# On the Efficiency of Finely Stratified Experiments [*]

Yuehao Bai

Department of Economics

University of Southern California

yuehao.bai@usc.edu

Jizhou Liu

HSBC Business School

Peking University

jizhou.liu@phbs.pku.edu.cn

Azeem M. Shaikh

Department of Economics

University of Chicago

amshaikh@uchicago.edu

Max Tabord-Meehan

Department of Economics

University of Toronto

m.tabordmeehan@utoronto.ca

September 19, 2025

**Abstract**

This paper studies the use of finely stratified designs for the efficient estimation of a large class of treatment effect parameters that arise in the analysis of experiments. 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 a binary treatment uniformly at random. The class of parameters considered are those that can be expressed as the solution to a set of moment conditions constructed using a known function of the observed data. They include, among other things, 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 this setting, we establish three results. First, we show that under a finely stratified design, the naïve method of moments estimator achieves the same asymptotic variance as what could typically be attained under alternative treatment assignment mechanisms only through *ex post* covariate adjustment. Second, we argue that the naïve method of moments estimator under a finely stratified design is asymptotically efficient by deriving a lower bound on the asymptotic variance of regular estimators of the parameter of interest in the form of a convolution theorem. In this sense, finely stratified experiments are attractive because they lead to efficient estimators of treatment effect parameters "by design." Finally, we strengthen this conclusion by establishing conditions under which a "fast-balancing" property of finely stratified designs is in fact necessary for the naïve method of moments estimator to attain the efficiency bound.

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

JEL classification codes: C12, C14

---

# 1 Introduction

This paper studies the use of finely stratified designs for the efficient estimation of a large class of treatment effect parameters that arise in the analysis of experiments. By a "finely stratified" design, we mean experiments in which units are divided into groups of a *fixed* size based on their covariate values and a proportion within each group is assigned to a binary treatment uniformly at random. The canonical example of such a design is a matched pairs design: here the fixed size of the groups equals two, and exactly one of the two units in each group is assigned to treatment at random, so that the marginal probability of treatment assignment is one half. More broadly, our use of the term "finely stratified" in this context contrasts with what we could call a "coarsely stratified" design, in which units are divided into a small set of *large* groups: see Example 4.2 below for a formal definition. The class of parameters considered are those that can be expressed as the solution to a set of moment conditions constructed using a known function of the observed data. This class of parameters includes many causal parameters of interest: average treatment effects (ATEs), 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 three results. First, we study the asymptotic properties of a naïve method of moments estimator under a finely stratified design. Here, by a naïve method of moments estimator, we mean an estimator constructed using a direct sample analog of the moment conditions. For example, in the case of the ATE, such an estimator is given by the Horvitz-Thompson estimator for the difference in means. We show that under a finely stratified design, the naïve method of moments estimator achieves the same asymptotic variance as what could typically be attained under alternative treatment assignment mechanisms only through *ex post* covariate adjustment using the same set of covariates. Such adjustment strategies 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). We further illustrate that this feature of finely stratified experiments stems from the way in which finely stratified designs balance treatment status across covariate values, a property we define formally below and refer to as "fast-balancing." Second, we derive a lower bound on the asymptotic variance of regular estimators of the parameter of interest in the form of a convolution theorem. This convolution theorem accommodates a large class of possible treatment assignment mechanisms, including covariate adaptive randomization (Efron, 1971; Wei, 1978; Zelen, 1974; Pocock and Simon, 1975; Hu and Hu, 2012; Bugni et al., 2018; Ye et al., 2022; Ma et al., 2020, 2024), re-randomization (Li and Ding, 2017; Li et al., 2018; Li and Ding, 2020; Li et al., 2020; Cytrynbaum, 2024), and fine stratification (Jiang et al., 2021; Bai et al., 2022; Cytrynbaum, 2023b). We show that the lower bound is attained by the naïve method of moments estimator under a finely stratified design. In this sense, the naïve method of moments estimator under a finely stratified design is asymptotically efficient. More succinctly, we say that finely stratified experiments lead to efficient estimators "by design." Finally, we strengthen this conclusion by characterizing all regular asymptotically linear estimators for a large class of treatment assignment mechanisms and use this characterization to establish conditions under which the fast-balancing property of finely stratified experiments is in fact a necessary condition for the naïve method of moments estimator to attain the efficiency bound.

Together, these results demonstrate that finely stratified experiments lead to efficient estimators that

prioritize transparency in that they preclude the researcher from "data snooping" associated with *ex post* nonparametric covariate adjustment. Importantly, concerns with this type of data snooping are not completely eliminated by typical pre-analysis plans because such adjustments involve choices, such as the choice of nonparametric estimator or tuning parameters, that are often not pre-registered prior to the experiment. The estimators are therefore attractive because they avoid performing nonparametric covariate adjustment in order to achieve efficiency and thereby remain "hands above the table" (Freedman, 2008; Lin, 2013).

Our paper builds upon two strands of literature. The first 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 derive the asymptotic behavior of the difference-in-means estimator of the ATE when treatment is assigned according to a matched pairs design, and Cytrynbaum (2023b), who develops related results for an experimental design referred to as "local randomization" that permits the proportion of units assigned to treatment to vary with the baseline covariates. Beyond settings that study estimation of the ATE, Bai et al. (2024b) develops 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. This paper, like those just mentioned, operates in a "super-population" framework, in which the outcomes and covariates are assumed to be drawn as an i.i.d. sample from a population distribution. This is in contrast to an alternative strand of the literature that studies finely stratified experiments from the design-based perspective (see, in particular, Imai, 2008; Imai et al., 2009; Fogarty, 2018a,b; Liu and Yang, 2020; Pashley and Miratrix, 2021; Bai et al., 2025b). To our knowledge, our paper is the first to analyze the properties of finely stratified experiments in a general framework that accommodates any parameter that can be characterized as the solution to a set of moment conditions involving a known function of the observed data. In more recent work, Cytrynbaum (2024) considers a similar framework to study finely stratified *re-randomized* experiments, which nest finely stratified experiments as a special case. However, his results assume that the moment functions which define the parameter of interest are continuous in a way that precludes parameters like the Quantile Treatment Effect. Moreover, we emphasize that none of the above papers formally establish the asymptotic efficiency of finely stratified experiments. The second strand of literature concerns bounds on the efficiency with which treatment effect parameters can be estimated in experiments. We note that, due to the potential for dependence in treatment assignments across individuals, we cannot immediately appeal to standard semi-parametric efficiency results (see, for example, van der Vaart, 1998; Chen et al., 2008). Two important recent papers in this literature studying efficiency bounds in the special case of estimating the ATE are Armstrong (2022) and Rafi (2023). Even in this special case, their results differ from ours in important and empirically relevant ways; Remark 4.3 provides an in-depth discussion of the connection between these results and ours. See also Bai (2022) for some finite-sample optimality properties of matched pairs designs for estimation of the ATE.

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. Section 3 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 studies estimation of the asymptotic variance. In Section 4.1, we develop our lower bound on the asymptotic variance of regular estimators of these parameters and show that it is achieved by the the naïve method of moments estimator in a finely stratified design. In Section 4.2, we characterize all regular asymptotically linear estimators and

argue that the fast-balancing property is necessary for the naïve method of moments estimator to attain the efficiency bound. In Section 5, we illustrate the practical relevance of our theoretical results through a simulation study. 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 $i$th 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 many 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 $X_i$ is some known constant for every $1 \le i \le n$, as is often the case in most experiments:

**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 $1 \le i \le n$.

Assumption 2.1 restricts the probability of assignment to be the fixed fraction $\eta$ across the entire experimental sample, but this restriction can be weakened so that $\eta$ is replaced by $\eta(X_i)$ for many of our subsequent results: see Remark 4.4 for a discussion. Given Assumption 2.1, it can be shown that $(X_i, A_i, R_i)$ are identically distributed for $1 \le i \le n$, and their marginal distribution does not change with $n$ (see Lemma A.7 in the Appendix). As a consequence, we denote the marginal distribution of $(X_i, A_i, R_i)$ by $P$. We consider parameters $\theta_0 \in \Theta \subset \mathbf{R}^{d_\theta}$ that can be defined as the solution to a set of moment equalities. As we show below, these parameters include a large class of causal parameters defined in terms of potential outcomes and potential treatments. Formally, let $m : \mathbb{R}^{d_x} \times \{0, 1\} \times \mathbf{R}^{d_r} \to \mathbf{R}^{d_\theta}$ be a known measurable function, then we consider parameters $\theta_0$ that uniquely solve the moment equality

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

We emphasize that $m(\cdot)$ is not a function of any unknown nuisance parameters, but may depend on the known value of $\eta$ in Assumption 2.1. We present five examples of well-known parameters that 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 $i$th 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, $\theta_0$ solves the moment condition in (3) with

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

For papers that consider estimators based on (4), see Hirano and Imbens (2001) and Hirano et al. (2003).  ∎

**Example 2.2** (Quantile Treatment Effect). Let $Y_i(a) = R_i(a)$ denote a scalar potential outcome for the $i$th 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

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

In other words, $\theta_0$ is defined to be the vector of $\tau$th 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 $Q\{Y(a) \le q_{Y(a)}(\tau) + \epsilon\} > Q\{Y(a) \le q_{Y(a)}(\tau)\}$ for all $\epsilon > 0$, then it follows from Assumption 2.1 and Lemma 1 in Firpo (2007) that $\theta_0$ solves the moment condition in (3) with

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

for $\theta = (\theta^{(1)}, \theta^{(0)})'$. Note that the quantile treatment effect $q_{Y(1)}(\tau) - q_{Y(0)}(\tau)$ can then be defined as $h(\theta_0)$ where $h : \mathbb{R}^2 \to \mathbb{R}$ 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 ($\tilde{Y}_i(a) \in \mathbb{R}$) and treatment take-up ($D_i(a) \in \{0, 1\}$) 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 $E_Q[D_i(1) - D_i(0)] \ne 0$ 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 $\theta_0$ solves the moment condition in (3) with

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

If we further assume instrument monotonicity (i.e., $P\{D_i(1) \ge D_i(0)\} = 1$) and instrument exclusion, then $\theta_0$ 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 $i$th 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 $\omega : \mathbb{R}^{d_x} \to \mathbb{R}$. It then follows from Assumption 2.1 that $\theta_0$ solves the moment condition in (3) with

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

By defining $Y_i$ to be the average outcome in the $i$th cluster, $\theta_0$ 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. (2024b) 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 $i$th 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

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

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

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

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

where $\mathrm{expit}(z) = \frac{\exp(z)}{1 + \exp(z)}$. The log-odds ratio can then be defined as $h(\theta_0)$ where $h : \mathbb{R}^2 \to \mathbb{R}$ is given by $h(s, t) = t$. This parameter appears in, for example, Zhang et al. (2008). ∎

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, an economic 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 $\theta_0$ which is constructed as a solution to the sample analogue of (3):

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

Because $\hat{\theta}_n$ is constructed directly using the moment function $m(\cdot)$, we call $\hat{\theta}_n$ the naïve method of moments estimator. Note that $\hat{\theta}_n$ as defined in (6) is closely related to standard estimators of the parameter $\theta_0$ 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 ATE. 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.

Before proceeding, in the remainder of this section, we provide a more detailed summary of the main contributions of our paper. To this end, first note that 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 the naïve method of moments estimator satisfies

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

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 show that if we assign $A^{(n)}$ using a finely stratified design (see Assumption 3.1 below for a formal definition) then, under appropriate assumptions so that we achieve "fast balance" of the treatment across covariate values (see Assumptions 3.2, 3.3 below),

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

where $\mathbb{V} \geq \mathbb{V}_*$ (see Theorem 3.1). Under i.i.d. assignment, the naïve method of moment estimator $\hat{\theta}_n$ cannot generally attain $\mathbb{V}_*$, but an estimator that attains $\mathbb{V}_*$ could instead be constructed by appropriately "augmenting" the moment function, and then considering an estimator which solves the augmented moment equation. For instance, if we consider the ATE in Example 2.1, then it is straightforward to show that the following augmented moment function identifies $\theta_0$:

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

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, in which $\mu_1(\cdot)$ and $\mu_0(\cdot)$ are non-parametrically estimated in a first step, attains $\mathbb{V}_*$ (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 covariate adjustment by exploiting the information contained in $X^{(n)}$ that may not have been captured in the original moment function $m(\cdot)$. Similar nonparametric covariate 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 this sense, we show that fine stratification can perform nonparametric covariate adjustment "by design" for the large class of parameters that can be expressed in terms of moment conditions of the form given in (3), thus generalizing similar observations made in Bai et al. (2022), Bai (2022), and Cytrynbaum (2023b) in the special case of estimating the ATE. As we explain in the discussion following Theorem 3.1, this feature of finely stratified experiments is possible because fine stratification leads to fast-balancing of the treatment across covariate values, as defined formally in (14) below.

Earlier work on efficient treatment effect estimation has noted that the variance $\mathbb{V}_*$ is in fact the efficiency bound for estimating $\theta_0$ under i.i.d. assignment (see, for instance, Cattaneo, 2010). A natural follow-up question is whether or not $\mathbb{V}_*$ continues to be the efficiency bound for estimating $\theta_0$ under a finely stratified design, or more generally for complex experimental designs which induce dependence in the treatment assignments across individuals in the experiment. In Section 4.1, we show that $\mathbb{V}_*$ continues to be the efficiency bound for estimating $\theta_0$ for a large class of treatment assignment mechanisms with a fixed marginal probability of treatment assignment, which includes finely stratified designs as a special case. We can thus conclude that, from the perspective of asymptotic efficiency, finely stratified designs are optimal experimental designs for a broad range of treatment effect estimation problems. In Section 4.2 we build on this result and establish conditions under which efficient estimation of $\theta_0$ using the naïve method of moments estimator can be achieved only if the experimental design leads to fast-balancing of the treatment across covariate values. In this sense, we show that the fast-balancing property of finely stratified experiments is in fact a *necessary* condition for achieving efficient estimation of $\theta_0$ "by design."

# 3    The Asymptotic Variance of Finely Stratified Experiments

In this section, we derive the asymptotic distribution of the method of moments estimator $\hat{\theta}_n$ when treatment is assigned by fine stratification over the baseline covariates $X^{(n)}$. Such assignment mechanisms use the covariates $X^{(n)}$ to group units with similar covariate values into blocks 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 $\ell$ and $k$ be arbitrary positive integers with $\ell < k$ 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. We note, however, that our framework is not intended to reflect settings in which the blocks themselves are randomly sampled from the population of interest (see, for instance, the discussion in Section 4.4 of Pashley and Miratrix, 2021). Given such a partition, we assume that treatment status is assigned as described in the following assumption:

**Assumption 3.1.** Treatment status is assigned so that $(R^{(n)}(1), R^{(n)}(0)) \perp\!\!\!\perp A^{(n)} | X^{(n)}$ 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 $(\underbrace{0, 0, \ldots, 0}_{k-\ell}, \underbrace{1, 1, \ldots, 1}_{\ell})$.

The assignment mechanism described in Assumptions 3.1 generalizes the definition of a matched pairs design. In particular, we recover a matched pairs design if we set $(\ell, k) = (1, 2)$, with $\eta = 1/2$. Indeed, 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 $X^{(n)}$, $\pi$ encompasses a broad variety of ways of pairing the $n$ units according to the observed, baseline covariates $X^{(n)}$. Given such a $\pi$, we assume that treatment status is assigned so that Assumption 3.1 holds 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).

**Remark 3.1.** Note that Assumption 3.1 generalizes matched pairs designs along two dimensions: first, it allows for treatment fractions other than $\eta = 1/2$. Second, it allows for choices of $\ell$ 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 3.1 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 Section 3.1 for details, and Imbens (2011) for an early 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 3.2.** 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 that satisfy Assumption 3.2. When $X_i \in \mathbb{R}$ and $E_Q[X_i^2] < \infty$, a simple algorithm that satisfies Assumption 3.2 is simply to order units from smallest to largest and then block adjacent units into blocks of size $k$. In the case of matched pairs, if $\dim(X_i) > 1$ and $E_Q[\|X_i\|^d] < \infty$ for $d \geq \dim(X_i) + 1$, then Assumption 3.2 is satisfied by the `nbpmatching` algorithm in R that minimizes the sum of squared distances of $X$ within pairs. See Appendix A of Bai et al. (2024b) for details. Beyond the case of pairs, Cytrynbaum (2023b) demonstrates that the optimal blocking satisfies the following bound

$$\frac{1}{n} \sum_{1 \leq j \leq n/k} \max_{i, i' \in \lambda_j} \|X_i - X_{i'}\|^2 = O_P(n^{2/d - 2/(\dim(X_i) + 1)}) ,$$

from which we can deduce that the rate of convergence depends on both the dimension of $X_i$ as well as the number of moments it possesses. Finally, we impose the following assumptions to derive the large-sample properties of $\hat{\theta}_n$. In what follows, when writing expectations and variances, we suppress the subscripts $P$ and $Q$ whenever doing so does not lead to confusion.

**Assumption 3.3.** Let $m(\cdot) = (m_s(\cdot) : 1 \le s \le d_\theta)'$. 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) $E[m(X_i, A_i, R_i, \theta)]$ is differentiable at $\theta_0$ with a nonsingular derivative $M = \frac{\partial}{\partial \theta'} E[m(X, A, R, \theta)]\big|_{\theta = \theta_0}$.

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

(d) For $1 \le s \le d_\theta$, $\{m_s(x, a, r, \theta) : \theta \in \Theta\}$ is pointwise measurable in the sense that there exists a countable set $\Theta^*$ 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 $x, a, r$.

(e) (i) $\sup_{\theta \in \Theta} E[\|m(X, a, R(a), \theta)\|] < \infty$ for $a \in \{0, 1\}$. (ii) $\{m_s(x, 1, r, \theta) : \theta \in \Theta^*\}$ and $\{m_s(x, 0, r, \theta) : \theta \in \Theta^*\}$ are $Q$-Donsker for $1 \le s \le d_\theta$.

(f) For $a \in \{0, 1\}$, $E[m_s(X, a, R(a), \theta_0) | X = x]$ is $C$-Lipschitz for $1 \le s \le d_\theta$, for some constant $C < \infty$.

Assumption 3.3(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 3.3(b) is a standard assumption used when deriving the properties of $Z$-estimators. 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 3.3(c) requires the moment function to be mean-square continuous in $\theta$. Assumption 3.3(d) 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 3.3(e) is a standard assumption which guarantees the existence of an integrable envelope and allows us to invoke a uniform law of large numbers and a uniform central limit theorem (see, for instance, page 81 of van der Vaart and Wellner (1996) for a definition of a Donsker class). In particular, this assumption can be verified for Examples 2.1–2.5. Assumption 3.3(f) is a common assumption which simplifies some arguments when studying finely stratified designs, and ensures units that are close in terms of the baseline covariates are also close in terms of their moments. Note that Assumption 3.3(f) could be dropped following the approximation arguments in Lemma C.5 of Cytrynbaum (2023b); see also Examples 4.4 and 4.9 for further discussion.

The following theorem establishes the asymptotic variance of the naïve method of moments estimator when the treatment assignment mechanism is finely stratified in the sense of satisfying Assumptions 3.1–3.2. Its proof relies on a crucial technical result in Han and Wellner (2021), which allows us to compare the empirical process that depends on the treatment assignments with the empirical process that only depends on i.i.d. quantities. As a consequence of us leveraging this result, Assumption 3.3 is comparable to the typical assumptions imposed to study the properties of method of moments estimators using i.i.d. data.

**Theorem 3.1.** *Suppose the treatment assignment mechanism satisfies Assumptions 3.1–3.2 and the moment functions satisfy Assumption 3.3. Let $\hat{\theta}_n$ be defined as in (6). Then,*

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = \frac{1}{\sqrt{n}} \sum_{1 \le i \le n} \psi^*(X_i, A_i, R_i, \theta_0) + o_P(1) \ . \tag{9}$$

*where*

$$
\begin{aligned}
&\psi^*(X_i, A_i, R_i, \theta_0) \\
&= -M^{-1}\Big( I\{A_i = 1\}(m(X_i, 1, R_i, \theta_0) - E[m(X_i, 1, R_i(1), \theta_0)|X_i]) \\
&\qquad + I\{A_i = 0\}(m(X_i, 0, R_i, \theta_0) - E[m(X_i, 0, R_i(0), \theta_0)|X_i]) \\
&\qquad + \eta E[m(X_i, 1, R_i(1), \theta_0)|X_i] + (1 - \eta)E[m(X_i, 0, R_i(0), \theta_0)|X_i] \Big) \ .
\end{aligned}
$$

*Further, we have that*

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \overset{d}{\to} N(0, \mathbb{V}_*) \ , \tag{10}$$

*where*

$$\mathbb{V}_* = \mathrm{Var}[\psi^*(X_i, A_i, R_i, \theta_0)] \ . \tag{11}$$

In order to make Theorem 3.1 useful for inference about $\theta_0$, we describe in Section 3.1 an estimator $\hat{\mathbb{V}}_n$ of $\mathbb{V}_*$. We now sketch an argument of the proof of Theorem 3.1 to highlight the fundamental role played by a "fast-balancing" property of finely stratified designs (see (14) below). In the proof of Theorem 3.1, we first establish that

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = -M^{-1}\frac{1}{\sqrt{n}} \sum_{1 \le i \le n} m(X_i, A_i, R_i, \theta_0) + o_P(1) \ . \tag{12}$$

To further establish (9), it thus suffices to show that, under a finely stratified design,

$$-M^{-1}\frac{1}{\sqrt{n}} \sum_{1 \le i \le n} m(X_i, A_i, R_i, \theta_0) = \frac{1}{\sqrt{n}} \sum_{1 \le i \le n} \psi^*(X_i, A_i, R_i, \theta_0) + o_P(1) \ . \tag{13}$$

To obtain this equivalence, consider the following decomposition of $m(\cdot)$:

$$
\begin{aligned}
&m(X_i, A_i, R_i, \theta_0) \\
&= \eta E[m(X_i, 1, R_i(1), \theta_0)|X_i] + (1 - \eta)E[m(X_i, 0, R_i(0), \theta_0)|X_i] \\
&\qquad + I\{A_i = 1\}(m(X_i, 1, R_i, \theta_0) - E[m(X_i, 1, R_i(1), \theta_0)|X_i]) \\
&\qquad + I\{A_i = 0\}(m(X_i, 0, R_i, \theta_0) - E[m(X_i, 0, R_i(0), \theta_0)|X_i]) \\
&\qquad + (A_i - \eta)(E[m(X_i, 1, R_i(1), \theta_0) - m(X_i, 0, R_i(0), \theta_0)|X_i]) \ .
\end{aligned}
$$

Then the equivalence follows if we can show that

$$\frac{1}{\sqrt{n}} \sum_{1 \le i \le n} (A_i - \eta)\left(E[m(X_i, 1, R_i(1), \theta_0) - m(X_i, 0, R_i(0), \theta_0)|X_i]\right) = o_P(1) \ . \tag{14}$$

We call (14) the fast-balancing condition for the function $m(\cdot)$. Intuitively, the fast-balancing condi-

tion imposes that the experimental design should balance the treatment across covariate values at a rate which is faster than sampling variation. To see why (14) holds for a finely stratified design, let $\Omega(X_i) = E[m(X_i, 1, R_i(1), \theta_0) - m(X_i, 0, R_i(0), \theta_0)|X_i]$ and first note that, by Assumption 2.1,

$$E\left[\frac{1}{\sqrt{n}} \sum_{1 \le i \le n} (A_i - \eta)\Omega(X_i)\Big| X^{(n)}\right] = 0 \ .$$

Next, for $1 \le s \le d_\theta$, let $\Omega^{(s)}(X_i)$ denote the $s$th component of $\Omega(X_i)$. Then it can be shown using Assumption 3.1 and 3.3(f) that for $1 \le s \le d_\theta$,

$$\text{Var}\left[\frac{1}{\sqrt{n}} \sum_{1 \le i \le n} (A_i - \eta)\Omega^{(s)}(X_i)\Big| X^{(n)}\right] \le C^2 \frac{\ell(k-\ell)}{k-1} \left(\frac{1}{n} \sum_{1 \le j \le n/k} \max_{i,i' \in \lambda_j} \|X_i - X_{i'}\|^2\right) \ ,$$

where $C$ denotes the Lipschitz constant in Assumption 3.3(f), and so the conditional variance converges in probability to zero under Assumption 3.2. The fast-balancing condition (14) then follows by an application of Chebyshev's inequality conditional on $X^{(n)}$ and the dominated convergence theorem. In Section 4.2, we further argue that the fast-balancing condition (14) is in fact a *necessary* condition which a given experimental design must satisfy to ensure (13).

**Remark 3.2.** Note it follows from (3) that

$$\eta E_Q[m(X_i, 1, R_i(1), \theta_0)] + (1 - \eta)E_Q[m(X_i, 0, R_i(0), \theta_0)] = E_P[m(X_i, A_i, R_i, \theta_0)] = 0 \ , \tag{15}$$

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

$$\begin{aligned}
\mathbb{V}_* &= \text{Var}[\psi^*(X_i, A_i, R_i, \theta_0)] \\
&= M^{-1}\big(E\big[\eta\,\text{Var}[m(X_i, 1, R_i(1), \theta_0)|X_i] + (1-\eta)\,\text{Var}[m(X_i, 0, R_i(0), \theta_0)|X_i]\big] \\
&\quad + \text{Var}\big[\eta E[m(X_i, 1, R_i(1), \theta_0)|X_i] + (1-\eta)E[m(X_i, 0, R_i(0), \theta_0)|X_i]\big]\big)(M^{-1})'
\end{aligned} \tag{16}$$

For instance, in the special case of the ATE (Example 2.1) we obtain that

$$\begin{aligned}
\text{Var}[\psi^*(X_i, A_i, R_i, \theta_0)] = E\bigg[&\frac{\text{Var}[Y_i(1)|X_i]}{\eta} + \frac{\text{Var}[Y_i(0)|X_i]}{1-\eta} \\
&+ (E[Y_i(1) - Y_i(0)|X_i] - E[Y_i(1) - Y_i(0)])^2\bigg] \ ,
\end{aligned} \tag{17}$$

which matches the asymptotic variance derived in Bai et al. (2022) for matched pairs. Theorem 3.1 however accommodates a much larger class of parameters, including those introduced in Examples 2.2–2.5. ∎

**Remark 3.3.** By comparing the variance expression in (7) to the variance expression for $\mathbb{V}_*$, we obtain

$$\mathbb{V} - \mathbb{V}_* = \eta(1-\eta)M^{-1}\,\text{Var}[E[m(X_i, 1, R_i(1), \theta_0) - m(X_i, 0, R_i(0), \theta_0)|X_i]](M^{-1})' \ , \tag{18}$$

which is positive semidefinite. From this, we conclude that the asymptotic variance of the naive method of moments estimator $\hat{\theta}_n$ is lower in a finely stratified design compared to i.i.d. assignment. In Section 4.1, we will further show that $\mathbb{V}_*$ is the lowest possible asymptotic variance among regular estimators for $\theta_0$

in a large class of treatment assignment mechanisms, including both i.i.d. assignment and finely stratified designs. When $d_\theta = 1$, we may express $\mathbb{V} - \mathbb{V}_*$ in terms of the "nonparametric $R^2$." In particular, $\mathbb{V} - \mathbb{V}_*$ is proportional to $E[R^2_{g,X}(g_i, X_i) \operatorname{Var}[g_i]]$, where

$$R^2_{g,X}(g_i, X_i) = \frac{\operatorname{Var}[E[g_i|X_i]]}{\operatorname{Var}[g_i]} \ , \tag{19}$$

and $g_i = m(X_i, 1, R_i(1), \theta_0) - m(X_i, 0, R_i(0), \theta_0)$. The quantity in (19) measures how much of the variation in $g_i$ can be explained nonparametrically by $X_i$. See, for instance, Chernozhukov et al. (2024). ∎

**Remark 3.4.** Note that finely stratified designs include as a special case completely randomized designs, i.e., experiments in which a fixed proportion of the entire sample is assigned to treatment uniformly at random. To see this, consider, for instance, simply matching on an exogenously generated covariate. In this special case, we find from (18) that $\mathbb{V}_* = \mathbb{V}$ as defined in Section 2. In this way, we see that completely randomized experiments are asymptotically no more efficient than i.i.d. assignment. ∎

## 3.1 Variance Estimation

In this subsection, we provide a consistent variance estimator for the asymptotic variance $\mathbb{V}_*$ in (16). We suppose that a consistent estimator $\widehat{M}_n$ for $M$ is available, i.e., $\widehat{M}_n \xrightarrow{P} M$. In examples where $m$ is differentiable in $\theta$, including Examples 2.1 and 2.3–2.5, the analog principle suggests that a natural estimator for $M$ is given by

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

In examples including Example 2.2 where $m$ is nonsmooth in $\theta$, $M$ may consist of components that require nonparametric estimators. See, for instance, Jiang et al. (2021).

It then suffices to construct a consistent estimator for the "meat" in (16). To motivate such an estimator, consider the expression in (16) when $d_\theta = 1$. By the law of total variance, this middle component equals $\Sigma_1 + \Sigma_2$, where

$$\Sigma_1 = \eta \operatorname{Var}[m(X_i, 1, R_i(1), \theta_0)] + (1 - \eta) \operatorname{Var}[m(X_i, 0, R_i(0), \theta_0)]$$

$$\Sigma_2 = -\eta(1 - \eta) E \big[ \big( E[m(X_i, 1, R_i(1), \theta_0)|X_i] - E[m(X_i, 1, R_i(1), \theta_0)]$$
$$- (E[m(X_i, 0, R_i(0), \theta_0)|X_i] - E[m(X_i, 0, R_i(0), \theta_0)]) \big)^2 \big]$$
$$= -\eta(1 - \eta) \Big( E[E[m(X_i, 1, R_i(1), \theta_0)|X_i]^2] + E[E[m(X_i, 0, R_i(0), \theta_0)|X_i]^2]$$
$$- 2E[E[m(X_i, 1, R_i(1), \theta_0)|X_i] E[m(X_i, 0, R_i(0), \theta_0)|X_i]]$$
$$- (E[m(X_i, 1, R_i(1), \theta_0)] - E[m(X_i, 0, R_i(0), \theta_0)])^2 \Big) \ .$$

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

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

where $\eta_1 = \eta$ and $\eta_0 = 1 - \eta$. The analog principle suggests that a natural estimator for $\Sigma_1$ is

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

To estimate $\Sigma_2$, we first define

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

and $\hat{\varsigma}_n(0,1)$ similarly. Next, define

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

Similarly, define

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

Finally, define

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

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. (2024b), and Cytrynbaum (2023b). The estimator $\hat{\varsigma}_n(0,0)$ is constructed similarly. A natural estimator for $\mathbb{V}_*$ is then given by

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

Note that for specific choices of $m(\cdot)$, $\hat{\mathbb{V}}_n$ recovers estimators which have been studied in prior work on inference in finely stratified experiments. For instance, in the case of matched pairs with $(\ell, k) = (1, 2)$ and

$m(\cdot)$ as in Example 2.1, so that $\theta_0$ is the ATE, $\hat{\mathbb{V}}_n$ exactly coincides with the estimator defined in equation (28) of Bai et al. (2022).

In addition to Assumption 3.2, we will now also require that the distances between units in *adjacent* blocks be "close" in terms of their baseline covariates:

**Assumption 3.4.** The blocks used in determining treatment status satisfy

$$\frac{1}{n} \sum_{1 \le j \le \lfloor n/2 \rfloor} \max_{i \in \lambda_{2j-1}, i' \in \lambda_{2j}} \|X_i - X_{i'}\|^2 \overset{P}{\to} 0 \ .$$

Note that given blocks which satisfy Assumption 3.2, it is always possible to re-order the blocks such that the pairs $\{(\lambda_{2j-1}, \lambda_{2j})\}_{1 \le j \le \lfloor n/2 \rfloor}$ satisfy Assumption 3.4, as long as we maintain the sufficient condition that $E[\|X_i\|^d] < \infty$ for $d \ge \dim(X_i) + 1$. This property could be achieved, for instance, by applying the `nbpmatching` algorithm to the block-means $\{\bar{X}_j\}_{1 \le j \le n/k}$, where $\bar{X}_j := \frac{1}{k} \sum_{i \in \lambda_j} X_i$: see Lemma A.6 in Cytrynbaum (2023a) for details.

To formally establish the consistency of $\hat{\mathbb{V}}_n$, we impose the following mild uniform integrability condition, as well as a Glivenko-Cantelli and uniform Lipschitz condition. All three conditions need only hold in an arbitrarily small neighborhood of $\theta_0$.

**Assumption 3.5.** There exists $\delta > 0$ such that

(a) For $a \in \{0, 1\}$,

$$\lim_{\lambda \to \infty} E \left[ \sup_{\theta \in \Theta : \|\theta - \theta_0\| < \delta} \|m(X_i, a, R_i(a), \theta)\|^2 I \left\{ \sup_{\theta \in \Theta} \|m(X_i, a, R_i(a), \theta)\| > \lambda \right\} \right] = 0 \ .$$

(b) $\{E[m_s(X_i, a, R_i(a), \theta)|X_i = x] : \|\theta - \theta_0\| < \delta\}$ and $\{E[m_s(X_i, a, R_i(a), \theta)m(X_i, a, R_i(a), \theta)'|X_i = x] : \|\theta - \theta_0\| < \delta\}$ are $Q$-Glivenko Cantelli for $1 \le s \le d_\theta$.

(c) For $a \in \{0, 1\}$, each component of $E[m(X, a, R(a), \theta)|X = x]$ and $E[m(X, a, R(a), \theta)m(X, a, R(a), \theta)'|X = x]$ is Lipschitz with a common Lipschitz constant across $\{\theta \in \Theta : \|\theta - \theta_0\| < \delta\}$.

Assumption 3.5(a) is a mild uniform integrability condition for the envelope function of the moment function in an arbitrarily small neighborhood of $\theta_0$. Assumption 3.5(b) is a mild condition that requires the conditional expectation of the moment functions to satisfy a uniform law of large numbers. See page 81 of van der Vaart and Wellner (1996) for a definition of a Glivenko-Cantelli class of functions. Assumption 3.5(c) strengthens Assumption 3.3(f) to hold uniformly in an arbitrarily small neighborhood of $\theta_0$.

The following theorem establishes the consistency of $\hat{\mathbb{V}}_n$ for $\mathbb{V}_*$:

**Theorem 3.2.** *Suppose the treatment assignment mechanism satisfies Assumptions 3.1, 3.2, and 3.4 and the moment functions satisfy Assumptions 3.3 and 3.5. Further suppose $\widehat{M}_n \overset{P}{\to} M$. Then, $\hat{\mathbb{V}}_n \overset{P}{\to} \mathbb{V}_*$.*

# 4    An Efficiency Bound and the Necessity of "Fast-Balancing"

In Section 4.1 we establish that $\mathbb{V}_*$ is the efficiency bound for a large class of experimental designs. As a consequence, we can conclude that finely stratified designs are asymptotically efficient "by design." Building on this result, in Section 4.2 we establish that a necessary condition for achieving the bound $\mathbb{V}_*$ when estimating $\theta_0$ using the naïve method of moments estimator is that the experimental design be fast-balancing, in the sense of (14).

## 4.1    Efficiency Bound

An inspection of the asymptotic variance in (11) reveals that $\mathbb{V}_*$ in fact coincides with the classical efficiency bound for estimating $\theta_0$ with i.i.d. assignment. For example, the variance derived in (17) coincides with the efficiency bound derived in Hahn (1998) for estimating the ATE with a known marginal treatment probability $\eta$. Therefore, another way to interpret our result in Theorem 3.1 is that the standard i.i.d. efficiency bound can be attained by a naïve method of moments estimator under a finely stratified design. On the other hand, because treatment status is not independent in a finely stratified design, a natural follow-up question is whether or not the efficiency bound for estimating $\theta_0$ changes relative to what can be obtained under i.i.d. assignment once we allow for more general assignment mechanisms. In this section, we show that $\mathbb{V}_*$ continues to be the efficiency bound for the class of parameters introduced in Section 2, while allowing for a more general class of treatment assignment mechanisms. The main restriction on treatment assignment is given by Assumption 2.1, which requires the marginal treatment probability to be known and equal to $\eta$. As mentioned earlier and explained in Remark 4.4 below, it is possible to relax this requirement so that $\eta$ can be replaced by a known function $\eta(X_i)$. For a discussion of how our efficiency bound compares with other results in the literature, see Remark 4.3.

We impose the following high-level assumption on the assignment mechanism:

**Assumption 4.1.** The treatment assignment mechanism is such that for any integrable function $\gamma : \mathbb{R}^{d_x} \to \mathbb{R}$,

$$\frac{1}{n} \sum_{1 \le i \le n} A_i \gamma(X_i) \overset{P}{\to} \eta E[\gamma(X_i)] \ .$$

In words, Assumption 4.1 requires that the assignment mechanism admits a law of large numbers for integrable functions of the covariate values. Examples 4.1–4.5 illustrate that the assumption holds for common treatment assignment mechanisms used in practice.

**Example 4.1** (i.i.d. assignment)**.** Let $A^{(n)}$ be assigned i.i.d., independently of $X^{(n)}$, such that $P\{A_i = 1\} = \eta$. Then it follows immediately by the law of large numbers that Assumption 4.1 is satisfied.    ∎

**Example 4.2** (Covariate-adaptive randomization (CAR))**.** Let $S : \mathbf{R}^{d_x} \to \mathcal{S} = \{1, \ldots, |\mathcal{S}|\}$ be a function that maps the covariates into a fixed, finite set of discrete strata. We call such a stratification "coarse", to distinguish it from "fine" stratification as defined in Section 3. Define $S_i = S(X_i)$ and 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 accommodates a large class of stratified assignment mechanisms, including stratified biased coin designs (Efron, 1971; Wei, 1978), minimization methods (Pocock and Simon, 1975; Hu and Hu, 2012) and stratified block randomization (see Zelen, 1974, for an early discussion). It follows from Lemma C.4 in Bugni et al. (2019) that for any integrable function $\gamma(\cdot)$,

$$\frac{1}{n} \sum_{1 \leq i \leq n} A_i \gamma(X_i) \xrightarrow{P} \eta \sum_{s \in \mathcal{S}} P\{S_i = s\} E[\gamma(X_i)|S_i = s] = \eta E[\gamma(X_i)] \ .$$

Therefore, Assumption 4.1 is satisfied. ∎

**Example 4.3** (CAR with general covariate features). Ma et al. (2024) propose a family of covariate adaptive randomization procedures which assign treatment sequentially based on an imbalance metric defined by "feature maps" of (potentially continuous) covariates. It follows by Theorem 3.5 of their paper that Assumption 4.1 is satisfied under appropriate conditions. ∎

**Example 4.4** (Matched pairs). Suppose $n$ is even and we assign treatment using a finely stratified design with $(\ell, k) = (1, 2)$. As discussed at the beginning of Section 3, such a design is also known as a matched pairs design. Assume that the pairing algorithm $\pi_n(X^{(n)})$ results in pairs that are close in the sense of Assumption 3.2. It then follows from the same argument used to establish (14) in the discussion following Theorem 3.1 that for any Lipschitz integrable function $\gamma(\cdot)$,

$$\frac{1}{n} \sum_{1 \leq i \leq n} A_i \gamma(X_i) \xrightarrow{P} \frac{1}{2} E[\gamma(X_i)] \ .$$

By approximating integrable functions by Lipschitz integrable functions as in Lemma A.1 in Hanneke et al. (2021), it can be shown that the convergence holds for any integrable function $\gamma(\cdot)$. Therefore, Assumption 4.1 is satisfied. ∎

**Example 4.5** (Re-randomization). Re-randomization is an assignment mechanism in which researchers specify a balance criterion for the covariates, and then repeatedly generate assignments using a completely randomized design until an assignment is found which achieves an acceptable covariate distribution according to the balance criterion. The properties of re-randomization procedures have been studied in Li and Ding (2017, 2020), Li et al. (2018, 2020), and Cytrynbaum (2024). It follows from Corollary 3.7 in Cytrynbaum (2024) that Assumption 4.1 holds for re-randomization designs, under appropriate assumptions. ∎

We now present an efficiency bound for the parameter $\theta_0$ introduced in Section 2. Formally, we characterize the bound via a convolution theorem that applies to all regular estimators of the parameter $\theta_0$. Following the definition on page 365 of van der Vaart (1998), by a regular estimator, we mean an estimator whose asymptotic distribution is invariant to "local" perturbations of the data generating process:

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

where $P_{t/\sqrt{n},g}$ represents a "local" perturbation of the distribution $P$ along a "path" with "score" $g$. We leave

the precise definition of regularity and related assumptions to Appendix A.2. In the paragraph following the statement of the theorem we provide some more details on the nature of our result.

**Theorem 4.1.** *Suppose Assumptions 2.1, 3.3(b), and 4.1 hold, as well as Condition A.1 described in Appendix A.2. Further suppose $\mathbb{V}_* < \infty$. Let $\tilde{\theta}_n$ be any regular estimator of the parameter $\theta_0$ in the sense of (S.16) in Appendix A.2. Then,*

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

*where*

$$L = N(0, \mathbb{V}_*) * B \ ,$$

*for $\mathbb{V}_*$ in (11) and some fixed probability measure $B$ which is specific to the estimator $\tilde{\theta}_n$.*

Given Theorem 4.1 we call $\mathbb{V}_* = \mathrm{Var}[\psi^*(X_i, A_i, R_i, \theta_0)]$ the efficiency bound for $\theta_0$, since our result shows that this is the lowest asymptotic variance attainable by any regular estimator under our assumptions. Indeed, it follows from Anderson's lemma (Lemma 8.5 in van der Vaart, 1998) that the asymptotic loss of any regular estimator is bounded below by the loss under $N(0, \mathbb{V}_*)$ for any "bowl-shaped" loss function (including, in particular, square loss). We note that our assumptions on the assignment mechanism preclude us from immediately appealing to standard semi-parametric convolution theorems (see, for instance, Theorem 25.20 in van der Vaart, 1989). Instead, we proceed by justifying an application of Theorem 3.1 in Armstrong (2022) combined with the convolution Theorem 3.11.2 in van der Vaart and Wellner (1996) to each $d_\theta$-dimensional parametric submodel separately, and then arguing that the supremum over all such submodels is attained by $\mathrm{Var}[\psi^*]$. A key observation is that in order to apply Theorem 3.1 in Armstrong (2022) to argue that the likelihood ratio process is locally asymptotically normal, the conditional information needs to settle down in the limit, which is guaranteed as long as Assumption 4.1 is satisfied.

**Remark 4.1.** Following similar arguments as those in Remark 3.2, we can deduce that our efficiency bound agrees with well-known bounds for common parameters (like those presented in Examples 2.1–2.3) in the setting of i.i.d. assignment. For example, we have noted in the case of the ATE (Example 2.1) that (17) matches the efficiency bound under i.i.d. assignment derived in Hahn (1998). See Rafi (2023) and Armstrong (2022) for related results in the context of stratified and response-adaptive experiments. Straightforward calculation also implies that, for the quantile treatment effect (Example 2.2), the efficiency bound is given by

$$E\left[\frac{1}{\eta}\frac{F_1\big(\theta_0(1)|X_i\big)\big(1 - F_1\big(\theta_0(1)|X_i\big)\big)}{f_1\big(\theta_0(1)\big)^2} + \frac{1}{1-\eta}\frac{F_0\big(\theta_0(0)|X_i\big)\big(1 - F_0\big(\theta_0(0)|X_i\big)\big)}{f_0\big(\theta_0(0)\big)^2}\right.$$
$$\left. + \left(\frac{F_1\big(\theta_0(1)|X_i\big) - \tau}{f_1\big(\theta_0(1)\big)} - \frac{F_0\big(\theta_0(0)|X_i\big) - \tau}{f_0\big(\theta_0(0)\big)}\right)^2\right] \ ,$$

which matches the efficiency bound under i.i.d. assignment derived in Firpo (2007) when the propensity score is set to $\eta$. ∎

**Remark 4.2.** The efficiency bound in Theorem 4.1 is attained by finely stratified experiments as in Theorem 3.1 if no additional covariates are available for estimation beyond the set of covariates $X_i$ used in the design. In practice, researchers may consider adjusting for additional baseline covariates in order to improve efficiency.

Suppose additional covariates $W^{(n)}$ are available and Assumption 3.1 is modified such that

$$(R^{(n)}(1), R^{(n)}(0), W^{(n)}) \perp\!\!\!\perp A^{(n)} \big| X^{(n)} \ .$$

It can be shown that the efficiency bound, allowing for additional covariate adjustment based on $X_i$ and $W_i$, is

$$\mathbb{V}_* - \eta(1-\eta)M^{-1}E[\mathrm{Var}[E[g_i|X_i, W_i]|X_i]](M^{-1})' \ , \tag{20}$$

where $g_i = m(X_i, 1, R_i(1), \theta_0) - m(X_i, 0, R_i(0), \theta_0)$. Then, as in Remark 3.3, the potential gain in efficiency from exploiting $W_i$ in addition to $X_i$ is proportional to

$$E[R^2_{g,X,W}(g_i, X_i, W_i)\,\mathrm{Var}[g_i|X_i]] \ ,$$

where

$$R^2_{g,X,W}(g_i, X_i, W_i) = \frac{\mathrm{Var}[E[g_i|X_i, W_i]|X_i]}{\mathrm{Var}[g_i|X_i]}$$

is the nonparametric $R^2$ from regressing $g_i$ on $X_i$ and $W_i$ conditional on $X_i$. As a result, the scope for improving efficiency by adjusting for additional covariates is limited if $R^2_{g,X,W}$ is small. In the case of estimating the ATE,

$$g_i = \frac{Y_i(1)}{\eta} + \frac{Y_i(0)}{1-\eta} \ ,$$

so the scope for improvement depends on how much additional variation in the weighted potential outcomes can be explained by $W_i$ beyond $X_i$. ∎

**Remark 4.3.** Here, we comment on how Theorem 4.1 relates to prior efficiency bounds in experiments with general assignment mechanisms. For the case of estimating the ATE, Armstrong (2022) derives an efficiency bound over a very large class of assignment mechanisms, including even response-adaptive designs, and shows that the bound is attained when units are assigned to treatment (control) with conditional probability proportional to the conditional standard deviation of the potential outcome under treatment (control). This type of assignment is sometimes referred to as the Neyman allocation. On the other hand, our results show that this bound may be quite loose whenever the assignment proportions are restricted to be anything not equal to the Neyman allocation, which is, of course, unknown. For example, the bound is not informative about what can be achieved if the assignment proportions were set to one half regardless of whether or not the conditional outcome variances across treatment and control are equal. Such settings frequently arise in practice due to logistical constraints or the absence of pilot data with which to estimate conditional variances of potential outcomes under treatment and control. Furthermore, as argued in Cai and Rafi (2022), even if pilot data is available, these quantities may be estimated so poorly that exogenously constraining the assignment proportions to one half leads to more efficient estimates of the ATE in practice. Motivated by such concerns, Rafi (2023) derives an efficiency bound for the ATE over the class of "coarsely-stratified" assignment mechanisms studied in Bugni et al. (2019), where the stratum-level assignment proportions are restricted *a priori* by the experimenter. This framework, however, rules out finely stratified designs. Finally, we once again emphasize that our analysis, unlike these other papers, applies to a general class of treatment effect parameters, including the ATE as a special case. ∎

**Remark 4.4.** 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 4.1 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.5 shows that the efficiency bound equals

$$
\begin{aligned}
\mathbb{V}_* &= \mathrm{Var}[\psi^*(X_i, A_i, R_i, \theta_0)] \\
&= M^{-1}\big(E\big[\eta(X_i)\,\mathrm{Var}[m(X_i, 1, R_i(1), \theta_0)|X_i] \\
&\qquad\qquad + (1 - \eta(X_i))\,\mathrm{Var}[m(X_i, 0, R_i(0), \theta_0)|X_i]\big] \\
&\quad + \mathrm{Var}\big[\eta(X_i)E[m(X_i, 1, R_i(1), \theta_0)|X_i] \\
&\qquad\qquad + (1 - \eta(X_i))E[m(X_i, 0, R_i(0), \theta_0)|X_i]\big]\big)(M^{-1})' \ ,
\end{aligned}
\tag{21}
$$

so that the only difference from (16) is that $\eta$ is replaced by $\eta(X_i)$.

If we additionally impose that $\eta(X_i)$ takes on a finite set of values $\{\eta_1, \ldots, \eta_S\}$, then this bound could be achieved by separately implementing 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 3.1–3.2 with $\ell/k = \eta_s$. For more general functions $\eta(\cdot)$, we conjecture one could employ the local randomization procedure in Cytrynbaum (2023b). ∎

## 4.2 The Necessity of "Fast-Balancing"

In this subsection, we provide conditions under which the fast-balancing condition described in (14) is a necessary condition for efficient estimation of $\theta_0$ "by design." As a supplement, we also provide necessary and sufficient conditions for an asymptotically linear estimator to be regular (in the sense of (S.16) in Appendix A.2) for a large class of treatment assignment mechanisms. Concretely, given an assignment mechanism, suppose $\tilde{\theta}_n$ is an asymptotically linear estimator for $\theta_0$ in the sense that

$$
\sqrt{n}(\tilde{\theta}_n - \theta_0) = \frac{1}{\sqrt{n}} \sum_{1 \leq i \leq n} \psi(X_i, A_i, R_i, \theta_0) + o_P(1) \ ,
\tag{22}
$$

where $E[\psi(X_i, A_i, R_i, \theta_0)] = 0$ and $\mathrm{Var}[\psi(X_i, A_i, R_i, \theta_0)] < \infty$. The results in this section derive necessary and sufficient conditions for $\tilde{\theta}_n$ to be regular, and further demonstrate that in order for $\tilde{\theta}_n$ to be regular and efficient, either $\psi = \psi^*$ or a fast-balancing condition involving $\psi(\cdot)$ needs to be satisfied. Therefore, finely stratified designs are not only sufficient to guarantee efficiency when estimating $\theta_0$ using the naïve method of moments estimator, but their fast-balancing property is also necessary.

In order to study the behavior of the estimator under local alternatives, we will impose the following high-level assumption on the "imbalance" of the treatment assignments. To describe the assumption, let $\rho$ denote any metric that metrizes weak convergence.

**Assumption 4.2.** The treatment assignment mechanism is such that for any square-integrable function $\gamma : \mathbf{R}^{d_x} \to \mathbf{R}^{d_\theta}$ with $E[\gamma(X_i)] = 0$ ,

$$
\rho\bigg(\frac{1}{\sqrt{n}} \sum_{1 \leq i \leq n} (A_i - \eta)\gamma(X_i), \ N(0, V_\gamma^{\mathrm{imb}})\bigg|X^{(n)}\bigg) \xrightarrow{P} 0
$$

for some deterministic variance $V_\gamma^{\text{imb}}$.

In the following examples, we discuss Assumption 4.2 in the context of some common treatment assignment mechanisms.

**Example 4.6.** Revisiting Example 4.1, let $A^{(n)}$ be assigned i.i.d., independently of $X^{(n)}$, such that $P\{A_i = 1\} = \eta$. Then, by verifying the conditions of the Lindeberg-Feller CLT conditional on $X^{(n)}$, it can be shown that Assumption 4.2 is satisfied with $V_\gamma^{\text{imb}} = \eta(1-\eta)\operatorname{Var}[\gamma(X_i)]$. ∎

**Example 4.7.** Revisiting Example 4.2, suppose treatment status is assigned using stratified block randomization, which is a special case of covariate-adaptive randomization where $A^{(n)}$ is such that

$$\sum_{1 \le i \le n} A_i I\{S_i = s\} = \left\lfloor \eta \sum_{1 \le i \le n} I\{S_i = s\} \right\rfloor \, ,$$

with all such assignments being drawn uniformly at random and independently across strata. It follows from Theorem 12.2.1 in Lehmann and Romano (2022) combined with a subsequencing argument that Assumption 4.2 is satisfied with $V_\gamma^{\text{imb}} = \eta(1-\eta)E[\operatorname{Var}[\gamma(X_i)|S_i]]$. ∎

**Example 4.8.** Revisiting Example 4.3, suppose treatment is assigned using the covariate adaptive randomization procedure described in Ma et al. (2024). Then it follows from Theorem 3.6 in their paper that, under appropriate assumptions,

$$\frac{1}{\sqrt{n}} \sum_{1 \le i \le n} (A_i - \eta)\gamma(X_i) \xrightarrow{d} N(0, \tilde{V}) \, ,$$

for some variance $\tilde{V}$. Note, however, that this result is *not* conditional on $X^{(n)}$ and thus does not immediately imply Assumption 4.2. We conjecture that a similar result could be established conditional on $X^{(n)}$ and thus Assumption 4.2 would be satisfied. ∎

**Example 4.9.** Revisiting Example 4.4, suppose $n$ is even and we assign treatment using a matched pairs design. It then follows by arguing as in the discussion following Theorem 3.1 that for any square-integrable Lipschitz function $\gamma(\cdot)$,

$$\operatorname{Var}\left[\frac{1}{\sqrt{n}} \sum_{1 \le i \le n} (A_i - \eta)\gamma(X_i) \middle| X^{(n)}\right] \xrightarrow{P} 0 \, .$$

Therefore, by Markov's inequality, Assumption 4.2 is satisfied with $V_\gamma^{\text{imb}} = 0$. By approximating square-integrable functions by square-integrable Lipschitz functions as in Lemma C.5 in Cytrynbaum (2023b), it can be shown that the convergence holds for any square-integrable function $\gamma(\cdot)$. ∎

**Example 4.10.** Revisiting Example 4.5, we note that, following Corollary 3.7 in Cytrynbaum (2024), we do not expect Assumption 4.2 to hold for re-randomization designs in general. ∎

We are now ready to state a theorem that characterizes all regular asymptotically linear estimators and establishes the necessity of the fast-balancing condition for efficient estimation using $\hat{\theta}_n$.

**Theorem 4.2.** *Suppose the treatment assignment mechanism satisfies Assumptions 2.1 and 4.1–4.2. Suppose $\tilde{\theta}_n$ is an asymptotically linear estimator for $\theta_0$ in the sense of (22). Then, $\tilde{\theta}_n$ is regular if and only if*

$$\psi(x, a, r, \theta_0) = \psi^*(x, a, r, \theta_0) + \psi^\perp(x, a, \theta_0) \, ,$$

for some function $\psi^\perp$ such that $E[\psi^\perp(X_i, A_i, \theta_0)|X_i] = \eta\psi^\perp(X_i, 1, \theta_0) + (1-\eta)\psi^\perp(X_i, 0, \theta_0) = 0$. Furthermore, if $\tilde{\theta}_n$ is regular, it attains the efficiency bound if and only if

$$\frac{1}{\sqrt{n}} \sum_{1 \leq i \leq n} (A_i - \eta)E[\psi(X_i, 1, R_i(1), \theta_0) - \psi(X_i, 0, R_i(0), \theta_0)|X_i] = o_P(1) \ . \tag{23}$$

To further understand condition (23), note that $E[\psi^*(X_i, 1, R_i(1))|X_i] = E[\psi^*(X_i, 0, R_i(0))|X_i]$, so

$$\frac{1}{\sqrt{n}} \sum_{1 \leq i \leq n} (A_i - \eta)E[\psi(X_i, 1, R_i(1), \theta_0) - \psi(X_i, 0, R_i(0), \theta_0)|X_i]$$

$$= \frac{1}{\sqrt{n}} \sum_{1 \leq i \leq n} (A_i - \eta)(\psi^\perp(X_i, 1, \theta_0) - \psi^\perp(X_i, 0, \theta_0))$$

$$= \frac{1}{\sqrt{n}} \sum_{1 \leq i \leq n} \psi^\perp(X_i, A_i, \theta_0) \ ,$$

where the last equality follows from the fact that $E[\psi^\perp(X_i, A_i, \theta_0)|X_i] = 0$. As a result, (23) holds either when the treatment assignment mechanism groups units with similar values of $\psi^\perp(X_i, 1, \theta_0) - \psi^\perp(X_i, 0, \theta_0)$, or when the estimator is based on the efficient influence function, so that $\psi^\perp = 0$. Recall that as an intermediate step in the proof of Theorem 3.1, we showed in (12) that for the naïve method of moments estimator $\hat{\theta}_n$, $\psi(\cdot) = -M^{-1}m(\cdot)$, so (23) coincides with the fast-balancing condition in (14). We can thus conclude from Theorem 4.2 that the fast-balancing condition is necessary to achieve efficient estimation based on the naïve method of moments estimator, when the class of assignment mechanisms satisfy Assumption 4.2.

**Example 4.11.** Revisiting Example 2.1, recall $\hat{\theta}_n$ estimates the ATE based on the moment conditions in (4). Direct calculation shows that for $\hat{\theta}_n$,

$$\psi^\perp(x, a, \theta_0) = (a - \eta)\left(\frac{\mu_1(x)}{\eta} + \frac{\mu_0(x)}{1 - \eta}\right) \ .$$

As a result, Theorem 4.2 demonstrates that $\hat{\theta}_n$ does not achieve the efficiency bound, unless

$$\frac{1}{\sqrt{n}} \sum_{1 \leq i \leq n} (A_i - \eta)\left(\frac{\mu_1(X_i)}{\eta} + \frac{\mu_0(X_i)}{1 - \eta}\right) = o_P(1) \ ,$$

which is indeed the case in finely stratified experiments when the treatment assignment mechanism satisfies Assumption 3.2. ∎

# 5  Simulations

In this section, we illustrate the theoretical results in Sections 3 and 4 through a simulation study. Throughout this section, we set $\eta = 1/2$, and compare the mean-squared error (MSE), bias, the length of the confidence interval and its coverage rate for the following combinations of treatment assignment mechanisms and estimators:

– i.i.d. treatment assignment and the naïve method of moments estimator

– i.i.d. treatment assignment and covariate adjusted estimators

– Matched pairs, i.e., a finely stratified design with $(\ell, k) = (1, 2)$, and the naïve method of moments estimator

– Matched tuples, i.e., a finely stratified design with $(\ell, k) = (2, 4)$, and the naïve method of moments estimator

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:

$$Y_i(a) = \mu_a(X_i) + \epsilon_i \ , \tag{24}$$

where $\mu_0(X_i) = \sum_{1 \leq l \leq 8} w_l \big(X_{i,l} + \frac{1}{3}(X_{i,l}^2 - 1)\big)$ for $w = (2, 1, 1, 0.5, 0.01, 0.001, 0.0001, 0.00001)$, $\mu_1(X_i) = 0.2 + \mu_0(X_i)$, $\epsilon_i \sim N(0, 4)$, $(X_i, \epsilon_i)$, $1 \leq i \leq n$ are i.i.d., and for each $1 \leq i \leq n$, $(X_i, \epsilon_i)$ are independent. In each of the simulations to follow, when forming pairs/tuples or performing regression adjustment, we use only a subvector of the covariates $X_i$ consisting of the first $T$ covariates, for $T \in \{2, 4, 8\}$.

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**

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

where $\hat{\mu}_a^Y(X_i)$ is the linear projection of $Y_i$ on $(1, (X_{i,l}, X_{i,l}^2 : 1 \leq l \leq T))$ in the subsample with $A_i = a$.

**Adjusted Estimator 2**

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

where $\hat{\mu}_a^Y(X_i)$ is the linear projection of $Y_i$ on on $(1, (X_{i,l}, X_{i,l}^2, X_{i,l}I\{X_{i,l} > \hat{t}_l\} : 1 \le l \le T))$ in the subsample with $A_i = a$, where $\hat{t}_l$ is the sample median of $X_{i,l}$, $1 \le i \le n$.

The first estimator $\hat{\theta}_n^{\mathrm{unadj}}$ is the naïve method of moments estimator given by the solution to (6). The second and third estimators $\hat{\theta}_n^{\mathrm{adj},1}$ and $\hat{\theta}_n^{\mathrm{adj},2}$ 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. $\hat{\theta}_n^{\mathrm{adj},1}$ and $\hat{\theta}_n^{\mathrm{adj},2}$ 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 $\hat{\theta}_n^{\mathrm{adj},1}$, $\hat{\theta}_n^{\mathrm{adj},2}$ 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 $\hat{\theta}_n^{\mathrm{adj},1}$, $\hat{\theta}_n^{\mathrm{adj},2}$ 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 $Y_i(d)$ is again given by (24). In addition, the take-up decision is determined as $D_i = A_i D_i(1) + (1 - A_i)D_i(0)$, where

$$D_i(0) = I\{\alpha_0 + \alpha(X_i) > \varepsilon_{1,i}\} ,$$

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

where $\alpha_0 = 0.2$, $\alpha_1 = 4$, $\alpha(X_i) = \sum_{1 \le l \le 8}\left(X_{i,l} + \frac{1}{3}(X_{i,l}^2 - 1)\right)$, $\varepsilon_{1,i}, \varepsilon_{2,i} \sim N(0, 4)$, $(X_i, \epsilon_i, \varepsilon_{1,i}, \varepsilon_{2,i})$, $1 \le i \le n$ are i.i.d., and for each $1 \le i \le n$, $(X_i, \epsilon_i, \varepsilon_{1,i}, \varepsilon_{2,i})$ are independent.

We consider the following three estimators for the LATE:

**Unadjusted Estimator**:
$$\hat{\theta}_n^{\mathrm{unadj}} = \frac{\sum\limits_{1 \le i \le n}(Y_i A_i - Y_i(1 - A_i))}{\sum\limits_{1 \le i \le n}(D_i A_i - D_i(1 - A_i))} .$$

**Adjusted Estimator 1**:
$$\hat{\theta}_n^{\mathrm{adj},1} = \frac{\sum\limits_{1 \le i \le n}\left(2A_i(Y_i - \hat{\mu}_1^Y(X_i)) - 2(1 - A_i)(Y_i - \hat{\mu}_0^Y(X_i)) + \hat{\mu}_1^Y(X_i) - \hat{\mu}_0^Y(X_i)\right)}{\sum\limits_{1 \le i \le n}\left(2A_i(D_i - \hat{\mu}_1^Y(X_i)) - 2(1 - A_i)(D_i - \hat{\mu}_0^Y(X_i)) + \hat{\mu}_1^D(X_i) - \hat{\mu}_0^D(X_i)\right)} ,$$

where $\hat{\mu}_a^Y(X_i)$ is the linear projection of $Y_i$ on $(1, (X_{i,l}, X_{i,l}^2 : 1 \leq l \leq T))$ in the subsample with $A_i = a$, and $\hat{\mu}_a^D(X_i)$ is estimated using from a logistic regression model with the same set of regressors in the subsample with $A_i = a$.

**Adjusted Estimator 2**: As in Adjusted Estimator 1, but in $\hat{\mu}_a^Y(X_i)$ and $\hat{\mu}_a^D(X_i)$ the regressors are instead $(1, (X_{i,l}, X_{i,l}^2, X_{i,l}1\{X_{i,l} > \hat{t}_l\} : 1 \leq l \leq T))$, where $\hat{t}_l$ is the sample median of $X_{i,l}$, $1 \leq i \leq n$.

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 1 displays the absolute value of the bias for each design and estimator pair computed across 4000 Monte Carlo replications. We find that all estimators have fairly low and comparable bias.

Table 2 displays the ratio of the empirical MSE for each design/estimator pair relative to the MSE of the unadjusted estimator under i.i.d. assignment, also computed across 4000 Monte Carlo replications. As expected given our theoretical results, we find that the empirical MSEs of the naïve unadjusted estimator under a matched pairs/tuples design closely match the empirical MSEs of the covariate adjusted estimators under i.i.d. assignment; this feature is particularly noteworthy given that the adjusted estimators are in fact correctly specified, and the correct specification would be unknowable in practice. We note that the MSE improvement of the unadjusted estimator with a matched pairs/tuples design relative to i.i.d. assignment is typically worse with 2 or 8 covariates than with 4 covariates: this phenomenon stems from the fact that the first 4 covariates are much stronger predictors of the control outcome than the last 4 covariates, which are almost uninformative. Although we have found in prior work (Bai et al., 2024c) that the matched pairs design delivers a lower MSE than the matched tuples design when the potential outcomes depend on the covariates linearly, we do not find that this is the case here with a nonlinear model. However, we do consistently find that the MSE with 8 covariates is smaller for the matched pairs design than the matched tuples design. This comparison illustrates that, although as discussed in Remark 3.1, our theoretical results imply matched pairs and matched tuples designs are not distinguishable asymptotically, their finite-sample properties may differ, especially when the number of covariates is large. In particular, note Assumption 3.2 is more stringent for matched tuples, for which $k = 4$, than matched pairs, for which $k = 2$.

Table 3 reports the coverage and average length of confidence intervals constructed using the estimator described in Section 3.1, again computed across 4000 Monte Carlo replications. We find that the confidence intervals have appropriate coverage at all sample sizes and for any number of covariates. The average length of the intervals follows a pattern similar to what was observed for the MSE, again because the first 4 covariates are much stronger predictors of the control outcome than the last 4 covariates, which are almost uninformative. Although we found in Table 2 that the MSE of the matched tuples design was worse than that of matched pairs with 8 covariates, this does not seem to translate to longer confidence intervals for matched tuples: this may be due to the fact that the matched tuples variance estimator is a "within block"

estimator, as described in Section 3.1. Similar findings have been reported by Bai et al. (2024c) for related problems in factorial designs.

Table 1: Absolute value of bias of estimators for different designs and estimators

|  | # of covariates ($T$) | i.i.d. assignment | | | Matched pairs | Matched tuples |
|---|---|---|---|---|---|---|
|  |  | Unadjusted | Adjusted 1 | Adjusted 2 | Unadjusted | Unadjusted |
| | | | | $n = 100$ | | |
| ATE | 2 | 0.0083 | 0.0065 | 0.0051 | 0.0212 | 0.0068 |
|  | 4 | 0.0083 | 0.0090 | 0.0090 | 0.0073 | 0.0124 |
|  | 8 | 0.0083 | 0.0134 | 0.0131 | 0.0010 | 0.0209 |
| LATE | 2 | 0.0905 | 0.0436 | 0.0396 | 0.0341 | 0.0076 |
|  | 4 | 0.0905 | 0.0276 | 0.0264 | 0.0301 | 0.0427 |
|  | 8 | 0.0905 | 0.0401 | 0.0376 | 0.0264 | 0.0297 |
| | | | | $n = 200$ | | |
| ATE | 2 | 0.0075 | 0.0068 | 0.0074 | 0.0084 | 0.0081 |
|  | 4 | 0.0075 | 0.0052 | 0.0067 | 0.0040 | 0.0040 |
|  | 8 | 0.0075 | 0.0045 | 0.0055 | 0.0145 | 0.0051 |
| LATE | 2 | 0.0545 | 0.0326 | 0.0339 | 0.0361 | 0.0328 |
|  | 4 | 0.0545 | 0.0182 | 0.0217 | 0.0123 | 0.0144 |
|  | 8 | 0.0545 | 0.0159 | 0.0180 | 0.0447 | 0.0216 |
| | | | | $n = 400$ | | |
| ATE | 2 | 0.0001 | 0.0011 | 0.0009 | 0.0041 | 0.0055 |
|  | 4 | 0.0001 | 0.0001 | 0.0001 | 0.0001 | 0.0016 |
|  | 8 | 0.0001 | 0.0006 | 0.0006 | 0.0056 | 0.0041 |
| LATE | 2 | 0.0142 | 0.0064 | 0.0059 | 0.0179 | 0.0198 |
|  | 4 | 0.0142 | 0.0012 | 0.0015 | 0.0004 | 0.0020 |
|  | 8 | 0.0142 | 0.0021 | 0.0019 | 0.0129 | 0.0064 |
| | | | | $n = 1000$ | | |
| ATE | 2 | 0.0021 | 0.0037 | 0.0039 | 0.0014 | 0.0025 |
|  | 4 | 0.0021 | 0.0021 | 0.0022 | 0.0020 | 0.0006 |
|  | 8 | 0.0021 | 0.0021 | 0.0023 | 0.0023 | 0.0027 |
| LATE | 2 | 0.0118 | 0.0120 | 0.0123 | 0.0019 | 0.0083 |
|  | 4 | 0.0118 | 0.0055 | 0.0057 | 0.0051 | 0.0007 |
|  | 8 | 0.0118 | 0.0054 | 0.0058 | 0.0057 | 0.0059 |
| | | | | $n = 2000$ | | |
| ATE | 2 | 0.0013 | 0.0010 | 0.0010 | 0.0011 | 0.0025 |
|  | 4 | 0.0013 | 0.0015 | 0.0015 | 0.0003 | 0.0021 |
|  | 8 | 0.0013 | 0.0015 | 0.0015 | 0.0001 | 0.0014 |
| LATE | 2 | 0.0062 | 0.0035 | 0.0035 | 0.0018 | 0.0050 |
|  | 4 | 0.0062 | 0.0037 | 0.0038 | 0.0011 | 0.0055 |
|  | 8 | 0.0062 | 0.0038 | 0.0037 | 0.0008 | 0.0032 |

Note: Entries report the absolute bias of each estimator under different sample sizes, designs, and covariate adjustment strategies. Results are averaged across 2000 replications.

Table 2: MSE ratios relative to unadjusted estimator under i.i.d. assignment

|  | # of covariates (T) | i.i.d. assignment | | | Matched pairs | Matched tuples |
|---|---|---|---|---|---|---|
|  |  | Unadjusted | Adjusted 1 | Adjusted 2 | Unadjusted | Unadjusted |
| | | | $n = 100$ | | | |
| ATE | 2 | 1.0000 | 0.7691 | 0.7809 | 0.7608 | 0.7253 |
|  | 4 | 1.0000 | 0.7166 | 0.7356 | 0.7123 | 0.6853 |
|  | 8 | 1.0000 | 0.7527 | 0.8102 | 0.7619 | 0.8083 |
| LATE | 2 | 1.0000 | 0.7550 | 0.7664 | 0.7527 | 0.7069 |
|  | 4 | 1.0000 | 0.6991 | 0.7165 | 0.6941 | 0.6682 |
|  | 8 | 1.0000 | 0.7209 | 0.7781 | 0.7312 | 0.7676 |
| | | | $n = 200$ | | | |
| ATE | 2 | 1.0000 | 0.7598 | 0.7619 | 0.7500 | 0.7254 |
|  | 4 | 1.0000 | 0.7084 | 0.7154 | 0.7016 | 0.7213 |
|  | 8 | 1.0000 | 0.7256 | 0.7420 | 0.7586 | 0.8042 |
| LATE | 2 | 1.0000 | 0.7562 | 0.7581 | 0.7219 | 0.7056 |
|  | 4 | 1.0000 | 0.7021 | 0.7095 | 0.6806 | 0.7068 |
|  | 8 | 1.0000 | 0.7081 | 0.7238 | 0.7317 | 0.7811 |
| | | | $n = 400$ | | | |
| ATE | 2 | 1.0000 | 0.7427 | 0.7441 | 0.7473 | 0.7698 |
|  | 4 | 1.0000 | 0.6828 | 0.6880 | 0.6936 | 0.7372 |
|  | 8 | 1.0000 | 0.6904 | 0.7013 | 0.7526 | 0.7986 |
| LATE | 2 | 1.0000 | 0.7420 | 0.7436 | 0.7528 | 0.7769 |
|  | 4 | 1.0000 | 0.6792 | 0.6845 | 0.6928 | 0.7379 |
|  | 8 | 1.0000 | 0.6864 | 0.6974 | 0.7501 | 0.7950 |
| | | | $n = 1000$ | | | |
| ATE | 2 | 1.0000 | 0.7526 | 0.7531 | 0.7853 | 0.7338 |
|  | 4 | 1.0000 | 0.6846 | 0.6862 | 0.7178 | 0.7123 |
|  | 8 | 1.0000 | 0.6889 | 0.6934 | 0.7722 | 0.7969 |
| LATE | 2 | 1.0000 | 0.7532 | 0.7537 | 0.7841 | 0.7318 |
|  | 4 | 1.0000 | 0.6842 | 0.6859 | 0.7145 | 0.7119 |
|  | 8 | 1.0000 | 0.6879 | 0.6925 | 0.7653 | 0.7916 |
| | | | $n = 2000$ | | | |
| ATE | 2 | 1.0000 | 0.7509 | 0.7512 | 0.7367 | 0.7316 |
|  | 4 | 1.0000 | 0.6926 | 0.6932 | 0.6847 | 0.6788 |
|  | 8 | 1.0000 | 0.6933 | 0.6934 | 0.6975 | 0.7253 |
| LATE | 2 | 1.0000 | 0.7501 | 0.7504 | 0.7377 | 0.7337 |
|  | 4 | 1.0000 | 0.6915 | 0.6922 | 0.6858 | 0.6788 |
|  | 8 | 1.0000 | 0.6914 | 0.6916 | 0.6966 | 0.7223 |

Note: For each specification, the MSE of the unadjusted estimator under i.i.d. assignment is normalized to one, and the other columns report the ratios of MSEs relative to this baseline.

Table 3: Coverage rates and lengths of 95% confidence intervals based on unadjusted estimator for matched pairs and matched tuples

| | # of covariates ($T$) | Matched pairs | | Matched tuples | |
|---|---|---|---|---|---|
| | | Coverage | CI length | Coverage | CI length |
| | | $n = 100$ | | | |
| | 2 | 0.9403 | 2.3191 | 0.9490 | 2.3224 |
| ATE | 4 | 0.9465 | 2.2744 | 0.9485 | 2.2830 |
| | 8 | 0.9487 | 2.3532 | 0.9395 | 2.3672 |
| | 2 | 0.9590 | 6.4075 | 0.9590 | 6.3865 |
| LATE | 4 | 0.9620 | 6.3184 | 0.9615 | 6.3176 |
| | 8 | 0.9610 | 6.5497 | 0.9527 | 6.5311 |
| | | $n = 200$ | | | |
| | 2 | 0.9495 | 1.6401 | 0.9497 | 1.6418 |
| ATE | 4 | 0.9485 | 1.5990 | 0.9440 | 1.6049 |
| | 8 | 0.9503 | 1.6529 | 0.9425 | 1.6617 |
| | 2 | 0.9550 | 4.4046 | 0.9563 | 4.4264 |
| LATE | 4 | 0.9557 | 4.3315 | 0.9530 | 4.3381 |
| | 8 | 0.9597 | 4.4833 | 0.9503 | 4.5038 |
| | | $n = 400$ | | | |
| | 2 | 0.9487 | 1.1602 | 0.9483 | 1.1614 |
| ATE | 4 | 0.9497 | 1.1255 | 0.9455 | 1.1285 |
| | 8 | 0.9495 | 1.1609 | 0.9445 | 1.1667 |
| | 2 | 0.9530 | 3.0955 | 0.9493 | 3.0963 |
| LATE | 4 | 0.9547 | 3.0122 | 0.9517 | 3.0110 |
| | 8 | 0.9555 | 3.1284 | 0.9505 | 3.1239 |
| | | $n = 1000$ | | | |
| | 2 | 0.9485 | 0.7349 | 0.9550 | 0.7352 |
| ATE | 4 | 0.9500 | 0.7086 | 0.9505 | 0.7102 |
| | 8 | 0.9417 | 0.7285 | 0.9430 | 0.7318 |
| | 2 | 0.9507 | 1.9468 | 0.9560 | 1.9452 |
| LATE | 4 | 0.9515 | 1.8836 | 0.9507 | 1.8857 |
| | 8 | 0.9467 | 1.9469 | 0.9463 | 1.9480 |
| | | $n = 2000$ | | | |
| | 2 | 0.9490 | 0.5194 | 0.9530 | 0.5193 |
| ATE | 4 | 0.9483 | 0.4998 | 0.9485 | 0.5004 |
| | 8 | 0.9505 | 0.5121 | 0.9483 | 0.5139 |
| | 2 | 0.9495 | 1.3714 | 0.9530 | 1.3721 |
| LATE | 4 | 0.9493 | 1.3250 | 0.9483 | 1.3258 |
| | 8 | 0.9533 | 1.3663 | 0.9505 | 1.3658 |

# 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 covariate adjustment for a large class of interesting parameters. This finding generalizes similar observations made by Bai et al. (2022), Bai (2022) and Cytrynbaum (2023b) for the special case of estimating the ATE.

Our simulation evidence suggests, however, that finely stratified experiments may produce less precise estimates than (correctly specified) covariate adjustment when the the dimension of $X_i$ is large relative to the sample size. For this reason, we recommend that practitioners construct their blocks using a subset of the baseline covariates that they believe have the highest explanatory power in terms of the nonparametric $R^2$ in (19); the pre-treatment measure of the outcomes of interest, for example, is typically believed to be one such covariate (see, in particular, Bruhn and McKenzie, 2009). The experimental data can then be analyzed efficiently using an unadjusted method-of-moments estimator.

If one wishes to perform covariate adjustment with additional covariates beyond those used for blocking, then this can be done *ex post*. As discussed in Remark 4.2, the scope for improvement from covariate adjustment is limited by the nonparametric $R^2$ from the regression of the moment functions on the additional covariates, conditional on the ones used for matching; if one has already matched on the covariates with the highest explanatory power, then the potential gain in efficiency from adjusting for these additional covariates may be limited. We further 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. (2024a) 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., 2024a,b, 2025a; 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.

# A    Proofs of Main Results

## A.1    Proof of Theorem 3.1

First note (10) follows from (9) and Lemma A.6. In particular, the second component of the decomposition therein is zero because $E[\psi^*|X, A] = E[\psi^*|X]$. To show (9), we first establish (12), i.e.,

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

By the proof of Theorem 5.21 in van der Vaart (1998), to show (12), it suffices to show

$$\mathbb{L}_n(\hat{\theta}_n) \xrightarrow{P} 0 \ , \tag{25}$$

where $\mathbb{L}_n(\theta) = (\mathbb{L}_n^{(1)}(\theta), \ldots, \mathbb{L}_n^{(d_\theta)}(\theta))'$ for

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

$$- \frac{1}{\sqrt{n}} \sum_{1 \leq i \leq n} (m_s(X_i, A_i, R_i, \theta_0) - E_P[m_s(X_i, A_i, R_i, \theta_0)]) \ .$$

To accomplish this, we study $\mathbb{L}_n^{(s)}(\theta)$ for $1 \leq s \leq d_\theta$ separately. It follows from Assumption 3.3(c)–(d), Proposition 8.11 in Kosorok (2008), and the arguments to establish (48) that

$$\sup_{\theta \in \Theta: \|\theta - \theta_0\| < \delta} |\mathbb{L}_n^{(s)}(\theta)| = \sup_{\theta \in \Theta^*: \|\theta - \theta_0\| < \delta} |\mathbb{L}_n^{(s)}(\theta)| \ .$$

Therefore, since $\hat{\theta}_n \xrightarrow{P} \theta_0$ by Lemma A.8, to show (25) it suffices to argue that for every $\epsilon > 0$ and every sequence $\delta_n \downarrow 0$ (p.89 of van der Vaart and Wellner, 1996),

$$\lim_{n \to \infty} P \left\{ \sup_{\theta \in \Theta^*: \|\theta - \theta_0\| < \delta_n} \left| \mathbb{L}_n^{(s)}(\theta) \right| > \epsilon \right\} = 0 \ . \tag{26}$$

Following the arguments in the proof of Lemma A.8, we decompose $\mathbb{L}_n^{(s)}(\theta) = \mathbb{L}_{n,1}^{(s)}(\theta) + \mathbb{L}_{n,0}^{(s)}(\theta)$, where

$$\mathbb{L}_{n,1}^{(s)}(\theta) = \frac{1}{\sqrt{n}} \sum_{1 \leq i \leq n} A_i(m_s(X_i, 1, R_i(1), \theta) - m_s(X_i, 1, R_i(1), \theta_0)$$

$$- E[m(X_i, 1, R_i(1), \theta) - m_s(X_i, 1, R_i(1), \theta_0)])$$

$$\mathbb{L}_{n,0}^{(s)}(\theta) = \frac{1}{\sqrt{n}} \sum_{1 \leq i \leq n} (1 - A_i)(m_s(X_i, 0, R_i(0), \theta) - m_s(X_i, 0, R_i(0), \theta_0)$$

$$- E[m_s(X_i, 0, R_i(0), \theta) - m_s(X_i, 0, R_i(0), \theta_0)]) \ .$$

Define

$$\rho_Q(\theta, \theta_0) = E_Q[(m_s(X, a, R(a), \theta) - m_s(X, a, R(a), \theta_0)$$

$$- E_Q[m_s(X, a, R(a), \theta) - m_s(X, a, R(a), \theta_0)])^2]^{1/2} .$$

Note by Assumption 3.3(c) that $\rho_Q(\theta, \theta_0)$ is continuous in $\theta$, i.e., as $\|\theta - \theta_0\| \to 0$,

$$\rho_Q(\theta, \theta_0) \le E_Q[(m_s(X, a, R(a), \theta) - m_s(X, a, R(a), \theta_0))^2]^{1/2} \to 0 .$$

Fix any sequence $\tilde{\delta}_n \downarrow 0$. For every $n$, there exists $n'$ such that $\{\theta \in \Theta^* : \|\theta - \theta_0\| < \delta_{n'}\} \subseteq \{\theta \in \Theta^* : \rho_Q(\theta, \theta_0) < \tilde{\delta}_n\}$. By Proposition C.1 in Han and Wellner (2021),

$$E\left[ \sup_{\rho_Q(\theta,\theta_0)<\tilde{\delta}_n} |\mathbb{L}_{n,a}^{(s)}(\theta)| \right] E\left[ \sup_{\rho_Q(\theta,\theta_0)<\tilde{\delta}_n} \left| \frac{1}{\sqrt{n}} \sum_{1 \le i \le n} (m_s(X_i, 1, R_i(1), \theta) \right. \right.$$
$$\left. \left. - m_s(X_i, 1, R_i(1), \theta_0) - E[(m_s(X_i, 1, R_i(1), \theta) - m_s(X_i, 1, R_i(1), \theta_0)]) \right| \right] \to 0 .$$

where the convergence follows from Assumption 3.3(e) and Corollary 2.3.12 in van der Vaart and Wellner (1996). We then obtain (26) by Markov's inequality.

Finally, we derive (9) from (12). Note that

$$\frac{1}{\sqrt{n}} \sum_{1 \le i \le n} m(X_i, A_i, R_i, \theta_0)$$
$$= \frac{1}{\sqrt{n}} \sum_{1 \le i \le n} \Big( \eta E[m(X_i, 1, R_i(1), \theta_0)|X_i] + (1 - \eta)E[m(X_i, 0, R_i(0), \theta_0)|X_i]$$
$$+ I\{A_i = 1\}(m(X_i, 1, R_i, \theta_0) - E[m(X_i, 1, R_i(1), \theta_0)|X_i])$$
$$+ I\{A_i = 0\}(m(X_i, 0, R_i, \theta_0) - E[m(X_i, 0, R_i(0), \theta_0)|X_i])$$
$$+ (A_i - \eta)(E[m(X_i, 1, R_i(1), \theta_0) - m(X_i, 0, R_i(0), \theta_0)|X_i]) \Big) .$$

Let $\Omega(X_i) = E[m(X_i, 1, R_i(1), \theta_0) - m(X_i, 0, R_i(0), \theta_0)|X_i]$ and note that by Assumption 2,

$$E\left[ \frac{1}{\sqrt{n}} \sum_{1 \le i \le n} (A_i - \eta)\Omega(X_i) \bigg| X^{(n)} \right] = 0 .$$

Recall $\Omega^{(s)}(X_i)$ is the $s$th component of $\Omega(X_i)$. Next, it follows from Assumption 3.1, 3.3(f), and equation (12.3) in Lehmann and Romano (2022) that for $1 \le s \le d_\theta$,

$$\text{Var}\left[ \frac{1}{\sqrt{n}} \sum_{1 \le i \le n} (A_i - \eta)\Omega^{(s)}(X_i) \bigg| X^{(n)} \right] = \frac{1}{n} \sum_{1 \le j \le n/k} \frac{\ell(k-\ell)}{k-1} \sum_{i \in \lambda_j} (\Omega_i^{(s)} - \overline{\Omega}_j^{(s)})^2$$
$$\le C^2 \frac{\ell(k-\ell)}{k-1} \frac{1}{n} \sum_{1 \le j \le n/k} \max_{i,i' \in \lambda_j} \|X_i - X_{i'}\|^2 ,$$

where $\overline{\Omega}_j^{(s)} = \frac{1}{k} \sum_{i \in \lambda_j} \Omega^{(s)}(X_i)$, and so the conditional variance converges in probability to zero under Assumption 3.2. It then follows from Markov's inequality and the fact that probabilities are bounded and hence uniformly integrable that

$$\frac{1}{\sqrt{n}} \sum_{1 \le i \le n} (A_i - \eta)\Omega(X_i) = o_P(1) .$$

Therefore,

$$-M^{-1}\frac{1}{\sqrt{n}}\sum_{1\le i\le n}m(X_i, A_i, R_i, \theta_0) = -M^{-1}\frac{1}{\sqrt{n}}\sum_{1\le i\le n}m^*(X_i, A_i, R_i, \theta_0) + o_P(1) \ ,$$

which, together with (12), implies the desired result in (9). ∎

## A.2 Proof of Theorem 3.2

By assumption $\widehat{M}_n \xrightarrow{P} M$. Therefore, it suffices to show that

$$\hat{\Sigma}_{1,n} \xrightarrow{P} \Sigma_1 \tag{27}$$

$$\hat{\Sigma}_{2,n} \xrightarrow{P} \Sigma_2 \ , \tag{28}$$

where

$$\Sigma_1 = \eta \operatorname{Var}[m(X_i, 1, R_i(1), \theta_0)] + (1-\eta)\operatorname{Var}[m(X_i, 0, R_i(0), \theta_0)]$$

$$\Sigma_2 = -\eta(1-\eta)\operatorname{Var}\big[E[m(X_i, 1, R_i(1), \theta_0)|X_i] - E[m(X_i, 1, R_i(1), \theta_0)]$$
$$- (E[m(X_i, 0, R_i(0), \theta_0)|X_i] - E[m(X_i, 0, R_i(0), \theta_0)])\big] \ .$$

In what follows, we will show (27). The proof of (28) will follow from similar steps, but with the calculations below replaced by the ones in the proof of Lemmas C.2–C.3 in Bai et al. (2024c), along with Assumptions 3.4–3.5. To that end, we show for $1 \le s \le d_\theta$,

$$\hat{\mu}_{1,n}^{(s)} := \frac{1}{\eta n}\sum_{1\le i\le n}m_s(X_i, 1, R_i, \hat{\theta}_n) \xrightarrow{P} E[m_s(X_i, 1, R_i(1), \theta_0)] =: \mu_1^{(s)} \ , \tag{29}$$

and similar arguments will establish the results for $a = 0$ as well as for the second moments and therefore (27). For $\theta \in \Theta$, define

$$\hat{\mu}_{1,n}^{(s)}(\theta) = \frac{1}{\eta n}\sum_{1\le i\le n}m_s(X_i, 1, R_i, \theta) \tag{30}$$

and note $\hat{\mu}_{1,n}^{(s)} = \hat{\mu}_{1,n}^{(s)}(\hat{\theta}_n)$. Suppose (29) doesn't hold. Then, there exists a subsequence $\{n_k\}_{k\ge 1}$ and $\epsilon_1, \epsilon_2 > 0$, such that

$$\lim_{k\to\infty}P\{|\hat{\mu}_{1,n}^{(s)} - \mu_1^{(s)}| > \epsilon_1\} \to \epsilon_2 \ . \tag{31}$$

Because $\hat{\theta}_{n_k} \xrightarrow{P} \theta_0$ by Theorem 3.1, there exists a further subsequence, which we still denote by $\{n_k\}_{k\ge 1}$ by an abuse of notation, along which $\hat{\theta}_{n_k} \to \theta_0$ with probability one. Along that subsequence, Lemma A.9 implies that $\hat{\mu}_{1,n}^{(s)} \xrightarrow{P} \mu_1^{(s)}$, in contradiction to (31). Therefore, (29) holds, and the theorem follows as discussed above. ∎

## A.3    Proofs for Section 4.1

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 $A^{(n)}$ satisfying Assumption 2.1 can be represented as a function of $X^{(n)}$ and some additional exogenous 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 $\mu^U$. In what follows, we consider a family $\{Q_t : t \in \mathbf{R}^{d_\theta}\}$ of marginal distributions indexed by $t$, and let $q_t^X$ denote the density function for $X_i$ with respect to a dominating measure $\mu^X$, $q_t^{R(a)|X}(r|x)$ denote the conditional density of $R_i(a)$ given $X_i$ with respect to a dominating measure $\mu^R$. With some abuse of notation, continue letting $P_{t,n}$ denote the distribution of $(U_n, X^{(n)}, R^{(n)})$. 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

$$\ell_n = p_n^U(U_n) \prod_{1 \le i \le n} q_t^X(X_i) \prod_{1 \le i \le n} \prod_{a \in \{0,1\}} q_t^{R(a)|X}(R_i|X_i)^{I\{A_i=a\}} . \tag{32}$$

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 denote the distribution $P_{t,n}$ by $P_t$.

We consider parametric submodels $\{P_t : t \in \mathbf{R}^{d_\theta}\}$, where $P_0 = P$, such that the following holds for some $g = (g^X, g^{R(1)|X}, g^{R(0)|X})$, each component of which is a $d_\theta$-dimensional function:

(a)  As $t \to 0$,
$$\int \frac{1}{\|t\|^2} \left( q_t^X(x)^{1/2} - q^X(x)^{1/2} - \frac{1}{2}q^X(x)^{1/2}t'g^X(x) \right)^2 d\mu^X(x) \to 0 . \tag{33}$$

(b)  For $a \in \{0,1\}$, as $t \to 0$,

$$\frac{1}{\|t\|^2} \iint \left( q_t^{R(a)|X}(r|x)^{1/2} - q^{R(a)|X}(r|x)^{1/2} - \frac{1}{2}q^{R(a)|X}(r|x)^{1/2}t'g^{R(a)|X}(r|x) \right)^2$$
$$\times d\mu^R(r)q^X(x)d\mu^X(x) \to 0 . \tag{34}$$

In what follows, we will 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 $q_{t,g}$. When writing expectations and variances, we suppress the subscripts $P$ and $Q$ whenever doing so does not lead to confusion. For completeness, we document the following properties of score functions which satisfy (33)–(34):

**Lemma A.1.** *For a parametric submodel $\{P_{t,g} : t \in \mathbf{R}^{d_\theta}\}$ with $P_{0,g} = P$ that satisfies (33)–(34),*

(a)  $E[g^X(X)g^X(X)'] < \infty$.

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

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

(d)  $E[g^{R(a)|X}(R(a)|X)|X] = 0$ 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 (34) and Markov's inequality imply that along a subsequence $t_{n_k}$,

$$\frac{1}{\|t_{n_k}\|^2} \int \left( q_{t_{n_k}}^{R(a)|X}(r|x)^{1/2} - q^{R(a)|X}(r|x)^{1/2} - \frac{1}{2} q^{R(a)|X}(r|x)^{1/2} t'_{n_k} g^{R(a)|X}(r|x) \right)^2$$
$$\times \, d\mu^R(r) \to 0$$

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

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

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

Further define $I = I^X + \eta E[I^{R(1)|X}(X)] + (1-\eta)E[I^{R(0)|X}(X)]$. We restrict ourselves to parametric submodels that satisfy (33)–(34) for a $g$ that satisfies the following conditions. These submodels exist by Lemma A.2 below.

**Condition A.1.** The function $g$ satisfies that

(a) $E[g^X(X)] = 0$ and $\mathrm{Var}[g^X(X)] < \infty$.

(b) For $a \in \{0, 1\}$, $E[g^{R(a)|X}(R(a)|X)|X] = 0$, and $\mathrm{Var}[g^{R(a)|X}(R(a)|X)] < \infty$ with probability one.

(c) $I$ is nonsingular.

**Lemma A.2.** *For any $g$ that satisfies Condition A.1, there exists a parametric submodel $\{P_{t,g} : t \in \mathbf{R}^{d_\theta}\}$ such that (33)–(34) hold.*

PROOF. We use a vector version of the construction in in Example 25.16 in van der Vaart (1998). Let $k(x)$ be any strictly positive function that is bounded from above and away from zero with a bounded derivative such that $k(0) = k'(0) = 1$; for example, take $k(x) = 2(1 + e^{-2x})^{-1}$. Define

$$q_t^X(x) = C(t)q^X(x)k(t'g^X(x)) \,,$$

where $C(t) = \left( \int q^X(x)k(t'g^X(x))d\mu^X(x) \right)^{-1}$, so that $q_t^X(x)$ is a probability density function. Differentiating both sides of $C(t) \int q^X(x)k(t'g^X(x))d\mu^X(x) = 1$ at $t = 0$, we get that $\frac{\partial}{\partial t}\big|_{t=0}C(t) = 0$. It can then be verified through direct calculation that

$$\frac{\partial}{\partial t}\bigg|_{t=0} \log q_t^X(x) = g^X(x) \,.$$

The quadratic mean differentiability requirement in (33) follows from Lemma 7.6 in van der Vaart (1998). Next, for each $x \in \mathbf{R}^{d_x}$, we define

$$q_t^{R(1)|X}(r|x) = C(t)q^{R(1)|X}(r|x)k(t'g^{R(1)|X}(r|x)) \,.$$

As above, it can be verified through direct calculation that

$$\frac{\partial}{\partial t}\bigg|_{t=0} \log q_t^{R(1)|X}(r|x) = g^{R(1)|X}(r|x) \ .$$

To show (34) for $a = 1$ (and symmetric arguments apply for $a = 0$), we modify the arguments in the proof of Lemma 7.6 in van der Vaart (1998). Define $s_t(r|x) = q_t^{R(1)|X}(r|x)^{1/2}$ and denote the $\ell$th component by $s_t^{(\ell)}(r|x)$ for $1 \le \ell \le d_\theta$. By the mean-value theorem, we have $s_t(r|x) - s(r|x) = \int_0^1 t' \dot{s}_{ut}(r|x) du$, where $\dot{s}_t = \frac{\partial}{\partial t} s_t$, so that it follows from Jensen's inequality that

$$\frac{1}{\|t\|^2} \iint (s_t(r|x) - s_0(r|x) - t'\dot{s}_0(r|x))^2 \, d\mu^R(r) q^X(x) d\mu^X(x)$$

$$\le \frac{1}{\|t\|^2} \iiint_0^1 \left(t'(\dot{s}_{ut}(r|x) - \dot{s}_0(r|x))\right)^2 du d\mu^R(r) q^X(x) d\mu^X(x)$$

$$\le \iiint_0^1 \|\dot{s}_{ut}(r|x) - \dot{s}_0(r|x)\|^2 du d\mu^R(r) q^X(x) d\mu^X(x) \tag{35}$$

where the first inequality follows from Jensen's inequality and the second by Cauchy-Schwarz. It then suffices to show (35) goes to zero as $t \to 0$. Analyzing componentwise, it suffices to show that for $1 \le \ell \le d_\theta$, as $t \to 0$,

$$\iiint_0^1 (\dot{s}_{ut}^{(\ell)}(r|x) - \dot{s}_0^{(\ell)}(r|x))^2 du d\mu^R(r) q^X(x) d\mu^X(x) \to 0 \ . \tag{36}$$

The integrand in (36) obviously converges to zero as $t \to 0$ by continuous differentiability of $q_t$. Recall that $\int \dot{s}_{ut}^{(\ell)}(r|x)^2 d\mu^R(r) = \frac{1}{4}[I_{ut}^{R(1)|X}(x)]_{(\ell,\ell)}$, where $I_{ut}^{R(1)|X}(x)$ is the conditional information for $a = 1$ given $x$ at $P_{ut,g}$. Therefore, it follows from Fubini's theorem that

$$\iiint_0^1 \dot{s}_0^{(\ell)}(r|x)^2 du d\mu^R(r) q^X(x) d\mu^X(x) = \frac{1}{4} \int [I_0^{R(1)|X}(x)]_{(\ell,\ell)} q^X(x) d\mu^X(x)$$

$$\iiint_0^1 \dot{s}_{ut}^{(\ell)}(r|x)^2 du d\mu^R(r) q^X(x) d\mu^X(x) = \frac{1}{4} \iint_0^1 [I_{ut}^{R(1)|X}(x)]_{(\ell,\ell)} du q^X(x) d\mu^X(x) \ .$$

To apply Vitali's theorem (Proposition 2.29 in van der Vaart, 1998), it suffices to show that

$$\iint_0^1 [I_{ut}^{R(1)|X}(x)]_{(\ell,\ell)} du q^X(x) d\mu^X(x) \to \int [I_0^{R(1)|X}(x)]_{(\ell,\ell)} q^X(x) d\mu^X(x)$$

as $t \to 0$. To do so, we fix any arbitrarily small $\delta > 0$ and note that at least for $t$ small enough, $\|ut\| \le \delta$ for $u \in [0,1]$, so we can apply the dominated convergence theorem with

$$[I_{ut}^{R(1)|X}(x)]_{(\ell,\ell)} \le \sup_{\|h\| \le \delta} [I_h^{R(1)|X}(x)]_{(\ell,\ell)} \ ,$$

as long as

$$\iint_0^1 \sup_{\|h\| \le \delta} [I_h^{R(1)|X}(x)]_{(\ell,\ell)} du q^X(x) d\mu^X(x)$$

$$= \int \sup_{\|h\| \le \delta} [I_h^{R(1)|X}(x)]_{(\ell,\ell)} q^X(x) d\mu^X(x) < \infty \ . \tag{37}$$

To show (37), we calculate the conditional information as

$$E\left[\left(\frac{\frac{\partial}{\partial h_\ell}C(h)}{C(h)} + \frac{k'(h'g^{R(1)|X}(R(1)|X))}{k(h'g^{R(1)|X}(R(1)|X))}g_\ell^{R(1)|X}(R(1)|X)\right)^2\bigg| X = x\right].$$

Note that $k'$ is bounded above and $k$ is bounded below, $C(h)$ is continuously differentiable with $C(0) = 1$, and so is bounded for $\|h\| \leq \delta$. Therefore, an application of the Cauchy-Schwarz inequality implies the previous expectation is bounded by a constant plus a constant multiple of $[I_0^{R(1)|X}(x)]_{(\ell,\ell)}$. The desired conclusion in (37) then follows because $E[I_0^{R(1)|X}(X)] < \infty$, and the proof is complete. ∎

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

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

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

**Lemma A.3.** *Suppose the treatment assignment mechanism satisfies Assumption 2.1 and the path satisfies (33)–(34) for $g$ satisfying Condition A.1. Then,*

$$L_{t,n}(g) = \frac{1}{\sqrt{n}}\sum_{1 \leq i \leq n} t's_g(X_i, A_i, R_i) - \frac{1}{2}t'I^Xt$$

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

*where*

$$s_g(x, a, r) = g^X(x) + I\{a = 1\}g^{R(1)|X}(r|x) + I\{a = 0\}g^{R(0)|X}(r|x) \tag{38}$$

*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 4.1, then, under $P_0$,*

$$L_{t,n}(g) \xrightarrow{d} N\left(-\frac{1}{2}t'It, t'It\right) ,$$

PROOF. The first result follows from Theorem 3.1 of Armstrong (2022). The second result follows from Lemma A.6 given Assumption 4.1 and the assumption that each component of $I^{R(a)|X}(x)$ is integrable, noting that $E[s_g(X, 1, R(1)) - s_g(X, 0, R(0))|X] = 0$. ∎

We emphasize that Lemma A.7 implies

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

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

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

$$\sqrt{n}(\theta(P_{t/\sqrt{n},g}) - \theta(P)) \to E_P[\psi^* s_g' t] \ . \tag{39}$$

In Lemma A.5 below, we provide explicit conditions which guarantee this is possible when $\theta(P)$ is defined by (3).

We call an estimator $\tilde{\theta}_n$ for $\theta(P)$ regular if for all $g$ satisfying Condition A.1 and $t \in \mathbf{R}^{d_\theta}$,

$$\sqrt{n}(\tilde{\theta}_n - \theta(P_{t/\sqrt{n},g})) \xrightarrow{P_{t/\sqrt{n},g}} L \tag{40}$$

for a fixed probability measure $L$.

The following lemma establishes a convolution theorem for regular estimators:

**Lemma A.4.** *Suppose $\theta$ satisfies* (39). *Let $\tilde{\theta}_n$ be a regular estimator for $\theta$. 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 $P$. To align our setting with Theorem 3.11.2 in van der Vaart and Wellner (1996), we first characterize the adjoint map (viewed as a mapping into $\mathcal{M}_g$) of the function $v \mapsto E[\psi^* v] \in \mathbb{R}^{d_\theta}$, where $v \in \mathcal{M}_g$. To that end, implicitly identifying each $b \in \mathbf{R}^{d_\theta}$ with the functional $b^* : \mathbb{R}^{d_\theta} \to \mathbb{R}$ given by $x \mapsto b'x$, we construct a $w(b) \in \mathbf{R}^{d_\theta}$ such that

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

for all $t \in \mathbf{R}^{d_\theta}$, where we note $w(b)' s_g \in \mathcal{M}_g$, and is thus the output of the adjoint map when applied to the functional $b^*$. Because $E[s_g s_g']$ is invertible, we immediately obtain

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

Here we use the assumption that $I$ is nonsingular. It then follows from the local asymptotic normality established in Lemma A.3 and Theorem 3.11.2 in van der Vaart and Wellner (1996) that

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

36

where $B_g$ is a fixed probability measure and $b'V_g b = E[(w(b)'s_g)^2]$, so that we have

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

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. $\blacksquare$

To apply Lemma A.4 to the setting in Section 4.1, we we establish the form of $\psi^*$ in (39) for the parameter $\theta_0 = \theta(P)$ defined by (3). Define $\eta(X_i) = P\{A_i = 1|X_i\}$. Note that

$$
\begin{aligned}
0 &= E_P[m(X_i, A_i, R_i, \theta(P))] \\
&= E_Q[m(X, 1, R(1), \theta(P))\eta(X)] + E_Q[m(X, 0, R(0), \theta(P))(1 - \eta(X))] \ .
\end{aligned}
\tag{41}
$$

**Lemma A.5.** *Suppose the treatment assignment mechanism satisfies Assumptions 2.1 and 4.1. Fix a function $g$ that satisfies Condition A.1. Suppose (33)–(34) holds. Fix $t \in \mathbf{R}^{d_\theta}$ and consider a one-dimensional submodel $\{P_{t/\sqrt{n},g}\}$ such that*

$$E_{Q_{t/\sqrt{n}}}[m(X, a, R(a), \theta(P))^2] = O(1) \tag{42}$$

$$E_{Q^X}[E_{Q_{t/\sqrt{n}}^{R(a)|X}}[m(X, a, R(a), \theta(P))^2 | X]] = O(1) \tag{43}$$

$$E_{Q_{t/\sqrt{n}}^X}[E_{Q^{R(a)|X}}[m(X, a, R(a), \theta(P))^2 | X]] = O(1) \tag{44}$$

*as $n \to \infty$ and $\theta(P_{t/\sqrt{n},g})$ is uniquely determined by (41). Then, $\theta(P_{t/\sqrt{n},g})$ defined by (41) satisfies*

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

*where*

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

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 (33)–(34) for some $g = (g^X, g^{R(1)|X}, g^{R(0)|X})$, each component of which is a one-dimensional function. (41) 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

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

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

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

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

It follows from the Cauchy-Schwarz inequality that

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

by the assumption that $E_P[m(X, a, R(a), \theta(P))^2] < \infty$, the facts that $0 \leq \eta(x) \leq 1$, $\int q^{R(1)|X}(r|x) d\mu^R(r) =$

38

1, and (33). 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 (43)–(44). Combining all previous results yields

$$\begin{aligned}
\frac{\partial}{\partial\nu}&E_{P_\nu}[m(X,A,R,\theta(P))]\Big|_{\nu=0} \\
&= E_Q[m(X,1,R(1),\theta(P))(g^X(X) + g^{R(1)|X}(R|X))\eta(X)] \\
&\quad + E_Q[m(X,0,R(0),\theta(P))(g^X(X) + g^{R(0)|X}(R|X))(1-\eta(X))] \\
&= E_P[m(X,A,R,\theta(P))(g^X(X) + I\{A=1\}g^{R(1)|X}(R) + I\{A=0\}g^{R(0)|X}(R))] \ .
\end{aligned}$$

On the other hand, by definition

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

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.7 together with Condition A.1. ∎

Finally, to prove Theorem 4.1 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 (42)–(44) hold as $n \to \infty$, and $\theta(P_{t/\sqrt{n},g})$ is uniquely determined by (41).

This assumption guarantees that every element satisfying Condition A.1 has a corresponding path for which we can apply Lemma A.5. 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.5, 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 4.1. First note $\theta$ satisfies (39) because of Lemma A.5 and Assumption A.1. The result then follows from Lemma A.4 upon noting that $\psi^* = s_g$ for some $g$ that satisfies Condition A.1. ∎

To study regular estimators, we need the following lemma.

**Lemma A.6.** *Suppose the treatment assignment mechanism satisfies Assumptions 2.1, 4.1, and 4.2. Let $f(x,a,r)$ be a vector-valued function such that $E[f(X,A,R)] = 0$ and $E[f^2(X,a,R(a))] < \infty$ for $a \in \{0,1\}$.*

*Then,*

$$\frac{1}{\sqrt{n}} \sum_{1 \leq i \leq n} f(X_i, A_i, R_i) \xrightarrow{d} N(0, V_f) \ ,$$

*where $V_f = V_{1,f} + V_{2,f} + V_{3,f}$ for*

$$V_{1,f} = \text{Var}[E[f(X, A, R)|X]] = \text{Var}[\eta E[f(X, 1, R(1))|X] + (1 - \eta)E[f(X, 0, R(0))|X]]$$
$$V_{2,f} = V_{E[f(X,1,R(1))-f(X,0,R(0))|X]}^{\text{imb}}$$
$$V_{3,f} = E[\eta \text{Var}[f(X, 1, R(1))|X] + (1 - \eta) \text{Var}[f(X, 0, R(0))|X]] \ .$$

PROOF. Note that

$$\mathbb{C}_n := \frac{1}{\sqrt{n}} \sum_{1 \leq i \leq n} f(X_i, A_i, R_i) = \mathbb{C}_{1,n} + \mathbb{C}_{2,n} + \mathbb{C}_{3,n} \ ,$$

where

$$\begin{aligned}
\mathbb{C}_{1,n} &= \frac{1}{\sqrt{n}} \sum_{1 \leq i \leq n} E[f(X_i, A_i, R_i)|X_i] \\
&= \frac{1}{\sqrt{n}} \sum_{1 \leq i \leq n} (\eta E[f(X_i, 1, R_i(1))|X_i] + (1 - \eta)E[f(X_i, 0, R_i(0))|X_i]) \\
\mathbb{C}_{2,n} &= \frac{1}{\sqrt{n}} \sum_{1 \leq i \leq n} (A_i - \eta)E[f(X_i, 1, R_i(1)) - f(X_i, 0, R_i(0))|X_i] \\
\mathbb{C}_{3,n} &= \frac{1}{\sqrt{n}} \sum_{1 \leq i \leq n} \big(A_i(f(X_i, 1, R_i(1)) - E[f(X_i, 1, R_i(1))|X_i]) \\
&\qquad\qquad + (1 - A_i)(f(X_i, 0, R_i(0)) - E[f(X_i, 0, R_i(0))|X_i])\big) \ .
\end{aligned}$$

Note that $\mathbb{C}_{1,n}$ has zero mean because $E[E[f(X_i, A_i, R_i)|X_i]] = E[f(X_i, A_i, R_i)] = 0$. Further note that $\mathbb{C}_{1,n} = E[\mathbb{C}_n|X^{(n)}]$, $\mathbb{C}_{2,n} = E[\mathbb{C}_n|X^{(n)}, A^{(n)}] - E[\mathbb{C}_n|X^{(n)}]$, and $\mathbb{C}_{3,n} = \mathbb{C}_n - E[\mathbb{C}_n|X^{(n)}, A^{(n)}]$. It follows from the central limit theorem and $E[f^2(X, a, R(a))] < \infty$ that

$$\mathbb{C}_{1,n} \xrightarrow{d} N(0, V_{1,f}) \ .$$

Next, it follows from Assumption 4.2 that

$$\rho(\mathbb{C}_{2,n}, N(0, V_{2,f})|X^{(n)}) \xrightarrow{P} 0 \ .$$

For $\mathbb{C}_{3,n}$, it follows from Assumption 4.1 that

$$\text{Var}[\mathbb{C}_{3,n}|X^{(n)}, A^{(n)}] \xrightarrow{P} V_{3,f} \ .$$

As a result, one can verify the Lindeberg condition conditional on $X^{(n)}$ and $A^{(n)}$ as in the proof of Lemma S.1.4 of Bai et al. (2022), and obtain

$$\rho(\mathbb{C}_{3,n}, N(0, V_{3,f})|X^{(n)}, A^{(n)}) \xrightarrow{P} 0 \ .$$

The proof can then be completed by the subsequencing argument at the end of the proof of Lemma S.1.4 of Bai et al. (2022). ∎

PROOF OF THEOREM 4.2. Recall from Lemma A.3 that $L_{t,n}(g)$ is also asymptotically linear. Because the treatment assignment mechanism satisfies Assumptions 2.1, 4.1, and 4.2, the path satisfies Condition A.1, and hence $\psi$ and $s_g$ jointly satisfy the conditions in Lemma A.6 (in particular, note that $E[s_g(X, 1, R(1)) - s_g(X, 0, R(0))|X] = 0$),

$$
\begin{pmatrix} \sqrt{n}(\tilde{\theta}_n - \theta(P)) \\ L_{t,n}(g) \end{pmatrix} = \begin{pmatrix} \frac{1}{\sqrt{n}} \sum_{1 \le i \le n} \psi(X_i, A_i, R_i, \theta(P)) \\ \frac{1}{\sqrt{n}} \sum_{1 \le i \le n} t' s_g(X_i, A_i, R_i) \end{pmatrix} - \begin{pmatrix} 0 \\ -\frac{1}{2} t' I t \end{pmatrix} + o_P(1)
$$

are jointly asymptotically normal. Because $E[s_g(X, 1, R(1)) - s_g(X, 0, R(0))|X] = 0$, the covariance in the second term in the joint variance in Lemma A.6 vanishes, and thus the overall covariance is given by

$$
\begin{aligned}
E[E[\psi|X] & g^X(X)']t \\
& + \eta E[(\psi(X, 1, R(1), \theta(P)) - E[\psi(X, 1, R(1), \theta(P))|X]) g^{R(1)|X}(R(1)|X)']t \\
& + (1 - \eta) E[(\psi(X, 0, R(0), \theta(P)) - E[\psi(X, 0, R(0), \theta(P))|X]) g^{R(0)|X}(R(0)|X)']t \\
= E[\psi s_g']t \ .
\end{aligned}
$$

It then follows from Le Cam's third lemma that under $P_{t/\sqrt{n},g}$, $\sqrt{n}(\tilde{\theta}_n - \theta_0)$ converges in distribution to a normal distribution with mean $E[\psi s_g']t$ and the same variance as in the limit under $P$. At the same time,

$$
\sqrt{n}(\tilde{\theta}_n - \theta(P_{t/\sqrt{n},g})) = \sqrt{n}(\tilde{\theta}_n - \theta(P)) - \sqrt{n}(\theta(P_{t/\sqrt{n},g}) - \theta(P)) \ .
$$

Therefore, (40) holds if and only if

$$
E[\psi s_g']t = E[\psi^* s_g']t \ .
$$

Furthermore, $\tilde{\theta}_n$ is regular if the equality holds for all $t$ and all $g$ satisfying Condition A.1, if and only if

$$
E[(\psi - \psi^*) s_g'] = 0
$$

for all $g$ satisfying Condition A.1. Note that

$$
\begin{aligned}
E[(\psi & - \psi^*) s_g'] \\
= E[E[\psi & (X, A, R, \theta(P)) - \psi^*(X, A, R, \theta(P))|X] g^X(X)'] \\
& + E[A_i(\psi(X, 1, R(1), \theta(P)) - \psi^*(X, 1, R(1), \theta(P)) \\
& \quad - E[\psi(X, 1, R(1), \theta(P)) - \psi^*(X, 1, R(1), \theta(P))|X]) g^{R(1)|X}(R(1)|X)'] \\
& + E[(1 - A_i)(\psi(X, 0, R(0), \theta(P)) - \psi^*(X, 0, R(0), \theta(P)) \\
& \quad - E[\psi(X, 0, R(0), \theta(P)) - \psi^*(X, 0, R(0), \theta(P))|X]) g^{R(0)|X}(R(0)|X)'] \ .
\end{aligned}
$$

By setting $g^{R(1)|X} = 0$, $g^{R(0)|X} = 0$, and $g^X(X) = E[\psi(X, A, R, \theta(P)) - \psi^*(X, A, R, \theta(P))|X]$, we get that

$$
E[\psi(X, A, R, \theta(P))|X] = E[\psi^*(X, A, R, \theta(P))|X] \ . \tag{45}
$$

Setting $g^X(X) = 0 = g^{R(0)|X}$ and $g^{R(1)|X} = \psi(X, 1, R(1), \theta(P)) - \psi^*(X, 1, R(1), \theta(P)) - E[\psi(X, 1, R(1), \theta(P)) - \psi^*(X, 1, R(1), \theta(P))|X]$, we get

$$\psi(X, 1, R(1), \theta(P)) - \psi^*(X, 1, R(1), \theta(P))$$
$$- E[\psi(X, 1, R(1), \theta(P)) - \psi^*(X, 1, R(1), \theta(P))|X] = 0 \ ,$$

which implies that $\psi(X, 1, R(1), \theta(P)) - \psi^*(X, 1, R(1), \theta(P))$ can only be a function of $X$. Denote it by $\psi^\perp(X, 1)$. Similarly, $\psi(X, 0, R(0), \theta(P)) - \psi^*(X, 0, R(0), \theta(P))$ can only be a function of $X$. Denote it by $\psi^\perp(X, 0)$. We have

$$\psi(X, A, R, \theta(P)) = A\psi(X, 1, R(1), \theta(P)) + (1 - A)\psi(X, 0, R(0), \theta(P))$$
$$= A(\psi^*(X, 1, R(1), \theta(P)) + \psi^\perp(X, 1, \theta(P)))$$
$$+ (1 - A)(\psi^*(X, 0, R(0), \theta(P)) + \psi^\perp(X, 0, \theta(P)))$$
$$= \psi^*(X, A, R, \theta(P)) + \psi^\perp(X, A, \theta(P)) \ ,$$

and it follows from (45) that $E[\psi^\perp(X, A, \theta(P))|X] = 0$. Finally, to show that $\tilde{\theta}_n$ is efficient if and only if (23) holds, we apply Lemma A.6 with $f = \psi$. If (23) holds, then $V_\psi = V_{1,\psi} + V_{3,\psi} = V_{\psi^*}$ because $V_{1,\psi} = V_{1,\psi^*}$, $V_{3,\psi} = V_{3,\psi^*}$ and $V_{2,\psi^*} = 0$. On the other hand, if $\tilde{\theta}_n$ is efficient, then $V_{2,\psi} = 0$, so (23) holds by Markov's inequality combined with the fact that probabilities are bounded and thus uniformly integrable. ∎

## A.4  Auxiliary Lemmas

**Lemma A.7.** *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 $C \in \mathbf{R}^{d_x}$.

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

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 expectations as well as the fact that $Q_n = Q^n$. 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. Condi-

tional 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 $Q_n = Q^n$. $\blacksquare$

**Lemma A.8.** *Suppose the treatment assignment mechanism satisfies Assumptions 3.1–3.2 and the moment functions satisfy Assumption 3.3. Then, $\hat{\theta}_n \xrightarrow{P} \theta_0$.*

PROOF. It follows from Assumption 3.3(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 3.3(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 $E_P[m_s(X_i, A_i, R_i, \theta_m)] \to E_P[m_s(X_i, A_i, R_i, \theta)]$. Here, the dominating function exists by Problem 2.4.1 in van der Vaart and Wellner (1996). Assumption 3.3(c) then implies

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

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

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.7 that

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

from which we obtain that

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

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

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

$$+ E\left[ \sup_{\theta \in \Theta^*} \left| \frac{1}{n} \sum_{1 \leq i \leq n} (1 - A_i)(m_s(X_i, 0, R_i(0), \theta) - (1 - \eta) E[m_s(X_i, 0, R_i(0), \theta)]) \right| \right]$$

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

$$+ E\left[\sup_{\theta \in \Theta^*} \left|\frac{1}{n} \sum_{1 \leq i \leq n} (1 - A_i)(m_s(X_i, 0, R_i(0), \theta) - E[m_s(X_i, 0, R_i(0), \theta)])\right|\right]$$

$$\lesssim E\left[\sup_{\theta \in \Theta^*} \left|\frac{1}{n} \sum_{1 \leq i \leq n} (m_s(X_i, 1, R_i(1), \theta) - E[m_s(X_i, 1, R_i(1), \theta)])\right|\right]$$

$$+ E\left[\sup_{\theta \in \Theta^*} \left|\frac{1}{n} \sum_{1 \leq i \leq n} (m_s(X_i, 0, R_i(0), \theta) - E[m_s(X_i, 0, R_i(0), \theta)])\right|\right] \to 0 .$$

where the first equality follows from (48), the first inequality follows from the triangle inequality, the second equality follows because $\sum_{1 \leq i \leq n} A_i = n\eta$ and $\sum_{1 \leq i \leq n}(1 - A_i) = n(1 - \eta)$, and the last inequality follows from Proposition C.1 in Han and Wellner (2021). The convergence follows from Assumption 3.3(e) and an application of the backward submartingale convergence theorem (see, for instance, Theorem 12.30 in Le Gall, 2022), as detailed in the proof of Theorem 3.1 in Han and Wellner (2021). The desired result in (47) then follows by Markov's inequality. ∎

**Lemma A.9.** *Suppose the treatment assignment mechanism satisfies Assumptions 3.1, 3.2, and 3.4, and the moment functions satisfy Assumptions 3.3 and 3.5. Then, for each deterministic sequence $\theta_n \to \theta_0$, $\hat{\mu}_{1,n}^{(s)}(\theta_n) \xrightarrow{P} \mu_1^{(s)}$ for $\mu_1^{(s)}$ in (29) and $\hat{\mu}_{1,n}^{(s)}(\theta_n)$ in (30).*

PROOF. The conclusion could follow from similar arguments to those in the proof of Lemma A.8 but we provide a different proof because it will apply generally to other components of the variance estimator. It suffices to prove that

$$\hat{\mu}_{1,n}^{(s)}(\theta_n) - E[\hat{\mu}_{1,n}^{(s)}(\theta_n)|X^{(n)}, A^{(n)}] \xrightarrow{P} 0 \tag{51}$$

$$E[\hat{\mu}_{1,n}^{(s)}(\theta_n)|X^{(n)}, A^{(n)}] \xrightarrow{P} \mu_1^{(s)} . \tag{52}$$

The desired result in (51) follows from similar arguments to those in the proof of equation (S.27) in Bai et al. (2022), where the uniform integrability condition is replaced by Assumption 3.5(a). To show (52), note

$$E[\hat{\mu}_{1,n}^{(s)}(\theta_n)|X^{(n)}, A^{(n)}] - \mu_1^{(s)}$$

$$= \frac{1}{\eta n} \sum_{1 \leq i \leq n} A_i E[m_s(X_i, 1, R_i(1), \theta_n)|X_i] - E[m_s(X_i, 1, R_i(1), \theta_0)]$$

$$= \frac{1}{\eta n} \sum_{1 \leq i \leq n} A_i E[m_s(X_i, 1, R_i(1), \theta_n)|X_i] - \frac{1}{n} \sum_{1 \leq i \leq n} E[m_s(X_i, 1, R_i(1), \theta_n)|X_i] \tag{53}$$

$$+ \frac{1}{n} \sum_{1 \leq i \leq n} E[m_s(X_i, 1, R_i(1), \theta_n)|X_i] - E[m_s(X_i, 1, R_i(1), \theta_0)] . \tag{54}$$

The difference in (54) converges in probability to zero by Assumption 3.5(b) and because $E[m(X_i, 1, R_i(1), \theta_n)] \to E[m(X_i, 1, R_i(1), \theta_0)]$ from $\theta_n \to \theta_0$ and Assumption 3.3(c). Next, note the absolute value of (53) can be written as

$$\left|\frac{k}{n} \sum_{1 \leq j \leq n/k} \frac{1}{\eta k} \sum_{i \in \lambda_j} A_i \left(E[m_s(X_i, 1, R_i(1), \theta_n)|X_i]\right.\right.$$

44

$$-\frac{1}{k}\sum_{i'\in\lambda_j}E[m_s(X_{i'},1,R_{i'}(1),\theta_n)|X_{i'}]\Bigg)\Bigg|$$

$$\leq\frac{k}{n}\sum_{1\leq j\leq n/k}\frac{1}{\eta k}\sum_{i\in\lambda_j}A_i\Big|E[m_s(X_i,1,R_i(1),\theta_n)|X_i]-E[m_s(X_{i'},1,R_{i'}(1),\theta_n)|X_{i'}]\Big|$$

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

where the inequality follows from the uniform Lipschitzness condition in Assumption 3.5(c) and the convergencef follows from Assumption 3.2. Therefore, (51) holds. ∎

# 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. (2025a). Inference in Experiments with Matched Pairs and Imperfect Compliance. *Journal of Business & Economic Statistics*, **43** 627–642.

BAI, Y., HUANG, X., ROMANO, J. P., SHAIKH, A. M. and TABORD-MEEHAN, M. (2025b). A new design-based variance estimator for finely stratified experiments. *arXiv preprint arXiv:2503.10851*.

BAI, Y., JIANG, L., ROMANO, J. P., SHAIKH, A. M. and ZHANG, Y. (2024a). Covariate adjustment in experiments with matched pairs. *Journal of Econometrics*, **241** 105740.

BAI, Y., LIU, J., SHAIKH, A. M. and TABORD-MEEHAN, M. (2024b). Inference in cluster randomized trials with matched pairs. *Journal of Econometrics*, **245** 105873.

BAI, Y., LIU, J. and TABORD-MEEHAN, M. (2024c). Inference for Matched Tuples and Fully Blocked Factorial Designs. *Quantitative Economics*, **15** 279–330.

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. (2018). Inference Under Covariate-Adaptive Randomization. *Journal of the American Statistical Association*, **113** 1784–1796.

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

CAI, Y. and RAFI, A. (2022). On the Performance of the Neyman Allocation with Small Pilots. ArXiv:2206.04643 [econ], URL http://arxiv.org/abs/2206.04643.

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

CATTANEO, M. D. (2010). Efficient semiparametric estimation of multi-valued treatment effects under ignorability. *Journal of Econometrics*, **155** 138–154.

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.

CHERNOZHUKOV, V., CINELLI, C., NEWEY, W., SHARMA, A. and SYRGKANIS, V. (2024). Long Story Short: Omitted Variable Bias in Causal Machine Learning. ArXiv:2112.13398 [cs, econ, stat], URL http://arxiv.org/abs/2112.13398.

CYTRYNBAUM, M. (2023a). Covariate Adjustment in Stratified Experiments. ArXiv:2302.03687 [econ, stat], URL http://arxiv.org/abs/2302.03687.

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

CYTRYNBAUM, M. (2024). Finely stratified rerandomization designs. *arXiv preprint arXiv:2407.03279*.

EFRON, B. (1971). Forcing a sequential experiment to be balanced. *Biometrika*, **58** 403–417.

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.

FOGARTY, C. B. (2018a). On mitigating the analytical limitations of finely stratified experiments. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **80** 1035–1056.

FOGARTY, C. B. (2018b). Regression-assisted inference for the average treatment effect in paired experiments. *Biometrika*, **105** 994–1000.

FREEDMAN, D. A. (2008). On regression adjustments to experimental data. *Advances in Applied Mathematics*, **40** 180–193.

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

HAN, Q. and WELLNER, J. A. (2021). Complex sampling designs: Uniform limit theorems and applications. *The Annals of Statistics*, **49** 459–485.

HANNEKE, S., KONTOROVICH, A., SABATO, S. and WEISS, R. (2021). Universal Bayes consistency in metric spaces. *The Annals of Statistics*, **49** 2129–2150.

HIRANO, K. and IMBENS, G. W. (2001). Estimation of Causal Effects using Propensity Score Weighting: An Application to Data on Right Heart Catheterization. *Health Services and Outcomes Research Methodology*, **2** 259–278.

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

HU, Y. and HU, F. (2012). Asymptotic properties of covariate-adaptive randomization. *The Annals of Statistics*, **40** 1794–1815.

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

IMAI, K., KING, G. and NALL, C. (2009). The Essential Role of Pair Matching in Cluster-Randomized Experiments, with Application to the Mexican Universal Health Insurance Evaluation. *Statistical Science*, **24** 29–53.

IMBENS, G. W. (2011). Experimental Design for Unit and Cluster Randomized Trials. Tech. rep.

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.

Le Gall, J.-F. (2022). *Measure theory, probability, and stochastic processes*. Springer.

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

Li, X. and Ding, P. (2017). General Forms of Finite Population Central Limit Theorems with Applications to Causal Inference. *Journal of the American Statistical Association*, **112** 1759–1769.

Li, X. and Ding, P. (2020). Rerandomization and Regression Adjustment. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, **82** 241–268.

Li, X., Ding, P. and Rubin, D. B. (2018). Asymptotic theory of rerandomization in treatment–control experiments. *Proceedings of the National Academy of Sciences*, **115** 9157–9162.

Li, X., Ding, P. and Rubin, D. B. (2020). Rerandomization in $2^{K}$ factorial experiments. *The Annals of Statistics*, **48** 43–63.

Lin, W. (2013). Agnostic notes on regression adjustments to experimental data: Reexamining Freedman's critique. *Annals of Applied Statistics*, **7** 295–318.

Liu, H. and Yang, Y. (2020). Regression-adjusted average treatment effect estimates in stratified randomized experiments. *Biometrika*, **107** 935–948.

Ma, W., Li, P., Zhang, L.-X. and Hu, F. (2024). A New and Unified Family of Covariate Adaptive Randomization Procedures and Their Properties. *Journal of the American Statistical Association*, **119** 151–162.

Ma, W., Qin, Y., Li, Y. and Hu, F. (2020). Statistical Inference for Covariate-Adaptive Randomization Procedures. *Journal of the American Statistical Association*, **115** 1488–1497.

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.

Pashley, N. E. and Miratrix, L. W. (2021). Insights on variance estimation for blocked and matched pairs designs. *Journal of Educational and Behavioral Statistics*, **46** 271–296.

Pocock, S. J. and Simon, R. (1975). Sequential treatment assignment with balancing for prognostic factors in the controlled clinical trial. *Biometrics* 103–115.

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.

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. 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.

WEI, L.-J. (1978). The adaptive biased coin design for sequential experiments. *The Annals of Statistics*, **6** 92–100.

YE, T., YI, Y. and SHAO, J. (2022). Inference on the average treatment effect under minimization and other covariate-adaptive randomization methods. *Biometrika*, **109** 33–47.

ZELEN, M. (1974). The randomization and stratification of patients to clinical trials. *Journal of Chronic Diseases*, **27** 365–375.

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