$. Note that the quantile treatment effect (QTE) $q_{Y(1)}(\tau) - q_{Y(0)}(\tau)$ can then be defined as $h(\theta_0)$ where $h : \mathbf{R}^2 \to \mathbf{R}$ is given by $h(s,t) = s - t$. ■

Example 2.3 (Local Average Treatment Effect)**.** Let $(\tilde{Y}_i(a), D_i(a)) = R_i(a)$ denote the vector of potential outcomes and treatment take-up under treatment $a \in \{0,1\}$, and let $(Y_i, D_i) = R_i$ denote the vector of observed outcomes and treatment take-up. Note here that $\tilde{Y}_i(a)$ corresponds to the potential outcome under assignment $a \in \{0,1\}$ and not to the potential outcome for a given take-up $D_i = d$. Suppose $E_Q[D_i(1) - D_i(0)] \neq 0$ and let

$$\theta_0 = \frac{E_Q[\tilde{Y}_i(1) - \tilde{Y}_i(0)]}{E_Q[D_i(1) - D_i(0)]} \ .$$

It then follows from Assumption 2.1 that θ_0 solves the moment condition in (3) with $m(\cdot)$ given by

$$m(X_i, A_i, R_i, \theta) = \frac{Y_i A_i}{\eta} - \frac{Y_i(1-A_i)}{1-\eta} - \theta\left(\frac{D_i A_i}{\eta} - \frac{D_i(1-A_i)}{1-\eta}\right) \ . \tag{5}$$

If we further assume instrument monotonicity (i.e., $P\{D_i(1) \geq D_i(0)\} = 1$), then θ_0 could be re-interpreted as the local average treatment effect (LATE) in the sense of Imbens and Angrist (1994). ■

Example 2.4 (Weighted Average Treatment Effect)**.** Let $Y_i(a) = R_i(a)$ denote a scalar potential outcome

for the ith unit under treatment $a \in \{0,1\}$, and let $Y_i = R_i$ denote the observed outcome. Let

$$\theta_0 = E_Q\left[\frac{\omega(X_i)}{E_Q[\omega(X_i)]}\left(Y_i(1) - Y_i(0)\right)\right] ,$$

for some known function $\omega : \mathbf{R}^{d_x} \to \mathbf{R}$. It then follows from Assumption 2.1 that θ_0 solves the moment condition in (3) with $m(\cdot)$ given by

$$m(X_i, A_i, R_i, \theta) = \omega(X_i)\left(\frac{Y_i A_i}{\eta} - \frac{Y_i(1-A_i)}{1-\eta}\right) - \omega(X_i)\theta .$$

Note that θ_0 defined in this way can accommodate the (cluster) size-weighted and equally-weighted average treatment effects considered in Bugni et al. (2022) and Bai et al. (2023c) in the context of cluster-level randomized controlled trials. ■

Example 2.5 (Log-Odds Ratio)**.** Let $Y_i(a) = R_i(a) \in \{0,1\}$ denote a binary potential outcome for the ith unit under treatment $a \in \{0,1\}$, and let $Y_i = R_i$ denote the observed outcome. Suppose $0 < P\{Y_i(a) = 0\} < 1$ for $a \in \{0,1\}$, and let $\theta_0 = (\theta_0(1), \theta_0(2))'$, where

$$\theta_0(1) = \text{logit}(E_Q[Y_i(0)]) ,$$

$$\theta_0(2) = \text{logit}(E_Q[Y_i(1)]) - \text{logit}(E_Q[Y_i(0)]) ,$$

with $\text{logit}(z) = \log(\frac{z}{1-z})$, so that $\theta_0(2)$ denotes the log-odds ratio of treatment 1 relative to treatment 0. It follows from Assumption 2.1 that θ_0 solves the moment condition in (3) with $m(\cdot)$ given by

$$m(X_i, A_i, R_i, \theta) = \begin{pmatrix} 1 - A_i \\ A_i \end{pmatrix}\left(Y_i - \text{expit}(\theta(1) + \theta(2)A_i)\right) ,$$

where $\text{expit}(z) = \frac{\exp(z)}{1+\exp(z)}$. The log-odds ratio can then be defined as $h(\theta_0)$ where $h : \mathbf{R}^2 \to \mathbf{R}$ is given by $h(s,t) = t$. ■

Additional examples could be obtained by considering combinations of Examples 2.1–2.5. For instance, combining the moment functions from Examples 2.3 and 2.4 would result in a weighted LATE parameter. Beyond these examples, certain treatment effect contrasts could also be related to the structural parameters in, for instance, a model of supply in demand: see, for example, the model estimated in Casaburi and Reed (2022).

Throughout the rest of the paper we consider the asymptotic properties of the method of moments estimator $\hat{\theta}_n$ for θ_0 which is constructed as a solution to the sample analogue of (3):

$$\frac{1}{n}\sum_{1 \le i \le n} m(X_i, A_i, R_i, \hat{\theta}_n) = 0 . \tag{6}$$

Note that $\hat{\theta}_n$ as defined in (6) is closely related to standard estimators of the parameter θ_0 in specific

examples. For instance, in Example 2.1,

$$\hat{\theta}_n = \frac{1}{\eta} \sum_{1 \leq i \leq n} Y_i A_i - \frac{1}{1-\eta} \sum_{1 \leq i \leq n} Y_i (1 - A_i) \ ,$$

so that $\hat{\theta}_n$ is a Horvitz-Thompson analogue of the standard difference-in-means estimator for the average treatment effect. In Example 2.3,

$$\hat{\theta}_n = \frac{\frac{1}{\eta} \sum_{1 \leq i \leq n} Y_i A_i - \frac{1}{1-\eta} \sum_{1 \leq i \leq n} Y_i (1 - A_i)}{\frac{1}{\eta} \sum_{1 \leq i \leq n} D_i A_i - \frac{1}{1-\eta} \sum_{1 \leq i \leq n} D_i (1 - A_i)} \ ,$$

so that $\hat{\theta}_n$ is a Horvitz-Thompson analogue of the standard Wald estimator for the local average treatment effect.

If $A^{(n)}$ were assigned i.i.d., independently of $X^{(n)}$, then it can be shown under mild conditions on $m(\cdot)$ (see, for instance, Theorem 5.1 in van der Vaart, 1998) that

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N(0, \mathbb{V}) \ ,$$

where

$$\mathbb{V} = M^{-1} E_P[m(X_i, A_i, R_i, \theta_0) m(X_i, A_i, R_i, \theta_0)'](M^{-1})' \ , \tag{7}$$

with $M = \frac{\partial}{\partial \theta} E_P[m(X, A, R, \theta)]\Big|_{\theta=\theta_0}$. In Section 3, we derive an efficiency bound $\mathbb{V}_*$ for estimating θ_0 for general assignment mechanisms which satisfy Assumption 2.1, and argue that in general $\mathbb{V} \geq \mathbb{V}_*$ (see Remark 3.2). For this reason, we call $\hat{\theta}_n$ the "naïve" method of moments estimator. It is now well understood that a more efficient estimator of θ_0 can be constructed by appropriately "augmenting" the moment function, and then considering an estimator which solves the augmented moment equation. For instance, if we consider Example 2.1, then it is straightforward to show that the following augmented moment function identifies θ_0:

$$m^*(X_i, A_i, R_i, \theta) = \left(\frac{A_i(Y_i - \mu_1(X_i))}{\eta} - \frac{(1 - A_i)(Y_i - \mu_0(X_i))}{1 - \eta} + \mu_1(X_i) - \mu_0(X_i) \right) - \theta \ , \tag{8}$$

where $\mu_a(X_i) = E_Q[Y_i(a)|X_i]$. This choice of $m^*(\cdot)$ produces the well known doubly-robust moment condition for estimating the ATE (Robins et al., 1995; Hahn, 1998). It can then be shown that an appropriately constructed two-step estimator will achieve the efficiency bound $\mathbb{V}_*$ (Tsiatis et al., 2008; Farrell, 2015; Chernozhukov et al., 2017; Rafi, 2023). Intuitively, the estimator obtained from the augmented moment function $m^*(\cdot)$ performs nonparametric regression adjustment by exploiting the information contained in $X^{(n)}$ which may not have been captured in the original moment function $m(\cdot)$. Similar nonparametric regression adjustments based on augmented moment equations have been developed for other parameters of interest (Zhang et al., 2008; Belloni et al., 2017; Jiang et al., 2022a,b). In Section 4, we show that if we assign $A^{(n)}$ using a "finely-stratified" design (i.e., a treatment assignment scheme which uses the covariates $X^{(n)}$ to block units into groups of fixed size: see Assumption 4.1 below for a formal definition), then it is possible to achieve the efficiency bound $\mathbb{V}_*$ that we derive in the subsequent section using the original "naïve" estimator $\hat{\theta}_n$. In this sense, we show that "fine stratification" can perform nonparametric regression adjustment "by

design" for the large class of parameters that can be expressed in terms of moment conditions of the form given in (3); this generalizes similar observations made in Bai et al. (2022), Bai (2022), and Cytrynbaum (2023b) in the special case of estimating the average treatment effect.

3 Efficiency Bound

In this section, we derive an efficiency bound for the class of parameters introduced in Section 2 under a general class of treatment assignment mechanisms. In what follows, when writing expectations and variances, we suppress the subscripts P and Q whenever doing so does not lead to confusion. We impose the following high-level assumption on the assignment mechanism:

Assumption 3.1. The treatment assignment mechanism is such that for any integrable Lipschitz functions $\gamma_0, \gamma_1 : \mathbf{R}^{d_x} \to \mathbf{R}$,

$$\frac{1}{n} \sum_{1 \leq i \leq n} \sum_{a \in \{0,1\}} I\{A_i = a\} \gamma_a(X_i) \xrightarrow{P} \eta E[\gamma_1(X_i)] + (1-\eta) E[\gamma_0(X_i)] \ .$$

In other words, Assumption 3.1 requires that the assignment mechanism admits a law of large numbers for "well-behaved" functions of the covariate values. Examples 3.1–3.2 illustrate that the assumption holds for a large class of treatment assignment mechanisms used in practice.

Example 3.1 (Covariate-adaptive randomization)**.** Let $S : \mathbf{R}^{d_x} \to \mathcal{S} = \{1, \ldots, |\mathcal{S}|\}$ be a function that maps the covariates into a set of discrete "strata." Assume that treatment status is assigned so that $(R^{(n)}(1), R^{(n)}(0), X^{(n)}) \perp\!\!\!\perp A^{(n)} \big| S^{(n)}$, and that for $s \in \mathcal{S}$,

$$\frac{\sum_{1 \leq i \leq n} I\{S_i = s, A_i = 1\}}{\sum_{1 \leq i \leq n} I\{S_i = s\}} \xrightarrow{P} \eta \ .$$

This high-level assumption subsumes stratified assignment mechanisms commonly used in empirical practice (see, for instance, Duflo et al., 2015; Dizon-Ross, 2019). It follows from Lemma C.4 in Bugni et al. (2019) that for any integrable functions γ_0, γ_1,

$$\frac{1}{n} \sum_{1 \leq i \leq n} \sum_{a \in \{0,1\}} I\{A_i = a\} \gamma_a(X_i) \xrightarrow{P} \sum_{s \in \mathcal{S}} P\{S_i = s\} (\eta E[\gamma_1(X_i) | S_i = s] + (1-\eta) E[\gamma_0(X_i) | S_i = s]) \ .$$

Therefore, Assumption 3.1 is satisfied. ■

Example 3.2 (Matched pairs)**.** Suppose n is even and consider pairing the experimental units into $n/2$ pairs, represented by the sets

$$\{\pi(2j-1), \pi(2j)\} \text{ for } j = 1, \ldots, n/2 \ ,$$

where $\pi = \pi_n(X^{(n)})$ is a permutation of n elements. Because of its possible dependence on $X^{(n)}$, π encompasses a broad variety of ways of pairing the n units according to the observed, baseline covariates $X^{(n)}$. Given such a π, we assume that treatment status is assigned so that $(R^{(n)}(1), R^{(n)}(0)) \perp\!\!\!\perp A^{(n)} \big| X^{(n)}$ and,

conditional on $X^{(n)}$, $(A_{\pi(2j-1)}, A_{\pi(2j)}), j = 1, \ldots, n/2$ are i.i.d. and each uniformly distributed over the values in $\{(0,1),(1,0)\}$. For some examples of such an assignment mechanism being used in practice, see, for instance, Angrist and Lavy (2009), Banerjee et al. (2015), and Bruhn et al. (2016). Assume that the pairing algorithm $\pi_n(X^{(n)})$ results in pairs that are "close" in the sense of Assumption 2.3 in Bai et al. (2022). It then follows from the proof of Lemma S.1.5 of of Bai et al. (2022) that

$$\frac{1}{n}\sum_{1\leq i\leq n}\sum_{a\in\{0,1\}} I\{A_i = a\}\gamma_a(X_i) \xrightarrow{P} \frac{1}{2}E[\gamma_0(X_i)] + \frac{1}{2}E[\gamma_1(X_i)] \ .$$

Therefore, Assumption 3.1 is satisfied. ■

Next, we impose the following high-level assumption on the distributions Q and P:

Assumption 3.2. The distributions Q and P are such that

(a) $\mathrm{Var}[m(X_i, a, R_i(a), \theta_0)|X_i = x]$ is a Lipschitz function.

(b) θ_0 is uniquely determined by (3) and $M = \frac{\partial}{\partial\theta}E[m(X_i, A_i, R_i, \theta)]\Big|_{\theta=\theta_0}$ is invertible.

Assumption 3.2(a) is a smoothness condition that is required in settings where X_i is continuous to ensure that the function $\psi^*(\cdot)$ we derive in Theorem 3.1 below is in fact the efficient influence function. Note that the assumption is trivially satisfied if the support of X_i is discrete. Assumption 3.2(b) is a standard assumption used when deriving the properties of Z-estimators (see, for instance, Theorem 5.1 in van der Vaart, 1989).

We now present the first main result of the paper: an efficiency bound for the parameter θ_0 introduced in Section 2. Formally, we characterize the bound via a convolution theorem which applies to all "regular" estimators of the parameter θ_0, where "regular" here should be understood in the standard sense necessary to rule out, for instance, super-efficient estimators (see, for instance, Example 8.1 in van der Vaart, 1998). In stating our theorem we leave the precise definition of "regular" and related assumptions to Appendix A.1. In the paragraph following the statement of the theorem we provide some more details on the nature of our result.

Theorem 3.1. *Suppose Assumptions 2.1 and 3.1–3.2 hold, and maintain the additional regularity conditions (16), (17) and Assumption A.1 described in Appendix A.1. Let $\tilde{\theta}_n$ be any "regular" estimator of the parameter θ_0 in the sense of (20) in Appendix A.1. Then,*

$$\sqrt{n}(\tilde{\theta}_n - \theta_0) \xrightarrow{d} L \ ,$$

where

$$L = N(0, E[\psi^*\psi^{*\prime}]) * B \ ,$$

for some fixed probability measure B which is specific to the estimator $\tilde{\theta}_n$, with

$$\begin{aligned}&\psi^*(X_i, A_i, R_i, \theta_0)\\ &= -M^{-1}\Big(I\{A_i = 1\}(m(X_i, 1, R_i, \theta_0) - E[m(X_i, 1, R_i(1), \theta_0)|X_i])\end{aligned}$$

$$+ I\{A_i = 0\}(m(X_i, 0, R_i, \theta_0) - E[m(X_i, 0, R_i(0), \theta_0)|X_i])$$
$$+ \eta E[m(X_i, 1, R_i(1), \theta_0)|X_i] + (1-\eta)E[m(X_i, 0, R_i(0), \theta_0)|X_i]\Big) \ ,$$

and $M = \frac{\partial}{\partial \theta} E[m(X, A, R, \theta)]\Big|_{\theta=\theta_0}$.

Given Theorem 3.1 we call $\mathbb{V}_* = \text{Var}[\psi^*(X_i, A_i, R_i, \theta_0)]$ the efficiency bound for θ_0, since our result shows that this is the lowest asymptotic variance attainable by any regular estimator under our assumptions. We note that our assumptions on the assignment mechanism preclude us from applying results based on "standard" arguments (see, for instance, van der Vaart, 1998). Specifically, if we define a tangent set as the collection of score functions of "smooth" one-dimensional parametric sub-models in an appropriate sense, then we are not able to guarantee that the resulting tangent set is linear (or even a convex cone) while *simultaneously* verifying that the likelihood ratio process is locally asymptotically normal for arbitrary assignment mechanisms which satisfy Assumption 3.1. Instead, we proceed by justifying an application of Corollary 3.1 in Armstrong (2022) combined with the convolution Theorem 3.11.2 in van der Vaart and Wellner (1996) to each d_θ-dimensional parametric submodel separately, and then arguing that the supremum over all such submodels is attained by $\text{Var}[\psi^*]$ under Assumption 3.2.

Remark 3.1. Note it follows from (3) that

$$\eta E_Q[m(X_i, 1, R_i(1), \theta_0)] + (1-\eta)E_Q[m(X_i, 0, R_i(0), \theta_0)] = E_P[m(X_i, A_i, R_i, \theta_0)] = 0 \ ,$$

so that $E[\psi^*(X_i, A_i, R_i, \theta_0)] = 0$. It is further straightforward to show using Assumption 2.1 that

$$\begin{aligned} \mathbb{V}_* &= \text{Var}[\psi^*(X_i, A_i, R_i, \theta_0)] \\ &= M^{-1}\big(E\big[\eta \,\text{Var}[m(X_i, 1, R_i(1), \theta_0)|X_i] + (1-\eta)\,\text{Var}[m(X_i, 0, R_i(0), \theta_0)|X_i]\big] \\ &\qquad + \text{Var}\left[\eta E[m(X_i, 1, R_i(1), \theta_0)|X_i] + (1-\eta)E[m(X_i, 0, R_i(0), \theta_0)|X_i]\right]\big)(M^{-1})' \end{aligned} \tag{9}$$

From this we can deduce that our efficiency bound recovers well-known bounds for common parameters (like those presented in Examples 2.1–2.3) in the setting of i.i.d. assignment. For instance, in the case of the ATE (Example 2.1) we obtain that

$$\text{Var}[\psi^*(X_i, A_i, R_i, \theta_0)] = E\left[\frac{\text{Var}[Y_i(1)|X_i]}{\eta} + \frac{\text{Var}[Y_i(0)|X_i]}{1-\eta} + (E[Y_i(1) - Y_i(0)|X_i] - E[Y_i(1) - Y_i(0)])^2\right] \ ,$$

which matches the efficiency bound under i.i.d. assignment derived in Hahn (1998). See Armstrong (2022) and Rafi (2023) for related results in the context of stratified and adaptive experiments. Straightforward calculation also implies for the QTE (Example 2.2) that

$$\text{Var}[\psi^*(X_i, A_i, R_i, \theta_0)] = E\Bigg[\frac{1}{\eta}\frac{F_1\big(\theta_0(1)|X\big)\big(1 - F_1\big(\theta_0(1)|X\big)\big)}{f_1\big(\theta_0(1)\big)^2} + \frac{1}{1-\eta}\frac{F_0\big(\theta_0(0)|X\big)\big(1 - F_0\big(\theta_0(0)|X\big)\big)}{f_0\big(\theta_0(0)\big)^2}$$
$$+ \left(\frac{F_1\big(\theta_0(1)|X\big) - \tau}{f_1\big(\theta_0(1)\big)} - \frac{F_0\big(\theta_0(0)|X\big) - \tau}{f_0\big(\theta_0(0)\big)}\right)^2\Bigg] \ ,$$

which matches the efficiency bound under i.i.d. assignment derived in Firpo (2007) when the propensity score is set to η. ■

Remark 3.2. Note that by comparing the variance expression in (7) to the variance expression for $\mathbb{V}_*$, we obtain

$$\mathbb{V} - \mathbb{V}_* = \eta(1-\eta)M^{-1}E[(E[m(X,1,R(1)|X] - E[m(X,0,R(0)|X]) \\ \times (E[m(X,1,R(1)|X] - E[m(X,0,R(0)|X])'](M^{-1})' \ ,$$

which is positive semidefinite. From this we conclude that the variance of the "naive" method of moments estimator $\hat{\theta}_n$ is generally inefficient when $A^{(n)}$ is assigned using i.i.d. assignment. ■

Remark 3.3. Although we focus on the case where $\eta_i(X_i) = P\{A_i = 1|X_i\} = \eta$ is a constant, the proof of Theorem 3.1 holds when $\eta_i(x) = \eta(x)$ for $1 \leq i \leq n$, where $\eta(x)$ is an arbitrary known and fixed function. In these settings, Lemma A.4 shows that the efficiency bound equals

$$\begin{aligned} \mathbb{V}_* &= \text{Var}[\psi^*(X_i, A_i, R_i, \theta_0)] \\ &= M^{-1}\big(E\big[\eta(X_i)\,\text{Var}[m(X_i,1,R_i(1),\theta_0)|X_i] + (1-\eta(X_i))\,\text{Var}[m(X_i,0,R_i(0),\theta_0)|X_i]\big] \\ &\quad + \text{Var}\left[\eta(X_i)E[m(X_i,1,R_i(1),\theta_0)|X_i] + (1-\eta(X_i))E[m(X_i,0,R_i(0),\theta_0)|X_i]\right]\big)(M^{-1})' \ , \end{aligned} \tag{10}$$

so that the only difference from (9) is that η is replaced by $\eta(X_i)$. Consider Example 2.1 and note the moment condition for the ATE is now given by

$$m(X_i, A_i, R_i, \theta) = \frac{Y_i A_i}{\eta(X_i)} - \frac{Y_i(1-A_i)}{1-\eta(X_i)} - \theta \ . \tag{11}$$

Straightforward calculation implies that in this example, the efficiency bound in (10) becomes

$$E\left[\frac{\text{Var}[Y_i(1)|X_i]}{\eta(X_i)} + \frac{\text{Var}[Y_i(0)|X_i]}{1-\eta(X_i)} + (E[Y_i(1) - Y_i(0)|X_i] - E[Y_i(1) - Y_i(0)])^2\right] \ , \tag{12}$$

which again matches the efficiency bound under i.i.d. assignment in Hahn (1998). If we additionally impose that $\eta(X_i) = \eta(S(X_i))$ for S taking on finitely many values as in Example 3.1, then this expression matches the bound derived in Theorem 3.1 in Rafi (2023). ■

Remark 3.4. Here, we comment on how Theorem 3.1 relates to prior efficiency bounds in experiments with general assignment mechanisms. Armstrong (2022) derives an efficiency bound for the average treatment effect over a very large class of assignment mechanisms, which includes for instance response-adaptive designs. However, his bound leaves the assignment proportions completely unrestricted. As a consequence, his bound is necessarily loose whenever the assignment proportions are exogenously constrained away from the Neyman allocation (for instance if the assignment proportions were set to one half regardless of whether or not the conditional outcome variances across treatment and control are equal). In contrast, Rafi (2023) derives an efficiency bound for the average treatment effect over the class of all stratified assignment mechanisms in the sense of Bugni et al. (2019), where the stratum-level assignment proportions are restricted *a priori* by the

experimenter. Our analysis, like Rafi (2023), exogenously constrains the marginal probability of assignment, but permits a richer class of assignment schemes, including the "finely stratified" designs that we consider in the next section. We note, however, that this comes as the cost of requiring modestly stronger assumptions on the class of data generating processes (via Assumption 3.1). Finally, we once again emphasize that our analysis applies to a general class of treatment effect parameters, including the average treatment effect as a special case. ■

4 The Asymptotic Variance of Finely Stratified Experiments

In this section, we derive the limiting distribution of the method of moments estimator $\hat{\theta}_n$ when treatment is assigned by fine stratification over the baseline covariates $X^{(n)}$. As mentioned previously, such assignment mechanisms use the covariates $X^{(n)}$ to block units with "similar" covariate values into groups of fixed size, and then assign treatment completely at random within each block. In order to describe this assignment mechanism formally, we require some further notation to define the blocks of units. Let ℓ and k be arbitrary positive integers with $\ell \leq k$ and set $\eta = \ell/k$. For simplicity, assume that n is divisible by k. We then represent blocks of units using a partition of $\{1, \ldots, n\}$ given by

$$\left\{\lambda_j = \lambda_j(X^{(n)}) \subseteq \{1, \ldots, n\}, 1 \leq j \leq n/k\right\} ,$$

with $|\lambda_j| = k$. Because of its possible dependence on $X^{(n)}$, $\{\lambda_j : 1 \leq j \leq n/k\}$ encompasses a variety of different ways of blocking the n units according to the observed, baseline covariates. Given such a partition, we assume that treatment status is assigned as described in the following assumption:

Assumption 4.1. Treatment status is assigned so that $(R^{(n)}(1), R^{(n)}(0)) \perp\!\!\!\perp A^{(n)} \big| X^{(n)}$ and, conditional on $X^{(n)}$,

$$\{(A_i : i \in \lambda_j) : 1 \leq j \leq n/k\}$$

are i.i.d. and each uniformly distributed over all permutations of $(\underbrace{0, 0, \ldots, 0}_{k-\ell}, \underbrace{1, 1, \ldots, 1}_{\ell})$.

Remark 4.1. The assignment mechanism described in Assumptions 4.1 generalizes the definition of a matched pair design (Example 3.2). In particular, we recover a matched pair design if we set $(\ell, k) = (1, 2)$, with $\eta = 1/2$. Note that Assumption 4.1 generalizes matched pair designs along two dimensions: first, it allows for treatment fractions other than $\eta = 1/2$. Second, it allows for choices of ℓ and k which are not relatively prime. For instance, if we set $(\ell, k) = (2, 4)$, then $\eta = 1/2$ as in matched pairs, but now the assignment mechanism blocks units into groups of size 4 and assigns two units to treatment, two units to control. Although Theorem 4.1 below establishes that allowing for this level of flexibility has no effect on the asymptotic properties of our estimator, in our experience we have found that designs which employ these treatment "replicates" in each block can simplify the construction of variance estimators in practice. See Remark 4.4 below for further discussion. ■

Our analysis will require some discipline on the way in which the blocks are formed. In particular, we

will require that the units in each block be "close" in terms of their baseline covariates in the sense described by the following assumption:

Assumption 4.2. The blocks used in determining treatment status satisfy

$$\frac{1}{n}\sum_{1\leq j\leq n/k}\max_{i,i'\in\lambda_j}\|X_i - X_{i'}\|^2 \xrightarrow{P} 0\ .$$

Bai et al. (2022) and Cytrynbaum (2023b) discuss blocking algorithms which satisfy Assumption 4.2. When $X_i \in \mathbf{R}$ and $E[X_i^2] < \infty$, a simple algorithm which satisfies Assumption 4.2 is to simply order units from smallest to largest and then block adjacent units into blocks of size k.

The next two sets of assumptions allow us to derive the large sample properties of $\hat{\theta}_n$. We impose Assumption 4.3 to establish the consistency of $\hat{\theta}_n$, and we further impose Assumption 4.4 to establish its limiting distribution.

Assumption 4.3. Let $m(\cdot) = (m_s(\cdot) : 1 \leq s \leq d_\theta)'$. Then the moment functions are such that

(a) For every $\epsilon > 0$, $\inf_{\theta\in\Theta:\|\theta-\theta_0\|>\epsilon}\|E[m(X_i, A_i, R_i, \theta)]\| > 0$.

(b) For $1 \leq s \leq d_\theta$, $\{m_s(x, a, r, \theta) : \theta \in \Theta\}$ with $a \in \{0, 1\}$ fixed is a VC-class of functions in (x, r).

(c) For $1 \leq s \leq d_\theta$, $\{m_s(x, a, r, \theta) : \theta \in \Theta\}$ is pointwise measurable in the sense that there exists a countable set Θ^* such that for each $\theta \in \Theta$, there exists a sequence $\{\theta_m\} \subset \Theta^*$ such that $m_s(x, a, r, \theta_m) \to m_s(x, a, r, \theta)$ as $m \to \infty$ for all x, a, r.

(d) $E\left[\sup_{\theta\in\Theta^*}\|m(X, a, R(a), \theta)\|\right] < \infty$ for $a \in \{0, 1\}$.

(e) For some $K < \infty$,

$$\sup_{\theta\in\Theta^*}\|E[m(X, a, R(a), \theta)|X = x] - E[m(X, a, R(a), \theta)|X = x']\| \leq K\|x - x'\|$$

for all $x, x' \in \mathbf{R}^{d_x}$.

(f) $E\left[\sup_{\theta\in\Theta^*}\left\|m(X, a, R(a), \theta)\right\|\middle| X = x\right]$ is Lipschitz for $a \in \{0, 1\}$.

Assumption 4.3(a) is a standard assumption to ensure the solution to (3) is "well separated." It appears as a condition, for instance, in Theorem 5.9 in van der Vaart (1998) . Assumption 4.3(b) can be readily verified in Examples 2.1–2.5 because the moment conditions are either constructed as linear functions in θ (multiplied or composed with fixed functions), or dependent on θ through indicator functions. Assumption 4.3(c) is a standard condition to guarantee the measurability of the supremum of a suitable class of functions. In particular, it allows us to define expectations of suprema without invoking outer expectations. See Example 2.3.4 in van der Vaart and Wellner (1996) for details. Assumption 4.3(d) guarantees the existence of an envelope function needed to establish a uniform law of large numbers. Assumptions 4.3(e)–(f) mirror common assumptions used when studying matched pairs designs to ensure units that are close in terms of the baseline covariates are also close in terms of their moments.

Assumption 4.4. Let $m(\cdot) = (m_s(\cdot) : 1 \leq s \leq d_\theta)'$. The moment functions are such that

(a) $E[m(X_i, A_i, R_i, \theta)]$ is differentiable at θ_0 with a nonsingular derivative M.

(b) For Θ^* in Assumption 4.3(c), $E\left[\sup_{\theta \in \Theta^*} \left\|m(X, a, R(a), \theta)\right\|^2\right] < \infty$ for $a \in \{0, 1\}$.

(c) For $1 \leq s \leq d_\theta$, $E[((m_s(X, a, R(a), \theta) - m_s(X, a, R(a), \theta_0))^2] \to 0$ as $\theta \to \theta_0$ for $a \in \{0, 1\}$.

(d) For $1 \leq s \leq d_\theta$, $\{E[(m_s(X, a, R(a), \theta)|X = x] : \theta \in \Theta\}$ is a VC-class of functions for $a \in \{0, 1\}$.

(e) For Θ^* in Assumption 4.3(c) and some $K < \infty$, for $1 \leq s \leq d_\theta$,

$$\sup_{\theta \in \Theta^*} |E[m_s^2(X, a, R(a), \theta)|X = x] - E[m_s^2(X, a, R(a), \theta)|X = x']| \leq K\|x - x'\|$$
$$\sup_{\theta \in \Theta^*} |E[m_s(X, a, R(a), \theta)m_s(X, a, R(a), \theta_0)|X = x] - E[m_s(X, a, R(a), \theta)m_s(X, a, R(a), \theta_0)|X = x']| \leq K\|x - x'\| ,$$

for all $x, x' \in \mathbf{R}^{d_x}$, $a \in \{0, 1\}$

(f) For Θ^* in Assumption 4.3(c), $E\left[\sup_{\theta \in \Theta^*} \left\|m(X, a, R(a), \theta)\right\|^2 \middle| X = x\right]$ is Lipschitz for $a \in \{0, 1\}$.

Assumption 4.4(a) subsumes Assumption 3.2(b). See, for instance, Theorem 3.1 in Newey and McFadden (1994) and Theorem 5.21 in van der Vaart (1998). Because differentiability is imposed on their expectations instead of the moment functions themselves, the moment functions are allowed to be nonsmooth as in Example 2.2. Assumption 4.4(b) guarantees the existence of an envelope function needed to establish a uniform law of large numbers. Assumption 4.4(c) implies the moment functions are mean-square continuous in θ. Assumptions 4.4(e)–(f) again mirror common assumptions used when studying matched pairs to ensure units that are close in terms of the baseline covariates are also close in terms of their moments. Assumption 4.4(d) is again readily verified in Examples 2.1, 2.3–2.5 because θ enters separably in these examples. To verify the assumption for Example 2.2, note that for any random variables $Y(a), X$, the subgraphs $\{(x, t) : t < P\{Y(a) \leq \theta(a)|X = x\}\}$ are linearly ordered in $\theta(a)$ because the conditional distribution function is increasing in $\theta(a)$. Therefore, the class of subgraphs is VC with index 2 (see, for instance, the last sentence of the proof of Lemma 2.6.16 in van der Vaart and Wellner, 1996).

Remark 4.2. We note that some of the assumptions imposed in Assumptions 4.3 and 4.4 are seemingly more stringent than the low-level conditions considered in previous papers which study inference for certain specific parameters of interest under matched pair designs (Bai et al., 2022; Jiang et al., 2021; Cytrynbaum, 2023b; Bai et al., 2023a). We suspect that, with more delicate arguments, some of these assumptions could be weakened for specific parameters of interest, but we do not pursue this in the paper. ■

The following theorem establishes that the "naïve" method of moments estimator attains the efficiency bound $\mathbb{V}_*$ when the treatment assignment mechanism is finely stratified in the sense of satisfying Assumptions 4.1–4.2. This finding contrasts with the discussion in Remark 3.2 which showed that the variance of the naive estimator is generally inefficient under i.i.d. treatment assignment.

Theorem 4.1. *Suppose the treatment assignment mechanism satisfies Assumptions 4.1–4.2 and the moment functions satisfy Assumptions 4.3–4.4. Let $\hat{\theta}_n$ be defined as in (6). Then,*

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = \frac{1}{\sqrt{n}} \sum_{1 \leq i \leq n} \psi^*(X_i, A_i, R_i, \theta_0) + o_P(1) \ . \tag{13}$$

Further, we have that

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N(0, \mathbb{V}_*) \ . \tag{14}$$

Remark 4.3. Although Theorem 4.1 is focused on the case where $\eta(X_i) = \eta$ is a constant, straightforward modifications of the treatment assignment mechanism described in Assumptions 4.1–4.2 can attain the efficiency bound in more general settings. For instance, suppose $\eta(X_i)$ takes on a fixed discrete set of values $\{\eta_1, \ldots, \eta_S\}$, we could then simply implement a finely stratified experiment over each set $\{i : \eta(X_i) = \eta_s\}$ for $1 \leq s \leq S$. In other words, separately within each stratum defined by the units for which $\eta(X_i) = \eta_s$, employ the assignment mechanism described in Assumptions 4.1–4.2 with $\ell/k = \eta_s$. For more general functions $\eta(X_i)$, we conjecture that the efficiency bound could be attained by employing the local randomization procedure proposed in Cytrynbaum (2023b). ■

Remark 4.4. It is possible to construct a consistent estimator for $\mathbb{V}_*$ based on the variance formula in (9). Here, we briefly describe a construction when θ_0 is a scalar and refer interested readers to Bai et al. (2022), Bai (2022), and Bai et al. (2023d) for formal arguments which could be used to prove its validity. First note that in certain examples (including Examples 2.1 and 2.3–2.5), the analog principle suggests that a natural estimator for M is given by

$$\widehat{M}_n = \frac{1}{n} \sum_{1 \leq i \leq n} \frac{\partial}{\partial \theta} m(X_i, A_i, R_i, \theta) \Big|_{\theta = \hat{\theta}_n} \ .$$

Under suitable conditions, it follows directly from following arguments in each of the papers mentioned above that $\widehat{M}_n \xrightarrow{P} M$.[2] Therefore, it suffices to construct a consistent estimator for the "meat" in (9). By the law of total variance, this middle component equals $\Sigma_1 + \Sigma_2$, where

$$\begin{aligned}
\Sigma_1 &= \eta \operatorname{Var}[m(X_i, 1, R_i(1), \theta_0)] + (1 - \eta) \operatorname{Var}[m(X_i, 0, R_i(0), \theta_0)] \\
\Sigma_2 &= -\eta(1-\eta) E\big[\big(E[m(X_i, 1, R_i(1), \theta_0)|X_i] - E[m(X_i, 1, R_i(1), \theta_0)] \\
&\qquad - (E[m(X_i, 0, R_i(0), \theta_0)|X_i] - E[m(X_i, 0, R_i(0), \theta_0)])\big)^2\big] \\
&= -\eta(1-\eta)\Big(E[E[m(X_i, 1, R_i(1), \theta_0)|X_i]^2] + E[E[m(X_i, 0, R_i(0), \theta_0)|X_i]^2] \\
&\qquad - 2E[E[m(X_i, 1, R_i(1), \theta_0)|X_i] E[m(X_i, 0, R_i(0), \theta_0)|X_i]] \\
&\qquad - (E[m(X_i, 1, R_i(1), \theta_0)] - E[m(X_i, 0, R_i(0), \theta_0)])^2\Big)
\end{aligned}$$

For $a \in \{0, 1\}$, define

$$\hat{\mu}_n(a) = \frac{1}{\eta_a n} \sum_{1 \leq i \leq n} I\{A_i = a\} m(X_i, A_i, R_i, \hat{\theta}_n) \ ,$$

[2] In examples including Example 2.2 where m is nonsmooth in θ, M may consist of components which require nonparametric estimators, and in such cases bootstrap procedures may be preferable. See, for instance, Jiang et al. (2021).

where $\eta_1 = \eta$ and $\eta_0 = 1 - \eta$. The analog principle again suggests that a natural estimator for Σ_1 is

$$\hat{\Sigma}_{1,n} = \frac{1}{n} \sum_{1 \leq i \leq n} I\{A_i = 1\}(m(X_i, A_i, R_i, \hat{\theta}_n) - \hat{\mu}_n(1))^2 + \frac{1}{n} \sum_{1 \leq i \leq n} I\{A_i = 0\}(m(X_i, A_i, R_i, \hat{\theta}_n) - \hat{\mu}_n(0))^2 \ .$$

To estimate Σ_2, we first define

$$\hat{\varsigma}_n(0,1) = \frac{k}{n} \sum_{1 \leq j \leq n/k} \frac{1}{\ell(k-\ell)} \sum_{i,i' \in \lambda_j : A_i = 1, A_{i'} = 0} m(X_i, A_i, R_i, \hat{\theta}_n) m(X_{i'}, A_{i'}, R_{i'}, \hat{\theta}_n) \ .$$

Next, define

$$\hat{\varsigma}_n(1,1) = \begin{cases} \frac{k}{n} \sum_{1 \leq j \leq n/k} \frac{1}{\binom{\ell}{2}} \sum_{i < i' \in \lambda_j : A_i = A_{i'} = 1} m(X_i, A_i, R_i, \hat{\theta}_n) m(X_{i'}, A_{i'}, R_{i'}, \hat{\theta}_n) & \text{if } \ell > 1 \\ \frac{k}{2n} \sum_{1 \leq j \leq \frac{n}{2k}} \sum_{i \in \lambda_{2j}, i' \in \lambda_{2j-1} : A_i = A_{i'} = 1} m(X_i, A_i, R_i, \hat{\theta}_n) m(X_{i'}, A_{i'}, R_{i'}, \hat{\theta}_n) & \text{if } \ell = 1 \ . \end{cases}$$

Similarly, define

$$\hat{\varsigma}_n(0,0) = \begin{cases} \frac{k}{n} \sum_{1 \leq j \leq n/k} \frac{1}{\binom{k-\ell}{2}} \sum_{i < i' \in \lambda_j : A_i = A_{i'} = 0} m(X_i, A_i, R_i, \hat{\theta}_n) m(X_{i'}, A_{i'}, R_{i'}, \hat{\theta}_n) & \text{if } k - \ell > 1 \\ \frac{k}{2n} \sum_{1 \leq j \leq \frac{n}{2k}} \sum_{i \in \lambda_{2j}, i' \in \lambda_{2j-1} : A_i = A_{i'} = 0} m(X_i, A_i, R_i, \hat{\theta}_n) m(X_{i'}, A_{i'}, R_{i'}, \hat{\theta}_n) & \text{if } k - \ell = 1 \ . \end{cases}$$

Finally, define

$$\hat{\Sigma}_{2,n} = -\eta(1-\eta)(\hat{\varsigma}_n(1,1) + \hat{\varsigma}_n(0,0) - 2\hat{\varsigma}_n(0,1) - (\hat{\mu}_n(1) - \hat{\mu}_n(0))^2) \ .$$

The estimator $\hat{\varsigma}_n(1,1)$ is constructed in one of two ways depending on the number of treated units in each block. If more than one unit in each block is treated, then we take the averages of all pairwise products of the treated units in each block, and average them across all blocks. We call this a "within block" estimator. If instead only one unit in each block is treated, then we take the product of two treated units in *adjacent* blocks. We call this a "between block" estimator, and note that similar constructions have been used previously in Abadie and Imbens (2008), Bai et al. (2022), Bai et al. (2023c), and Cytrynbaum (2023b). The estimator $\hat{\varsigma}_n(0,0)$ is constructed similarly. Bai et al. (2023d) compare the finite-sample properties of the "within block" and "between block" variance estimators via simulation. Their findings are that experimental designs which allow for a "within block" variance estimator have better small sample inferential performance, at the cost of slightly increasing the mean-squared error of the estimator $\hat{\theta}_n$, relative to experimental designs which require the use of the "between block" variance estimator. Under suitable assumptions, it follows from similar arguments to those in Bai (2022) and Bai et al. (2023d) that $\hat{\Sigma}_{1,n} \xrightarrow{P} \Sigma_1$ and $\hat{\Sigma}_{2,n} \xrightarrow{P} \Sigma_2$. A natural estimator for $\mathbb{V}_*$ is therefore given by

$$\hat{\mathbb{V}}_n = \widehat{M}_n^{-2} \left(\hat{\Sigma}_{1,n} + \hat{\Sigma}_{2,n} \right) \ .$$

Thus, provided M is invertible, we have that $\hat{\mathbb{V}}_n \xrightarrow{P} \mathbb{V}_*$. ■

5 Simulations

In this section, we illustrate the results in Sections 3 and 4 with a simulation study. Specifically, we set $\eta = 1/2$, and compare the mean-squared errors (MSE) obtained from the "naive" estimator $\hat{\theta}_n$ and various adjusted estimators, for i.i.d. treatment assignment versus matched pairs assignment (see Example 3.2 and Remark 4.1). In Section 5.1, we present the model specifications and estimators for estimating the ATE as in Example 2.1. In Section 5.2, we present the model specifications and estimators for estimating the LATE as in Example 2.3. Section 5.3 reports the simulation results.

5.1 Average Treatment Effect

In this section, we present model specifications and estimators for estimating the ATE as in Example 2.1. Recall that in this case the moment function we consider is given by

$$m(X_i, R_i, A_i, \theta) = \frac{Y_i A_i}{\eta} - \frac{Y_i(1 - A_i)}{1 - \eta} - \theta \ ,$$

with $R_i = Y_i$. For $a \in \{0, 1\}$ and $1 \leq i \leq n$, the potential outcomes are generated according to the equation:

$$Y_i(a) = \mu_a(X_i) + \sigma_a(X_i)\epsilon_i \ .$$

In each of the specifications, $((X_i, \epsilon_i) : 1 \leq i \leq n)$ are i.i.d; for $1 \leq i \leq n$, X_i and ϵ_i are independent.

Model 1: $\mu_0(X_i) = X_i + (X_i^2 - 1)/3$, $\mu_1(X_i) = 0.2 + \mu_0(X_i)$, $\epsilon_i \sim N(0, 1)$, $X_i \sim N(0, 1)$ and $\sigma_a(X_i) = 2$.

Model 2: As in Model 1, but $\mu_a(X_i) = 0.2I\{a = 1\} + \gamma_a(\sin(X_i) + X_i) + (X_i^2 - 1)/3$ where $\gamma_1 = 1$ and $\gamma_0 = -1$, and $\sigma_a(X_i) = (1 + a)X_i^2$.

Model 3: As in Model 2, but $\mu_1(X_i) = 0.2 + 3(X_i^2 - 1)$ and $\mu_0(X_i) = 0$.

We consider the following three estimators for the ATE:

Unadjusted Estimator:

$$\hat{\theta}_n^{\text{unadj}} = \frac{1}{n/2} \sum_{1 \leq i \leq n} (Y_i A_i - Y_i(1 - A_i)) \ .$$

Adjusted Estimator 1:

$$\hat{\theta}_n^{\text{adj},1} = \frac{1}{n} \sum_{1 \leq i \leq n} \left(2A_i(Y_i - \hat{\mu}_1^Y(X_i)) - 2(1 - A_i)(Y_i - \hat{\mu}_0^Y(X_i)) + \hat{\mu}_1^Y(X_i) - \hat{\mu}_0^Y(X_i)\right) \ ,$$

where $\hat{\mu}_a^Y(X_i)$ is constructed by running a least squares regression of Y_i on $(1, X_i, X_i^2)$ using the sample from $A_i = a$.

Adjusted Estimator 2:

$$\hat{\theta}_n^{\text{adj},2} = \frac{1}{n} \sum_{1 \le i \le n} \left(2A_i(Y_i - \hat{\mu}_1^Y(X_i)) - 2(1-A_i)(Y_i - \hat{\mu}_0^Y(X_i)) + \hat{\mu}_1^Y(X_i) - \hat{\mu}_0^Y(X_i) \right) ,$$

where $\hat{\mu}_a^Y(X_i)$ is constructed by running a least squares regression of Y_i on $(1, X_i, X_i^2, X_i 1\{X_i > t\})$ where t is the sample median using the sample from $A_i = a$.

The first estimator $\hat{\theta}_n^{\text{unadj}}$ is the method of moments estimator given by the solution to (6). The second and third estimators $\hat{\theta}_n^{\text{adj},1}$ and $\hat{\theta}_n^{\text{adj},2}$ are covariate-adjusted estimators which can be obtained as two-step method of moments estimators from solving the "augmented" moment equation (8) described in the discussion at the end of Section 2. $\hat{\theta}_n^{\text{adj},1}$ and $\hat{\theta}_n^{\text{adj},2}$ differ in the choice of basis functions used in the construction of the estimators $\hat{\mu}_a(x)$. Note that by the double-robustness property of the augmented estimating equation (8), it can be shown that the adjusted estimators $\hat{\theta}_n^{\text{adj},1}$, $\hat{\theta}_n^{\text{adj},2}$ are consistent and asymptotically normal regardless of the choice of estimators $\hat{\mu}_a(x)$, but consistency of $\hat{\mu}_a(x)$ to $\mu_a(x)$ would ensure that $\hat{\theta}_n^{\text{adj},1}$, $\hat{\theta}_n^{\text{adj},2}$ are efficient under i.i.d. assignment (Robins et al., 1995; Tsiatis et al., 2008; Chernozhukov et al., 2017).

5.2 Local Average Treatment Effect

In this section, we present the model specifications and estimators for estimating the LATE as in Example 2.3. Recall that in this case the moment condition we consider is given by

$$m(X_i, A_i, R_i, \theta) = \frac{Y_i A_i}{\eta} - \frac{Y_i(1-A_i)}{1-\eta} - \theta \left(\frac{D_i A_i}{\eta} - \frac{D_i(1-A_i)}{1-\eta} \right) ,$$

with $R_i = (Y_i, D_i)$. The outcome is determined by the relationship $Y_i = D_i Y_i(1) + (1-D_i)Y_i(0)$, where $Y_i(d) = \mu_d(X_i) + \sigma_a(X_i)\epsilon_i$ follows the same outcome model as in the ATE setup of Section 5.1. In addition, we have $D_i = A_i D_i(1) + (1-A_i)D_i(0)$, where

$$D_i(0) = I\left\{\alpha_0 + \alpha\left(X_i\right) > \varepsilon_{1,i}\right\} ,$$
$$D_i(1) = \begin{cases} I\left\{\alpha_1 + \alpha\left(X_i\right) > \varepsilon_{2,i}\right\} & \text{if } D_i(0) = 0 \\ 1 & \text{otherwise} \end{cases} .$$

For each outcome model, we set $\alpha_0 = 0.5$, $\alpha_1 = 1$, $\alpha(X_i) = X_i + (X_i^2 - 1)/3$ and $\varepsilon_{1,i}, \varepsilon_{2,i} \sim N(0, 4)$.

We consider the following three estimators for the LATE:

Unadjusted Estimator:

$$\hat{\theta}_n^{\text{unadj}} = \frac{\sum_{1 \le i \le n} \left(Y_i A_i - Y_i(1-A_i)\right)}{\sum_{1 \le i \le n} \left(D_i A_i - D_i(1-A_i)\right)} .$$

Adjusted Estimator 1:

$$\hat{\theta}_n^{\text{adj},1} = \frac{\sum_{1 \le i \le n} \left(2A_i(Y_i - \hat{\mu}_1^Y(X_i)) - 2(1-A_i)(Y_i - \hat{\mu}_0^Y(X_i)) + \hat{\mu}_1^Y(X_i) - \hat{\mu}_0^Y(X_i) \right)}{\sum_{1 \le i \le n} \left(2A_i(D_i - \hat{\mu}_1^Y(X_i)) - 2(1-A_i)(D_i - \hat{\mu}_0^Y(X_i)) + \hat{\mu}_1^D(X_i) - \hat{\mu}_0^D(X_i) \right)} ,$$

where $\hat{\mu}_a^Y(X_i)$ is estimated by running a least squares regression of Y_i on $(1, X_i, X_i^2)$ using the sample from $A_i = a$, and $\hat{\mu}_a^D(X_i)$ is estimated using logistic regressions using the same set of regressors using the sample from $A_i = a$.

Adjusted Estimator 2: As in Adjusted Estimator 1, but $\hat{\mu}_a^Y(X_i)$ and $\hat{\mu}_a^D(X_i)$ are estimated respectively by running a least squares and logistic regressions of Y_i on $(1, X_i, X_i^2, X_i 1\{X_i > t\})$ where t is the sample median.

Similarly to Section 5.1, $\hat{\theta}_n^{\text{unadj}}$ solves (6) for the moment condition given in (5). The second and third estimators are covariate adjusted estimators which can be obtained as two-step method of moments estimators from solving an "augmented" version of the moment condition (5) (see, for instance, Chernozhukov et al., 2018; Jiang et al., 2022a).

5.3 Simulation Results

Table 5.3 displays the ratio of the MSE for each design/estimator pair relative to the MSE of the unadjusted estimator under i.i.d. assignment, computed across 2000 Monte Carlo replications. As expected given our theoretical results, we find that the empirical MSEs of the naive unadjusted estimator under a matched pair design closely match the empirical MSEs of the covariate adjusted estimators under i.i.d. assignment.

6 Recommendations for Empirical Practice

We conclude with some recommendations for empirical practice based on our theoretical results. Overall, our findings highlight the general benefit of fine stratification for designing efficient experiments: finely stratified experiments "automatically" perform fully-efficient regression adjustment for a large class of interesting parameters. This generalizes similar observations made by Bai et al. (2022), Bai (2022) and Cytrynbaum (2023b) for the special case of estimating the average treatment effect.

One caveat to this result, however, is that it crucially hinges on the assumption that units within each block are sufficiently "close" (Assumption 4.2), and such a condition becomes difficult to satisfy as the dimension of X_i increases. For this reason, we recommend that practitioners construct their blocks using a small subset of the baseline covariates which they feel are the most relevant (for instance, the baseline level of the experimental outcome, as suggested by Bruhn and McKenzie, 2009). Regression adjustment with additional covariates beyond those used for blocking can then be done *ex post*, although we caution that care must be taken to ensure that the adjustment is performed in such a way that it guarantees a gain in efficiency: see Bai et al. (2023b) and Cytrynbaum (2023a) for related discussions. Recent work has developed such methods of covariate adjustment for specific parameters of interest (see, for instance, Bai et al., 2023b,c,a; Cytrynbaum, 2023a), but we leave the development of a method of covariate adjustment which applies at the level of generality considered in this paper to future work.

Table 1: MSE ratios relative to unadjusted estimator under i.i.d. assignment

			I.I.D. assignment			**Matched pairs**
		Model	Unadjusted	Adjusted 1	Adjusted 2	Unadjusted
$n = 200$	ATE	1	1.0	0.4574	0.4629	0.4760
		2	1.0	1.0046	0.9688	0.8629
		3	1.0	0.7299	0.7217	0.7542
	LATE	1	1.0	0.3866	0.3970	0.3753
		2	1.0	0.8065	0.8758	0.7767
		3	1.0	0.5175	0.5216	0.5212
$n = 400$	ATE	1	1.0	0.4465	0.4699	0.4532
		2	1.0	0.9732	0.9780	0.9697
		3	1.0	0.7156	0.6834	0.6949
	LATE	1	1.0	0.4389	0.4402	0.4184
		2	1.0	0.8710	0.8683	0.7195
		3	1.0	0.5328	0.5173	0.5357
$n = 1000$	ATE	1	1.0	0.4341	0.4703	0.4670
		2	1.0	0.9787	0.9132	0.8977
		3	1.0	0.7223	0.7147	0.7195
	LATE	1	1.0	0.4307	0.4459	0.4369
		2	1.0	0.8583	0.8536	0.8567
		3	1.0	0.5147	0.5093	0.4901
$n = 2000$	ATE	1	1.0	0.4379	0.4444	0.4357
		2	1.0	0.9805	0.9659	0.9854
		3	1.0	0.6912	0.7235	0.6941
	LATE	1	1.0	0.4392	0.4378	0.4378
		2	1.0	0.8329	0.8499	0.8717
		3	1.0	0.5386	0.5913	0.5453

Note: For each model, the MSE of the unadjusted estimator under i.i.d. assignment are normalized to one and the other columns contain the ratios of the MSEs against that of the unadjusted estimator under i.i.d. assignment. MSEs are calculated across 2000 replications.

A Proofs of Main Results

A.1 Proofs for Section 3

Recall that P_n denotes the distribution of the observed data $(X^{(n)}, A^{(n)}, R^{(n)})$, and Q denotes the marginal distribution of the vector $(R_i(1), R_i(0), X_i)$. Note that any treatment assignment mechanism satisfying Assumption 2.1 can be represented as a function of $X^{(n)}$ and some additional randomization device $U_n \in \mathbf{R}$. Let $p_n^{U_n}$ denote the density function for U_n with respect to a dominating measure μ^U. In what follows, we consider a family $\{Q_t : t \in \mathbb{R}^{d_\theta}\}$ of marginal distributions indexed by t, and let q_t^X denote the density function for X_i with respect to a dominating measure μ^X, $q_t^{R(a)|X}(r|x)$ denote the conditional density of $R_i(a)$ given X_i with respect to a dominating measure μ^R. Further let $P_{t,n}$ denote the distribution of $Z^{(n)}$ and note it is jointly determined by Q_t and the distribution of U_n. We require that $Q_0 = Q$ and $P_{0,n} = P_n$ and define $q^X = q_0^X$ and $q^{R(a)|X} = q_0^{R(a)|X}$. As a consequence, the density function of $P_{t,n}$ is given by

$$\ell_n = p_n^U(U_n) \prod_{1 \le i \le n} q_t^X(X_i) \prod_{1 \le i \le n} \prod_{a \in \{0,1\}} q_t^{R(a)|X}(R_i|X_i)^{I\{A_i = a\}} \ . \tag{15}$$

Because the density $p_n^{U_n}$ does not depend on t, and in general we will only concern ourselves with the ratio of likelihoods at different values of t (so that $p_n^{U_n}$ in the ratio will cancel), in what follows we suppress the dependence on n and simply identify the distribution $P_{t,n}$ with its corresponding marginal distribution P_t. We consider a parametric submodel $\{P_t : t \in \mathbf{R}^{d_\theta}\}$, where $P_0 = P$, such that the following holds for $g = (g^X, g^{R(1)|X}, g^{R(0)|X})$, each component of which is a d_θ-dimensional function:

(a) As $t \to 0$,

$$\int \frac{1}{\|t\|^2} \Big(q_t^X(x)^{1/2} - q^X(x)^{1/2} - \frac{1}{2} q^X(x)^{1/2} t' g^X(x) \Big)^2 d\mu^X(x) \to 0 \ . \tag{16}$$

(b) For $a \in \{0,1\}$, $E_Q[g^{R(a)|X}(R(a)|X) g^{R(a)|X}(R(a)|X)'|X = x]$ is Lipschitz and for Q-almost every x, as $t \to 0$,

$$\frac{1}{\|t\|^2} \iint \Big(q_t^{R(a)|X}(r|x)^{1/2} - q^{R(a)|X}(r|x)^{1/2} - \frac{1}{2} q^{R(a)|X}(r|x)^{1/2} t' g^{R(a)|X}(r|x) \Big)^2 d\mu^R(r) q^X(x) d\mu^X(x) \to 0 \ . \tag{17}$$

In what follows, we will further index a parametric submodel by its associated function g, denoted by $P_{t,g}$, to emphasize the role of g. Similarly we denote the density of $Q_{t,g}$ by $q_{t,g}$.

Define the information of X as $I^X = E_Q[g^X(X) g^X(X)']$. Define the conditional information of $R(a)$ given $X = x$ as

$$I^{R(a)|X}(x) = E_Q[g^{R(a)|X}(R(a)|X) g^{R(a)|X}(R(a)|X)'|X = x] \ .$$

Lemma A.1. *For a parametric submodel $\{P_{t,g} : t \in \mathbf{R}^{d_\theta}\}$ with $P_{0,g} = P$ that satisfies (16)–(17),*

(a) $I^X < \infty$.

(b) $E_Q[g^X(X)] = 0$.

(c) $E_Q[g^{R(a)|X}(R(a)|X)g^{R(a)|X}(R(a)|X)'] < \infty$ *and hence* $I^{R(a)|X}(X) < \infty$ *with probability one under* Q.

(d) $E_Q[g^{R(a)|X}(R(a)|X)|X] = 0$ *with probability one under* Q.

PROOF. (a) and (b) follow from Lemma 14.2.1 in Lehmann and Romano (2022). (c) follows from the same lemma. In order to show (d), fix $t_n \to 0$. Note (17) and Markov's inequality imply that along a subsequence t_{n_k},

$$\frac{1}{\|t_{n_k}\|^2} \int \left(q_{t_{n_k}}^{R(a)|X}(r|x)^{1/2} - q^{R(a)|X}(r|x)^{1/2} - \frac{1}{2} q^{R(a)|X}(r|x)^{1/2} t_{n_k}' g^{R(a)|X}(r|x) \right)^2 d\mu^R(r) \to 0$$

for Q-almost every x. Along that subsequence, another application of Lemma 14.2.1 in Lehmann and Romano (2022) implies (d). ■

For $t \in \mathbf{R}^{d_\theta}$, the log-likelihood ratio between $P_{t/\sqrt{n},g}$ and $P_0 = P$ is

$$L_{n,t}(g) = \frac{1}{n} \sum_{1 \le i \le n} \log \frac{q_{t/\sqrt{n},g}^X(X_i)}{q^X(X_i)} + \frac{1}{n} \sum_{1 \le i \le n} \sum_{a \in \{0,1\}} I\{A_i = a\} \log \frac{q_{t/\sqrt{n},g}^{R(a)|X}(R_i|X_i)}{q^{R(a)|X}(R_i|X_i)} .$$

The following lemma establishes an expansion of the log-likelihood ratio and local asymptotic normality of $\{P_{t/\sqrt{n},g}\}$.

Lemma A.2. *Suppose the treatment assignment mechanism satisfies Assumption 2.1 and* g *satisfies* (16)–(17). *Then,*

$$L_{n,t}(g) = \frac{1}{\sqrt{n}} \sum_{1 \le i \le n} t' s_g(X_i, A_i, R_i) - \frac{1}{2} t' I^X t - \frac{1}{2n} \sum_{1 \le i \le n} \sum_{a \in \{0,1\}} I\{A_i = a\} t' I^{R(a)|X}(X_i) t + o_P(1) ,$$

where

$$s_g(x, a, r) = g^X(x) + I\{a = 1\} g^{R(1)|X}(r|x) + I\{a = 0\} g^{R(0)|X}(r|x) \tag{18}$$

and $I = I^X + \eta E_Q[I^{R(1)|X}(X_i)] + (1 - \eta) E_Q[I^{R(0)|X}(X_i)]$. *If in addition the assignment mechanism satisfies Assumption 3.1, then, under* P_0,

$$L_{n,t}(g) \xrightarrow{d} N\Big(-\frac{1}{2} t' I t, t' I t \Big) ,$$

PROOF. The first result follows from Theorem 3.1 of Armstrong (2022). The second result follows from Corollary 3.1 of Armstrong (2022) given Assumption 3.1 and the assumption that $I^{R(a)|X}(x)$ is Lipschitz. ■

We emphasize that Lemma A.5 implies

$$\sum_{1 \le i \le n} s_g(X_i, A_i, R_i)$$

is the sum of n identically distributed, despite possibly dependent, random variables. Therefore, in what follows, quantities like $E_P[s_g]$ are well defined.

Let the following condition collect the properties of the functions g that are of interest to us:

Condition A.1. The function g satisfies that $E_P[g^X(X)] = 0$, $E_P[g^X(X)g^X(X)'] < \infty$, $E_P[g^{R(a)|X}(R|X)$ $g^{R(a)|X}(R|X)'] < \infty$, $E_P[g^{R(a)|X}(R|X)|X] = 0$ with probability one, and $E_P[g^{R(a)|X}(R|X)g^{R(a)|X}(R|X)'|X = x]$ is Lipschitz for $a \in \{0, 1\}$. In addition, I is nonsingular.

We note that for any g that satisfies Condition A.1, there exists a parametric submodel $\{P_{t,g} : t \in \mathbf{R}^{d_\theta}\}$ such that (16)–(17) hold. Such a construction follows from the construction on p.69 in Tsiatis (2006) and can be done separately for $g_1(x)$ and $g_2^a(r|x)$ for each x separately so that they satisfy (16)–(17).

Let $\theta(P) \in \mathbf{R}^{d_\theta}$ be a parameter of interest. Further suppose that for each g satisfying Condition A.1, there exists a $d_\theta \times 1$ vector of functions $\psi^* \in L^2(P)$ such that for all $t \in \mathbf{R}^{d_\theta}$, as $n \to \infty$,

$$\sqrt{n}(\theta(P_{t/\sqrt{n},g}) - \theta(P)) \to E_P[\psi^* s_g' t] \ . \tag{19}$$

We provide explicit conditions which guarantee this is possible when $\theta(P)$ is defined by (3), in Lemma A.4 below.

We recall an estimator $\tilde{\theta}_n$ for $\theta(P)$ is regular if for all g and $t \in \mathbf{R}^{d_\theta}$,

$$\sqrt{n}(\tilde{\theta}_n - \theta(P_{t/\sqrt{n},g})) \xrightarrow{P_{t/\sqrt{n},g}} L \tag{20}$$

for a fixed probability measure L.

The following lemma establishes a convolution theorem for regular estimators:

Lemma A.3. *Suppose θ satisfies (19). Let $\tilde{\theta}_n$ be a regular estimator for θ. Further suppose that $\psi^* = s_g$ for some function g satisfying Condition A.1. Then,*

$$L = N(0, E_P[\psi^* \psi^{*\prime}]) * B \ ,$$

where B is a fixed probability measure.

Proof. In what follows, for each g satisfying Condition A.1, we consider the linear subspace given by

$$\mathcal{M}_g = \{t' s_g : t \in \mathbf{R}^{d_\theta}\} \ .$$

Note that $t' s_g$ appears in the expansion of the log-likelihood ratio between $P_{t/\sqrt{n},g}$ and P. We first derive the Riesz representer along the parametric subspace $\mathcal{M}_g$. In particular, for each $b \in \mathbf{R}^{d_\theta}$, we solve for $w(b) \in \mathbf{R}^{d_\theta}$ via the property that,

$$b' E_P[\psi^* s_g' t] = E_P[w(b)' s_g s_g' t]$$

needs to hold for all $t \in \mathbf{R}^{d_\theta}$ and get

$$w(b) = E[s_g s_g']^{-1} E[s_g \psi^{*\prime}] b \ .$$

Therefore, the Riesz representer is

$$E_P[\psi^* s_g'] E[s_g s_g']^{-1} s_g \ .$$

It then follows from the local asymptotic normality established in Lemma A.2 and Theorem 3.11.2 in van der Vaart and Wellner (1996) that

$$L = N(0, V_g) * B_g \ ,$$

where

$$V_g = E_P[\psi^* s_g'] E_P[s_g s_g']^{-1} E[s_g \psi^{*\prime}]$$

and B_g is a fixed probability measure. Furthermore, by a standard projection argument, in particular the fact that the second moment of $\psi^* - E_P[\psi^* s_g'] E_P[s_g s_g']^{-1} s_g$ is positive semi-definite, it can be shown that V_g is maximized in the matrix sense when $s_g = \psi^*$. Note this maximum is attained by our assumption that $\psi^* = s_g$ for some g satisfying Condition A.1. The conclusion then follows. ■

To apply Lemma A.3 to the setting in Section 3, we we establish the form of ψ^* in (19) for the parameter $\theta_0 = \theta(P)$ defined by (3). Define $\eta(X_i) = P\{A_i = 1 | X_i\}$. Note that

$$0 = E_P[m(X_i, A_i, R_i, \theta(P))] = E_Q[m(X, 1, R(1), \theta(P))\eta(X)] + E_Q[m(X, 0, R(0), \theta(P))(1 - \eta(X))] \ . \quad (21)$$

Lemma A.4. *Suppose the treatment assignment mechanism satisfies Assumptions 2.1 and 3.1. Fix a function g that satisfies Condition A.1. Suppose (16)–(17) holds. Fix $t \in \mathbf{R}^{d_\theta}$ and consider a one-dimensional submodel $\{P_{t/\sqrt{n},g}\}$ such that*

$$\begin{aligned} E_{Q_{t/\sqrt{n}}}[m(X, a, R(a), \theta(P))^2] &= O(1) \\ E_{Q^X}[E_{Q^{R(a)|X}_{t/\sqrt{n}}}[m(X, a, R(a), \theta(P))^2 | X]] &= O(1) \\ E_{Q^X_{t/\sqrt{n}}}[E_{Q^{R(a)|X}}[m(X, a, R(a), \theta(P))^2 | X]] &= O(1) \end{aligned} \quad (22)$$

as $n \to \infty$ and $\theta(P_{t/\sqrt{n},g})$ is uniquely determined by (21). Then, $\theta(P_{t/\sqrt{n},g})$ defined by (21) satisfies

$$\begin{aligned} &\sqrt{n}(\theta(P_{t/\sqrt{n},g}) - \theta(P)) \\ &\to M^{-1} E_P[m(X_i, A_i, R_i, \theta(P))(g^X(X_i) + I\{A_i = 1\} g^{R(1)|X}(R_i|X_i) + I\{A_i = 0\} g^{R(0)|X}(R_i|X_i))'] t \\ &= E_P[\psi^*(X_i, A_i, R_i, \theta(P))(g^X(X_i) + I\{A_i = 1\} g^{R(1)|X}(R_i|X_i) + I\{A_i = 0\} g^{R(0)|X}(R_i|X_i))'] t \ , \end{aligned}$$

where

$$\begin{aligned} &\psi^*(X_i, A_i, R_i, \theta(P)) \\ &= M^{-1} \Big(\eta(X_i) E_Q[m(X_i, 1, R_i(1), \theta(P)) | X_i] + (1 - \eta(X_i)) E_Q[m(X_i, 0, R_i(0), \theta(P)) | X_i] \\ &\qquad + I\{A_i = 1\}(m(X_i, 1, R_i, \theta(P)) - E_Q[m(X_i, 1, R_i(1), \theta(P)) | X_i]) \\ &\qquad + I\{A_i = 0\}(m(X_i, 0, R_i, \theta(P)) - E_Q[m(X_i, 0, R_i(0), \theta(P)) | X_i]) \Big) \ . \end{aligned}$$

PROOF. In what follows, we only use the property that the quadratic mean derivative of $P_{t/\sqrt{n},g}$ is given by $s'_g t$. Therefore, for ease of notation we consider a generic one-dimensional submodel $\{P_\nu : \nu \in [-\epsilon, \epsilon]\}$ that satisfies (16)–(17) for some $g = (g^X, g^{R(1)|X}, g^{R(0)|X})$, each component of which is a one-dimensional function. (21) implies

$$0 = \int m(x,1,r,\theta(P_\nu)) q_\nu^{R(1)|X}(r|x) d\mu^R(r) \eta(x) q_\nu^X(x) d\mu^X(x) \\ + \int m(x,0,r,\theta(P_\nu)) q_\nu^{R(0)|X}(r|x) d\mu^R(r)(1-\eta(x)) q_\nu^X(x) d\mu^X(x)$$

Note that

$$\int m(x,1,r,\theta(P)) q_\nu^{R(1)|X}(r|x) d\mu^R(r) \eta(x) q_\nu^X(x) d\mu^X(x) \\ - \int m(x,1,r,\theta(P)) q^{R(1)|X}(r|x) d\mu^R(r) \eta(x) q^X(x) d\mu^X(x) = \gamma_1(\nu) + \gamma_2(\nu) + \gamma_3(\nu) + \gamma_4(\nu) \ ,$$

where

$$\gamma_1(\nu) = \int m(x,1,r,\theta(P)) \big(q_\nu^{R(1)|X}(r|x)^{1/2} - q^{R(1)|X}(r|x)^{1/2}\big) q_\nu^{R(1)|X}(r|x)^{1/2} d\mu^R(r) \eta(x) q_\nu^X(x) d\mu^X(x)$$
$$\gamma_2(\nu) = \int m(x,1,r,\theta(P)) \big(q_\nu^{R(1)|X}(r|x)^{1/2} - q^{R(1)|X}(r|x)^{1/2}\big) q^{R(1)|X}(r|x)^{1/2} d\mu^R(r) \eta(x) q_\nu^X(x) d\mu^X(x)$$
$$\gamma_3(\nu) = \int m(x,1,r,\theta(P)) q^{R(1)|X}(r|x) d\mu^R(r) \eta(x) \big(q_\nu^X(x)^{1/2} - q^X(x)^{1/2}\big) q_\nu^X(x)^{1/2} d\mu^X(x)$$
$$\gamma_4(\nu) = \int m(x,1,r,\theta(P)) q^{R(1)|X}(r|x) d\mu^R(r) \eta(x) \big(q_\nu^X(x)^{1/2} - q^X(x)^{1/2}\big) q^X(x)^{1/2} d\mu^X(x) \ .$$

It follows from the Cauchy-Schwarz inequality that

$$\frac{1}{\nu}\gamma_4(\nu) - \int m(x,1,r,\theta(P)) q^{R(1)|X}(r|x) d\mu^R(r) \eta(x) \frac{1}{2} g^X(x) q^X(x)^{1/2} \times q^X(x)^{1/2} d\mu^X(x) \\ \le \int \Big(m(x,1,r,\theta(P))^2 q^{R(1)|X}(r|x) d\mu^R(r) \eta(x)^2 q^X(x) d\mu^X(x) \Big)^{1/2} \\ \times \Big(\int q^{R(1)|X}(r|x) d\mu^R(r) \Big(\frac{1}{\nu}(q_\nu^X(x)^{1/2} - q^X(x)^{1/2}) - \frac{1}{2} g^X(x) q^X(x)^{1/2} \Big)^2 d\mu^X(x) \Big)^{1/2} \to 0$$

by the assumption that $E_P[m(X,a,R(a),\theta(P))^2] < \infty$, the facts that $0 \le \eta(x) \le 1$, $\int q^{R(1)|X}(r|x) d\mu^R(r) = 1$, and (16). Similar arguments implies as $\nu \to 0$,

$$\frac{1}{\nu}\gamma_1(\nu) - \int m(x,1,r,\theta(P)) \frac{1}{2} g^{R(1)|X}(r|x) q^{R(1)|X}(r|x) d\mu^R(r) \eta(x) q^X(x) d\mu^X(x) \to 0$$

because $E_{P_\nu}[m(X,a,R(a),\theta(P))^2] = O(1)$ as $\nu \to 0$. The limits of $\gamma_2(\nu)$ and $\gamma_3(\nu)$ can be derived following similar arguments using the last two conditions in (22). Combining all previous results yields

$$\frac{\partial}{\partial \nu} E_{P_\nu}[m(X,A,R,\theta(P))]\Big|_{\nu=0} \\ = E_Q[m(X,1,R(1),\theta(P))(g^X(X) + g^{R(1)|X}(R|X))\eta(X)]$$

$$+ E_Q[m(X, 0, R(0), \theta(P))(g^X(X) + g^{R(0)|X}(R|X))(1 - \eta(X))]$$
$$= E_P[m(X, A, R, \theta(P))(g^X(X) + I\{A = 1\}g^{R(1)|X}(R) + I\{A = 0\}g^{R(0)|X}(R))] \ .$$

On the other hand, by definition

$$M_{\theta(P)} = \frac{\partial}{\partial \theta} E_P[m(X, A, R, \theta)]\Big|_{\theta=\theta(P)} \ .$$

The formula for the derivative therefore follows from the implicit function theorem (in particular, because we have assumed the existence of $\theta(P_\nu)$ along the path, it follows from the last part of the proof of Theorem 3.2.1 in Krantz and Parks (2013)). The second equality follows from Lemma A.5 together with Condition A.1. ■

Finally, to prove Theorem 3.1 we require the following additional regularity condition:

Assumption A.1. For every function g satisfying Condition A.1 and every $t \in \mathbf{R}^{d_\theta}$ there exists a submodel $P_{t/\sqrt{n},g}$ for which (22) holds as $n \to \infty$, and $\theta(P_{t/\sqrt{n},g})$ is uniquely determined by (21).

This assumption guarantees that every element satisfying Condition A.1 has a corresponding path for which we can apply Lemma A.4. A similar assumption appears in Chen and Santos (2018) (see their Assumption 4.1(iv)). Note that a simple sufficient condition for the first part of Assumption A.1 is that $m(x, a, r, \theta(P))$ is a bounded function in (x, r) on the support of $(X, R(a))$. The second part of Assumption A.1 can be verified easily in specific examples (see, for instance, Examples 2.1–2.5 in the main text). Alternatively, Assumption A.1 could be avoided by assuming that we can differentiate under the integral in the final step of the proof of Lemma A.4, from which we would immediately obtain the expression for the pathwise derivative. See, for instance, Newey (1994) and Chen et al. (2008).

Proof of Theorem 3.1. First note θ satisfies (19) because of Lemma A.4 and Assumption A.1. The result then follows from Lemma A.3 upon noting that $\psi^* = s_g$ for some g that satisfies Condition A.1 because of Assumption 3.2. ■

A.2 Proof of Theorem 4.1

First note (14) follows from (13) and the same proof as that of Lemma B.3 in Bai (2022). To establish (13), we follow the proof of Theorem 5.21 in van der Vaart (1998). We start by noting that because Assumptions 4.1–4.2, 4.3(e), and 4.4(b) hold, it follows from the same proof as that of Lemma B.3 in Bai (2022) that

$$\begin{aligned}\frac{1}{\sqrt{n}} \sum_{1 \le i \le n} m(X_i, A_i, R_i, \theta_0) = \frac{1}{\sqrt{n}} \sum_{1 \le i \le n} \sum_{a \in \{0,1\}} I\{A_i = a\}(m(X_i, a, R_i(a), \theta_0) - E_Q[m(X_i, a, R_i(a), \theta_0)|X_i]) \\ + \frac{\eta}{\sqrt{n}} \sum_{1 \le i \le n} (E_Q[m(X_i, 1, R_i(1), \theta_0)|X_i] - E_Q[m(X_i, 1, R_i(1), \theta_0)]) \\ + \frac{(1 - \eta)}{\sqrt{n}} \sum_{1 \le i \le n} (E_Q[m(X_i, 0, R_i(0), \theta_0)|X_i] - E_Q[m(X_i, 0, R_i(0), \theta_0)]) + o_P(1) \ .\end{aligned}$$

where we note $\eta E_Q[m(X_i, 1, R_i(1), \theta_0)] + (1-\eta)E_Q[m(X_i, 0, R_i(0), \theta_0)] = E_P[m(X_i, A_i, R_i, \theta_0)] = 0$ by (3). Therefore, by the proof of Theorem 5.21 in van der Vaart (1998), it suffices to show

$$\mathbb{L}_n(\hat{\theta}_n) \xrightarrow{P} 0 \ , \tag{23}$$

where

$$\mathbb{L}_n(\theta) = \frac{1}{\sqrt{n}} \sum_{1 \le i \le n} (m(X_i, A_i, R_i, \theta) - E_P[m(X_i, A_i, R_i, \theta)]) \\ - \frac{1}{\sqrt{n}} \sum_{1 \le i \le n} (m(X_i, A_i, R_i, \theta_0) - E_P[m(X_i, A_i, R_i, \theta_0)]) \ .$$

To accomplish this, we study m_s for $1 \le s \le d_\theta$ separately. It follows from Assumption 4.3(c), (d), and the proof of Proposition 8.11 in Kosorok (2008) that

$$\begin{aligned} &\sup_{\theta \in \Theta : \|\theta - \theta_0\| < \delta} \Big| \frac{1}{\sqrt{n}} \sum_{1 \le i \le n} (m_s(X_i, A_i, R_i, \theta) - E_P[m_s(X_i, A_i, R_i, \theta)]) \\ &\qquad - \frac{1}{\sqrt{n}} \sum_{1 \le i \le n} (m_s(X_i, A_i, R_i, \theta_0) - E_P[m_s(X_i, A_i, R_i, \theta_0)]) \Big| \\ &= \sup_{\theta \in \Theta^* : \|\theta - \theta_0\| < \delta} \Big| \frac{1}{\sqrt{n}} \sum_{1 \le i \le n} (m_s(X_i, A_i, R_i, \theta) - E_P[m_s(X_i, A_i, R_i, \theta)]) \\ &\qquad - \frac{1}{\sqrt{n}} \sum_{1 \le i \le n} (m_s(X_i, A_i, R_i, \theta_0) - E_P[m_s(X_i, A_i, R_i, \theta_0)]) \Big| \ , \end{aligned}$$

and thus since $\hat{\theta}_n \xrightarrow{P} \theta_0$ by Lemma A.6, to show (23) it suffices to argue that for every $\epsilon > 0$,

$$\lim_{\delta \downarrow 0} \limsup_{n \to \infty} P \left\{ \sup_{\theta \in \Theta^* : \|\theta - \theta_0\| < \delta} \left| \mathbb{L}_n^{(s)}(\theta) \right| > \epsilon \right\} = 0 \ , \tag{24}$$

where

$$\mathbb{L}_n^{(s)}(\theta) = \frac{1}{\sqrt{n}} \sum_{1 \le i \le n} (m_s(X_i, A_i, R_i, \theta) - E_P[m_s(X_i, A_i, R_i, \theta)]) \\ - \frac{1}{\sqrt{n}} \sum_{1 \le i \le n} (m_s(X_i, A_i, R_i, \theta_0) - E_P[m_s(X_i, A_i, R_i, \theta_0)]) \ .$$

Consider the following decomposition:

$$\left| \mathbb{L}_n^{(s)}(\theta) \right| \le \sum_{a \in \{0,1\}} \left(\mathbb{L}_{1,a,n}^{(s)}(\theta) + \mathbb{L}_{2,a,n}^{(s)}(\theta) + \mathbb{L}_{3,a,n}^{(s)}(\theta) \right) \ ,$$

where

$$\mathbb{L}_{1,a,n}^{(s)}(\theta) = \Big| \frac{1}{\sqrt{n}} \sum_{1 \le i \le n} I\{A_i = a\}(m_s(X_i, a, R_i(a), \theta) - E_Q[m_s(X_i, a, R_i(a), \theta) | X_i])$$

$$- \frac{1}{\sqrt{n}} \sum_{1 \le i \le n} I\{A_i = a\}(m_s(X_i, a, R_i(a), \theta_0) - E_Q[m_s(X_i, a, R_i(a), \theta_0)|X_i])\Big|$$

$$\mathbb{L}^{(s)}_{2,a,n}(\theta) = \Big| \frac{1}{\sqrt{n}} \sum_{1 \le i \le n} \Big(I\{A_i = a\} - \eta_a \Big)$$
$$\times (E_Q[m_s(X_i, a, R_i(a), \theta)|X_i] - E_Q[m_s(X_i, a, R_i(a), \theta_0)|X_i])\Big|$$

$$\mathbb{L}^{(s)}_{3,a,n}(\theta) = \Big| \frac{\eta_a}{\sqrt{n}} \sum_{1 \le i \le n} \big(E_Q[m_s(X_i, a, R_i(a), \theta)|X_i] - E_Q[m_s(X, a, R(a), \theta)]\big)$$
$$- \big(E_Q[m_s(X_i, a, R_i(a), \theta_0)|X_i] - E_Q[m_s(X, a, R(a), \theta_0)]\big)\Big| \ ,$$

where $\eta_1 = \eta$ and $\eta_0 = 1 - \eta$. Then to establish (24), it suffices to establish that

$$\lim_{\delta \downarrow 0} \limsup_{n \to \infty} P\left\{ \sup_{\theta \in \Theta^* : \|\theta - \theta_0\| < \delta} \mathbb{L}^{(s)}_{\ell,a,n}(\theta) > \epsilon \right\} = 0 \ . \tag{25}$$

for $\ell \in \{1, 2, 3\}$ and $a \in \{0, 1\}$.

Step 1. First we consider $\mathbb{L}^{(s)}_{3,a,n}$. It follows from Assumption 4.4(d) and Theorems 2.5.2 and 2.6.7 in van der Vaart and Wellner (1996) that the class of functions

$$\{E_Q[m_s(X, a, R(a), \theta)|X = x] : \theta \in \Theta^*\} \ ,$$

is Donsker, and thus we obtain by Theorem 3.34 in Dudley (2014) that

$$\lim_{\delta \downarrow 0} \limsup_{n \to \infty} P\left\{ \sup_{\theta \in \Theta^* : \rho_Q(\theta, \theta_0) < \delta} \mathbb{L}^{(s)}_{3,a,n}(\theta) > \epsilon \right\} = 0 \ ,$$

where $\rho_Q(\theta, \theta_0) = E_Q[(E_Q[m_s(X, a, R(a), \theta)|X] - E_Q[m_s(X, a, R(a), \theta_0)|X])^2]$. We then obtain (25) for $\ell = 3$ since, by Assumption 4.4(c) as $\theta \to \theta_0$,

$$\rho_Q(\theta, \theta_0) = E_Q[(E_Q[m_s(X, a, R(a), \theta)|X] - E_Q[m_s(X, a, R(a), \theta_0)|X])^2]$$
$$\le E_Q[(m_s(X, a, R(a), \theta) - m_s(X, a, R(a), \theta_0))^2] \to 0 \ . \tag{26}$$

Step 2. Next, we study $\mathbb{L}^{(s)}_{2,a,n}$. Define

$$f(X, \theta) = E[m_s(X, a, R(a), \theta)|X] - E[m_s(X, a, R(a), \theta_0)|X] \ .$$

Note

$$\mathbb{L}^{(s)}_{2,a,n}(\theta) = C \Big| \frac{1}{\sqrt{n/k}} \sum_{1 \le j \le n/k} \alpha_j(\theta) \Big| \ ,$$

for some constant $C > 0$, where $\alpha_j(\theta) \in \{\frac{1}{\ell} \sum_{i \in I} f(X_i, \theta) - \frac{1}{k - \ell} \sum_{i \in \lambda_j \setminus I} f(X_i, \theta) : I \subset \lambda_j, |I| = \ell\}$, $E[\alpha_j(\theta)|X^{(n)}] = 0$, and $\alpha_j(\theta), 1 \le j \le n/k$ are independent conditional on $X^{(n)}$.

Define

$$h(x_1,\ldots,x_k,\theta) = \sup_{I\subset\{1,\ldots,k\},|I|=\ell}\Big(\frac{1}{\ell}\sum_{i\in I} f(x_i,\theta) - \frac{1}{k-\ell}\sum_{i\in\{1,\ldots,k\}\setminus I} f(x_i,\theta)\Big) - \inf_{I\subset\{1,\ldots,k\},|I|=\ell}\Big(\frac{1}{\ell}\sum_{i\in I} f(x_i,\theta) - \frac{1}{k-\ell}\sum_{i\in\{1,\ldots,k\}\setminus I} f(x_i,\theta)\Big)$$

and the classes of functions

$$\mathcal{H}_\delta = \{h(x_1,\ldots,x_k,\theta) : \theta\in\Theta^*, \|\theta-\theta_0\|<\delta\}$$
$$\mathcal{H}_\infty = \{h(x_1,\ldots,x_k,\theta) : \theta\in\Theta^*\}\ .$$

Let $P_n^\dagger$ denote a measure that puts mass $\frac{k}{n}$ on each of $(X_i : i\in\lambda_j), 1\le j\le n/k$. It follows from a generalized version of Hoeffding's inequality (see, for instance, Theorem 2.2.6 in Vershynin, 2018) that conditional on $X^{(n)}$,

$$\Big\{\frac{1}{\sqrt{n/k}}\sum_{1\le j\le n/k}\alpha_j(\theta) : \theta\in\Theta^*, \|\theta-\theta_0\|<\delta\Big\}$$

is sub-Gaussian for the seminorm

$$\|h\|_{P_n^\dagger} = \Big(\int h^2 dP_n^\dagger\Big)^{1/2}\ .$$

Let $N(\epsilon,\mathcal{H}_\delta,L_2(P_n^\dagger))$ denote the covering number of $\mathcal{H}_\delta$ with respect to $\|\cdot\|_{P_n^\dagger}$. Let $\delta_n\downarrow 0$ be an arbitrary decreasing sequence. It follows from the maximal inequality in Corollary 2.2.8 in van der Vaart and Wellner (1996) (note $\alpha_j(\theta_0)=0$) that

$$E\Big[\sup_{\theta\in\Theta^*:\|\theta-\theta_0\|<\delta_n}\mathbb{L}_{2,a,n}^{(s)}(\theta)\Big|X^{(n)}\Big] \lesssim \int_0^\infty \sqrt{\log N(\epsilon,\mathcal{H}_{\delta_n},L_2(P_n^\dagger))}d\epsilon\ . \tag{27}$$

The upper limit of the integral is in fact c_n, where

$$c_n^2 = \sup_{\theta\in\Theta^*:\|\theta-\theta_0\|<\delta_n}\frac{4k}{n}\sum_{1\le j\le n/k}\sup_{I\in\lambda_j,|I|=\ell}\Big(\frac{1}{\ell}\sum_{i\in I}f(X_i,\theta) - \frac{1}{k-\ell}\sum_{i\in\lambda_j\setminus I}f(X_i,\theta)\Big)^2 \tag{28}$$
$$\lesssim \frac{1}{n}\sum_{1\le j\le n/k}\max_{i,i'\in\lambda_j}\|X_i-X_{i'}\|^2 \xrightarrow{P} 0$$

by Assumptions 4.2 and 4.3(e) and the inequality $(a+b)^2\le 2(a^2+b^2)$. Moreover,

$$H(x_1,\ldots,x_k) = \sum_{1\le i\le k}E\Big[\sup_{\theta\in\Theta^*}|m_s(X,a,R(a),\theta)| + |m_s(X,a,R(a),\theta_0)|\Big|X=x_i\Big]$$

is an envelope function for $\mathcal{H}_\infty$ (and thus $\mathcal{H}_\delta$ for all δ) and $E[H^2]<\infty$ by Assumption 4.4(b). A change of variable in (27) implies

$$E\Big[\sup_{\theta\in\Theta^*:\|\theta-\theta_0\|<\delta_n}\mathbb{L}_{2,a,n}^{(s)}(\theta)\Big|X^{(n)}\Big] \lesssim \int_0^{\frac{c_n}{\|H\|_{P_n^\dagger}}}\sqrt{\log N(\epsilon\|H\|_{P_n^\dagger},\mathcal{H}_{\delta_n},L_2(P_n^\dagger))}d\epsilon\|H\|_{P_n^\dagger}$$

$$\leq \int_0^{\frac{c_n}{\|H\|_{P_n^\dagger}}} \sup_\nu \sqrt{\log N(\epsilon\|H\|_\nu, \mathcal{H}_\infty, L_2(\nu))}d\epsilon\|H\|_{P_n^\dagger} ,$$

where the supremum for ν is over probability measures with discrete support such that $\|H\|_\nu > 0$. Also note that if $\|H\|_{P_n^\dagger} = 0$ then the conditional expectation on the left-hand side is trivially zero, so we can without loss of generality assume $\|H\|_{P_n^\dagger} > 0$. The Cauchy-Schwarz inequality implies

$$E\left[\int_0^{\frac{c_n}{\|H\|_{P_n^\dagger}}} \sup_\nu \sqrt{\log N(\epsilon\|H\|_\nu, \mathcal{H}_\infty, L_2(\nu))}d\epsilon\|H\|_{P_n^\dagger}\right]$$
$$\leq E\left[\left(\int_0^{\frac{c_n}{\|H\|_{P_n^\dagger}}} \sup_\nu \sqrt{\log N(\epsilon\|H\|_\nu, \mathcal{H}_\infty, L_2(\nu))}d\epsilon\right)^2\right]^{1/2} E\left[\|H\|_{P_n^\dagger}^2\right]^{1/2} ,$$

where the supremum is over all measures ν with discrete support such that $\|H\|_\nu > 0$. It follows from Assumption 4.4(b), the inequality $(a+b)^2 \leq 2(a^2+b^2)$, and the conditional Jensen's inequality that

$$E\left[\|H\|_{P_n^\dagger}^2\right] \lesssim E\left[\frac{1}{n}\sum_{1\leq i\leq n} E\left[\sup_{\theta\in\Theta^*} |m_s(X_i, a, R_i(a), \theta)|\middle| X_i\right]^2\right] \leq E\left[\sup_{\theta\in\Theta^*} |m_s(X, a, R(a), \theta)|^2\right] < \infty .$$

On the other hand,

$$\|H\|_{P_n^\dagger}^2 \geq \frac{1}{n}\sum_{1\leq i\leq n} E\left[\sup_{\theta\in\Theta^*} |m_s(X_i, a, R_i(a), \theta)|\middle| X_i\right]^2 \xrightarrow{P} E\left[E\left[\sup_{\theta\in\Theta^*} |m_s(X, a, R(a), \theta)|\middle| X_i\right]^2\right] , \tag{29}$$

the right-hand side of which can be assumed to be strictly positive, because otherwise $\sup_{\theta\in\Theta^*:\|\theta-\theta_0\|<\delta_n} \mathbb{L}_{2,a,n}^{(s)}(\theta) = 0$. Therefore, it follows from (28) and (29) that

$$\frac{c_n}{\|H\|_{P_n^\dagger}} \xrightarrow{P} 0 . \tag{30}$$

From Assumption 4.4(d), Lemma A.7, Theorem 2.6.7 in van der Vaart and Wellner (1996), and Lemma 9.13 in Kosorok (2008), we know

$$\int_0^1 \sup_\nu \sqrt{\log N(\epsilon\|H\|_\nu, \mathcal{H}_\infty, L_2(\nu))}d\epsilon < \infty .$$

Therefore,

$$E\left[\left(\int_0^{\frac{c_n}{\|H\|_{P_n^\dagger}}} \sup_\nu \sqrt{\log N(\epsilon\|H\|_\nu, \mathcal{H}_\infty, L_2(\nu))}d\epsilon\right)^2\right] \to 0$$

by Lemma A.8 combined with (30) and the continuous mapping theorem. Therefore, it follows from Markov's inequality that

$$\sup_{\theta\in\Theta^*:\|\theta-\theta_0\|<\delta_n} \mathbb{L}_{2,a,n}^{(s)}(\theta) \xrightarrow{P} 0 ,$$

as $n \to \infty$, from which (25) follows (see, for instance, Section 2.1.2 in van der Vaart and Wellner, 1996).

Step 3. Finally, we study $\mathbb{L}_{1,a,n}^{(s)}(\theta)$. Define

$$\mathbb{B}_n(\theta) = \frac{1}{\sqrt{\eta_a n}} \sum_{1 \le i \le n} I\{A_i = a\}(m_s(X_i, a, R_i(a), \theta) - E_Q[m_s(X_i, a, R_i(a), \theta)|X_i]) \ ,$$

Let $\delta_n \downarrow 0$ be an arbitrary decreasing sequence. To establish our result we will show

$$\sup_{\theta \in \Theta^* : \|\theta - \theta_0\| < \delta_n} |\mathbb{B}_n(\theta) - \mathbb{B}_n(\theta_0)| \xrightarrow{P} 0 \ , \tag{31}$$

as $n \to \infty$. As in the proof of Lemma A.6, we define

$$\tilde{P}_n = \frac{1}{\eta_a n} \sum_{1 \le i \le n : A_i = a} \delta_{(X_i, R_i(a))} \ .$$

Define the classes of functions

$$\mathcal{F}_{\theta_0, \infty} = \{m_s(x, a, r(a), \theta) : \theta \in \Theta^*\} \ .$$

Pick an envelope function for $\mathcal{F}_{\theta_0,\infty}$ as

$$F = \sup_{\theta \in \Theta^*} |m_s(X, a, R(a), \theta)| \ .$$

and define

$$\zeta_n^2 = \sup_{\theta \in \Theta^* : \|\theta - \theta_0\| < \delta_n} \int (m_s(X, a, R(a), \theta) - m_s(X, a, R(a), \theta_0))^2 d\tilde{P}_n \ .$$

Step 3(a). Our next goal is to show for every $\xi > 0$,

$$P\{\zeta_n^2 > \xi | X^{(n)}, A^{(n)}\} \xrightarrow{P} 0 \ . \tag{32}$$

To do so, first note by triangle inequality that $\zeta_n^2 \le \mathbb{C}_{1,n} + \mathbb{C}_{2,n} + \mathbb{C}_{3,n}$, where

$$\begin{aligned}
\mathbb{C}_{1,n} = \sup_{\theta \in \Theta^*} \Bigg| & \int (m_s(X, a, R(a), \theta) - m_s(X, a, R(a), \theta_0))^2 d\tilde{P}_n \\
& - E\left[\int (m_s(X, a, R(a), \theta) - m_s(X, a, R(a), \theta_0))^2 d\tilde{P}_n \middle| X^{(n)}, A^{(n)} \right] \Bigg| \\
\mathbb{C}_{2,n} = \sup_{\theta \in \Theta^*} \Bigg| & E\left[\int (m_s(X, a, R(a), \theta) - m_s(X, a, R(a), \theta_0))^2 d\tilde{P}_n \middle| X^{(n)}, A^{(n)} \right] \\
& - E[(m_s(X, a, R(a), \theta) - m_s(X, a, R(a), \theta_0))^2] \Bigg| \\
\mathbb{C}_{3,n} = \sup_{\theta \in \Theta^* : \|\theta - \theta_0\| < \delta_n} & E[(m_s(X, a, R(a), \theta) - m_s(X, a, R(a), \theta_0))^2] \ .
\end{aligned}$$

Assumption 4.4(c) implies

$$\mathbb{C}_{3,n} = \sup_{\theta \in \Theta^* : \|\theta - \theta_0\| < \delta_n} E[(m_s(X, a, R(a), \theta) - m_s(X, a, R(a), \theta_0))^2] \to 0 \ . \tag{33}$$

Next, Assumption 4.4(b), (f) and similar arguments to those used to show (52) and (53) are $o_P(1)$ imply

$$\mathbb{C}_{2,n} = \sup_{\theta \in \Theta^*} \left| E\Big[\int (m_s(X,a,R(a),\theta) - m_s(X,a,R(a),\theta_0))^2 d\tilde{P}_n \Big| X^{(n)}, A^{(n)} \Big] \right.$$
$$\left. - E[(m_s(X,a,R(a),\theta) - m_s(X,a,R(a),\theta_0))^2] \right| \xrightarrow{P} 0 \ . \quad (34)$$

Further define

$$\mathcal{G} = \{(m_s(x,a,r(a),\theta) - m_s(x,a,r(a),\theta_0))^2 : \theta \in \Theta^*\} \ .$$

We then study $\mathbb{C}_{1,n}$. We will establish for every $\xi > 0$,

$$P\Big\{ \sup_{f \in \mathcal{G}} \Big| \int f d\tilde{P}_n - E\Big[\int f d\tilde{P}_n \Big| X^{(n)}, A^{(n)} \Big] \Big| > \xi \Big| X^{(n)}, A^{(n)} \Big\} \xrightarrow{P} 0 \quad (35)$$

as $n \to \infty$. It follows the symmetrization Lemma 6.2 in Ledoux and Talagrand (1991) applied conditional on $X^{(n)}, A^{(n)}$ for the distribution

$$\bigotimes_{1 \le i \le n : A_i = 1} P\{X_i, R_i(1) | X_i\}$$

that

$$E\Big[\sup_{f \in \mathcal{G}} \Big| \int f d\tilde{P}_n - E\Big[\int f d\tilde{P}_n \Big| X^{(n)}, A^{(n)} \Big] \Big| \Big| X^{(n)}, A^{(n)} \Big]$$
$$\le 2 E_P \Big[E_\tau \Big[\sup_{f \in \mathcal{G}} \Big| \frac{1}{\eta_a n} \sum_{1 \le i \le n} \tau_i f(X_i, R_i(a)) \Big| \Big] \Big| X^{(n)}, A^{(n)} \Big] \ , \quad (36)$$

where $E_\tau[\cdot]$ should be understood as the expectation with respect to $(\tau_i, 1 \le i \le n)$, holding all else fixed. Note Assumption 4.3(b) and Theorem 2.6.7 in van der Vaart and Wellner (1996) imply $\mathcal{F}_{\theta_0,\infty}$ is totally bounded in $L_2(\tilde{P}_n)$. Accordingly, for $\epsilon > 0$, let $N(\epsilon, \mathcal{F}_{\theta_0,\infty}, L_2(\tilde{P}_n))$ denote the covering number of $\mathcal{F}_{\theta_0,\infty}$ with respect to $L_2(\tilde{P}_n)$. Let f_1, f_2 be any pair of functions in $\mathcal{G}$, where we denote

$$f_j = (m_s(x,a,r(a),\theta_j) - m_s(x,a,r(a),\theta_0))^2, \ j = 1,2 \ ,$$

then the Cauchy-Schwarz inequality implies

$$\int |f_1 - f_2| d\tilde{P}_n \le \int |m_s(x,a,r(a),\theta_1) - m_s(x,a,r(a),\theta_2)| 2F d\tilde{P}_n \le \|m_s(\cdot,\theta_1) - m_s(\cdot,\theta_2)\|_{\tilde{P}_n} 2 \|F\|_{\tilde{P}_n}$$

where $\|\cdot\|_{\tilde{P}_n}$ denotes the $L_2(\tilde{P}_n)$-norm. Therefore

$$N(2\epsilon \|F\|^2_{\tilde{P}_n}, \mathcal{G}, L_1(\tilde{P}_n)) \le N(\epsilon \|F\|_{\tilde{P}_n}, \mathcal{F}_{\theta_0,\infty}, L_2(\tilde{P}_n)). \quad (37)$$

For every $\epsilon > 0$, the right-hand side is uniformly bounded across n by Assumption 4.3(b) and Theorem 2.6.7 in van der Vaart and Wellner (1996). Note it follows from Assumptions 4.1–4.2, 4.4(b), (f), and similar

arguments to those in the first part of the proof of Lemma S.1.5 in Bai et al. (2022) that

$$E\big[\|F\|^2_{\tilde{P}_n}|X^{(n)},A^{(n)}\big]=\frac{1}{\eta_a n}\sum_{1\le i\le n} I\{A_i=a\}E[F^2|X_i]\xrightarrow{P}E[F^2]\ . \tag{38}$$

We can assume without loss of generality $E[F^2]>0$ because otherwise $m_s(x,a,r(a),\theta)\equiv 0$. Therefore,

$$P\Big\{E\big[\|F\|^2_{\tilde{P}_n}|X^{(n)},A^{(n)}\big]\le\frac{1}{2}E[F^2]\Big\}\to 0\ . \tag{39}$$

On the other hand, Assumptions 4.1–4.2, 4.4(b), (f), and similar arguments to those in the last part of the proof of Lemma S.1.5 in Bai et al. (2022) that

$$P\Big\{\Big|\|F\|^2_{\tilde{P}_n}-E\big[\|F\|^2_{\tilde{P}_n}|X^{(n)},A^{(n)}\big]\Big|>\frac{1}{4}E[F^2]\Big|X^{(n)},A^{(n)}\Big\}\xrightarrow{P}0\ . \tag{40}$$

(35) now follows from (39)–(40) and similar arguments to those used in the last step of the proof of Lemma A.6.

To conclude (32) holds, note $\mathbb{C}_{3,n}$ is a sequence of constants and $\mathbb{C}_{2,n}$ is a function of $X^{(n)},A^{(n)}$, and hence

$$\begin{aligned}
&P\big\{P\big\{\zeta_n^2>\xi|X^{(n)},A^{(n)}\big\}>\epsilon\big\}\\
&\le P\Big\{\mathbb{C}_{2,n}>\frac{\xi}{3}\Big\}+P\Big\{\mathbb{C}_{3,n}>\frac{\xi}{3}\Big\}\\
&\qquad+P\Big\{P\big\{\mathbb{C}_{1,n}+\mathbb{C}_{2,n}+\mathbb{C}_{3,n}>\xi|X^{(n)},A^{(n)}\big\}>\epsilon,\mathbb{C}_{2,n}\le\frac{\xi}{3},\mathbb{C}_{3,n}\le\frac{\xi}{3}\Big\}\\
&\le P\Big\{\mathbb{C}_{2,n}>\frac{\xi}{3}\Big\}+P\Big\{\mathbb{C}_{3,n}>\frac{\xi}{3}\Big\}+P\Big\{P\Big\{\mathbb{C}_{1,n}>\frac{\xi}{3}\Big|X^{(n)},A^{(n)}\Big\}>\epsilon\Big\}\\
&\xrightarrow{P}0\ ,
\end{aligned}$$

where the convergence follows from (33), (34), and (35).

Step 3(b). Next, we show for every $\xi>0$,

$$P\Big\{\frac{\zeta_n^2}{\|F\|^2_{\tilde{P}_n}}>\xi\Big|X^{(n)},A^{(n)}\Big\}\xrightarrow{P}0\ . \tag{41}$$

For every $\epsilon>0$,

$$\begin{aligned}
&P\Big\{P\Big\{\frac{\zeta_n^2}{\|F\|^2_{\tilde{P}_n}}>\xi\Big|X^{(n)},A^{(n)}\Big\}>\epsilon\Big\}\\
&\le P\Big\{E\big[\|F\|^2_{\tilde{P}_n}|X^{(n)},A^{(n)}\big]\le\frac{1}{2}E[F^2]\Big\}\\
&\qquad+P\Big\{P\Big\{\frac{\zeta_n^2}{\|F\|^2_{\tilde{P}_n}}>\xi\Big|X^{(n)},A^{(n)}\Big\}>\epsilon,E\big[\|F\|^2_{\tilde{P}_n}|X^{(n)},A^{(n)}\big]>\frac{1}{2}E[F^2]\Big\}\\
&\le P\Big\{E\big[\|F\|^2_{\tilde{P}_n}|X^{(n)},A^{(n)}\big]\le\frac{1}{2}E[F^2]\Big\}
\end{aligned}$$

$$
\begin{aligned}
&+ P\Big\{ P\Big\{ \{\zeta_n^2 > \frac{1}{4}\xi E[F^2]\} \cup \{ |\|F\|^2_{\tilde{P}_n} - E\big[\|F\|^2_{\tilde{P}_n}|X^{(n)}, A^{(n)}\big]| > \frac{1}{4}E[F^2]\} \Big| X^{(n)}, A^{(n)} \Big\} > \epsilon \Big\} , \\
&\qquad\qquad E\big[\|F\|^2_{\tilde{P}_n}|X^{(n)}, A^{(n)}\big] > \frac{1}{2}E[F^2] \Big\} \\
&\le P\Big\{ E\big[\|F\|^2_{\tilde{P}_n}|X^{(n)}, A^{(n)}\big] \le \frac{1}{2}E[F^2] \Big\} \\
&\qquad + P\Big\{ P\Big\{ \zeta_n^2 > \frac{1}{4}\xi E[F^2] \Big| X^{(n)}, A^{(n)} \Big\} > \frac{\epsilon}{2} \Big\} \\
&\qquad + P\Big\{ P\Big\{ \Big| \|F\|^2_{\tilde{P}_n} - E\big[\|F\|^2_{\tilde{P}_n}|X^{(n)}, A^{(n)}\big] \Big| > \frac{1}{4}E[F^2] \Big| X^{(n)}, A^{(n)} \Big\} > \frac{\epsilon}{2} \Big\} \xrightarrow{P} 0 ,
\end{aligned}
$$

where we use the fact that $E\big[\|F\|^2_{\tilde{P}_n}|X^{(n)}, A^{(n)}\big]$ is a function of $X^{(n)}, A^{(n)}$ and the convergence follows from (32) and (39)–(40).

Step 3(c). Fix $\epsilon > 0$. Following almost verbatim the first part of the proof of Theorem 2.5.2 in van der Vaart and Wellner (1996), with $\tilde{P}_n$ replacing the empirical measure, we obtain

$$
\begin{aligned}
&P\Big\{ \sup_{\theta\in\Theta^*:\|\theta-\theta_0\|<\delta_n} |\mathbb{B}_n(\theta) - \mathbb{B}_n(\theta_0)| > \epsilon \Big| X^{(n)}, A^{(n)} \Big\} \\
&\le \frac{1}{\epsilon} E\Big[\Big(\int_0^{\frac{\zeta_n}{\|F\|_{\tilde{P}_n}}} \sup_\nu \sqrt{\log N(\epsilon\|F\|_\nu, \mathcal{F}_{\theta_0,\infty}, L_2(\nu))}d\epsilon \Big)^2 \Big| X^{(n)}, A^{(n)} \Big]^{1/2} E\big[\|F\|^2_{\tilde{P}_n}||X^{(n)}, A^{(n)}\big]^{1/2} , \quad (42)
\end{aligned}
$$

where the supremum for ν is over probability measures with discrete supports. Also note that if $\|F\|_{\tilde{P}_n} = 0$ then the conditional expectation on the left-hand side is trivially zero, so we can without loss of generality assume $\|F\|_{\tilde{P}_n} > 0$. Assumption 4.3(b) implies

$$
\begin{aligned}
&E\Big[\Big(\int_0^{\frac{\zeta_n}{\|F\|_{\tilde{P}_n}}} \sup_\nu \sqrt{\log N(\epsilon\|F\|_\nu, \mathcal{F}_{\theta_0,\infty}, L_2(\nu))}d\epsilon \Big)^2 \Big| X^{(n)}, A^{(n)} \Big] \\
&\qquad\qquad \le \Big(\int_0^{\infty} \sup_\nu \sqrt{\log N(\epsilon\|F\|_\nu, \mathcal{F}_{\theta_0,\infty}, L_2(\nu))}d\epsilon \Big)^2 < \infty . \quad (43)
\end{aligned}
$$

We now argue

$$
E\Big[\Big(\int_0^{\frac{\zeta_n}{\|F\|_{\tilde{P}_n}}} \sup_\nu \sqrt{\log N(\epsilon\|F\|_\nu, \mathcal{F}_{\theta_0,\infty}, L_2(\nu))}d\epsilon \Big)^2 \Big| X^{(n)}, A^{(n)} \Big] \xrightarrow{P} 0 . \quad (44)
$$

Note the last inequality in (43) and the dominated convergence theorem implies that for every $\epsilon > 0$, there exists a $\xi > 0$ such that

$$
\Big(\int_0^{\xi} \sup_\nu \sqrt{\log N(\epsilon\|F\|_\nu, \mathcal{F}_{\theta_0,\infty}, L_2(\nu))}d\epsilon \Big)^2 < \epsilon . \quad (45)
$$

Then consider the following decomposition:

$$
\begin{aligned}
&E\Big[\Big(\int_0^{\frac{\zeta_n}{\|F\|_{\tilde{P}_n}}} \sup_\nu \sqrt{\log N(\epsilon\|F\|_\nu, \mathcal{F}_{\theta_0,\infty}, L_2(\nu))}d\epsilon \Big)^2 \Big| X^{(n)}, A^{(n)} \Big] \\
&= E\Big[\Big(\int_0^{\frac{\zeta_n}{\|F\|_{\tilde{P}_n}}} \sup_\nu \sqrt{\log N(\epsilon\|F\|_\nu, \mathcal{F}_{\theta_0,\infty}, L_2(\nu))}d\epsilon \Big)^2 I\Big\{ \frac{\zeta_n}{\|F\|_{\tilde{P}_n}} \le \xi \Big\} \Big| X^{(n)}, A^{(n)} \Big]
\end{aligned}
$$

$$+ E\left[\left(\int_0^{\frac{\zeta_n}{\|F\|_{\tilde{P}_n}}} \sup_\nu \sqrt{\log N(\epsilon\|F\|_\nu, \mathcal{F}_{\theta_0,\infty}, L_2(\nu))}d\epsilon\right)^2 I\left\{\frac{\zeta_n}{\|F\|_{\tilde{P}_n}} > \xi\right\}\middle| X^{(n)}, A^{(n)}\right]$$
$$\lesssim \epsilon + P\left\{\frac{\zeta_n}{\|F\|_{\tilde{P}_n}} > \xi\middle| X^{(n)}, A^{(n)}\right\} \xrightarrow{P} \epsilon \ ,$$

where the inequality follows from (43) and (45) and the convergence follows from (41). Because $\epsilon > 0$ was arbitrary, (44) follows.

It thus follows from (38), (42), and (44) that

$$P\left\{\sup_{\theta\in\Theta^*:\|\theta-\theta_0\|<\delta_n} |\mathbb{B}_n(\theta) - \mathbb{B}_n(\theta_0)| > \epsilon\middle| X^{(n)}, A^{(n)}\right\} \xrightarrow{P} 0 \ .$$

By the law of iterated expectations and the dominated convergence theorem we thus obtain

$$P\left\{\sup_{\theta\in\Theta^*:\|\theta-\theta_0\|<\delta_n} |\mathbb{B}_n(\theta) - \mathbb{B}_n(\theta_0)| > \epsilon\right\} \to 0 \ ,$$

as desired. ■

A.3 Auxiliary Lemmas

Lemma A.5. *Suppose (2) holds and* $\Pr\{A_i = 1|X_i = x\}$ *as a function is identical across* $1 \le i \le n$. *Then,*

$$(R_i(1), R_i(0)) \perp\!\!\!\perp A_i|X_i \ . \tag{46}$$

Moreover, (X_i, A_i, R_i) *is identically distributed across* $1 \le i \le n$.

Proof. Fix $a \in \{0, 1\}$ and any Borel sets $B \in \mathbf{R}^{d_r} \times \mathbf{R}^{d_r}$ and $C \in \mathbf{R}^{d_x}$.

$$\begin{aligned}
&E[\Pr\{(R_i(1), R_i(0)) \in B, A_i = a|X_i\}I\{X_i \in C\}] \\
&= E[E[\Pr\{(R_i(1), R_i(0)) \in B, A_i = a|X^{(n)}\}|X_i]I\{X_i \in C\}] \\
&= E[E[\Pr\{(R_i(1), R_i(0)) \in B|X^{(n)}\}\Pr\{A_i = a|X^{(n)}\}|X_i]I\{X_i \in C\}] \\
&= E[\Pr\{(R_i(1), R_i(0)) \in B|X_i\}\Pr\{A_i = a|X_i\}I\{X_i \in C\}] \ ,
\end{aligned}$$

where the first equality follows from the law of iterated expectations, the second equality follows from (2), the third equality follows from the law of iterated expectation as well as the facts that $Q_n = Q^n$ and $\Pr\{A_i = 1|X_i = x\}$ as a function is identical across $1 \le i \le n$. The first statement of the lemma then follows from the definition of a conditional expectation.

To prove the second statement, fix units i and i'. Clearly X_i and $X_{i'}$ are identically distributed. Conditional on X_i, for any Borel set $C \in \mathbf{R}^{d_r}$ and $a \in \{0, 1\}$, it follows (a) that

$$\Pr\{R_i \in C, A_i = a|X_i\} = \Pr\{A_i = a|X_i\}\Pr\{R_i(a) \in C|X_i\} \ .$$

The conclusion then follows because $\Pr\{A_i = 1|X_i = x\}$ is identical across $1 \leq i \leq n$ and $Q_n = Q^n$. ■

Lemma A.6. *Suppose the treatment assignment mechanism satisfies Assumptions 4.1–4.2 and the moment functions satisfy Assumption 4.3. Then,* $\hat{\theta}_n \xrightarrow{P} \theta_0$.

Proof of Lemma A.6. It follows from Assumption 4.3(a) and Theorem 5.9 in van der Vaart (1998) that we only need to establish for each $1 \leq s \leq d_\theta$,

$$\sup_{\theta\in\Theta} \left|\frac{1}{n}\sum_{1\leq i\leq n} m_s(X_i, A_i, R_i, \theta) - E_P[m_s(X_i, A_i, R_i, \theta)]\right| \xrightarrow{P} 0 \ . \tag{47}$$

To begin, note it follows from Assumption 4.3(d) and the dominated convergence theorem that if $m_s(x, a, r, \theta_m) \to m_s(x, a, r, \theta)$ as $m \to \infty$ for $\{\theta_m\} \subset \Theta^*$, then $E_P[m_s(X_i, A_i, R_i, \theta_m)] \to E_P[m_s(X_i, A_i, R_i, \theta)]$. Assumption 4.3(c) then implies

$$\begin{aligned}\sup_{\theta\in\Theta} \left|\frac{1}{n}\sum_{1\leq i\leq n} m_s(X_i, A_i, R_i, \theta) - E_P[m_s(X_i, A_i, R_i, \theta)]\right| \\ = \sup_{\theta\in\Theta^*} \left|\frac{1}{n}\sum_{1\leq i\leq n} m_s(X_i, A_i, R_i, \theta) - E_P[m_s(X_i, A_i, R_i, \theta)]\right| ,\end{aligned} \tag{48}$$

which is measurable. Next, note that

$$m(X_i, A_i, R_i, \theta) = A_i m(X_i, 1, R_i(1), \theta) + (1 - A_i)m(X_i, 0, R_i(0), \theta) \ . \tag{49}$$

and it follows from Lemma A.5 that

$$E_P[m(X_i, A_i, R_i, \theta)] = \frac{\ell}{k}E_Q[m(X_i, 1, R_i(1), \theta)] + \frac{k-\ell}{k}E_Q[m(X_i, 0, R_i(0), \theta)] \ , \tag{50}$$

which implies that

$$\begin{aligned}&\sup_{\theta\in\Theta^*} \left|\frac{1}{n}\sum_{1\leq i\leq n} m_s(X_i, A_i, R_i, \theta) - E_P[m_s(X_i, A_i, R_i, \theta)]\right| \\ &\leq \sup_{\theta\in\Theta^*} \left|\frac{1}{n}\sum_{1\leq i\leq n} A_i m_s(X_i, 1, R_i(1), \theta) - \frac{\ell}{k}E[m_s(X_i, 1, R_i(1), \theta)]\right| \\ &\qquad + \sup_{\theta\in\Theta^*} \left|\frac{1}{n}\sum_{1\leq i\leq n} (1 - A_i)m_s(X_i, 0, R_i(0), \theta) - \frac{k-\ell}{k}E[m_s(X_i, 0, R_i(0), \theta)]\right| .\end{aligned}$$

We study the first term on the right-hand side and similar arguments apply to the second term.

$$\begin{aligned}&\sup_{\theta\in\Theta^*} \left|\frac{1}{n}\sum_{1\leq i\leq n} A_i m_s(X_i, 1, R_i(1), \theta) - \frac{\ell}{k}E[m_s(X_i, 1, R_i(1), \theta)]\right| \\ &\leq \sup_{\theta\in\Theta^*} \left|\frac{1}{n}\sum_{1\leq i\leq n} A_i(m_s(X_i, 1, R_i(1), \theta) - E[m_s(X_i, 1, R_i(1), \theta)|X_i])\right|\end{aligned} \tag{51}$$

$$+ \sup_{\theta\in\Theta^*} \left| \frac{1}{n} \sum_{1\le i\le n} (A_i - \frac{\ell}{k}) E[m_s(X_i, 1, R_i(1), \theta)|X_i] \right| \tag{52}$$

$$+ \sup_{\theta\in\Theta^*} \left| \frac{\ell}{kn} \sum_{1\le i\le n} (E[m_s(X_i, 1, R_i(1), \theta)|X_i] - E[m_s(X_i, 1, R_i(1), \theta)]) \right| . \tag{53}$$

We study each term separately. First note (52) is bounded by

$$\sup_{\theta\in\Theta^*} \frac{1}{n} \sum_{1\le j\le n/k} |\zeta_j^*(\theta)| ,$$

where

$$\zeta_j^*(\theta) = \sup_{I\subset\lambda_j} \left\{ \frac{k-\ell}{k} \sum_{i\in I} E[m_s(X_i, 1, R_i(1), \theta)|X_i] - \frac{\ell}{k} \sum_{i\in\lambda_j\setminus I} E[m_s(X_i, 1, R_i(1), \theta)|X_i] : |I| = \ell \right\} .$$

By Assumption 4.3(e) and Assumption 4.2 we then obtain

$$\sup_{\theta\in\Theta^*} \frac{1}{n} \sum_{1\le j\le n/k} |\zeta_j^*(\theta)| \lesssim \frac{1}{n} \sum_{1\le j\le n/k} \max_{i,i'\in\lambda_j} \|X_i - X_{i'}\| \xrightarrow{P} 0 .$$

For (53), note the class of functions

$$\{E[m_s(X, 1, R(1), \theta)|X = x] : \theta \in \Theta^*\}$$

are Lipschitz continuous in x with a uniform Lipschitz constant. It therefore follows from Corollary 4.1 in van der Vaart (1994), applied with I_js being hypercubes, that (53) converges in probability to 0.

To analyze (51), we apply the arguments in the proof of Theorem 2.4.3 in van der Vaart and Wellner (1996) conditional on $X^{(n)}, A^{(n)}$. Define $F = \sup_{\theta\in\Theta^*} m_s(X, 1, R(1), \theta)$, which is measurable because Θ^* is countable by Assumption 4.3(c). Define for any $K > 0$

$$\mathcal{F}_s^K(1) = \{m_s(X, 1, R(1), \theta) I\{F \le K\} : \theta \in \Theta^*\} ,$$

and

$$\mathcal{F}_s(1) = \{m_s(X, 1, R(1), \theta) : \theta \in \Theta^*\} .$$

Next, let $\tau_i, 1 \le i \le n$ be a sequence of i.i.d. Rademacher random variables independent of all other variables. It follows from Markov's inequality and the symmetrization Lemma 6.3 in Ledoux and Talagrand (1991) applied conditional on $X^{(n)}, A^{(n)}$ for the distribution

$$\bigotimes_{1\le i\le n: A_i=1} P\{X_i, R_i(1)|X_i\}$$

that for every $\epsilon > 0$,

$$P\left\{ \sup_{\theta\in\Theta^*} \left|\frac{1}{n}\sum_{1\le i\le n} A_i(m_s(X_i,1,R_i(1),\theta) - E[m_s(X_i,1,R_i(1),\theta)|X_i])\right| > \epsilon \middle| X^{(n)}, A^{(n)} \right\} \tag{54}$$

$$\le \frac{1}{\epsilon} E\left[\sup_{\theta\in\Theta^*} \left|\frac{1}{n}\sum_{1\le i\le n} A_i(m_s(X_i,1,R_i(1),\theta) - E[m_s(X_i,1,R_i(1),\theta)|X_i])\right| \middle| X^{(n)}, A^{(n)} \right]$$

$$\le \frac{2}{\epsilon} E_P\left[E_\tau\left[\sup_{\theta\in\Theta^*} \left|\frac{1}{n}\sum_{1\le i\le n} \tau_i A_i m_s(X_i,1,R_i(1),\theta)\right|\right] \middle| X^{(n)}, A^{(n)} \right]$$

$$\le \frac{2}{\epsilon} E_P\left[E_\tau\left[\sup_{\theta\in\Theta^*} \left|\frac{1}{n}\sum_{1\le i\le n} \tau_i A_i \min\{m_s(X_i,1,R_i(1),\theta), K\}\right|\right] \middle| X^{(n)}, A^{(n)} \right]$$

$$+ 2E[FI\{F > K\}] \,, \tag{55}$$

where $E_\tau[\cdot]$ should be understood as the expectation with respect to $(\tau_i, 1 \le i \le n)$, holding all else fixed. The last term could be made as small as possible by choosing K large because of Assumption 4.3(d). Next, define

$$\tilde{P}_n = \frac{1}{\eta n} \sum_{1\le i\le n: A_i = 1} \delta_{(X_i, R_i(1))} \,,$$

where δ denotes the Dirac measure. Note that $\mathcal{F}_s(1)$ is a VC class by Assumption 4.3(b) and so $\mathcal{F}_s^K(1)$ is a VC-class by Lemma 2.6.18(vi) in van der Vaart and Wellner (1996), and thus both totally bounded in $L_1(\tilde{P}_n)$ by Theorem 2.6.7 in van der Vaart and Wellner (1996) (note that if $\|F\|_{L_1(\tilde{P}_n)} = 0$ then the conditional expectation immediately below is trivially zero, so we can without loss of generality assume $\|F\|_{L_1(\tilde{P}_n)} > 0$). Accordingly, define $\mathcal{G}$ to be an ϵ-net in $L_1(\tilde{P}_n)$ for $\mathcal{F}_s^K(1)$ with cardinality $N(\epsilon, \mathcal{F}_s^K(1), L_1(\tilde{P}_n))$. We have

$$E_\tau\left[\sup_{\theta\in\Theta^*} \left|\frac{1}{\eta n}\sum_{1\le i\le n} \tau_i A_i \min\{m_s(X_i,1,R_i(1),\theta), K\}\right|\right]$$

$$\le E_\tau\left[\sup_{f\in\mathcal{G}} \left|\frac{1}{\eta n}\sum_{1\le i\le n} \tau_i A_i f(X_i, R_i(1))\right|\right] + \epsilon$$

$$\le \sqrt{1 + \log N(\epsilon, \mathcal{F}_s^K(1), L_1(\tilde{P}_n))}\sqrt{\frac{12}{n}} \sup_{f\in\mathcal{G}} \left(\int f^2 d\tilde{P}_n\right)^{1/2} + \epsilon$$

$$\le (\sqrt{1 + \log N(\epsilon, \mathcal{F}_s^K(1), L_1(\tilde{P}_n))}\sqrt{\frac{12}{n}} K + \epsilon$$

$$\le (\sqrt{1 + \log N(\epsilon, \mathcal{F}_s(1), L_1(\tilde{P}_n))}\sqrt{\frac{12}{n}} K + \epsilon$$

$$\lesssim \left(\sqrt{1 + (V-1)\log\left(\frac{\|F\|_{L_1(\tilde{P}_n)}}{\epsilon}\right)}\sqrt{\frac{12}{n}} K + \epsilon\right) \wedge K \,, \tag{56}$$

where the second inequality follows from Lemma 2.2.2 in van der Vaart and Wellner (1996) applied with $\exp(x^2) - 1$ and Hoeffding's lemma, the third follows because $|f| \le K$ for $f \in \mathcal{G}$, the fourth inequality follows from the fact that $N(\epsilon, \mathcal{F}_s^K(1), L_1(\tilde{P}_n)) \le N(\epsilon, \mathcal{F}_s(1), L_1(\tilde{P}_n))$ because

$$\int |f_1 I\{F \le K\} - f_2 I\{F \le K\}| d\tilde{P}_n = \int |f_1 - f_2| I\{F \le K\} d\tilde{P}_n \le \int |f_1 - f_2| d\tilde{P}_n \,, \tag{57}$$

and the last inequality follows from Theorem 2.6.7 in van der Vaart and Wellner (1996) and Assumption 4.3(b).

Note it follows from Assumptions 4.1–4.2, 4.3(d)–(f), and similar arguments to those in the first part of the proof of Lemma S.1.5 in Bai et al. (2022) that

$$E\big[\|F\|_{L_1(\tilde{P}_n)}|X^{(n)},A^{(n)}\big]=\frac{1}{\eta n}\sum_{1\le i\le n}I\{A_i=1\}E[|F||X_i]\xrightarrow{P}E[|F|]\ . \tag{58}$$

We can assume without loss of generality $E[|F|]>0$ because otherwise $m_s(x,1,r(1),\theta)\equiv 0$. Therefore,

$$P\Big\{E\big[\|F\|_{L_1(\tilde{P}_n)}|X^{(n)},A^{(n)}\big]>E[|F|]+\frac{E[|F|]}{2}\Big\}\to 0\ . \tag{59}$$

On the other hand, Assumptions 4.1–4.2, 4.3(d)–(f) and similar arguments to those in the last part of the proof of Lemma S.1.5 in Bai et al. (2022) that

$$P\Big\{\Big|\|F\|_{L_1(\tilde{P}_n)}-E\big[\|F\|_{L_1(\tilde{P}_n)}|X^{(n)},A^{(n)}\big]\Big|>\frac{E[|F|]}{2}\Big|X^{(n)},A^{(n)}\Big\}\xrightarrow{P}0\ . \tag{60}$$

Let

$$\mathbb{L}_n=P\left\{\sup_{\theta\in\Theta^*}\Big|\frac{1}{n}\sum_{1\le i\le n}A_i(m_s(X_i,1,R_i(1),\theta)-E[m_s(X_i,1,R_i(1),\theta)|X_i])\Big|>\epsilon\Big|X^{(n)},A^{(n)}\right\}\ .$$

To conclude the proof, note for every $\eta>0$, for n large enough,

$$\begin{aligned}
&P\{\mathbb{L}_n>\eta\}\\
&\le P\Big\{E\big[\|F\|_{L_1(\tilde{P}_n)}|X^{(n)},A^{(n)}\big]>\frac{3}{2}E[|F|]\Big\}\\
&\qquad+P\Big\{\mathbb{L}_n>\eta,E\big[\|F\|_{L_1(\tilde{P}_n)}|X^{(n)},A^{(n)}\big]\le\frac{3}{2}E[|F|]\Big\}\\
&\le P\Big\{E\big[\|F\|_{L_1(\tilde{P}_n)}|X^{(n)},A^{(n)}\big]>\frac{3}{2}E[|F|]\Big\}\\
&\qquad+P\Bigg\{E\left[\left(\sqrt{1+(V-1)\log\left(\frac{\|F\|_{L_1(\tilde{P}_n)}}{\epsilon}\right)}\sqrt{\frac{12}{n}}K+\epsilon\right)\wedge K\Big|X^{(n)},A^{(n)}\right]>\eta'\ ,\\
&\qquad\qquad E\big[\|F\|_{L_1(\tilde{P}_n)}|X^{(n)},A^{(n)}\big]\le\frac{3}{2}E[|F|]\Bigg\}\\
&\le P\Big\{E\big[\|F\|_{L_1(\tilde{P}_n)}|X^{(n)},A^{(n)}\big]>\frac{3}{2}E[|F|]\Big\}\\
&\qquad+P\Big\{P\Big\{\Big|\|F\|_{L_1(\tilde{P}_n)}-E\big[\|F\|_{L_1(\tilde{P}_n)}|X^{(n)},A^{(n)}\big]\Big|>\frac{E[|F|]}{2}\Big|X^{(n)},A^{(n)}\Big\}>\eta'',\\
&\qquad\qquad E\big[\|F\|_{L_1(\tilde{P}_n)}|X^{(n)},A^{(n)}\big]\le\frac{3}{2}E[|F|]\Big\}\xrightarrow{P}0\ ,
\end{aligned}$$

where η',η'' are suitably chosen constants, the last line follows from the law of total expectation combined with the fact that the quantity in the expectation is bounded by K, and the convergence follows from (59)–(60). ■

Lemma A.7. *For $f : \mathbf{R}^{d_x} \to \mathbf{R}$, define*

$$h_f(x_1, \ldots, x_k) = \sup_{I \subset \{1,\ldots,k\}, |I|=\ell} \Big(\frac{1}{\ell} \sum_{i \in I} f(x_i) - \frac{1}{k-\ell} \sum_{i \in \{1,\ldots,k\} \setminus I} f(x_i) \Big) \\ - \inf_{I \subset \{1,\ldots,k\}, |I|=\ell} \Big(\frac{1}{\ell} \sum_{i \in I} f(x_i) - \frac{1}{k-\ell} \sum_{i \in \{1,\ldots,k\} \setminus I} f(x_i) \Big) \, .$$

Then,

$$|h_f(x_1, \ldots, x_k) - h_g(x_1, \ldots, x_k)|^2 \lesssim \sum_{1 \le i \le k} |f(x_i) - g(x_i)|^2$$

Proof. Suppose the supremum and infimum in the definition of h_f are attained at I^* and I_*. Then,

$$h_f - h_g \le \Big(\frac{1}{\ell} \sum_{i \in I^*} f(x_i) - \frac{1}{k-\ell} \sum_{i \in \{1,\ldots,k\} \setminus I^*} f(x_i) \Big) - \Big(\frac{1}{\ell} \sum_{i \in I^*} g(x_i) - \frac{1}{k-\ell} \sum_{i \in \{1,\ldots,k\} \setminus I^*} g(x_i) \Big) \\ + \Big(\frac{1}{\ell} \sum_{i \in I_*} f(x_i) - \frac{1}{k-\ell} \sum_{i \in \{1,\ldots,k\} \setminus I_*} f(x_i) \Big) - \Big(\frac{1}{\ell} \sum_{i \in I_*} g(x_i) - \frac{1}{k-\ell} \sum_{i \in \{1,\ldots,k\} \setminus I_*} g(x_i) \Big) \, ,$$

and the result follows from repeated applications of the inequality $(a+b)^2 \le 2(a^2+b^2)$. ■

Lemma A.8. *If $X_n \xrightarrow{P} 0$ and $|X_n| \le X$ with $E[X] < \infty$, then $E[X_n] \to 0$.*

Proof. Suppose not. Then along a subsequence $\{n_k\}$, $E[|X_{n_k}|] \to \delta > 0$. Because $X_n \xrightarrow{P} 0$, there exists a further subsequence along which $X_{n_{k_\ell}} \to 0$ with probability one, and by the dominated convergence theorem $E[X_{n_{k_\ell}}] \to 0$, a contradiction. ■

References

Abadie, A. and Imbens, G. W. (2008). Estimation of the Conditional Variance in Paired Experiments. *Annales d'Économie et de Statistique*, **1** 175–187.

Angrist, J. and Lavy, V. (2009). The Effects of High Stakes High School Achievement Awards: Evidence from a Randomized Trial. *American Economic Review*, **99** 1384–1414.

Armstrong, T. B. (2022). Asymptotic Efficiency Bounds for a Class of Experimental Designs. ArXiv:2205.02726 [stat], URL http://arxiv.org/abs/2205.02726.

Bai, Y. (2022). Optimality of Matched-Pair Designs in Randomized Controlled Trials. *American Economic Review*, **112** 3911–3940.

Bai, Y., Guo, H., Shaikh, A. M. and Tabord-Meehan, M. (2023a). Inference in experiments with matched pairs and imperfect compliance.

Bai, Y., Jiang, L., Romano, J. P., Shaikh, A. M. and Zhang, Y. (2023b). Covariate Adjustment in Experiments with Matched Pairs. ArXiv:2302.04380 [econ], URL http://arxiv.org/abs/2302.04380.

Bai, Y., Liu, J., Shaikh, A. M. and Tabord-Meehan, M. (2023c). Inference in Cluster Randomized Trials with Matched Pairs. ArXiv:2211.14903 [econ, stat], URL http://arxiv.org/abs/2211.14903.

Bai, Y., Liu, J. and Tabord-Meehan, M. (2023d). Inference for Matched Tuples and Fully Blocked Factorial Designs. ArXiv:2206.04157 [econ, math, stat], URL http://arxiv.org/abs/2206.04157.

Bai, Y., Romano, J. P. and Shaikh, A. M. (2022). Inference in Experiments With Matched Pairs. *Journal of the American Statistical Association*, **117** 1726–1737.

Banerjee, A., Duflo, E., Glennerster, R. and Kinnan, C. (2015). The Miracle of Microfinance? Evidence from a Randomized Evaluation. *American Economic Journal: Applied Economics*, **7** 22–53.

Belloni, A., Chernozhukov, V., Fernandez-Val, I. and Hansen, C. (2017). Program evaluation and causal inference with high-dimensional data. *Econometrica*, **85** 233–298.

Bruhn, M., Leão, L. d. S., Legovini, A., Marchetti, R. and Zia, B. (2016). The Impact of High School Financial Education: Evidence from a Large-Scale Evaluation in Brazil. *American Economic Journal: Applied Economics*, **8** 256–295.

Bruhn, M. and McKenzie, D. (2009). In pursuit of balance: Randomization in practice in development field experiments. *American economic journal: applied economics*, **1** 200–232.

Bugni, F., Canay, I., Shaikh, A. and Tabord-Meehan, M. (2022). Inference for Cluster Randomized Experiments with Non-ignorable Cluster Sizes. ArXiv:2204.08356 [econ, stat], URL http://arxiv.org/abs/2204.08356.

Bugni, F. A., Canay, I. A. and Shaikh, A. M. (2019). Inference under covariate-adaptive randomization with multiple treatments. *Quantitative Economics*, **10** 1747–1785.

Casaburi, L. and Reed, T. (2022). Using Individual-Level Randomized Treatment to Learn about Market Structure. *American Economic Journal: Applied Economics*, **14** 58–90.

Chen, X., Hong, H. and Tarozzi, A. (2008). Semiparametric Efficiency in GMM Models with Auxiliary Data. *The Annals of Statistics*, **36** 808–843.

Chen, X. and Santos, A. (2018). Overidentification in Regular Models. *Econometrica*, **86** 1771–1817.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C. and Newey, W. (2017). Double/Debiased/Neyman Machine Learning of Treatment Effects. *American Economic Review*, **107** 261–265.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W. and Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, **21** C1–C68.

Cytrynbaum, M. (2023a). Covariate adjustment in stratified experiments.

Cytrynbaum, M. (2023b). Designing representative and balanced experiments by local randomization. *arXiv preprint arXiv:2111.08157*.

Dizon-Ross, R. (2019). Parents' beliefs about their children's academic ability: Implications for educational investments. *American Economic Review*, **109** 2728–65.

Dudley, R. M. (2014). *Uniform Central Limit Theorems*. 2nd ed. Cambridge Studies in Advanced Mathematics, Cambridge University Press, Cambridge.

Duflo, E., Dupas, P. and Kremer, M. (2015). Education, hiv, and early fertility: Experimental evidence from kenya. *American Economic Review*, **105** 2757–2797.

Duflo, E., Glennerster, R. and Kremer, M. (2007). Using randomization in development economics research: A toolkit. *Handbook of development economics*, **4** 3895–3962.

Farrell, M. H. (2015). Robust inference on average treatment effects with possibly more covariates than observations. *Journal of Econometrics*, **189** 1–23.

Firpo, S. (2007). Efficient Semiparametric Estimation of Quantile Treatment Effects. *Econometrica*, **75** 259–276.

Frolich, M. (2007). Nonparametric IV estimation of local average treatment effects with covariates. *Journal of Econometrics*, **139** 35–75.

Hahn, J. (1998). On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects. *Econometrica*, **66** 315–331.

Heckman, J. J., Ichimura, H. and Todd, P. (1998). Matching as an econometric evaluation estimator. *The review of economic studies*, **65** 261–294.

Hirano, K., Imbens, G. W. and Ridder, G. (2003). Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score. *Econometrica*, **71** 1161–1189.

Imbens, G. W. and Angrist, J. D. (1994). Identification and Estimation of Local Average Treatment Effects. *Econometrica*, **62** 467–475.

Jiang, L., Linton, O. B., Tang, H. and Zhang, Y. (2022a). Improving Estimation Efficiency via Regression-Adjustment in Covariate-Adaptive Randomizations with Imperfect Compliance. ArXiv:2201.13004 [econ, stat], URL http://arxiv.org/abs/2201.13004.

Jiang, L., Liu, X., Phillips, P. C. and Zhang, Y. (2021). Bootstrap inference for quantile treatment effects in randomized experiments with matched pairs. *Review of Economics and Statistics* 1–47.

Jiang, L., Phillips, P. C. B., Tao, Y. and Zhang, Y. (2022b). Regression-Adjusted Estimation of Quantile Treatment Effects under Covariate-Adaptive Randomizations. ArXiv:2105.14752 [econ, stat], URL http://arxiv.org/abs/2105.14752.

Kosorok, M. R. (2008). *Introduction to Empirical Processes and Semiparametric Inference*. Springer Series in Statistics, Springer-Verlag, New York.

Krantz, S. G. and Parks, H. R. (2013). *The Implicit Function Theorem: History, Theory, and Applications*. Springer, New York, NY.

Ledoux, M. and Talagrand, M. (1991). *Probability in Banach Spaces: Isoperimetry and Processes*. Classics in Mathematics, Springer-Verlag, Berlin Heidelberg.

Lehmann, E. and Romano, J. P. (2022). *Testing Statistical Hypotheses*. Springer Texts in Statistics, Springer International Publishing, Cham.

Newey, W. K. (1994). The Asymptotic Variance of Semiparametric Estimators. *Econometrica*, **62** 1349–1382.

Newey, W. K. and McFadden, D. (1994). Large sample estimation and hypothesis testing. *Handbook of econometrics*, **4** 2111–2245.

Rafi, A. (2023). Efficient semiparametric estimation of average treatment effects under covariate adaptive randomization. *arXiv preprint arXiv:2305.08340*.

Robins, J. M., Rotnitzky, A. and Zhao, L. P. (1995). Analysis of Semiparametric Regression Models for Repeated Outcomes in the Presence of Missing Data. *Journal of the American Statistical Association*, **90** 106–121.

Rosenberger, W. F. and Lachin, J. M. (2015). *Randomization in clinical trials: theory and practice*. John Wiley & Sons.

Tsiatis, A. A. (2006). *Semiparametric Theory and Missing Data*. Springer Series in Statistics, Springer, New York, NY.

Tsiatis, A. A., Davidian, M., Zhang, M. and Lu, X. (2008). Covariate adjustment for two-sample treatment comparisons in randomized clinical trials: a principled yet flexible approach. *Statistics in Medicine*, **27** 4658–4677.

van der Vaart, A. (1989). On the Asymptotic Information Bound. *The Annals of Statistics*, **17**.

van der Vaart, A. (1994). Bracketing smooth functions. *Stochastic Processes and their Applications*, **52** 93–105.

van der Vaart, A. W. (1998). *Asymptotic statistics*, vol. 3 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge.

van der Vaart, A. W. and Wellner, J. (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer Series in Statistics, Springer-Verlag, New York.

Vershynin, R. (2018). *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press, Cambridge.

Zhang, M., Tsiatis, A. A. and Davidian, M. (2008). Improving efficiency of inferences in randomized clinical trials using auxiliary covariates. *Biometrics*, **64** 707–715.