# Finely Stratified Rerandomization Designs

Max Cytrynbaum*

January 9, 2025

**Abstract**

We study estimation and inference on causal parameters under finely stratified rerandomization designs, which use baseline covariates to match units into groups (e.g. matched pairs), then rerandomize within-group treatment assignments until a balance criterion is satisfied. We show that finely stratified rerandomization does partially linear regression adjustment "by design," providing nonparametric control over the stratified covariates and linear control over the rerandomized covariates. We introduce several new forms of rerandomization, allowing for imbalance metrics based on nonlinear estimators, and proposing a minimax scheme that minimizes the computational cost of rerandomization subject to a bound on estimation error. While the asymptotic distribution of GMM estimators under stratified rerandomization is generically non-normal, we show how to restore asymptotic normality using ex-post linear adjustment tailored to the stratification. We derive new variance bounds that enable conservative inference on finite population causal parameters, and provide asymptotically exact inference on their superpopulation counterparts.

*Yale Department of Economics. Correspondence: max.cytrynbaum@yale.edu

# 1 Introduction

Stratified randomization is commonly used to increase statistical precision in experimental research.[1] Recent theoretical work (e.g. Bai et al. (2021)) has shown that fine stratification, which randomizes treatment assignments within small groups of tightly matched units, makes unadjusted estimators like difference of means automatically semiparametrically efficient.[2] In finite samples, however, the performance of such designs can deteriorate rapidly with the dimension of the covariates used for stratification, due to a curse of dimensionality in matching.[3] This motivates the search for alternative designs that insist upon nonparametric balance for a few important covariates, but only attempt to balance linear functions of the remaining variables. In this paper, we study finely stratified rerandomization designs, which first tightly match the units into groups using a small set of important covariates, then rerandomize within-group treatment assignments until a balance criterion on the remaining covariates is satisfied.

Our first contribution is to derive the asymptotic distribution of generalized method of moments (GMM) estimators under stratified rerandomization, allowing for estimation of generic causal parameters defined by moment equalities. We consider both superpopulation and finite population parameters, the latter of which may be more appropriate for experiments run in a convenience sample (Abadie et al. (2014)), as is the case for the vast majority of experiments in economics (Niehaus and Muralidharan (2017)). We introduce novel finite population parameters, studying a finite population local average treatment effect heterogeneity parameter in an application to Angrist et al. (2013). As in previous work on rerandomization (e.g. Li et al. (2018)), the asymptotic distribution of GMM estimators is an independent sum of a normal and a truncated normal term. We show that, modulo this truncated term, unadjusted GMM under stratified rerandomization behaves like semiparametrically adjusted GMM (e.g. Graham (2011)) under an iid design, with automatic nonparametric control over the stratification covariates and linear control over the rerandomization covariates. Intuitively, stratified rerandomization implements partially linear regression adjustment "by design."

Our second contribution is to introduce novel forms of rerandomization based on nonlinear balance criteria. For example, we allow acceptance or rejection of

[1] For example, Cytrynbaum (2024a) reports a survey of 50 experimental papers in the AER and AEJ from 2018-2023, where 57% used some form of stratified randomization.

[2] See Cytrynbaum (2024b), Armstrong (2022), and Bai et al. (2024b) for more discussion.

[3] Under regularity conditions, the convergence rate of finite sample variance to asymptotic variance is $O(n^{-2/(d+1)})$ for dimension $d$ covariates, see Cytrynbaum (2024b).

an allocation based on the difference of estimated covariate densities between treatment and control units. We also study a design that rerandomizes until an estimated propensity score is approximately constant, forcing the covariates to have no predictive power for treatment assignments in our realized sample. We prove that the designs in a general family of nonlinear rerandomization methods are all asymptotically equivalent to standard rerandomization based on a difference of covariate means, with an implicit choice of covariates and balance criterion, which we characterize.

Our third contribution is to study optimization of the balance criterion itself. We propose a novel minimax scheme that allows the researcher to specify prior information about the relationship between covariates and outcomes, then rerandomizes until the worst case covariate imbalance consistent with this prior is small. We prove that this design minimizes the (asymptotic) computational cost of rerandomization, subject to a strict bound on estimation error over the set of models consistent with the prior. If our prior information set contains the truth, then this design bounds the asymptotic variance of stratified rerandomization within a small additive factor of the optimal semiparametrically adjusted variance. If the information set is instead a Wald region estimated from pilot data, we show that our minimax design bounds the asymptotic variance in the main experiment with high probability.

Our fourth contribution is to provide simple inference methods for generic causal parameters under stratified rerandomization designs. To do this, we first derive the optimal ex-post linear adjustment for GMM estimation, which depends on the stratification.[4] Optimal adjustment makes the asymptotic distribution of GMM insensitive to the rerandomization acceptance criterion, removing the truncated normal term from the limiting distribution and restoring asymptotic normality. We also show that combining rerandomization with ex-post linear adjustment provides a form of double robustness to covariate imbalances, which helps explain the strong performance of this method in our simulations. For finite population causal parameters, the asymptotic variance is generically not identified (Neyman (1990)). We derive novel identified variance bounds for general finite population causal parameters, enabling asymptotically conservative inference that still exploits the efficiency gains from both stratified rerandomization and optimal adjustment. For superpopulation parameters, we present new inference methods

[4]This extends recent work on optimal adjustment under pure stratified randomization for ATE estimation, e.g. see Cytrynbaum (2024a), Bai et al. (2024a), or Liu and Yang (2020).

that are asymptotically exact.

Finally, we provide simulations and an empirical application to estimating treatment effect heterogeneity among compliers in Angrist et al. (2013), which both show the value of adding a rerandomization step to finely stratified designs. This effect can be seen clearly in Figure 3, which compares stratified rerandomization to stratification plus ex-post adjustment, and Figure 4, which compares stratified rerandomization to other designs like pure fine stratification. See the relevant sections below for more detailed discussion.

## 1.1 Related Literature

This paper builds on the literature on fine stratification in econometrics as well as the literature on rerandomization in statistics. Stratified randomization has a long history in statistics, see Cochran (1977) for a survey. Recent work on fine stratification in econometrics includes Bai et al. (2021), Bai (2022), Cytrynbaum (2024b), Armstrong (2022), and Bai et al. (2024b). A sample of some recent work in the statistics literature on rerandomization includes Morgan and Rubin (2012) and Li et al. (2018), Wang et al. (2021), and Wang and Li (2022). We build on both of these literatures, studying the consequence of rerandomizing treatments within data-adaptive fine strata. We show that finely stratified rerandomization does semiparametric (partially linear) regression adjustment "by design," providing nonparametric control over a few important variables and linear control over the rest.

For our main asymptotic theory (Section 3), the most closely related previous work is Wang et al. (2021) and Bai et al. (2024b). Wang et al. (2021) study estimation of the sample average treatment effect (SATE) under stratified rerandomization, with quadratic imbalance metrics based on the Mahalanobis norm. We study rerandomization within data-adaptive fine strata, providing asymptotic theory for generic superpopulation and finite population causal parameters defined by moment equalities. We also allow for essentially arbitrary rerandomization acceptance criteria, not necessarily based on quadratic forms. Bai et al. (2024b) study estimation of superpopulation parameters defined by moment equalities under pure stratified randomization, without rerandomization. We extend these results to stratified rerandomization as well as generic finite population parameters, providing "SATE-like" versions of the parameters in Bai et al. (2024b).[5] In concurrent

[5]These parameters can be seen as causal versions of the conditional estimand defined in Abadie et al. (2014).

work, Wang and Li (2024a) study GMM estimation of univariate superpopulation parameters under stratified rerandomization with fixed, discrete strata. We study significantly more general forms of stratification and rerandomization criteria than considered in their work, allowing for both finite and superpopulation parameters of arbitrary dimension and fine stratification with continuous covariates.

For nonlinear rerandomization (Section 4), the closest related results are Ding and Zhao (2024) and Li et al. (2021). Ding and Zhao (2024) rerandomize based on the p-value of a logistic regression coefficient, while we rerandomize until a general smooth propensity estimate is close to constant in $L_2$ norm. Li et al. (2021) simulate a density based rerandomization design, but provide limited theoretical results. To the best of our knowledge, we present the first asymptotic theory for rerandomization based on the difference of nonlinear (e.g. density) estimates. For acceptance region optimization (Section 5), the closest related results are Schindl and Branson (2024), who study the optimal choice of norm for quadratic rerandomization, while Liu et al. (2023) chooses an optimal quadratic rerandomization design using a Bayesian criterion, in both cases for rerandomization without stratification. We propose a novel minimax approach that accepts or rejects based on the value of a convex penalty function, tailored to prior information provided by the researcher or estimated from a pilot.

Our work on optimal adjustment (Section 6) extends recent work on adjustment for stratified designs, e.g. Liu and Yang (2020), Cytrynbaum (2024a), Bai et al. (2024a), to stratified rerandomization and GMM parameters. Finally our work on inference under data-adaptive fine stratification (Section 7) builds on previous work by Abadie and Imbens (2008), Bai et al. (2021), and Cytrynbaum (2024b). Other recent work that has considered variance bounds for finite population causal parameters includes Aronow et al. (2014), Fogarty (2018), Ding et al. (2019), Abadie et al. (2020), and Xu (2021).

## 2 Framework and Designs

Consider data $W_i = (R_i, S_i(1), S_i(0))$ with $(W_i)_{i=1}^n \overset{\text{iid}}{\sim} F$. The $S_i(d) \in \mathbb{R}^{d_S}$ denote potential outcome vectors for a binary treatment $d \in \{0, 1\}$, while $R_i$ denote other pre-treatment variables, such as covariates. For treatment assignments $D_i \in \{0, 1\}$, the realized outcome $S_i = S_i(D_i) = D_i S_i(1) + (1 - D_i) S_i(0)$. In what follows, for any array $(a_i)_{i=1}^n$ we denote $E_n[a_i] = n^{-1} \sum_{i=1}^n a_i$, with $\bar{a}_1 = E_n[a_i | D_i = 1] = E_n[a_i D_i]/E_n[D_i]$ and similarly $\bar{a}_0 = E_n[a_i | D_i = 0]$. Next, we define stratified

rerandomization designs.

**Definition 2.1** (Stratified Rerandomization)**.** Let treatment proportions $p = l/k$ and suppose that $n$ is divisible by $k$ for notational simplicity.

(1) (Stratification). Partition the experimental units into $n/k$ disjoint groups (strata) $s$ with $\{1, \ldots, n\} = \bigcup_s s$ disjointly and $|s| = k$. Let $\psi = \psi(R)$ with $\psi \in \mathbb{R}^{d_\psi}$ denote a vector of stratification variables, which may be continuous or discrete. Suppose the groups satisfy the matching condition[6]

$$\frac{1}{n} \sum_s \sum_{i,j \in s} |\psi_i - \psi_j|_2^2 = o_p(1). \tag{2.1}$$

Require that the groups only depend on the stratification variables $\psi_{1:n}$ and data-independent randomness $\pi_n$, so that $s = s(\psi_{1:n}, \pi_n)$ for each $s$.

(2) (Randomization). Independently for each $|s| = k$, draw treatment variables $(D_i)_{i \in s}$ by setting $D_i = 1$ for exactly $l$ out of $k$ units, uniformly at random.

(3) (Check Balance). For rerandomization covariates $h = h(R)$, consider an imbalance metric $\mathcal{I}_n = \sqrt{n}(\bar{h}_1 - \bar{h}_0) + o_p(1)$.[7] For an acceptance region $A \subseteq \mathbb{R}^{d_h}$, check if the balance criterion $\mathcal{I}_n \in A$ is satisfied. If so, accept $D_{1:n}$. If not, repeat from the beginning of (2).

Intuitively, steps (1) and (2) describe a data-driven "matched k-tuples" design, while step (3) rerandomizes within k-tuples until the balance criterion is satisfied. Equation 2.1 is a tight-matching condition, requiring that the groups are clustered locally in $\psi$ space. Cytrynbaum (2024b) provides algorithms to match units into groups that satisfy this condition for any fixed $k$.

**Example 2.2** (Pure Stratification)**.** Stratification without rerandomization can be obtained by setting $A = \mathbb{R}^{d_h}$ in Definition 2.1. Treatment effect estimation under such designs was studied in Bai (2022), Cytrynbaum (2024b), and Bai et al. (2024b). Definition 2.1 allows for fine stratification (also known as matched k-tuples), with the number of data-dependent groups $s = s(\psi_{1:n}, \pi_n)$ growing with $n$. It also allows for coarse stratification with strata $x \in \{1, \ldots, m\}$ and fixed $m$, studied e.g. in Bugni et al. (2018). This can be obtained in our framework by setting $\psi = x$ and matching units into groups $s$ at random within each $\{i : x_i = k\}$.

[6]The matching condition in Equation 2.1 was introduced by Bai et al. (2021) for matched pairs randomization ($k = 2$). See Bai (2022) and Cytrynbaum (2024b) for generalizations.

[7]In particular, we require that $\mathcal{I}_n = \sqrt{n}(\bar{h}_1 - \bar{h}_0) + o_p(1)$ under the law induced by "pure" stratified randomization, the design in steps (1) and (2) only, studied e.g. in Cytrynbaum (2024b).

**Example 2.3** (Complete Randomization)**.** For $p = l/k$, we say that $D_{1:n}$ are completely randomized with probability $p$ if $P(D_{1:n} = d_{1:n}) = 1/\binom{n}{np}$ for all $d_{1:n}$ with $\sum_i d_i = np$.[8] Equivalently, complete randomization is coarse stratification with $m = 1$ above. This can be obtained by setting $\psi = 1$ and $A = \mathbb{R}^{d_h}$ in Definition 2.1, matching units into groups at random.

Next, we discuss a convenient rerandomization scheme that allows the researcher to select the approximate number of draws until acceptance.

**Example 2.4** (Mahalanobis Rerandomization)**.** Consider matched k-tuples rerandomization as in Equation 2.1. Define within-tuple demeaned covariates $\check{X}_i = X_i - k^{-1}\sum_{j\in s(i)} X_j$ and set $\Sigma_n = \mathrm{Var}(D)^{-1}\frac{k}{k-1}E_n[\check{X}_i\check{X}_i']$. Consider rerandomizing until

$$n(\bar{X}_1 - \bar{X}_0)'\Sigma_n^{-1}(\bar{X}_1 - \bar{X}_0) \leq \epsilon^2. \tag{2.2}$$

This scheme was studied e.g. in Wang et al. (2021) for the case without data-adaptive strata. Equation 2.2 is equivalent to $\mathcal{I}_n \in A$ for $\mathcal{I}_n = \Sigma_n^{-1/2}\sqrt{n}(\bar{X}_1 - \bar{X}_0)$ and $A = \{x : |x|_2 \leq \epsilon\}$. Work in Cytrynbaum (2024a) implies that under matched k-tuples randomization, $\Sigma_n \xrightarrow{p} \Sigma = \mathrm{Var}(D)^{-1}E[\mathrm{Var}(X|\psi)]$, so $\mathcal{I}_n = \sqrt{n}(\bar{h}_1 - \bar{h}_0) + o_p(1)$ for $h = \Sigma^{-1/2}X$. Then this design satisfies Definition 2.1. One can show that $n(\bar{X}_1 - \bar{X}_0)'\Sigma_n^{-1}(\bar{X}_1 - \bar{X}_0) \Rightarrow \chi^2_r$ for $r = \dim(X)$ under pure stratification.[9] If $\epsilon(\alpha)$ is chosen as the $\alpha$ quantile of $\chi^2_r$, $P(\chi^2_r \leq \epsilon(\alpha)^2) = \alpha$, then $P(\mathcal{I}_n \in A) = \alpha + o(1)$. Setting $\alpha = 1/m$, gives approximately $m$ expected rerandomizations until acceptance for large enough $n$.

Mahalanobis rerandomization uses a convenient choice of acceptance criterion, but the variance normalization and implicit acceptance region in Equation 2.4 are not generally efficient for estimating causal parameters. We provide alternative designs that optimize the shape of acceptance region $A$ in Section 5 below.

**Causal Estimands.** Next, we introduce a generic family of causal estimands defined by moment equalities. Let $g(D, R, S, \theta) \in \mathbb{R}^{d_g}$ be a score function for generalized method of moments (GMM) estimation. Recall $W = (R, S(1), S(0))$ and for $D|W \sim \text{Bernoulli}(p)$ define $\phi(W, \theta) = E[g(D, R, S, \theta)|W] = pg(1, R, S(1), \theta) + (1-p)g(0, R, S(0), \theta)$. By construction, we have $E[\phi(W, \theta)] = 0 \iff E[g(D, R, S, \theta)] = 0$. The function $\phi(W, \theta)$ provides a convenient parameterization to define our paired finite population and superpopulation causal estimands.

[8] For notational simplicity, we may assume that $n = lk$ for some $l \in \mathbb{N}$.

[9] For instance, this follows from Lemma A.8 in Cytrynbaum (2024a) and Corollary 3.6 below.

**Definition 2.5** (Causal Estimands)**.** The *superpopulation* estimand $\theta_0$ is the unique solution to $E[\phi(W,\theta)] = 0$. The *finite population* estimand $\theta_n$ is the unique solution to $E_n[\phi(W_i,\theta)] = 0$.

In what follows, we study GMM estimation of both $\theta_0$ and $\theta_n$ under stratified rerandomization designs, showing an asymptotic equivalence between stratified rerandomization and partially linear covariate adjustment. In particular, this framework allows us to introduce several useful finite population estimands $\theta_n$ that do not appear to have been previously considered in the literature, such as Example 2.7 below. The estimand $\theta_n$ may be a more appropriate target for experiments run in a convenience sample, as is the case for the vast majority of experiments reported in the economics literature (Niehaus and Muralidharan (2017)). Inference on $\theta_n$, provided in Section 7, is generically more powerful than for $\theta_0$, since we only have to account for estimation uncertainty due to random assignment, without extra variability from sampling into the experiment. Note GMM estimation of $\theta_0$ under pure stratification was studied in Bai et al. (2024b).

**Example 2.6** (ATE and SATE)**.** Define the Horvitz-Thompson weights $H = \frac{D-p}{p-p^2}$ and let $g(D,Y,\theta) = HY - \theta$, so that $\phi(W,\theta) = E[HY|W] - \theta = Y(1) - Y(0) - \theta$. Then $\theta_0 = E[Y(1) - Y(0)] = \text{ATE}$, the average treatment effect, and $\theta_n = E_n[Y_i(1) - Y_i(0)] = \text{SATE}$, the sample average treatment effect.

Consider a setting where the researcher wants to estimate a parametric model of treatment effect heterogeneity in an experiment with noncompliance and randomized binary instrument $Z$. In the next example, we define finite population and superpopulation local average treatment effect (LATE) *heterogeneity* parameters, applying them in our empirical application to Angrist et al. (2013) below.

**Example 2.7** (LATE Heterogeneity)**.** Let $D(z)$ be potential treatments for a binary instrument $z \in \{0,1\}$. Let $Y(d)$ be the potential outcomes, with realized outcome $Y = Y(D(Z))$. Suppose $D(1) \geq D(0)$, and define compliance indicator $C = \mathbb{1}(D(1) > D(0))$, assuming $E[C] > 0$. Imbens and Angrist (1994) define the $\text{LATE} = E[Y(1) - Y(0)|C = 1]$. Let $H = (Z-p)/(p-p^2)$ and consider the score function $g(Z,D,Y,X,\theta) = (HY - HD \cdot f(X,\theta))\nabla_\theta f(X,\theta)$. A calculation shows

$$\phi(W,\theta) = E[g(Z,D,Y,X,\theta)|W] = C \cdot (Y(1) - Y(0) - f(X,\theta))\nabla_\theta f(X,\theta).$$

The moment condition $E[\phi(W,\theta)] = 0$ is the FOC of a treatment effect prediction problem in the complier population $C = 1$. In particular, for $\tau \equiv Y(1) - Y(0)$, the

parameter $\theta_0$ is the best parametric predictor $\theta_0 = \text{argmin}_\theta E[(\tau - f(X,\theta))^2|C=1]$ of treatment effects for compliers.[10] Specializing to $f(X,\theta) = X'\theta$, this is the best linear predictor (BLP) of treatment effect heterogeneity among the compliers $\theta_0 = \text{argmin}_\theta E[(\tau - X'\theta)^2|C=1]$, while $f(X,\theta) = \theta$ recovers the LATE $= E[\tau|C=1]$. Setting $E_n[\phi(W_i,\theta)] = 0$, the corresponding finite population parameter is

$$\theta_n = \operatorname*{argmin}_\theta E_n[(\tau_i - f(X_i,\theta))^2|C_i=1]. \tag{2.3}$$

We can also specialize to $f(X,\theta) = X'\theta$ for a finite population version of the BLP of LATE. The finite population LATE was studied in Ren (2023) under complete randomization, but the more general heterogeneity parameters here appear to be novel. We consider both $\theta_0$ and $\theta_n$ when studying treatment effect heterogeneity among compliers in the empirical application to Angrist et al. (2013) in Section 9.

**GMM Estimation.** For positive-definite weighting matrix $M_n \in \mathbb{R}^{d_g \times d_g}$ with $M_n \xrightarrow{p} M \succ 0$ and sample moment $\widehat{g}(\theta) \equiv E_n[g(D_i,R_i,S_i,\theta)]$, the GMM estimator[11] is

$$\widehat{\theta} = \operatorname*{argmin}_{\theta\in\Theta} \widehat{g}(\theta)' M_n' \widehat{g}(\theta). \tag{2.4}$$

We are mostly interested in the exactly identified case, where $\widehat{\theta}$ solves $\widehat{g}(\widehat{\theta}) = 0$. In what follows, we study the properties of generalized method of moments (GMM) estimation of the causal parameters $\theta_0$ and $\theta_n$ under stratified rerandomization.

# 3 Asymptotics for GMM Estimation

In this section, we characterize the asymptotic distribution of the GMM estimator $\widehat{\theta}$ under the stratified rerandomization designs in Definition 2.1. We show that the asymptotic variance of $\widehat{\theta}$ is proportional to the residuals of a partially linear regression model, up to a remainder term due to slackness in the rerandomization criterion. In this sense, stratified rerandomization does partially linear regression adjustment "by design." First, we state some technical conditions that are needed for the following results.

**Assumption 3.1** (Acceptance Region)**.** *Suppose $A \subseteq \mathbb{R}^{d_h}$ has non-empty interior and* $\text{Leb}(\partial A) = 0$,[12] *and require $E[\text{Var}(h|\psi)] \succ 0$ and $E[|\psi|_2^2 + |h|_2^2] < \infty$.*

[10] For example, if $Y$ is binary then $Y(1) - Y(0) \in \{-1,0,1\}$, so the link function model $f(X,\theta) = 2L(X'\theta) - 1$ for $L = \text{Logit}$ may be appropriate.

[11] In our examples, we will mainly be concerned with the exactly identified case. However, the theory for the over identified case is almost identical, so we include this as well.

[12] Note that $\partial A$ denotes the boundary of $A$, the limit points of both $A$ and $A^c$.

Next we state the technical conditions needed for GMM estimation. Define the matrix $G = E[(\partial/\partial\theta')\phi(W,\theta)]|_{\theta=\theta_0} \in \mathbb{R}^{d_g \times d_\theta}$ and let $g_d(W,\theta) = g(d,R,S(d),\theta)$ for $d \in \{0,1\}$. Recall the Frobenius norm $|B|_F^2 = \sum_{ij} B_{ij}^2$ for any matrix $B$.

**Assumption 3.2** (GMM)**.** *The following conditions hold for $d \in \{0,1\}$:*

*(a) (Identification). The matrix $G$ is full rank, and $g_0(\theta) = 0$ iff $\theta = \theta_0$.*

*(b) We have $E[g_d(W,\theta_0)^2] < \infty$ and $E[\sup_{\theta\in\Theta} |g_d(W,\theta)|_2] < \infty$. Also $\theta \to g_d(W,\theta)$ is continuous almost surely, and $\Theta$ is compact.*[13]

*(c) There exists a neighborhood $\theta_0 \in U \subseteq \Theta$ such that $G_d(W,\theta) \equiv \partial/\partial\theta' g_d(W,\theta)$ exists and is continuous. Also $E[\sup_{\theta\in U} |\partial/\partial\theta' g_d(W,\theta)|_F] < \infty$.*

Compactness could likely be relaxed using concavity assumptions or a VC class condition, but we do not pursue this here. In what follows it will be conceptually useful to reparameterize the score function.

**Sampling and Assignment Expansion.** Recall $\phi(W,\theta) = E[g(D,R,S,\theta)|W]$ for $W = (R, S(1), S(0))$. Define $a(W,\theta) \equiv \text{Var}(D)(g_1(W,\theta) - g_0(W,\theta))$, which we refer to as the "assignment function." For Horvitz-Thompson weights $H = (D-p)/(p-p^2)$, a calculation shows we can expand

$$g(D,R,S,\theta) = \phi(W,\theta) + Ha(W,\theta). \tag{3.1}$$

Our results below show that $a(W,\theta)$ parameterizes estimator variance due to random *assignment*, while $\phi(W,\theta)$ parameterizes the variance due to random *sampling* for the superpopulation estimand $\theta_0$.

**Example 3.3** (ATE and SATE)**.** Continuing Example 2.6 above, define $\bar{Y} = (1-p)Y(1) + pY(0)$. This is a convex combination that summarizes both potential outcomes, which we view as the unit's "outcome level." Then for the score $g(D,Y,\theta) = HY - \theta$, we have $a(W,\theta) = \text{Var}(D)(Y(1)/p - (-Y(0)/(1-p))) = \bar{Y}$. Another simple calculation[14] shows that for difference of means $\widehat{\theta} = E_n[H_iY_i]$ and estimands $\theta_n = \text{SATE}$, $\theta_0 = \text{ATE}$

$$\begin{aligned}\widehat{\theta} - \theta_0 &= (\widehat{\theta} - \theta_n) + (\theta_n - \theta_0) = E_n[H_i a(W_i)] + E_n[\phi(W_i,\theta_0)] \\ &= (E_n[\bar{Y}_i|D_i=1] - E_n[\bar{Y}_i|D_i=0]) + (E_n[Y_i(1) - Y_i(0)] - \theta_0)\end{aligned}$$

[13] We can formally resolve measurability issues with the sup expressions by either (1) explicitly working with outer probability (e.g. van der Vaart and Wellner (1996)) or (2) requiring that $\{g_d(\cdot,\theta), \theta \in \Theta\}$ is universally separable for $d = 0, 1$ (Pollard (1984), p.38). To focus on the practical design issues, we avoid this formalism, implicitly assuming that all quantities are appropriately measurable.

[14] For stratified designs $E_n[D_i] = p$, so $E_n[H_iY_i] = \bar{Y}_1 - \bar{Y}_0$. This is not true for iid designs.

The assignment term $E_n[H_i a(W_i)]$ from Equation 3.1 isolates the estimation error due to chance imbalances in the outcome levels $\bar{Y}_i$ between treatment and control during random assignment. By contrast, the term $E_n[\phi(W_i, \theta_0)]$ for sampling function $\phi(W, \theta) = Y(1) - Y(0) - \theta$ isolates the estimation error due to random sampling of heterogeneous units.

## 3.1 Finite Population Estimand

Our first theorem studies GMM estimation of the finite population estimand $\theta_n$, which solves $E_n[\phi(W_i, \theta_n)] = 0$. We extend these results to $\theta_0$ in Corollary 3.7 below. To state the theorem, define the GMM linearization matrix $\Pi = -(G'MG)^{-1}G'M \in \mathbb{R}^{d_\theta \times d_g}$. Note that in the exactly identified case $d_g = d_\theta$, we just have $\Pi = -G^{-1}$. We also denote the constant $v_D = \text{Var}(D) = p - p^2$.

Before stating the main result, we first derive the influence function for GMM estimation of $\theta_n$ under stratified rerandomization.

**Lemma 3.4** (Linearization)**.** *Suppose $D_{1:n}$ as in Definition 2.1 and require Assumption 3.1, 3.2. Then $\sqrt{n}(\widehat{\theta} - \theta_n) = \sqrt{n}E_n[H_i \Pi a(W_i, \theta_0)] + o_p(1)$.*

Lemma 3.4 generalizes Example 3.3 above, showing that

$$\widehat{\theta} - \theta_n = \Pi \left( E_n[a(W_i, \theta_0) | D_i = 1] - E_n[a(W_i, \theta_0) | D_i = 0] \right) + o_p(n^{-1/2}).$$

This implies that the errors in estimating any finite population GMM parameter $\theta_n$ are driven by random imbalances in the assignment function $a(W_i, \theta_0)$ between treatment and control units, at least to first-order. Our main theorem shows that, by balancing $\psi$ and $h$ ex-ante, stratified rerandomization reduces these imbalances, improving precision.

**Theorem 3.5** (GMM)**.** *Suppose $D_{1:n}$ as in Definition 2.1. Require Assumption 3.1, 3.2. Then $\sqrt{n}(\widehat{\theta} - \theta_n) | W_{1:n} \Rightarrow \mathcal{N}(0, V_a) + R_A$, independent RV's with*

$$V_a = \min_{\gamma \in \mathbb{R}^{d_h \times d_\theta}} v_D^{-1} E[\text{Var}(\Pi a(W, \theta_0) - \gamma' h | \psi)]. \tag{3.2}$$

*Let $\gamma_0$ be optimal in Equation 3.2. The term $R_A$ is a truncated Gaussian vector*

$$R_A \sim \gamma_0' Z_h \,|\, Z_h \in A, \qquad Z_h \sim \mathcal{N}(0, v_D^{-1} E[\text{Var}(h | \psi)]). \tag{3.3}$$

Note the variance $V_a$ is a matrix $V_a \in \mathbb{R}^{d_\theta \times d_\theta}$, so the minimum should be interpreted in the positive semidefinite sense, $V(\gamma_0) = \min_\gamma V(\gamma)$ if $V(\gamma_0) \preceq V(\gamma)$ for

all $\gamma \in \mathbb{R}^{d_h \times d_\theta}$. Theorem 3.5 shows that $\sqrt{n}(\widehat{\theta} - \theta_n)$ is asymptotically distributed as an independent sum of a normal $\mathcal{N}(0, V_a)$ and truncated normal vector $R_A$. Both terms only depend on the "assignment" influence function component $\Pi a(W, \theta_0)$.

**Partially Linear Adjustment.** Let $\mathcal{L}(\psi) = L_2^{d_\theta}(\psi)$ be the $d_\theta$-fold Cartesian product of $L_2(\psi)$, the space of square-integrable functions. Then the variance $V_a$ in Theorem 3.5 is can be written in terms of the residuals of the assignment function $\Pi a(W, \theta_0)$ in a partially linear regression on $\psi$ and $h$:

$$V_a = \min_{\substack{\gamma \in \mathbb{R}^{d_h \times d_\theta} \\ t \in \mathcal{L}(\psi)}} v_D^{-1} \operatorname{Var}\left(\Pi a(W, \theta_0) - \gamma' h - t(\psi)\right). \tag{3.4}$$

This shows stratified rerandomization does partially linear regression adjustment "by design," providing nonparametric control over $\psi$ and linear control over $h$.

**Residual Imbalance.** The truncated Gaussian $R_A \sim \gamma_0' Z_h \,|\, Z_h \in A$ arises from residual covariate imbalances due to slackness in the acceptance criterion, since $A \neq \{0\}$. If $A$ is symmetric about zero, i.e. $x \in A$ iff $-x \in A$, then $E[R_A] = 0$, so the GMM estimator $\widehat{\theta}$ is first-order unbiased, as usual. In principle, $R_A$ can be made negligible relative to $\mathcal{N}(0, V_a)$ in large enough samples by choosing very small $A$. For example, if $A = B(0, \epsilon)$ then $R_{B(0,\epsilon)} \sim \{\gamma_0' Z_h \,|\, |Z_h|_2 \leq \epsilon\} \xrightarrow{p} 0$ as $\epsilon \to 0$. However, in finite samples this may be computationally infeasible and could even invalidate our first-order asymptotic approximation.[15] We develop a minimax criterion to choose an efficient region $A$ in Section 5 below.

To isolate the precision gains due to rerandomization, the following corollary specializes Theorem 3.5 to the case of stratification without rerandomization ($A = \mathbb{R}^{d_h}$), as well as complete randomization, defined in Examples 2.2 and 2.3.

**Corollary 3.6** (Pure Stratification)**.** *Suppose $D_{1:n}$ as in Definition 2.1 with $A = \mathbb{R}^{d_h}$. Require Assumption 3.1. Then $\sqrt{n}(\widehat{\theta} - \theta_n) | W_{1:n} \Rightarrow \mathcal{N}(0, V_a)$ with $V_a = v_D^{-1} E[\operatorname{Var}(\Pi a(W, \theta_0) | \psi)]$. In particular, if $D_{1:n}$ is completely randomized $\psi = 1$, then $V_a = v_D^{-1} \operatorname{Var}(\Pi a(W, \theta_0))$.*

Corollary 3.6 shows that fine stratification reduces the variance of GMM estimation of $\theta_n$ to $V_a = v_D^{-1} E[\operatorname{Var}(\Pi a(W, \theta_0) | \psi)] \leq v_D^{-1} \operatorname{Var}(\Pi a(W, \theta_0))$, a nonparametric improvement. Rerandomization as in Definition 2.1 provides a further linear variance reduction to $V_a = \min_{\gamma \in \mathbb{R}^{d_h \times d_\theta}} E[\operatorname{Var}(\Pi a(W, \theta_0) - \gamma' h | \psi)]$, up to the residual imbalance term $R_A$.

[15] See Wang and Li (2022) for a detailed analysis of complete rerandomization, where $\epsilon_n$ can change with sample size.

## 3.2 Superpopulation Estimand

This section extends the asymptotics above to the superpopulation estimand $\theta_0$ solving $E[\phi(W, \theta_0)] = 0$. We show that by targeting $\theta_0$ we incur additional sampling variance that is invariant to the distribution of treatment assignments $D_{1:n}$.

**Corollary 3.7** (Superpopulation Estimand)**.** *Suppose $D_{1:n}$ is as in Definition 2.1. Require Assumption 3.1, 3.2.*

*(a) We have $\sqrt{n}(\widehat{\theta} - \theta_0) \Rightarrow \mathcal{N}(0, V_\phi) + \mathcal{N}(0, V_a) + R_A$, independent RV's with $V_\phi = \text{Var}(\Pi\phi(W, \theta_0))$ and $V_a$, $R_A$ exactly as in Theorem 3.5.*

*(b) (Pure Stratification). If $A = \mathbb{R}^{d_h}$, this is $\sqrt{n}(\widehat{\theta} - \theta_0) \Rightarrow \mathcal{N}(0, V)$ with*

$$V = \text{Var}(\Pi\phi(W, \theta_0)) + v_D^{-1} E[\text{Var}(\Pi a(W, \theta_0)|\psi)].$$

The corollary shows that targeting $\theta_0$ instead of $\theta_n$ adds an extra independent $\mathcal{N}(0, V_\phi)$ term to the asymptotic distribution. The variance $V_\phi$ arises due to iid random sampling of the sampling function $\Pi\phi(W, \theta_0)$. Notice that stratified rerandomization only reduces the variance due to imbalances in the assignment function $\Pi a(W, \theta_0)$, while the variance due to sampling $\Pi\phi(W, \theta_0)$ is irreducible. In this sense, the statistical consequences of different designs and adjustment strategies all happen at the level of the finite population estimand $\theta_n$, while targeting the superpopulation $\theta_0$ just adds extra sampling noise. Note that for pure stratification, Bai et al. (2024b) were the first to derive an analogue of part (b) of Corollary 3.7, under different GMM regularity conditions than we use here.[16]

**Example 3.8** (SATE)**.** Theorem 3.5 and Corollary 3.7 show $\sqrt{n}(\widehat{\theta} - \text{SATE})|W_{1:n} \Rightarrow \mathcal{N}(0, V_a) + R_A$ and $\sqrt{n}(\widehat{\theta} - \text{ATE}) \Rightarrow \mathcal{N}(0, V_\phi + V_a) + R_A$ with

$$V_\phi = \text{Var}(Y(1) - Y(0)) \qquad V_a = \min_{\gamma \in \mathbb{R}^{d_h}} v_D^{-1} E[\text{Var}(\bar{Y} - \gamma' h|\psi)]. \tag{3.5}$$

The term $V_\phi$ reflects sampling variance due to treatment effect heterogeneity. The term $V_a$ is the variance due to random assignment, caused by random imbalances in outcome levels $\bar{Y}$ between $D_i = 1$ and $D_i = 0$. Balanced randomization and adjustment can be used to reduce $V_a$, while $V_\phi$ is an irreducible sampling variance.

**Remark 3.9.** Wang et al. (2021) study SATE estimation under stratified rerandomization in the sequence of finite populations framework. By contrast, here

[16] In particular, Bai et al. (2024b) allow for non-smooth GMM scores and impose a VC dimension condition on $g_d(W, \theta)$. We restrict to the smooth case, using compactness of $\Theta$ to avoid entropy conditions.

we allow for data-adaptive strata $s = s(\psi_{1:n}, \pi_n)$, endogenizing the process of fine stratification. Using the tight-matching condition 2.1, we are able to derive a simple closed form for the asymptotic variance, providing a novel connection between stratified rerandomization and partially linear regression adjustment.

**Example 3.10** (CATE)**.** Specializing Example 2.7, consider estimating the best linear predictor of treatment effect heterogeneity in an experiment with perfect compliance. We can use the slightly simpler score $g(D, X, Y, \theta) = (HY - X'\theta)X$. For $\tau = Y(1) - Y(0)$ we have $\phi(W, \theta) = (\tau - X'\theta)X$, so the parameters $\theta_n, \theta_0$ are

$$\theta_n = \operatorname*{argmin}_\theta E_n[(\tau_i - X_i'\theta)^2], \qquad \theta_0 = \operatorname*{argmin}_\theta E[(\tau - X'\theta)^2].$$

The parameter $\theta_n$ was studied in Ding et al. (2019) under complete randomization. A simple calculation shows that assignment function $a(W, \theta_0) = \bar{Y}X$ and $\Pi = E[XX']^{-1}$. Then for residual $e = \tau - X'\theta_0$, the variances in Corollary 3.7 are

$$V_\phi = \mathrm{Var}(\Pi e X), \qquad V_a = \min_{\gamma \in \mathbb{R}^{d_h \times d_x}} v_D^{-1} E[\mathrm{Var}(\Pi \bar{Y} X - \gamma' h | \psi)].$$

**Efficient Design.** The expression for $V_a$ shows that to precisely estimate heterogeneity parameters $\theta_n$ and $\theta_0$, it is important to include not only variables that predict outcome levels $\bar{Y}$ in $\psi$ and $h$, but also their interactions with the desired heterogeneity variable $X$. We consider such interacted designs in our simulations and empirical application to Angrist et al. (2013) below.

# 4 Nonlinear Rerandomization

In this section, we introduce several novel "nonlinear" rerandomization criteria, proving that in many cases such designs are first-order equivalent to linear rerandomization (Definition 2.1), with an implicit choice of covariates $h$ and acceptance region $A$. This shows that our asymptotics and inference methods apply to a larger class of asymptotically linear rerandomization schemes.

## 4.1 GMM Rerandomization

First, we generalize the imbalance metric $\mathcal{I}_n$ in Definition 2.1, allowing rejection of $D_{1:n}$ based on potentially nonlinear features of the in-sample distribution of treatments and covariates $(D_i, X_i)_{i=1}^n$. Let $m(X_i, \beta)$ be a GMM score function, separate from the score $g(D, R, S)$ defining the estimands above. We can define a

large class of designs by stratifying and rerandomizing until $\sqrt{n}(\widehat{\beta}_1 - \widehat{\beta}_0) \approx 0$ for within-arm GMM estimators

$$E_n[D_i m(X_i, \widehat{\beta}_1)] = 0, \quad E_n[(1 - D_i)m(X_i, \widehat{\beta}_0)] = 0. \tag{4.1}$$

**Definition 4.1** (GMM Rerandomization)**.** Define $\mathcal{I}_n^m = \sqrt{n}(\widehat{\beta}_1 - \widehat{\beta}_0)$ as above, where $m(X, \beta)$ is a score satisfying Assumption 3.2. Suppose $d_\beta = d_m$ (exact identification) and let $A$ be a symmetric acceptance region. Do the following: (1) form groups as in Definition 2.1. (2) Draw $D_{1:n}$ by stratified randomization and estimate Equation 4.1. (3) If imbalance $\mathcal{I}_n^m = \sqrt{n}(\widehat{\beta}_1 - \widehat{\beta}_0) \in A$, accept $D_{1:n}$. Otherwise, repeat from (2).

If $m(X_i, \beta) = X_i - \beta$, then $\widehat{\beta}_d = \bar{X}_d$ for $d = 0, 1$ and $\mathcal{I}_n^m = \mathcal{I}_n$, so linear rerandomization is a special case. However, Definition 4.1 also allows for novel designs, such as rerandomizing until the estimated densities of $X_i|D_i = 1$ among treated and $X_i|D_i = 0$ among control are similar. To the best of our knowledge, we provide the first asymptotic theory for such a design.

**Example 4.2** (Density Rerandomization)**.** Let $f(X, \beta)$ be a possibly misspecified parametric likelihood. Draw $D_{1:n}$ and form MLE $\widehat{\beta}_1 \in \operatorname{argmax}_\beta E_n[D_i \log f(X_i, \beta)]$ and $\widehat{\beta}_0 \in \operatorname{argmax}_\beta E_n[(1 - D_i) \log f(X_i, \beta)]$, rerandomizing until the estimated parameters $\sqrt{n}|\widehat{\beta}_1 - \widehat{\beta}_0|_2 \leq \epsilon$. Under regularity conditions,[17] $\widehat{\beta}_d$ are GMM estimators with score $m(X_i, \beta) = \nabla_\beta \log f(X_i, \beta)$, so this procedure is a GMM rerandomization with acceptance region $A = \{x : |x|_2 \leq \epsilon\}$.

Let $\beta^*$ be the unique solution to $E[m(X, \beta^*)] = 0$ and define the Jacobian $G_m = E[(\partial/\partial\beta')m(X_i, \beta^*)]$. Our next result shows that GMM rerandomization with acceptance criterion $\mathcal{I}_n^m \in A$ is equivalent to linear rerandomization (Definition 2.1) with an implicit choice of rerandomization covariates $h_i = m_i^* \equiv m(X_i, \beta^*)$ and linearly transformed acceptance region.

**Theorem 4.3** (GMM Rerandomization)**.** *Suppose $D_{1:n}$ is as in Definition 4.1 and Assumption 3.2 holds. Then $\sqrt{n}(\widehat{\theta} - \theta_n)|W_{1:n} \Rightarrow \mathcal{N}(0, V_a) + R$, independent RV's with*

$$V_a = \min_{\gamma \in \mathbb{R}^{d_m \times d_\theta}} v_D^{-1} E[\mathrm{Var}(\Pi a(W, \theta_0) - \gamma' m_i^* | \psi)]. \tag{4.2}$$

*The residual $R \sim [\gamma_0' Z_m \,|\, Z_m \in G_m A]$ for $Z_m \sim \mathcal{N}(0, v_D^{-1} E[\mathrm{Var}(m_i^*|\psi)])$, where $\gamma_0$ is optimal in Equation 4.2.*

[17] For example, if $\beta \to \log f(X, \beta)$ is a.s. strictly concave, the key identification condition in Assumption 3.2 will be satisfied.

Theorem 4.3 shows that by rerandomizing until $\sqrt{n}(\widehat{\beta}_1 - \widehat{\beta}_0) \in A$, we implicitly balance the influence function $-G_m^{-1} m(X_i, \beta^*)$ for the difference of GMM estimators above. In particular, this shows that all GMM rerandomization designs are first-order equivalent to linear rerandomization (Definition 2.1) for some choice of $h_i$ and acceptance region $A$.

For completeness, we provide a feasible linear rerandomization that exactly mimics the behavior in Theorem 4.3. To do so, let $\widehat{h}_i = m(X_i, \widehat{\beta})$ for $E_n[m(X_i, \widehat{\beta})] = 0$ solving the pooled GMM problem, and rerandomize until $\sqrt{n}(E_n[\widehat{h}_i | D_i = 1] - E_n[\widehat{h}_i | D_i = 0]) \in \widehat{G}_m A$ for $\widehat{G}_m \xrightarrow{p} G_m$.

**Corollary 4.4** (Feasible Equivalence)**.** *Suppose Assumption 3.1, 3.2 and let $m(X, \beta)$ as in Definition 4.1. Let $D_{1:n}$ be rerandomized as in Definition 2.1 with $\widehat{h}_i = m(X_i, \widehat{\beta})$ and acceptance region $\widehat{G}_m A$. Then $\sqrt{n}(\widehat{\theta} - \theta_n) | W_{1:n} \Rightarrow \mathcal{N}(0, V_a) + R$, with both variables identical to those in Theorem 4.3.*

One consequence of Theorem 4.3 is that density based rerandomization for likelihood in an exponential family with sufficient statistic $r(X_i)$ is asymptotically equivalent to linear rerandomization setting $h_i = r(X_i)$.

**Example 4.5** (Density Rerandomization)**.** Define $f(x, \beta) = \exp(\beta' r(x) - t(\beta))$, with sufficient statisitc $r(x)$ for some measure $\nu$ on $x \in \mathcal{X}$. If the $(r_j(x))_{j=1}^k$ are $\nu$-a.s. linearly independent, then $\beta \to \log f(x, \beta)$ is strictly concave for all $x$.[18] Then the score $m(X, \beta) = \nabla_\beta \log f(X, \beta)$ has a unique solution $E[m(X, \beta^*)] = 0$, showing that quasi-MLE in this family can be formulated as a GMM problem. By Theorem 4.3, density rerandomization using $f(x, \beta)$ is asymptotically equivalent to linear rerandomization with $h_i^* = \nabla_\beta \log f(X_i, \beta^*) = r(X_i) - \nabla_\beta t(\beta^*)$. Since $E_n[\nabla_\beta t(\beta^*) | D_i = 1] - E_n[\nabla_\beta t(\beta^*) | D_i = 0] = 0$, this design is equivalent to setting $h_i = r(X_i)$. For example, if $x \in \{\pm 1\}^k$ are binary variables, consider density-based rerandomization using the graphical model[19]

$$f(x, \beta) = \exp\left( \sum_j x_j \beta_j + \sum_{j<l} x_j x_l \beta_{jl} - t(\beta) \right).$$

The parameters $\beta_{jl}$ model correlation between the binary variables $x_j$ and $x_l$. For $x \in \{\pm 1\}^k$ with $k$ large, this is a tractable alternative to nonparametricallly

[18] This holds since the log partition function $t(\beta) = \log \int_{\mathcal{X}} \exp(\beta' r(x)) d\nu(x)$ is strictly convex for $\beta$ s.t. $t(\beta) < \infty$ in this case. See e.g. Wainwright and Jordan (2008) Chapter 3 for an introduction to the properties of the log partition function $t(\beta)$.

[19] This is known as the Ising model. Categorical variables with $l \geq 2$ levels and higher interactions can be added. See Wainwright and Jordan (2008) for MLE algorithms in this family.

modeling the full joint distribution, or e.g. stratifying on all $2^k$ cells. Corollary 4.4 shows that rerandomizing based on the difference of quasi-MLE density estimates in this family[20] is asymptotically equivalent to a simpler linear rerandomization design with $h = r(x) = ((x_j)_j, (x_j x_l)_{j<l})$.

## 4.2 Propensity Score Rerandomization

To motivate a propensity score based rerandomization procedure, note that despite $E[D_i|X_i] = p$ for all units, in finite samples the *realized propensity* $\widehat{p}(B) = E_n[D_i|X_i \in B]$ may significantly diverge from $p$ in certain regions $B \subseteq \mathbb{R}^{d_X}$ of the covariate space. This implies that covariates $X_i$ are predictive of treatment assignments $D_i$ ex-post, a form of "in-sample confounding," which vanishes as $n \to \infty$ but affects precision. To prevent this, we could reject allocations for which $|\widehat{p}(B) - p| > \epsilon$ for some collection of sets $B$. To make this idea tractable, set $X = (1, h)$ and consider a propensity model $p(X, \beta) = L(X'\beta)$ for smooth link function $L$ (e.g. Logit), and define the MLE estimator

$$\widehat{\beta} \in \underset{\beta \in \mathbb{R}^{d_\beta}}{\operatorname{argmax}} E_n[D_i \log L(X_i'\beta) + (1 - D_i) \log(1 - L(X_i'\beta))]. \tag{4.3}$$

The average gap between the realized and ex-ante propensity score can be measured by

$$\mathcal{J}_n = n E_n[(p - L(X_i'\widehat{\beta}))^2]. \tag{4.4}$$

Intuitively, if $\mathcal{J}_n$ is large, then the covariates $X$ are predictive of treatment status in some parts of the covariate space. To avoid this, we propose rerandomizing until the imbalance metric $\mathcal{J}_n$ is below a threshold:

**Definition 4.6** (Propensity Rerandomization)**.** Do the following: (1) form groups as in Definition 2.1. (2) Draw $D_{1:n}$ by stratified randomization and estimate the propensity model in Equation 4.3. (3) If imbalance $\mathcal{J}_n \leq \epsilon^2$, accept. Otherwise, repeat from (2).

This design is illustrated in Figure 1. Note that the covariate distribution is approximately balanced between $D = 1$ and $D = 0$ after acceptance. Our next result shows that propensity rerandomization as in Definition 4.6 is equivalent to a simpler linear rerandomization design, with an implicit choice of ellipsoidal

[20] This is well-motivated when $\psi$ is expected to be more important than $(x_j)_j$. We don't want to stratify on both, since this could radically decrease match quality on $\psi$.

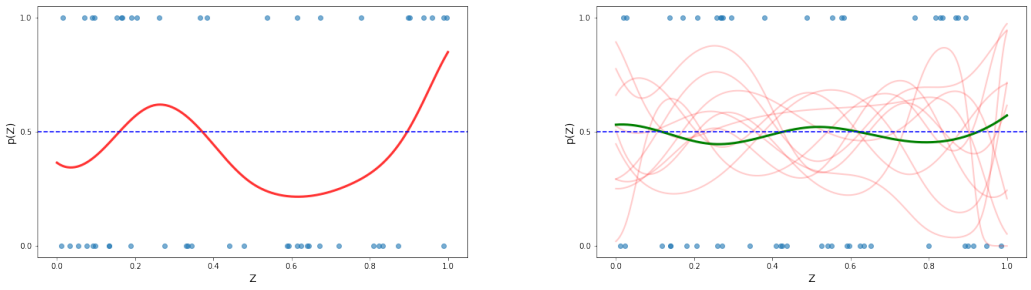


Figure 1: Propensity rerandomization (Definition 4.6) with $p = 1/2$ for $Z \sim \text{Unif}[0,1]$ and $X = B(Z)$ a B-spline basis. LHS: $D_{1:n}$ and estimated propensity with $\widehat{p}(Z) \ll 1/2$, for $Z \in [0.4, 0.9]$, showing imbalance. RHS: Accepted allocation $D_{1:n}$ with $\mathcal{J}_n \leq \epsilon^2$.

acceptance region. We require some extra regularity conditions on the link function $L$, which for brevity we state in Appendix C.3.

**Theorem 4.7** (Propensity Rerandomization)**.** *Suppose $D_{1:n}$ is as in Definition 4.6. Require Assumptions 3.2, C.3. Then $\sqrt{n}(\widehat{\theta} - \theta_n)|W_{1:n} \Rightarrow \mathcal{N}(0, V_a) + R$.*

$$V_a = \min_{\gamma \in \mathbb{R}^{d_h \times d_\theta}} v_D^{-1} E[\text{Var}(\Pi a(W, \theta_0) - \gamma' h | \psi)].$$

*The residual $R \sim \gamma_0' Z_h \,|\, Z_h' \,\text{Var}(h)^{-1} Z_h \leq \epsilon v_D^{-2}$ for $Z_h \sim \mathcal{N}(0, v_D^{-1} E[\text{Var}(h|\psi)])$ and $\gamma_0$ optimal in the equation above.*

Theorem 4.7 shows that for any sufficiently regular link function,[21] propensity rerandomization is asymptotically equivalent to Mahalanobis rerandomization in Example 2.4, with acceptance criterion $n(\bar{h}_1 - \bar{h}_0)' \,\text{Var}_n(h_i)^{-1}(\bar{h}_1 - \bar{h}_0) \leq \epsilon v_D^{-2}$. Equivalently, propensity rerandomization behaves like linear rerandomization with $\mathcal{I}_n = \sqrt{n}(\bar{h}_1 - \bar{h}_0)$ and ellipsoidal acceptance region $A = \text{Var}(h)^{1/2} B(0, \epsilon v_D^{-2})$.[22]

This section showed that a large family rerandomization methods based on nonlinear estimation were asymptotically equivalent to standard linear rerandomization. To obtain new designs with better first-order properties, we may require more stringent acceptance criteria, such as rerandomizing based on the magnitude of a nonparametric two-sample test statistic. We leave such difficult extensions to future work.

[21] Theorem 4.7 uses MLE estimation of $\widehat{\beta}$, though we conjecture the result would be identical for inverse probability tilting (Graham (2012)) or tailored loss function (Zhao (2019)) estimation.

[22] A related result was found by Ding and Zhao (2024), who study rerandomizing until the p-value of a logistic regression coefficient is above a threshold.

# 5 Optimizing Acceptance Regions

In this section, we study efficient choice of the acceptance region $A \subseteq \mathbb{R}^{d_h}$. We propose a minimax rerandomization scheme and show that it minimizes the computational cost of rerandomization subject to a strict lower bound on statistical efficiency. This can be viewed as a form of dimension reduction, increasing rerandomization acceptance probability by downweighting less important directions in the covariate space $h$.

For simplicity, we restrict to the case of estimating $\theta_n = \text{SATE}$. Example 3.8 showed that $\sqrt{n}(\widehat{\theta} - \text{SATE})|W_{1:n} \Rightarrow \mathcal{N}(0, V(\gamma_0)) + \gamma_0' Z_{hA}$, independent RV's with $Z_{hA} = Z_h | Z_h \in A$ and variance $V(\gamma_0)$ that does not depend on $A$. The term $Z_{hA}$ arises from residual imbalances in $h$ due to slackness in the acceptance region, $A \neq \{0\}$. The coefficient $\gamma_0$ comes from the partially linear regression[23]

$$\bar{Y} = \gamma_0' h + t_0(\psi) + e, \qquad E[e|\psi] = 0,\ E[eh] = 0. \tag{5.1}$$

All together, the residual imbalance term $\gamma_0' Z_{hA}$ is the limiting distribution under rerandomization of $\gamma_0' \sqrt{n}(\bar{h}_1 - \bar{h}_0)$, the projection of covariate imbalances in $h$ along the direction $\gamma_0$. This suggests an oracle acceptance criterion that rerandomizes until the imbalance $|\gamma_0' \sqrt{n}(\bar{h}_1 - \bar{h}_0)| \leq \epsilon$, with $A = \{x : |\gamma_0' x| \leq \epsilon\}$, reducing the problem to one dimension from arbitrary $\dim(h)$. Of course, this oracle design is infeasible since $\gamma_0$ is unknown when designing the experiment.

## 5.1 Minimax Rerandomization

Since $\gamma_0$ is unknown at design-time, we instead take a minimax approach that incorporates prior information about the coefficient $\gamma_0$. For belief set $B \subseteq \mathbb{R}^{d_h}$ specified by the researcher, consider rerandomizing until the worst case imbalance consistent with $B$ is small enough,

$$\sup_{\gamma \in B} |\gamma' \sqrt{n}(\bar{h}_1 - \bar{h}_0)| \leq \epsilon. \tag{5.2}$$

Equivalently, for imbalance $\mathcal{I}_n = \sqrt{n}(\bar{h}_1 - \bar{h}_0)$ we rerandomize until $p_B(\mathcal{I}_n) \leq \epsilon$ for the convex penalty function $p_B(x) = \sup_{\gamma \in B} |\gamma' x|$. This significantly generalizes the quadratic imbalance penalty $p(x) = x' \operatorname{Var}(h)^{-1} x$ implicitly used by Mahalanobis rerandomization (Example 2.4). Our next result shows that Equation 5.2 is a linear rerandomization design, characterizing the induced region $A$.

[23] This expansion is without loss. We do not impose well-specification $E[e|\psi, h] = 0$.

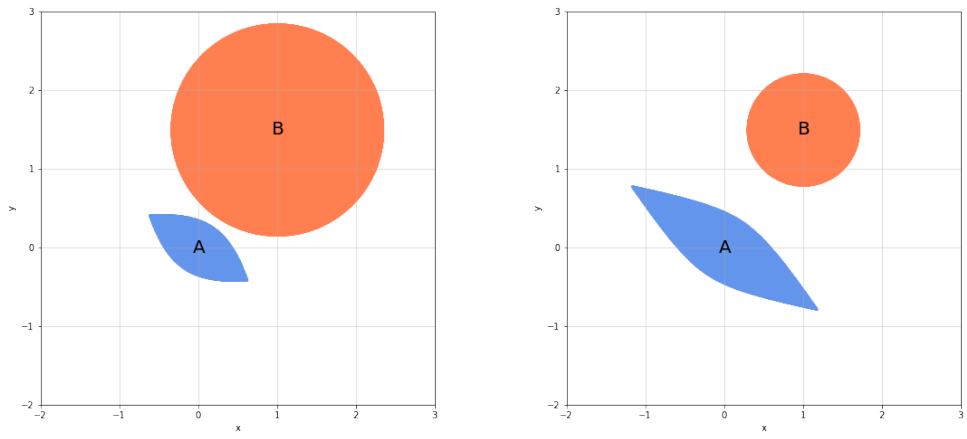


Figure 2: Prior information $B$ and $A_0 = \epsilon B^\circ$ for Example 5.2.

**Proposition 5.1** (Acceptance Region)**.** *The criterion* $p_B(\mathcal{I}_n) \leq \epsilon \iff \mathcal{I}_n \in A_0$ *for* $A_0 = \epsilon B^\circ$ *with* $B^\circ = \{x : \sup_{\gamma \in B} |\gamma' x| \leq 1\} \subseteq \mathbb{R}^{d_h}$*, the absolute polar set of* $B$*. The set* $A_0$ *is symmetric and convex. If* $B$ *is bounded,* $A_0$ *is closed and has non-empty interior.*[24]

Note that since $A_0$ is symmetric, the discussion after Theorem 3.5 implies that the asymptotic distribution of $\widehat{\theta}$ under the design in Equation 5.2 is centered at zero. We let $B$ be totally bounded in what follows. The proposition shows that in this case $A_0$ is a "nice" set: symmetric, convex, and with non-empty interior, satisfying the conditions of Assumption 3.1.

**Dimension Reduction.** The oracle region $A = \{x : |\gamma_0' x| \leq \epsilon\}$ reduced the rerandomization problem to one dimension for arbitrary $\dim(h)$. Similarly, the minimax acceptance region $A_0 = \epsilon B^\circ$ can be viewed as a "soft" form of dimension reduction. To see this, note that the region $A_0$ is very stringent about imbalances $\sqrt{n}(\bar{h}_1 - \bar{h}_0)$ aligned with our belief set $B$, but can allow large imbalances in directions approximately orthogonal to $B$, effectively downweighting these directions in the space of covariates $h$. This effect can be seen in the following example, depicted in Figure 2.

**Example 5.2** (Ball)**.** One natural belief specification is to set $B = \bar{\gamma} + B_2(0, u)$, for an uncertainty parameter $u$ and a priori coefficient guess $\bar{\gamma} \approx \gamma_0$. Lemma 5.3 below derives the corresponding acceptance region $A_0 = \{x : |x'\bar{\gamma}| + u|x|_2 \leq \epsilon\}$. For small $u$, acceptance region $A_0$ mimics the oracle, allowing very large imbalances

[24] If $\operatorname{int} B \neq \emptyset$ then $A_0$ is bounded. See Aliprantis and Border (2006) for more on polar sets.

$\sqrt{n}(\bar{h}_1 - \bar{h}_0)$ as long as $\bar{\gamma}'\sqrt{n}(\bar{h}_1 - \bar{h}_0) \approx 0$. For larger $u$, $A_0$ penalizes imbalances in all directions, with a slight extra penalty for being aligned with the coefficient guess $\bar{\gamma}$. This provides a sliding scale of dimension reduction, allowing us to continuously transition between full-dimensional $h$ and one-dimensional $\bar{\gamma}'h$ depending on the uncertainty level $u$.

More generally, the following lemma provides a useful characterization of the acceptance region $A_0 = \epsilon B^\circ$ from Theorem 5.5 for a large family of specifications of the belief set $B$. To state the lemma, recall that $|x|_p = (\sum_j |x_j|^p)^{1/p}$ for $p \in [1, \infty)$ and $|x|_\infty = \max_j |x_j|$. For $p \in [1, \infty]$, denote $B_p(0, 1) = \{x : |x|_p \leq 1\}$.

**Lemma 5.3** (Belief Specification)**.** *For $p \in [1, \infty]$, let $1/p + 1/q = 1$. Suppose beliefs $B = \bar{\gamma} + UB_p(0, 1)$, for $\bar{\gamma} \in \mathbb{R}^{d_h}$ and $U$ invertible. Then the acceptance region $A_0 = \{x : |x'\bar{\gamma}| + |U'x|_q \leq \epsilon\}$.*

**Example 5.4** (Rectangle)**.** Assume $\gamma_{0j} \in [a_j, b_j]$ for each $1 \leq j \leq d_h$, so $B = \prod_{j=1}^{d_h}[a_j, b_j]$. This allows for sign and magnitude constraints, e.g. $0 \leq \gamma_{0j} \leq m$. Lemma 5.3 shows that the acceptance region has form $A_0 = \epsilon B^\circ = \{x : |x'(a + b)/2| + (1/2)\sum_j |x_j|b_j - |x_j|a_j \leq \epsilon\}$, for $a = (a_j)_j$, $b = (b_j)_j$.

## 5.2 Minimizing Computational Cost

Intuitively, by ignoring imbalances $\mathcal{I}_n = \sqrt{n}(\bar{h}_1 - \bar{h}_0)$ approximately orthogonal to our beliefs $B$, we can "stretch" the acceptance region $A_0$ in directions unlikely to cause large estimation errors, increasing the probability of acceptance $P(\mathcal{I}_n \in A)$. Since the expected number of independent randomizations until acceptance is $P(\mathcal{I}_n \in A)^{-1}$, we can view this as minimizing the computational cost of rerandomization, subject to a bound on estimation error. This intuition is formalized in Theorem 5.5 below. To state the theorem, we first define the family of possible limiting distributions of $\widehat{\theta}$ consistent with our beliefs $\gamma_0 \in B$ and choice of acceptance region $A \subseteq \mathbb{R}^{d_h}$.

**Limiting Distributions.** We showed above that $\sqrt{n}(\widehat{\theta} - \text{SATE})|W_{1:n} \Rightarrow L_0$ for $L_0 = \mathcal{N}(0, V(\gamma_0)) + \gamma_0' Z_{hA}$. Since $\gamma_0$ is unknown, define a family of possible limiting distributions of $\widehat{\theta}$ by $\mathcal{L}_B = \{L_{\gamma A} : \gamma \in B, A \subseteq \mathbb{R}^{d_h}\}$, with each $L_{\gamma A} = \mathcal{N}(0, V(\gamma)) + \gamma' Z_{hA}$ a sum of independent RV's. For any distribution in this family, the conditional asymptotic bias of $\widehat{\theta}$ given realized covariate imbalances $Z_{hA}$ is $\text{bias}(L_{\gamma A}|Z_{hA}) \equiv E[L_{\gamma A}|Z_{hA}]$. Our main result shows that the polar acceptance region $A_0 = \epsilon B^\circ$ minimizes asymptotic computational cost $P(Z_h \in A)^{-1}$, subject

to a strict constraint on conditional bias, uniformly over all limiting distributions consistent with our beliefs.

**Theorem 5.5** (Minimax). *The acceptance region* $A_0 = \epsilon B^\circ$ *solves*[25]

$$A_0 = \underset{A \subseteq \mathbb{R}^{d_h}}{\operatorname{argmin}} P(Z_h \in A)^{-1} \quad s.t. \quad \sup_{\gamma \in B} |\operatorname{bias}(L_{\gamma A} | Z_{hA})| \leq \epsilon. \tag{5.3}$$

*In particular, if* $\gamma_0 \in B$ *(well-specification) then* $|\operatorname{bias}(L_0|Z_{hA_0})| \leq \epsilon$ *and* $\operatorname{Var}(L_0) \leq V_a + \epsilon^2$*, where* $V_a$ *is the partially linear variance in Equation 3.4.*

The final statement of the theorem shows that if $B$ is well-specified ($\gamma_0 \in B$), setting $A_0 = \epsilon B^\circ$ bounds the magnitude of the conditional asymptotic bias $E[L_0|Z_{hA_0}]$ of the GMM estimator $\widehat{\theta}$ above by $\epsilon$. By the law of total variance, this implies that the variance $\operatorname{Var}(L_0)$ of the asymptotic distribution $\sqrt{n}(\widehat{\theta} - \theta_n) \Rightarrow L_0 = \mathcal{N}(0, V_a) + \gamma_0' Z_{hA_0}$ is within $\epsilon^2$ of the optimal partially linear variance $V_a$.

Results closely related to Theorem 5.5 can also be found in the previous work of Liu et al. (2023), who derive optimal Mahalanobis-style completely rerandomized designs under a Bayesian criterion, with Gaussian prior on $\gamma_0$.

**Beliefs from Pilot Data.** In Section B.1 in the appendix, we extend this framework to accommodate belief sets specified as Wald regions learned from pilot data. In that case, $\operatorname{Var}(L_0|\mathcal{D}_{pilot}) \leq V_a + \epsilon^2$ with high probability, without assuming well-specification in the second theorem statement.

# 6 Restoring Normality

In this section, we study optimal linearly adjusted GMM estimation under stratified rerandomization. We show that optimal linear adjustment tailored to the stratification $\psi$ removes the impact of acceptance region $A$ to first-order, restoring asymptotic normality. This enables standard t-statistic and Wald-test based inference on the parameters $\theta_n$ and $\theta_0$ under stratified rerandomization designs, provided in Section 7 below. We also describe a novel form of double robustness to covariate imbalances from combining rerandomization with ex-post adjustment. Let $w$ denote the covariates used for ex-post adjustment and suppose $E[|w|_2^2] < \infty$.

[25] Implicitly, we maximize only over Borel-measurable sets $A \in \mathcal{B}(\mathbb{R}^{d_h})$. The solution $A_0$ is unique up to the equivalence class $\{A \in \mathcal{B}(\mathbb{R}^{d_h}) : \operatorname{Leb}(A \triangle A_0) = 0\}$, where $\triangle$ denotes symmetric difference.

**Definition 6.1** (Adjusted GMM)**.** Suppose that $\widehat{\alpha} \xrightarrow{p} \alpha \in \mathbb{R}^{d_w \times d_g}$. For $H_i = \frac{D_i - p}{p - p^2}$ Define the linearly adjusted GMM estimator $\widehat{\theta}_{adj} = \widehat{\theta} - E_n[H_i \widehat{\alpha}' w_i]$. We refer to $\widehat{\alpha}$ as the *adjustment coefficient* matrix.

First, we extend Corollary 3.6 to provide asymptotics for the adjusted GMM estimator under pure stratification ($A = \mathbb{R}^{d_h}$).

**Proposition 6.2** (Linear Adjustment)**.** *Suppose $D_{1:n}$ as in Definition 2.1 with $A = \mathbb{R}^{d_h}$. Require Assumption 3.2. Then we have $\sqrt{n}(\widehat{\theta}_{adj} - \theta_n) | W_{1:n} \Rightarrow \mathcal{N}(0, V_a(\alpha))$ with $V_a(\alpha) = v_D^{-1} E[\mathrm{Var}(\Pi a(W, \theta_0) - \alpha' w | \psi)]$ and $\sqrt{n}(\widehat{\theta}_{adj} - \theta_0) \Rightarrow \mathcal{N}(0, V_\phi + V_a(\alpha))$.*

A version of this result was given in Cytrynbaum (2024a) for the special case $\theta_0 = \mathrm{ATE}$. Motivated by Proposition 6.2, we define the optimal linear adjustment coefficient as the minimizer of the asymptotic variance $V_a(\alpha)$, in the positive semidefinite sense.

**Optimal Adjustment Coefficient.** Define the coefficient

$$\alpha_0 \in \underset{\alpha \in \mathbb{R}^{d_w \times d_\theta}}{\mathrm{argmin}} \; E[\mathrm{Var}(\Pi a(W, \theta_0) - \alpha' w | \psi)]. \tag{6.1}$$

Note that if $w = h$ then $\alpha_0 = \gamma_0$ in Theorem 3.5. If $E[\mathrm{Var}(w|\psi)] \succ 0$, then the unique minimizer of Equation 6.1 is the partially linear regression coefficient matrix $\alpha_0 = E[\mathrm{Var}(w|\psi)]^{-1} E[\mathrm{Cov}(w, \Pi a(W, \theta_0) | \psi)]$. Observe that $\alpha_0$ varies with the stratification variables $\psi$, as observed in Cytrynbaum (2024b) and Bai et al. (2024a) for the case of ATE estimation. The main result of this section shows that adjustment by a consistent estimate of $\alpha_0$ restores asymptotic normality.

**Theorem 6.3** (Restoring Normality)**.** *Suppose $D_{1:n}$ is rerandomized as in Definition 2.1. Require Assumption 3.1, 3.2. Let $h \subseteq w$ and suppose $\widehat{\alpha} \xrightarrow{p} \alpha_0$. Then $\sqrt{n}(\widehat{\theta}_{adj} - \theta_n) | W_{1:n} \Rightarrow \mathcal{N}(0, V_a^{adj})$ and $\sqrt{n}(\widehat{\theta}_{adj} - \theta_0) \Rightarrow N(0, V_\phi + V_a^{adj})$.*

$$V_\phi = \mathrm{Var}(\Pi \phi(W, \theta_0)) \qquad V_a^{adj} = \min_{\alpha \in \mathbb{R}^{d_w \times d_g}} v_D^{-1} E[\mathrm{Var}(\Pi a(W, \theta_0) - \alpha' w | \psi)].$$

**Two-step Adjustment.** The optimal coefficient $\alpha_0$ may depend on the unknown parameter $\theta_0$. This suggests a two-step adjustment strategy:

(1) Use the unadjusted GMM estimator $\widehat{\theta}$ to consistently estimate $\widehat{\alpha} \xrightarrow{p} \alpha_0$.

(2) Report the adjusted estimator $\widehat{\theta}_{adj} = \widehat{\theta} - E_n[H_i \widehat{\alpha}' w_i]$.

Similar to two-step efficient GMM, this process can be iterated until convergence to improve finite sample properties. One feasible estimator $\widehat{\alpha} \xrightarrow{p} \alpha_0$ is given

in the following theorem. To state the result, define the within-group partialled covariates $\check{w}_i = w_i - \sum_{j \in s(i)} w_j$, where $s(i)$ is the group containing unit $i$ in Definition 2.1. Let $\widehat{\Pi} \xrightarrow{p} \Pi$ estimate the linearization matrix and denote the score evaluation $\widehat{g}_i \equiv g(D_i, R_i, S_i, \widehat{\theta})$. Define the adjustment coefficient estimator

$$\widehat{\alpha} = v_D E_n[\check{w}_i \check{w}_i']^{-1} \left[ \mathrm{Cov}_n(\check{w}_i, \widehat{\Pi}\widehat{g}_i | D_i = 1) - \mathrm{Cov}_n(\check{w}_i, \widehat{\Pi}\widehat{g}_i | D_i = 0) \right]. \tag{6.2}$$

**Theorem 6.4** (Feasible Adjustment)**.** *Suppose $D_{1:n}$ is as in Definition 2.1. Require Assumption 3.1, 3.2. Assume that $E[\mathrm{Var}(w|\psi)] \succ 0$. Then $\widehat{\alpha} = \alpha_0 + o_p(1)$.*

In some cases, $\alpha_0$ may not depend on $\theta_0$. For example, if $a(W, \theta) = a_1(\psi, \theta) + a_2(W)$ then $\alpha_0 = E[\mathrm{Var}(w|\psi)]^{-1} E[\mathrm{Cov}(w, \Pi a_2(W)|\psi)]$. In such cases, one-step optimal adjustment is possible.

**Corollary 6.5** (One-step Adjustment)**.** *Suppose $a(W, \theta) = a_1(\psi, \theta) + a_2(W)$. Then for any $\theta \in \Theta$, substituting $g_i = g(D_i, R_i, S_i, \theta)$ for $\widehat{g}_i$ in $\widehat{\alpha}$ above, $\widehat{\alpha} = \alpha_0 + o_p(1)$.*

One-step adjustment is possible in many linear GMM problems, including the best linear predictor of treatment effect heterogeneity parameter in Example 3.10.

**Example 6.6** (Adjusting CATE Estimate)**.** Continuing Example 3.10, suppose we want to estimate treatment effect heterogeneity relative to a small vector of important covariates $X$, while adjusting optimally for larger set of measured covariates $w$ to both improve precision and restore asymptotic normality under rerandomization. For GMM score $g(Y, D, X, \theta) = (HY - X'\theta)X$ we have $\theta_n = \mathrm{argmin}_\theta E_n[(Y_i(1) - Y_i(0) - X_i'\theta)^2]$. Then $a(W, \theta) = \bar{Y}X$ and $\Pi = E[XX']^{-1}$. Letting $\theta = 0$ gives $g(Y, D, X, 0) = HYX$. After some algebra, Corollary 6.5 shows that $\widehat{\alpha} = \alpha_0 + o_p(1)$ for $\widehat{\alpha} = E_n[\check{w}_i \check{w}_i']^{-1}[(1-p)\,\mathrm{Cov}_n(\check{w}_i, Y_i X_i | D_i = 1) + p\,\mathrm{Cov}_n(\check{w}_i, Y_i X_i | D_i = 0)] E_n[X_i X_i']^{-1}$.

## 6.1 Double Robustness from Rerandomization

Theorem 6.3 shows that stratified rerandomization behaves like optimal ex-post adjustment tailored to both the stratification and GMM problem.[26] However, we find in our simulations and empirical application that stratified rerandomization can perform significantly better than ex-post adjustment in finite samples, and further efficiency gains are possible by combining both methods. In this subsection, we provide a brief theoretical justification for this phenomenon, showing

[26] For the case without stratification, this equivalence was originally shown in Li et al. (2018).

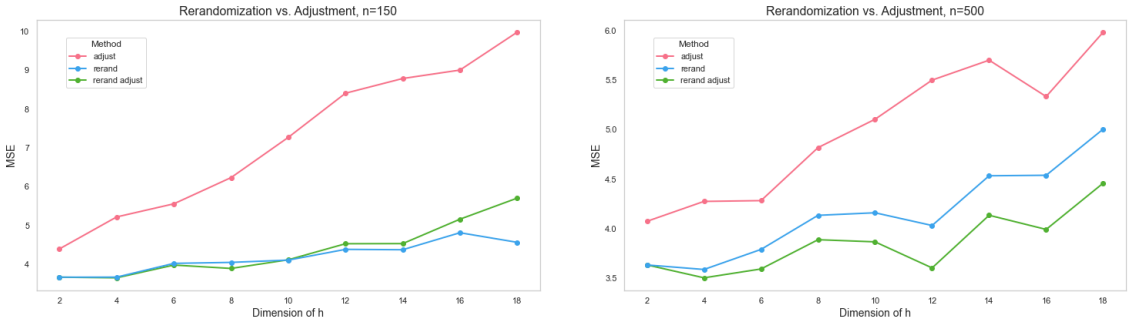


Figure 3: Adjustment vs. Rerandomization, $\theta_n = \text{SATE}$ and $n \in \{150, 500\}$.

that combining rerandomization and adjustment provides a novel form of double robustness to covariate imbalances.

Let $h = w$ so $\alpha_0 = \gamma_0$. Then we can denote $\widehat{\alpha} = \widehat{\gamma}$. Consider the partially linear projection $\Pi a(W, \theta_0) = \gamma_0' h + t(\psi) + e$ with $e \perp h$ and $E[e|\psi] = 0$.[27] Define $\bar{t}_d = E_n[t(\psi_i)|D_i = d]$ and note that $\bar{t}_1 - \bar{t}_0 = o_p(n^{-1/2})$ by fine stratification on $\psi$. Then by Lemma 3.4, adjusted GMM $\widehat{\theta}_{adj} = \widehat{\theta} - \widehat{\gamma}'(\bar{h}_1 - \bar{h}_0)$ has

$$\begin{aligned}\sqrt{n}(\widehat{\theta}_{adj} - \theta_n) &= \sqrt{n}[(\gamma_0 - \widehat{\gamma})'(\bar{h}_1 - \bar{h}_0) + (\bar{t}_1 - \bar{t}_0) + (\bar{e}_1 - \bar{e}_0)] + o_p(1) \\ &= (\gamma_0 - \widehat{\gamma})'\sqrt{n}(\bar{h}_1 - \bar{h}_0) + \sqrt{n}(\bar{e}_1 - \bar{e}_0) + o_p(1).\end{aligned}$$

This shows a form of double robustness from combining rerandomization with ex-post adjustment. If the estimation error $\gamma_0 - \widehat{\gamma}$ is large, then the first imbalance term above may still be negligible as long as we rerandomized until $\sqrt{n}(\bar{h}_1 - \bar{h}_0)$ is small enough. By contrast, without rerandomization adjustment will perform poorly, as shown in the LHS of Figure 3 where $\gamma_0$ is not estimated well for small $n$ and large $\dim(w)$. This effect is exacerbated by stratification, since the within-stratum partialling operation $\check{w}_i = w_i - \sum_{j \in s(i)} w_j$ tends to decrease the variance of the regressors $w_i$, making estimation of $\alpha_0$ more difficult.[28]

The product structure shows that even when both $\widehat{\gamma} - \gamma_0$ and $\bar{h}_1 - \bar{h}_0$ are small, we get an extra benefit from combining the two methods. This is evident from the RHS of Figure 3 with $n = 500$, where adjustment is competitive, but rerandomization still performs better, and combining the two methods is best. This effect is especially important in regimes where the optimal adjustment coefficient $\alpha_0$ is poorly estimated, such as for small $n$ and large $\dim(w)$, or otherwise ill-conditioned design matrix $E_n[\check{w}_i \check{w}_i'] \approx E[\text{Var}(w|\psi)]$. A full theory of high-dimensional strat-

[27] This is the partially linear projection of $\Pi a(W, \theta_0)$ on $\psi, h$ and is without loss of generality.
[28] The condition number of $E_n[\check{w}_i \check{w}_i]$ generally increases as we stratify more finely.

ification, rerandomization, and ex-post adjustment is beyond the scope of the current work, but this is an interesting area for future research.[29]

# 7 Variance Bounds and Inference Methods

In this section, we provide methods for inference on generic causal parameters under stratified rerandomization designs. We make crucial use of asymptotic normality of the optimally adjusted GMM estimator $\widehat{\theta}_{adj}$ developed in the previous section. The asymptotic variance for estimating the finite population parameter $\theta_n$ is generally not identified. To enable inference, we provide novel identified upper bounds on the variance, allowing for conservative inference that still reflects the precision gains from stratified rerandomization. The asymptotic variance for estimating the superpopulation parameter $\theta_0$ is identified, and in this case we provide asymptotically exact inference methods.

## 7.1 Variance Bounds

First, we briefly review the classical variance bounds for $\theta_n = \text{SATE}$ estimation under completely randomized assignment. In this case, we have $\sqrt{n}(\widehat{\theta} - \text{SATE}) \Rightarrow \mathcal{N}(0, V_a)$ with $V_a = \text{Var}(D)^{-1}\text{Var}(\bar{Y})$ for $\bar{Y} = (1-p)Y(1) + pY(0)$. The variance $\text{Var}(\bar{Y}) \propto \text{Cov}(Y(1), Y(0))$. Since $Y(1)$ and $Y(0)$ are never simultaneously observed, $V_a$ is not identified. Let $\sigma_d^2 = \text{Var}(Y(d))$ and $\tau = Y(1) - Y(0)$. The Cauchy-Schwarz inequality $|\text{Cov}(Y(1), Y(0))| \leq \sigma_1\sigma_0$ and some algebra produces

$$V_a = \frac{\sigma_1^2}{p} + \frac{\sigma_0^2}{1-p} - \text{Var}(\tau) \leq \frac{\sigma_1^2}{p} + \frac{\sigma_0^2}{1-p} - (\sigma_1 - \sigma_0)^2 \leq \frac{\sigma_1^2}{p} + \frac{\sigma_0^2}{1-p}. \tag{7.1}$$

Both upper bounds were proposed in Neyman (1990). Theorem 7.1 below extends the sharper bound to generic finite population causal parameters, accounting for both design-time stratified rerandomization and optimal ex-post adjustment.

To develop the bounds, recall that $\sqrt{n}(\widehat{\theta}_{adj} - \theta_n) \Rightarrow N(0, V_a^{adj})$ with $V_a^{adj} = v_D^{-1}E[\text{Var}(\Pi a(W, \theta_0) - \alpha_0' w|\psi)]$, where $\alpha_0 = E[\text{Var}(w|\psi)]^{-1}E[\text{Cov}(w, \Pi a(W, \theta_0)|\psi)]$ was the optimal adjustment coefficient. By definition, $\Pi a(W, \theta_0) = v_D\Pi(g_1(W, \theta_0) - g_0(W, \theta_0))$. Then the adjustment coefficient may be expanded as $\alpha_0 = \beta_1 - \beta_0$ for coefficients $\beta_d = E[\text{Var}(w|\psi)]^{-1}E[\text{Cov}(w, v_D\Pi g_d(W, \theta_0)|\psi)]$. Denote $g_d =$

[29] There are analytical complications from conditioning on $\sqrt{n}(\bar{h}_1 - \bar{h}_0) \in A$ e.g. with $\dim(h)$ growing. A recent breakthrough on this question was achieved by the careful analysis of Wang and Li (2022) for the case of complete rererandomization.

$g_d(W, \theta_0)$ and define the "within-arm" influence functions $m_d \equiv v_D \Pi g_d - \beta_d' w$. Tighter bounds are possible by targeting a fixed scalar contrast $c'\theta_n$ for some $c \in \mathbb{R}^{d_\theta}$. From Theorem 6.3, we have $\sqrt{n}(c'\widehat{\theta}_{adj} - c'\theta_0) \Rightarrow N(0, V_a^{adj}(c))$ for $V_a^{adj}(c) = c' V_a^{adj} c$. In terms of $m_d$, this is

$$\begin{aligned} V_a^{adj}(c) &= c' v_D^{-1} E[\mathrm{Var}(v_D \Pi(g_1 - g_0) - (\beta_1 - \beta_0)' w | \psi)] c \\ &= v_D^{-1} E[\mathrm{Var}(c' m_1 - c' m_0 | \psi)]. \end{aligned}$$

Similarly to above, $V_a^{adj}(c) \propto E[\mathrm{Cov}(c' m_1, c' m_0 | \psi)]$ where the cross-term is generically not identified, since $m_1$ and $m_0$ are not simultaneously observed. However, denoting $\tilde{\sigma}_d^2(c) = E[\mathrm{Var}(c' m_d | \psi)]$ we have the following simple upper bound:

**Theorem 7.1** (Variance Bounds)**.** *Under the conditions of Theorem 6.3, we have*

$$V_a^{adj}(c) \leq v_D^{-1}(\tilde{\sigma}_1(c) + \tilde{\sigma}_0(c))^2 = v_D^{-1} \left( \frac{\tilde{\sigma}_1^2(c)}{1-p} + \frac{\tilde{\sigma}_0^2(c)}{p} \right) - \left( \frac{\tilde{\sigma}_1(c)}{1-p} - \frac{\tilde{\sigma}_0(c)}{p} \right)^2.$$

We provide a consistent estimator of the bound $\bar{V}_a^{adj}(c)$ in Section 7.2 below. The next example shows how Theorem 7.1 generalizes the classical Neyman bounds for the simple case of inference on $\theta_n = \mathrm{SATE}$ under pure stratified randomization ($A = \mathbb{R}^{d_h}$) and optimal ex-post adjustment.

**Example 7.2** (Pure Stratification)**.** Let $\theta_n = \mathrm{SATE}$ so $c = 1$. Then for $H = \frac{D-p}{p-p^2}$ and GMM score $g(D, Y, \theta) = HY - \theta$ have $\Pi = 1$ and $v_D \Pi g_1 = (p - p^2)Y(1)/p = (1-p)Y(1)$. Then $\beta_1 = (1-p)\delta_1$ for $\delta_1 = \mathrm{argmin}_\delta E[\mathrm{Var}(Y(1) - \delta' w | \psi)]$ we have $m_1 = (1-p)(Y(1) - \delta_1' w)$. Similarly, $m_0 = p(Y(0) - \delta_0' w)$ with $\delta_0 = \mathrm{argmin}_\delta E[\mathrm{Var}(Y(0) - \delta' w | \psi)]$. Plugging into the second expression in Theorem 7.1, the variance $V_a^{adj} = \min_\gamma v_D^{-1} E[\mathrm{Var}(\bar{Y} - \gamma' w | \psi)]$ is bounded above by

$$\begin{aligned} \bar{V}_a^{adj} = &\frac{E[\mathrm{Var}(Y(1) - \delta_1' w | \psi)]}{p} + \frac{E[\mathrm{Var}(Y(0) - \delta_0' w | \psi)]}{1-p} \\ &- (E[\mathrm{Var}(Y(1) - \delta_1' w | \psi)]^{1/2} - E[\mathrm{Var}(Y(0) - \delta_0' w | \psi)]^{1/2})^2. \end{aligned}$$

For unadjusted complete randomization ($\psi = 1$, $w = 0$), we recover the sharper Neyman bound in Equation 7.1. If $\psi \neq 1$ and $w = 0$, we get a "finely stratified" bound, tighter for large difference in expected residual variance.

**Remark 7.3** (Covariate-Assisted Bounds by Design)**.** In some contexts, it is possible to use covariate information to tighten finite population variance bounds, e.g. as in Abadie et al. (2020). For example, under complete randomization

with $\theta_n$ = SATE, the non-identified $\text{Cov}(Y(1), Y(0)) = E[\text{Cov}(Y(1), Y(0)|\psi)] + \text{Cov}(E[Y(1)|\psi], E[Y(0)|\psi]) \equiv v_1 + v_2$ by law of total covariance. Only $v_1$ is non-identified, while $v_2$ can be consistently estimated using $\psi$. In our context, however, the term $v_2$ is already removed from the asymptotic variance due to stratified randomization of $D_{1:n}$. More generally, under stratified rerandomization with adjustment, $V_a \propto v_1 = E[\text{Cov}(Y(1)-\delta_1' w, Y(0)-\delta_0' w|\psi)]$, so covariate-assisted tightening happens "automatically" by design. Relative to the papers above, our work provides a tighter upper bound on $v_1$ even after covariate-assistance, corresponding to the sharper Neyman bound in Equation 7.1.

**Remark 7.4** (Sharp Bounds)**.** For $\theta_n$ = SATE estimation under completely randomized assignment, Aronow et al. (2014) derive sharp upper bounds on the variance $V_a = v_D^{-1} \text{Var}(\bar{Y})$. In principle, such bounds could be extended to the more general designs and estimators in our current setting. However, this construction and the associated variance estimators are quite involved, so we leave this significant extension to future work.[30]

## 7.2 Inference on the Finite Population Parameter

Building on the previous section, we construct a consistent estimator of the variance upper bound $\bar{V}_a^{adj}(c)$, enabling asymptotically conservative inference on linear contrasts of the finite population parameter $c'\theta_n$ under general designs.

Let $\mathcal{S}_n$ denote the set of groups (strata) constructed in Definition 2.1. For $s \in \mathcal{S}_n$, denote number of treated $a(s) = \sum_{i \in s} D_i$ and group size $k(s) = |s|$. For any $\widehat{\Pi} \xrightarrow{p} \Pi$ define estimators of the optimal within-arm adjustment coefficients $\beta_d$ above by $\widehat{\beta}_d = v_D E_n[\check{w}_i \check{w}_i']^{-1} \text{Cov}_n(\check{w}_i, \widehat{\Pi}\widehat{g}_i | D_i = d)$. Note that $\widehat{\beta}_1 - \widehat{\beta}_0 = \widehat{\alpha}$, our estimator of the optimal adjustment coefficient in Section 6. For $\widehat{g}_i \equiv g(D_i, X_i, S_i, \widehat{\theta}_{adj})$, define $\widehat{m}_i \equiv v_D \widehat{\Pi}\widehat{g}_i - D_i \widehat{\beta}_1' w_i - (1 - D_i)\widehat{\beta}_0' w_i$. First, suppose each group has at least two treated and control units,

$$\widehat{v}_1 = n^{-1} \sum_{s \in \mathcal{S}_n} \frac{1}{a(s) - 1} \sum_{i \neq j \in s} \widehat{m}_i \widehat{m}_j' D_i D_j / p$$

$$\widehat{v}_0 = n^{-1} \sum_{s \in \mathcal{S}_n} \frac{1}{(k-a)(s) - 1} \sum_{i \neq j \in s} \widehat{m}_i \widehat{m}_j' (1 - D_i)(1 - D_j)/(1 - p)$$

[30] Alternatively, note $E[\text{Cov}(Y(1), Y(0)|\psi)] \leq E[\sigma_1(\psi)\sigma_0(\psi)] \leq E[\sigma_1^2(\psi)]^{1/2} E[\sigma_0^2(\psi)]^{1/2}$. Theorem 7.1 uses the second bound, which we prefer since it can be naturally estimated using the stratification. The first bound could be tighter for large heteroskedasticity, but requires additional nonparametric estimation.

Note that this requires $2 \leq a(s) \leq k(s) - 2 \; \forall s \in \mathcal{S}_n$.

**Collapsed Strata.** If number of treated units $a(s) = 1$ or $a(s) = k(s) - 1$, as in matched pairs designs, the estimators above do not exist. In this case, we follow[31] the method of collapsed strata (Hansen et al. (1953)), first agglomerating the original groups $s \in \mathcal{S}_n$ into larger groups satisfying $2 \leq a(s) \leq k(s) - 2$. For example, in a matched triples design with $p = 1/3$, we agglomerate two triples into a larger group $s'$ of 6 units with $a(s') = 2$. To do so, for each $s \in \mathcal{S}_n$ define the centroid $\bar{\psi}_s = |s|^{-1} \sum_{i \in s} \psi_i$. Let $\nu : \mathcal{S}_n \to \mathcal{S}_n$ be a bijective matching between groups satisfying $\nu(s) \neq s$, $\nu^2 = \text{Id}$, and matching condition $\frac{1}{n} \sum_{s \in \mathcal{S}_n} |\bar{\psi}_s - \bar{\psi}_{\nu(s)}|_2^2 = o_p(1)$. In practice, $\nu$ is obtained by matching the group centroids $\bar{\psi}_s$ into pairs using the Derigs (1988) non-bipartite matching algorithm. Define $\mathcal{S}_n^\nu = \{s \cup \nu(s) : s \in \mathcal{S}_n\}$ to be the enlarged groups. If $a(s) = 1$ or $a(s) = k(s) - 1$, we replace $\mathcal{S}_n$ with the larger groups $\mathcal{S}_n^\nu$ in the definitions of $\widehat{v}_1$ and $\widehat{v}_0$.

**Variance Estimator.** Finally, define $\widehat{u}_1 = E_n[\frac{D_i}{p} \widehat{m}_i \widehat{m}_i'] - \widehat{v}_1$ and $\widehat{u}_0 = E_n[\frac{1-D_i}{1-p} \widehat{m}_i \widehat{m}_i'] - \widehat{v}_0$. The proof of Theorem 7.6 below shows that $c' \widehat{u}_d c \xrightarrow{p} \tilde{\sigma}_d^2(c)$ from Theorem 7.1, suggesting the variance estimator

$$\widehat{V}_a^{adj}(c) = v_D^{-1}([c' \widehat{u}_1 c]^{1/2} + [c' \widehat{u}_0 c]^{1/2})^2. \tag{7.2}$$

To formalize this, we require a slight strengthening of GMM Assumption 3.2.

**Assumption 7.5.** *Exists $\theta_0 \in U \subseteq \Theta$ open s.t. $E[\sup_{\theta \in U} |\partial / \partial \theta' g_d(W, \theta)|_F^2] < \infty$.*

**Theorem 7.6** (Inference)**.** *Suppose $D_{1:n}$ as in Definition 2.1 and impose Assumptions 3.1, 3.2, 7.5. Then $\widehat{V}_a^{adj}(c) \xrightarrow{p} \bar{V}_a^{adj}(c) \geq V_a^{adj}(c)$.*

Then the confidence interval $\widehat{C}_{fin} \equiv [c' \widehat{\theta}_{adj} \pm z_{1-\alpha/2} \widehat{V}_a^{adj}(c)^{1/2} / \sqrt{n}]$ has coverage $P(c' \theta_n \in \widehat{C}_{fin}) \geq 1 - \alpha - o(1)$ by Theorem 6.3 and Theorem 7.6.

The main result is stated for adjusted GMM estimation under stratified rerandomization, with ex-post adjustment to restore normality. For the case of pure stratification (no rerandomization) without adjustment, we can just set $w = 0$ in the formulas above, obtaining a specialization $\widehat{V}_a(c)$ of $\widehat{V}_a^{adj}(c)$:

**Corollary 7.7** (Pure Stratification)**.** *Impose Assumptions 3.1, 3.2, 7.5 and suppose that $A = \mathbb{R}^{d_h}$ and $w = 0$. Then $\widehat{V}_a(c) \xrightarrow{p} \bar{V}_a(c) \geq V_a(c) = c' V_a c$.*

[31] See Abadie and Imbens (2008), Bai et al. (2021), Cytrynbaum (2024b), Bai et al. (2024b) for recent use of this method for inference on superpopulation parameters. In particular, Bai et al. (2021) showed asymptotic exactness of the collapsed strata method for matched pairs designs under the matching condition above.

### 7.3 Inference on the Superpopulation Parameter

The asymptotic variance $V = V_\phi + V_a^{adj}$ for adjusted estimation of $\theta_0$ under stratified rerandomization (Theorem 6.3) is identified. In this case, we can modify the approach above to provide asymptotically exact inference methods. Additionally define

$$\widehat{v}_{10} = n^{-1} \sum_{s \in \mathcal{S}_n} \frac{k}{a(k-a)}(s) \sum_{i,j \in s} \widehat{m}_i \widehat{m}_j' D_i (1 - D_j).$$

With this extra definition in hand, set $\widehat{V} = \mathrm{Var}_n(\widehat{m}_i) - v_D(\widehat{v}_1 + \widehat{v}_0 - \widehat{v}_{10} - \widehat{v}_{10}')$.

**Theorem 7.8** (Superpopulation)**.** *Suppose $D_{1:n}$ is as in Definition 2.1, and impose Assumptions 3.1, 3.2, 7.5. Then $\widehat{V} \xrightarrow{p} V_\phi + V_a^{adj}$.*

By Theorem 6.3, $\sqrt{n}(\widehat{\theta}_{adj} - \theta_0) \Rightarrow N(0, V_\phi + V_a^{adj})$, so the result above allows for asymptotically exact joint inference on $\theta_0$ e.g. using standard Wald-test based confidence regions. For example, the interval $\widehat{C}_{pop} \equiv [c'\widehat{\theta}_{adj} \pm z_{1-\alpha/2}(c'\widehat{V}c)^{1/2}/\sqrt{n}]$ has $P(c'\theta_0 \in \widehat{C}_{pop}) = 1 - \alpha - o(1)$. Similarly to above, this CI can be specialized to pure stratification without adjustment by setting $w = 0$.

## 8 Simulations

In this section, we use simulations to study the finite-sample properties of various designs and estimators analyzed above. We consider data generated as $Y(d) = m_d(r) + e_d$ for observables $r$, varying the covariates $\psi$, $h$, and $w$ used for stratification, rerandomization, and adjustment respectively. In models 1-3, we consider quadratic outcome models of the form $Y(d) = c_d + r'\beta_d + r'Q_d r + e_d$. We vary $m = \dim(r)$, setting parameters $Q_d$ and $\beta_d$ as follows:

**Model** 1: $\beta_1 = \mathbb{1}_m/\sqrt{m}$, $\beta_0 = 0$ and $Q_d = 0$, $c_d = 0$ for $d \in \{0, 1\}$.

**Model** 2: As in Model 1, but with $\beta_{1,1} = 4$, $\beta_{0,1} = 0$, $\beta_{d,2:m} = \mathbb{1}_{m-1}/\sqrt{m-1}$.

**Model** 3: As in 2, but $Q_1 = \mathrm{Diag}(\alpha_1)$ for $\alpha_{1,1} = 2$ and $\alpha_{1,2:m} = 1/(2\sqrt{m-1})$.

**Model** 4: As in 2, but with $Y(d) = 2\arctan(r'\beta_d) + e_d$.

In Model 1, all covariates have equal importance. In Models 2-4, we think of $r_1$ as a baseline outcome with more importance than $r_{2:m}$. This asymmetric structure arises frequently in practice due to the relatively high predictive power of baseline outcomes for endline outcomes. The covariates are generated $r \sim \mathcal{N}(0, I_m)$. The

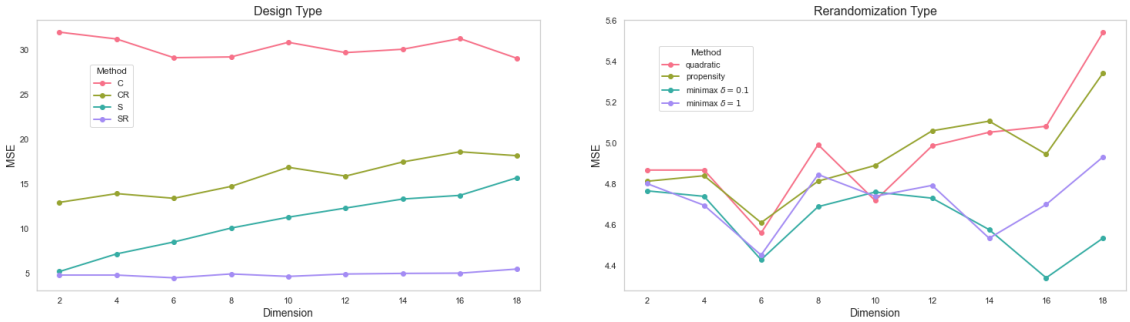


Figure 4: Designs and rerandomization types for $n = 150$, varying $\dim(r)$.

residuals $(e_1, e_0) \sim \mathcal{N}(0, \tilde{\Sigma})$ with $\text{Var}(e_d) = 4$, $\text{Corr}(e_1, e_0) = 0.8$, and $(e_1, e_0) \perp\!\!\!\perp r$. We set $p = 1/2$ in all simulations, corresponding to matched pairs rerandomization for $\psi, h$ non-constant.

In Table 1, we compare the efficiency and inference properties of various designs for estimating $\theta_n = \text{SATE}$. The design **C** refers to complete randomization. Design **S** is full stratification: for model 1, we set $\psi = r$, while for models 2-4, we let $\psi_1 = \sqrt{2}r_1$ and $\psi_{2:m} = r_{2:m}$ in the matching algorithm, putting more weight on the covariate believed to be important a priori.[32] Design **SR** is stratified rerandomization, with univariate $\psi = r_1$ and $h = r_{2:m}$. In this first simulation, we use simple Mahalanobis-style rerandomization (Example 2.4), with acceptance probability $\alpha = 1/500$. $\widehat{\theta}$ is the unadjusted GMM estimator of Definition 2.4, while $\widehat{\theta}_{adj}$ is the optimally adjusted GMM estimator of Theorem 6.4 with adjustment covariates $w = h$. For each model, we normalize the MSE of $\widehat{\theta}$ under complete randomization **C** to 1. All inference results are based on the adjusted estimator $\widehat{\theta}_{adj}$, comparing performance across different designs. In particular, Cover Fin. refers to coverage of $\theta_n$ using the (conservative) finite population variance bound estimator $\widehat{V}_a(c)$ in Section 7.2 and confidence interval $\widehat{C}_{fin}$. Cover Pop. presents coverage of $\theta_0$ for $\widehat{C}_{pop}$, using asymptotically exact variance estimator $\widehat{V}$ from Section 7.3. CI Width Fin. and Pop. report the width confidence intervals, normalized so that the width of $\widehat{C}_{pop}$ is 1 for $\widehat{\theta}_{adj}$ and design **C**.

We summarize a few important findings from Table 1. Stratified rerandomization **SR** is the most efficient design across all specifications and for both estimators $\widehat{\theta}$ and $\widehat{\theta}_{adj}$. While ex-post optimal adjustment and rerandomization have (approximately) the same effect asymptotically (Theorem 6.3), there is an additional finite sample efficiency gain from combining rerandomization and adjustment (**SR** and

[32] We match using the algorithms in Bai et al. (2021) for $p = 1/2$ and Cytrynbaum (2024b) for $p \neq 1/2$.

| | | | $n = 300$ | | | | | | $n = 600$ | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | MSE | | Cover | | CI Width | | MSE | | Cover | | CI Width | |
| dim($r$) | Mod. | Design | $\widehat{\theta}$ | $\widehat{\theta}_{adj}$ | Pop. | Fin. | Pop. | Fin. | $\widehat{\theta}$ | $\widehat{\theta}_{adj}$ | Pop. | Fin. | Pop. | Fin. |
| 5 | 1 | C | 1.00 | 0.89 | 0.94 | 0.94 | 1.00 | 0.70 | 1.00 | 0.86 | 0.96 | 0.96 | 1.00 | 0.69 |
| | | S | 0.85 | 0.87 | 0.94 | 0.98 | 1.03 | 0.82 | 0.87 | 0.88 | 0.95 | 0.97 | 1.02 | 0.77 |
| | | SR | 0.81 | 0.81 | 0.96 | 0.97 | 1.01 | 0.73 | 0.86 | 0.86 | 0.95 | 0.96 | 1.01 | 0.70 |
| | 2 | C | 1.00 | 0.62 | 0.94 | 0.94 | 1.00 | 0.67 | 1.00 | 0.61 | 0.95 | 0.97 | 1.00 | 0.66 |
| | | S | 0.62 | 0.62 | 0.95 | 0.97 | 1.04 | 0.80 | 0.64 | 0.63 | 0.95 | 0.97 | 1.02 | 0.74 |
| | | SR | 0.55 | 0.55 | 0.95 | 0.97 | 1.03 | 0.71 | 0.62 | 0.61 | 0.96 | 0.97 | 1.01 | 0.68 |
| | 3 | C | 1.00 | 0.73 | 0.94 | 0.97 | 1.00 | 0.76 | 1.00 | 0.75 | 0.96 | 0.98 | 1.00 | 0.76 |
| | | S | 0.60 | 0.64 | 0.95 | 0.98 | 0.98 | 0.75 | 0.62 | 0.62 | 0.96 | 0.98 | 0.96 | 0.68 |
| | | SR | 0.53 | 0.53 | 0.96 | 0.98 | 0.94 | 0.61 | 0.59 | 0.59 | 0.96 | 0.97 | 0.92 | 0.57 |
| | 4 | C | 1.00 | 0.80 | 0.93 | 0.95 | 1.00 | 0.86 | 1.00 | 0.81 | 0.94 | 0.97 | 1.00 | 0.86 |
| | | S | 0.73 | 0.74 | 0.95 | 0.98 | 1.02 | 0.92 | 0.79 | 0.79 | 0.96 | 0.97 | 1.01 | 0.88 |
| | | SR | 0.70 | 0.71 | 0.95 | 0.97 | 1.00 | 0.85 | 0.79 | 0.78 | 0.96 | 0.97 | 0.99 | 0.84 |
| 20 | 1 | C | 1.00 | 0.93 | 0.94 | 0.95 | 1.00 | 0.73 | 1.00 | 0.85 | 0.94 | 0.96 | 1.00 | 0.71 |
| | | S | 0.95 | 0.97 | 0.93 | 0.98 | 1.07 | 0.93 | 0.93 | 0.95 | 0.93 | 0.97 | 1.03 | 0.84 |
| | | SR | 0.88 | 0.87 | 0.95 | 0.98 | 1.04 | 0.83 | 0.85 | 0.83 | 0.95 | 0.97 | 1.02 | 0.77 |
| | 2 | C | 1.00 | 0.63 | 0.93 | 0.95 | 1.00 | 0.70 | 1.00 | 0.65 | 0.95 | 0.96 | 1.00 | 0.68 |
| | | S | 0.69 | 0.68 | 0.94 | 0.99 | 1.09 | 0.97 | 0.74 | 0.71 | 0.94 | 0.98 | 1.04 | 0.83 |
| | | SR | 0.59 | 0.61 | 0.96 | 0.99 | 1.11 | 0.87 | 0.65 | 0.64 | 0.96 | 0.98 | 1.06 | 0.77 |
| | 3 | C | 1.00 | 0.75 | 0.92 | 0.96 | 1.00 | 0.76 | 1.00 | 0.78 | 0.95 | 0.97 | 1.00 | 0.76 |
| | | S | 0.69 | 0.75 | 0.94 | 0.98 | 1.06 | 0.93 | 0.76 | 0.76 | 0.94 | 0.99 | 1.01 | 0.82 |
| | | SR | 0.53 | 0.57 | 0.96 | 0.99 | 1.02 | 0.76 | 0.59 | 0.60 | 0.95 | 0.98 | 0.96 | 0.66 |
| | 4 | C | 1.00 | 0.82 | 0.92 | 0.94 | 1.00 | 0.86 | 1.00 | 0.84 | 0.96 | 0.97 | 1.00 | 0.86 |
| | | S | 0.83 | 0.84 | 0.94 | 0.98 | 1.08 | 1.05 | 0.94 | 0.91 | 0.95 | 0.96 | 1.04 | 0.95 |
| | | SR | 0.75 | 0.75 | 0.95 | 0.98 | 1.05 | 0.94 | 0.83 | 0.82 | 0.96 | 0.97 | 1.02 | 0.88 |

Table 1: Design Comparison

$\widehat{\theta}_{adj}$), due to the double robustness property discussed in Section 6. This effect is especially pronounced for small $n$ and large $\dim(r)$, as shown previously in Figure 3, due to poor estimation of the optimal adjustment coefficient $\gamma_0$. For inference, CI Width is slightly larger for **S**, **SR** than for **C**, despite **SR** being the most efficient. Under design **C**, the estimators $\widehat{V}$ and $\widehat{V}_a$ tend to be too small, leading to undercoverage.[33] By contrast, coverage is approximately nominal for designs **S** and **SR**. Note $\widehat{C}_{fin}$ is often smaller than $\widehat{C}_{pop}$, showing that experimenters only interested in covering $\theta_n$ can report smaller confidence intervals.

**Figures.** We provide additional results for Model 4 in Figure 4, letting $n = 150$ and varying $\dim(r)$. **CR** refers to complete rerandomization and "quadratic" refers to the Mahalanobis design in Example 2.4. On the LHS, we see that pure fine stratification on all variables is competitive for small $\dim(r)$, while stratified rerandomization is preferred for $\dim(r) > 2$. On the RHS, we compare different types of rerandomization. The figure shows how **Opt1** and **Opt2** reduce the curse of dimensionality for rerandomization, since we are able to downweight less important dimensions of $h$. For a more detailed simulation comparing Mahalanobis vs. propensity vs. optimized rerandomization, see Section B.2 in the appendix.

# 9 Empirical Application

In this section, we apply our methods to data from the "Opportunity Knocks" experiment in Angrist et al. (2013). The authors randomized eligibility to receive payment for academic performance to first and second year students at a large Canadian university. They estimated the effect of the program on future student GPA, graduation, and other outcomes. They measured baseline covariates including high school GPA, sex, age, native language, and parent's education. Randomization was coarsely stratified on year in college, sex, and quartiles of high school GPA within year-sex cells, with approximately $p = 3/10$ of $n = 1203$ students assigned to receive incentives.

Some students assigned treatment $Z = 1$ (viewed as an instrument) did not engage with the program either by checking their earnings or making contact with the program advisor. The authors view this as noncompliance with the instrument $Z$ and estimate both intention-to-treat (ITT) effects and effects on compliers (LATE). Let $D \in \{0, 1\}$ denote endogeneous decision to engage with the program,

[33] This could be fixed by a sample-splitting or jackknife approach for GMM variance estimation under (non-iid) completely randomized treatment assignment, but this is not our focus here.

with $D(z)$ the potential treatments, $Y(d)$ the potential outcomes, and $T(z) = Y(D(z))$ the ITT potential outcomes with realized outcome $T = Y(D(Z)) = Y$. Angrist et al. (2013) also estimate ITT-style treatment effect heterogeneity along several dimensions, such as gender and student reported financial need.

In what follows, we use this data to study the efficiency and inference properties of various designs and estimators, including complete randomization, fine stratification on different variable sets, and coarse stratification as in the original study, also including rerandomized versions of each. To do so, we follow the common approach (e.g. Li et al. (2018), Bai (2022)) of imputing the missing potential outcomes, which allows us to simulate the MSE, coverage properties, and CI width under various counterfactual designs. In particular, we set $\widehat{T}(z) = T = Y$ if $Z = z$ in the observed data, and impute $\widehat{T}(z) = \widehat{m}_z^T(X) + \widehat{\sigma}_z^T(X)\epsilon_z$ if $Z = 1 - z$, where $\widehat{m}_z(X)$, $\widehat{\sigma}_z(X)$ are estimated using cross-validated LASSO and random forests applied to 11 baseline covariates their full pairwise interactions. The residual $\epsilon_z \sim \mathcal{N}(0, 1)$. We similarly impute missing potential treatments $\widehat{D}(z)$ for all units with $\widehat{D}(z) = D$ if $Z = z$. See Section B.3 for more details on this procedure.

Given imputed data $(X_i, \widehat{T}_i(z), \widehat{D}_i(z))$ for units $i = 1, \ldots, 1203$, we simulate an experiment of size $n$ as follows: (1) sample $(X_i, \widehat{T}_i(z), \widehat{D}_i(z))_{i=1}^n$ with replacement, (2) draw instrument assignments $\tilde{Z}_{1:n}$ e.g. by stratified rerandomization with covariates $\psi_i, h_i \subseteq X_i$. Then (3) observe realized treatments $\tilde{D}_i = \widehat{D}_i(\tilde{Z}_i)$ and outcomes $\tilde{Y}_i = \tilde{T}_i = \widehat{T}_i(\tilde{Z}_i)$ and (4) form estimators $\widehat{\theta}$ and $\widehat{\theta}_{adj}$ and confidence intervals $\widehat{C}_{fin}$ and $\widehat{C}_{pop}$ for the parameters LATE and CLATE described below.

We let rerandomization and adjustment sets $h$, $w$ include all 11 covariates above, as well as the pairwise interactions of HS GPA, sex, year, and mother and father's education with both financial need $F \in \{0, 1\}$ and HS GPA $G \in \mathbb{R}$, for a total of 21 adjustment covariates. The interactions are motivated by our desire to estimate treatment effect heterogeneity along the dimensions $F$ and $G$, as discussed in Example 3.10. We simulate the following designs: **C** is complete randomization, and **CR** is rerandomization. **S** is the original study design (coarse stratification), and **SR** is its rerandomized version using covariates $h$ above. **F** is fine stratification on HS GPA, and **FR** is finely stratified rerandomization. **F**+ is fine stratification on HS GPA, sex, and year and similarly for the rerandomized version **FR**+.[34] We let $p = 3/10$ and $n = 1200$ for all.

Table 2 presents efficiency and inference results for LATE-style treatment ef-

[34]For the last four designs **F**-**FR**+, we remove covariates included in $\psi$ from $w$ and $h$, to ensure that $E[\mathrm{Var}(w|\psi)] \succ 0$, as discussed in Section 6. This does not affect first-order efficiency.

| | | MSE | | Cover | | CI Width | |
|---|---|---|---|---|---|---|---|
| $\theta_n$ (LATE) | Design | $\widehat{\theta}$ | $\widehat{\theta}_{adj}$ | Pop. | Fin. | Pop. | Fin. |
| LATE | C | 1.19 | 1.00 | 0.94 | 0.98 | 1.00 | 0.95 |
| | CR | 1.00 | 0.97 | 0.95 | 0.99 | 1.00 | 0.94 |
| | S | 1.02 | 1.02 | 0.94 | 0.97 | 0.99 | 0.93 |
| | SR | 1.01 | 1.02 | 0.94 | 0.97 | 0.99 | 0.94 |
| | F | 1.04 | 0.97 | 0.95 | 0.98 | 0.98 | 0.91 |
| | FR | 0.96 | 0.94 | 0.94 | 0.99 | 0.98 | 0.91 |
| | F+ | 0.96 | 0.98 | 0.95 | 0.98 | 1.01 | 0.94 |
| | FR+ | 0.98 | 0.99 | 0.95 | 0.98 | 1.01 | 0.94 |
| CLATE (Fin.) | C | 3.30 | 1.00 | 0.93 | 0.98 | 1.00 | 1.01 |
| | CR | 1.97 | 0.89 | 0.95 | 0.98 | 0.98 | 0.97 |
| | S | 3.19 | 0.97 | 0.95 | 0.98 | 0.98 | 0.99 |
| | SR | 1.94 | 0.87 | 0.96 | 0.99 | 0.98 | 0.98 |
| | F | 3.20 | 1.05 | 0.94 | 0.98 | 1.04 | 1.04 |
| | FR | 1.95 | 1.00 | 0.95 | 0.98 | 1.02 | 1.01 |
| | F+ | 3.01 | 1.02 | 0.95 | 0.98 | 1.07 | 1.07 |
| | FR+ | 1.57 | 0.98 | 0.95 | 0.98 | 1.06 | 1.06 |
| CLATE (GPA) | C | 3.06 | 1.00 | 0.92 | 0.98 | 1.00 | 1.02 |
| | CR | 1.76 | 0.85 | 0.95 | 0.99 | 0.97 | 0.97 |
| | S | 1.42 | 0.98 | 0.94 | 0.98 | 0.97 | 1.01 |
| | SR | 1.07 | 0.89 | 0.94 | 0.99 | 0.97 | 0.99 |
| | F | 0.86 | 0.92 | 0.97 | 0.98 | 1.10 | 0.97 |
| | FR | 0.79 | 0.83 | 0.97 | 0.99 | 1.05 | 0.94 |
| | F+ | 1.41 | 1.44 | 0.96 | 0.95 | 1.39 | 1.32 |
| | FR+ | 1.32 | 1.34 | 0.96 | 0.97 | 1.38 | 1.32 |

Table 2: LATE Parameters

fects on compliers. In particular, if $C_i = \mathbb{1}(D_i(1) - D_i(0) > 0)$ is a compliance indicator then LATE $= E_n[Y_i(1) - Y_i(0)|C_i = 1]$ and CLATE (Example 2.7) is the coefficient on $x_i$ in the infeasible regression

$$\theta_n = \underset{\theta}{\operatorname{argmin}}\, E_n[(Y_i(1) - Y_i(0) - \theta'(1, x_i))^2|C_i = 1].$$

We consider heterogeneity variables $x_i = F_i \in \{0, 1\}$, an indicator for student financial stress, and $x_i = G_i \in \mathbb{R}$, the student's HS GPA. For $x_i = F_i$, the CLATE has a simple interpretation as the difference in treatment effects for compliers with and without financial stress:

$$\text{CLATE} = E_n[Y_i(1) - Y_i(0)|F_i = 1, C_i = 1] - E_n[Y_i(1) - Y_i(0)|F_i = 0, C_i = 1].$$

Cover Pop. and CI Width Pop. refer to inference on the superpopulation estimands $\theta_0$ corresponding to $\theta_n$, i.e. $\theta_0 = \operatorname{argmin}_\theta E[(Y(1) - Y(0) - \theta'(1, x))^2|C = 1]$ for $\theta_n = \text{CLATE}$ and $\theta_0 = E[Y(1) - Y(0)|C = 1]$ for LATE. The MSE of $\widehat{\theta}_{adj}$ and the CI width of $\widehat{C}_{pop}$ are normalized to 1 under design **C**.

We briefly summarize our main findings from the tables. The efficiency differences between designs are more pronounced for the CLATE heterogeneity variables than for the LATE. Finely stratified rerandomization **FR** is efficient for the majority of estimands, while **SR** is slightly more efficient for estimating treatment effect heterogeneity along the financial need variable $F \in \{0, 1\}$. Confidence intervals broadly have correct coverage. The width of $\widehat{C}_{fin}$ for inference on $\theta_n$ is slightly smaller than $\widehat{C}_{pop}$ for inference on $\theta_0$ on average, with the largest improvements for estimating CLATE (GPA).

# 10 Discussion and Recommendations for Practice

At a high level, we recommend experimenters finely stratify on a few variables expected to be most predictive of outcomes,[35] while rerandomizing to balance the remaining baseline covariates. This can be done using the stratified Mahalanobis design in Example 2.4 or the optimized designs in Section 5, if the researcher has a strong prior. Separating the baseline covariates into stratification and rerandomization tiers is an easy way to balance linear functions of the less important covariates, without degrading match quality when finely stratifying on the most important covariates like baseline outcomes.

[35] More generally, "highly predictive" is defined in the estimand-specific sense of Equation 3.2.

Our work in Section 6.1 showed that combining stratified rerandomization with optimal ex-post adjustment provides a form of double robustness to covariate imbalances between treatment groups, which seemed to matter in our simulations and empirical application. We recommend experimenters adopt this doubly-robust approach, using the stratification-tailored adjustment coefficients in Section 6. In Section 7, we provide the first valid methods for inference on both finite population and superpopulation GMM parameters under stratified rerandomization, enabling inference in new settings. Our work also provides new tools even for some settings with existing inference methods. For example, experimenters can use the finite population methods in Section 7.2 for more powerful inference than is currently available in the setting of stratification without rerandomization, e.g. in experiments in a convenience sample where we only require coverage of the finite population parameter.

This discussion also touches on several practical questions for which the theory does not give concrete guidance. For example, exactly which and how many covariates should we finely stratify on and which should we rerandomize in a given experiment to maximize finite sample efficiency? It may be possible to formally develop a high dimensional theory of stratified rerandomization in future work. However, even with such new technical results, optimizing the partition of covariates into stratification vs. rerandomization sets would likely require knowledge of DGP-specific constants that are not estimable at design-time before outcomes are observed, and may be difficult to specify beliefs over.[36] Providing practically useful and implementable theoretical guidance for such design issues remains a difficult open question for future work.

# A Proof of Main Asymptotic Results

Below, we carefully distinguish between $P_n$, the law of the data $(W_{1:n}, D_{1:n})$ under "pure" stratified randomization, and $Q_n$, the law under rerandomized stratification. We suppress the $n$ subscript in what follows.

**Definition A.1** (Pure Stratification)**.** For $(W_i)_{i=1}^n \overset{\text{iid}}{\sim} F$, let $P$ denote the law of $(W_{1:n}, D_{1:n})$ under the design in steps (1) and (2) of Definition 2.1.

We slightly generalize the rerandomization design introduced in Definition 2.1, which will be useful for our results on nonlinear rerandomization in Section 4.

[36] For example, this would likely require researchers to specify a prior on objects like the Lipschitz coefficient of the function $\psi \to E[a(W, \theta_0)|\psi]$.

**Definition A.2** (Rerandomization)**.** Consider the following:

(a) (Acceptance Regions). Let $\tau_n = \tau + o_p(1)$ for $\tau \in \mathbb{R}^{d_\tau}$ under $P$. Define sample acceptance region $T_n = \{x : b(x, \tau_n) \leq 0\}$ and population region $T = \{x : b(x, \tau) \leq 0\}$ for $b(x, y)$ a measurable function. Accept $D_{1:n}$ if $\mathcal{I}_n \in T_n$ for $\mathcal{I}_n = \sqrt{n}(\bar{h}_1 - \bar{h}_0) + o_p(1)$ under $P$.

(b) (Rerandomization Distribution). Let $\mathcal{F}_n = \sigma(W_{1:n}, \pi_n)$, where $\pi_n \perp\!\!\!\perp W_{1:n}$ is possibly used to break ties in matching (Equation 2.1). For any event $B$ and $P$ as in Definition A.1, define the rerandomization distribution $Q(B|\mathcal{F}_n) = P(B|\mathcal{F}_n, \mathcal{I}_n \in T_n)$ and $Q(B) = E[Q(B|\mathcal{F}_n)]$.

(c) (Assumptions). Assume $P(b(Z_h, \tau) = 0) = 0$ for $Z_h \sim \mathcal{N}(0, E[\text{Var}(h|\psi)])$. Require $P(Z_h \in T) > 0$. Suppose $E[|\phi|_2^2 + |h|_2^2] < \infty$.

The proof of Theorem 3.5 shows that Definition 2.1 specializes the above to $b(x, y) = b(x) = d(x, A) - d(x, A^c)$ for $d(x, A) = \inf_{z \in \mathbb{R}^{d_h}} |x - z|_2$. Next, the following essential lemma shows that the high level properties (e.g. convergence in probability) of $P$ are inherited by the rerandomized version $Q$.

**Lemma A.3** (Dominance)**.** *Let $(B_n)_{n\geq 1}$ and $(R_n)_{n\geq 1}$ events and random variables. Suppose that the rerandomization measure $Q$ is as in Definition A.2.*

*(a) If $B_n \in \mathcal{F}_n$ then $P(B_n) = Q(B_n)$. If $R_n$ is $\mathcal{F}_n$-measurable then $R_n = o_p(1)/O_p(1)$ under $P \iff R_n = o_p(1)/O_p(1)$ under $Q$.*

*(b) $Q(B_n) = o(1)$ if $P(B_n) = o(1)$. If $R_n = o_p(1)/O_p(1)$ under $P$ then $R_n = o_p(1)/O_p(1)$ under $Q$.*

*Proof of Lemma A.3.* (a) follows since $Q = P$ on $\mathcal{F}_n$ by definition. Let $c = P(Z_h \in T) > 0$ by assumption. Define $S_n = \{P(\mathcal{I}_n \in T_n|\mathcal{F}_n) \geq c/2\}$. Then by Lemma A.5, $P(\mathcal{I}_n \in T_n|\mathcal{F}_n) \xrightarrow{p} P(Z_h \in T) = c$, so $P(S_n) \to 1$. We have the upper bound

$$\mathbb{1}(S_n)Q(B_n|\mathcal{F}_n) = \mathbb{1}(S_n)P(B_n|\mathcal{I}_n \in T_n, \mathcal{F}_n) = \mathbb{1}(S_n)\frac{P(B_n, \mathcal{I}_n \in T_n|\mathcal{F}_n)}{P(\mathcal{I}_n \in T_n|\mathcal{F}_n)}$$
$$\leq (c/2)^{-1}\mathbb{1}(S_n)P(B_n, \mathcal{I}_n \in T_n|\mathcal{F}_n) \leq (c/2)^{-1}P(B_n|\mathcal{F}_n).$$

The first equality by definition of $Q$. The first inequality by the definition of $S_n$. The final inequality by additivity of measures. Then for $r_n \equiv (1-\mathbb{1}(S_n))Q(B_n|\mathcal{F}_n)$, we have $Q(B_n|\mathcal{F}_n) = \mathbb{1}(S_n)Q(B_n|\mathcal{F}_n) + r_n$. Note that $|r_n| \leq 1$ and $r_n \xrightarrow{p} 0$, so

$E_Q[r_n] = o(1)$ by modes of convergence. Then expand $Q(B_n)$ as

$$E_Q[Q(B_n|\mathcal{F}_n)] = E_Q[\mathbb{1}(S_n)Q(B_n|\mathcal{F}_n)] + E_Q[r_n] \leq (c/2)^{-1}E_Q[P(B_n|\mathcal{F}_n)] + o(1)$$
$$= (c/2)^{-1}E_P[P(B_n|\mathcal{F}_n)] + o(1) = (c/2)^{-1}P(B_n) + o(1).$$

The second equality follows from part (a), and the final equality by tower law. The $o_p(1)$ results follow by setting $B_n = \{R_n > \epsilon\}$. The $O_p(1)$ results follow by the $o_p(1)$ statement and Lemma C.11. □

**Proof Strategy.** Equipped with this lemma, we will take the following approach to prove Theorem 3.5: (1) show linearization of the GMM estimator $\widehat{\theta}$ about $\theta_n$ and $\theta_0$ under $P$, (2) invoke Lemma A.3 to show these properties still hold under $Q$, then (3) prove distributional convergence of the simpler linearized quantities directly under $Q$.

## A.1 Rerandomization Asymptotics

In this subsection, we develop the necessary tools for step (3) in our proof strategy. First, we state a conditional CLT for pure fine stratification, conditional on the data $W_{1:n}$ and tie-breaking randomness in the matching procedure $\pi_n$.

**Theorem A.4** (CLT). *Suppose $E[|a(W)|_2^2] < \infty$. Define $\mathcal{F}_n = \sigma(W_{1:n}, \pi_n)$. Let $D_{1:n}$ as in Definition A.1. Then $X_n \equiv \sqrt{n}E_n[H_i a(W_i)]$ has $X_n|\mathcal{F}_n \Rightarrow \mathcal{N}(0, V)$. In particular, for each $t \in \mathbb{R}^{d_a}$ we have $E[e^{it'X_n}|\mathcal{F}_n] = \phi(t) + o_p(1)$ with $\phi(t) = e^{-t'Vt/2}$ and $V = v_D^{-1}E[\mathrm{Var}(a|\psi)]$.*

The proof is in Section C.1 below. We use this CLT to establish the following key lemma, which will shortly allow us to to compute the asymptotic distribution of $\sqrt{n}E_n[H_i a(W_i)]$ directly under the rerandomization law $Q$.

**Lemma A.5.** *Let Definition A.2 hold. Let $\widehat{\Delta}_a = E_n[H_i a_i]$ and $\rho = (a, h)$. Fix $t \in \mathbb{R}^{d_a}$. Let $(Z_a, Z_h) \sim \mathcal{N}(0, \Sigma)$ for $\Sigma = v_D^{-1}E[\mathrm{Var}(\rho|\psi)]$. Then under $P$ in Definition A.1, have $E\left[e^{it'\sqrt{n}\widehat{\Delta}_a}\mathbb{1}\left(\mathcal{I}_n \in T_n\right)|\mathcal{F}_n\right] = E\left[e^{it'Z_a}\mathbb{1}\left(Z_h \in T\right)\right] + o_p(1)$.*

*Proof.* (1). Define $B_n = (\sqrt{n}\widehat{\Delta}_a, \mathcal{I}_n, \tau_n)$. Fix $t = (t_1, t_2, t_3) \in \mathbb{R}^{d_g+d_h+d_\tau}$ and consider the characteristic function

$$\phi_{B_n}(t) = E[e^{it_1'\sqrt{n}\widehat{\Delta}_a + it_2'\mathcal{I}_n + it_3'\tau_n}|\mathcal{F}_n] = e^{it_3'\tau}E[e^{it_1'\sqrt{n}\widehat{\Delta}_a + it_2'\mathcal{I}_n}|\mathcal{F}_n] + o_p(1)$$
$$= e^{it_3'\tau}E[e^{it_1'\sqrt{n}\widehat{\Delta}_a + it_2'\sqrt{n}\widehat{\Delta}_h}|\mathcal{F}_n] + o_p(1) = e^{it_3'\tau}e^{-t'\Sigma t/2} + o_p(1) = \phi_B(t) + o_p(1).$$

For the second equality, note that $e^{it_3'\tau_n} \xrightarrow{p} e^{it_3'\tau}$ by continuous mapping. Then $R_n = e^{it_1'\sqrt{n}\widehat{\Delta}_a + it_2'\sqrt{n}\widehat{\Delta}_h}(e^{it_3'\tau_n} - e^{it_3'\tau}) = o_p(1)$. Clearly $|R_n| \leq 2$, so $E[|R_n||\mathcal{F}_n] = o_p(1)$ by Lemma C.9. The third equality is identical, noting that $e^{it_2'\mathcal{I}_n} \xrightarrow{p} e^{it_2'\sqrt{n}\widehat{\Delta}_h}$ again by continuous mapping. The fourth equality is Theorem A.4 applied to $\sqrt{n}E_n[H_i\rho_i]$. The final expression is the characteristic function of $B = (Z_a, Z_h, \tau)$ with $(Z_a, Z_h) \sim \mathcal{N}(0, \Sigma)$. Then we have shown that $B_n|\mathcal{F}_n \Rightarrow B$ in the sense of Proposition C.8. Fix $t \in \mathbb{R}$ and define $G(z_1, z_2, x) = e^{it'z_1}\mathbb{1}(b(z_2, x) \leq 0)$, so

$$G(B_n) = e^{it'\sqrt{n}\widehat{\Delta}_a}\mathbb{1}(b(\mathcal{I}_n, \tau_n) \leq 0) = e^{it'\sqrt{n}\widehat{\Delta}_a}\mathbb{1}(\mathcal{I}_n \in T_n).$$

Define $E_G = \{w : G(\cdot) \text{ not continuous at } w\}$. By Proposition C.8, if $P(B \in E_G) = 0$ then $E[G(B_n)|\mathcal{F}_n] = E[G(B)] + o_p(1) = E[G(Z_a, Z_h, \tau)] + o_p(1)$, which is the required claim.

To finish the proof, we show that that $P(B \in E_G) = 0$. Write $G(z_1, z_2, x) = f(z_1)g(z_2, x)$ for $f(z_1) = e^{it'z_1}$ and $g(z_2, x) = \mathbb{1}(b(z_2, x) \leq 0)$ and define discontinuity point sets $E_f$ and $E_g$ as for $E_G$ above. By continuity of multiplication for bounded functions, if $z_1 \in E_f^c$ and $(z_2, x) \in E_g^c$ then $(z_1, z_2, x) \in E_G^c$. By contrapositive,

$$E_G \subseteq (E_f \times \mathbb{R}^{d_h + d_\tau}) \cup (\mathbb{R} \times E_g).$$

Clearly $E_f = \emptyset$, so $P(B \in E_G) = P((Z_h, \tau) \in E_g)$. Let $E_g^1 = \{z_h : (z_h, \tau) \in E_g\}$. We have $(Z_h, \tau) \in \mathbb{R}^{d_h} \times \{\tau\}$. Then $P((Z_h, \tau) \in E_g) = P(Z_h \in E_g^1)$. Since $z_h \to b(z_h, \tau)$ is continuous, $\{z_h : b(z_h, \tau) > 0\}$ is open. Let $z_h \in \{z_h : b(z_h, \tau) > 0\}$. Then for small enough $r$, if $z' \in B(z_h, r)$ then $b(z', \tau) > 0$ and $g(z', \tau) = 0$, so $g(z', \tau) - g(z_h, \tau) = 0$, so $z_h$ is a continuity point. A similar argument applied to $z_h \in \{z_h : b(z_h, \tau) < 0\}$ shows that the discontinuities $E_g^1 \subseteq \{z_h : b(z_h, \tau) = 0\}$. □

Finally, we come to the core asymptotic result for step (3) above.

**Theorem A.6** (Asymptotic Distribution)**.** *Let Definition A.2 hold. Suppose that* $(Z_a, Z_h) \sim v_D^{-1}E[\mathrm{Var}((a, h)|\psi)]$. *Then under* $Q$ *in Definition A.2:*

*(a) We have* $\sqrt{n}E_n[H_i a(W_i)]|\mathcal{F}_n \Rightarrow [Z_a|Z_h \in T] \sim \mathcal{N}(0, V_a) + R$, *independent RV's s.t.* $V_a = v_D^{-1}E[\mathrm{Var}(a(W) - \gamma_0'h|\psi)] = \min_{\gamma \in \mathbb{R}^{d_h \times d_\theta}} v_D^{-1}E[\mathrm{Var}(a(W) - \gamma'h|\psi)]$. *The residual term* $R \sim \gamma_0'Z_h \,|\, Z_h \in T$.

*(b) Let* $X_n = E_n[\phi(W_i)] + E_n[H_i a(W_i)]$. *Then we have* $\sqrt{n}(X_n - E[\phi(W)]) \Rightarrow [Z_\phi + Z_a|Z_h \in T] \sim \mathcal{N}(0, V_\phi) + \mathcal{N}(0, V_a) + R$ *The RV's are independent with* $V_\phi = \mathrm{Var}(\phi(W))$.

*Proof.* Consider (a). Let $\widehat{\Delta}_a = E_n[H_i a(W_i)]$. Let $t \in \mathbb{R}^{d_a}$. By definition of $Q$

$$E_Q\left[e^{it'\sqrt{n}\widehat{\Delta}_a}|\mathcal{F}_n\right] = E\left[e^{it'\sqrt{n}\widehat{\Delta}_a}|\,\mathcal{I}_n \in T_n, \mathcal{F}_n\right] = \frac{E\left[e^{it'\sqrt{n}\widehat{\Delta}_a}\mathbb{1}(\mathcal{I}_n \in T_n)|\mathcal{F}_n\right]}{P(\mathcal{I}_n \in T_n|\mathcal{F}_n)} \equiv \frac{a_n}{b_n}.$$

Define $a_\infty = E\left[e^{it'Z_a}\mathbb{1}(Z_h \in T)\right]$ and $b_\infty = P\left(Z_h \in T\right)$. By Lemma A.5, $a_n \xrightarrow{p} a_\infty$ and $b_n \xrightarrow{p} b_\infty$, with $b_\infty > 0$ by assumption in Definition A.2. Then we have $b_n^{-1} = O_p(1)$. Then $|a_n/b_n - a_\infty/b_\infty|$ may be expanded as $\left|\frac{a_n b_\infty - a_\infty b_n}{b_n b_\infty}\right| = O_p(1)|(a_n - a_\infty)b_\infty + a_\infty(b_\infty - b_n)| \lesssim_P |a_n - a_\infty| + |b_\infty - b_n| = o_p(1)$. The final equality by Lemma A.5. Then we have shown

$$E_Q\left[e^{itA_n}|\mathcal{F}_n\right] = \frac{a_\infty}{b_\infty} + o_p(1) = \frac{E\left[e^{it'Z_a}\mathbb{1}(Z_h \in T)\right]}{P\left(Z_h \in T\right)} = E[e^{it'Z_a}|Z_h \in T] + o_p(1).$$

This proves the first statement. Next, we characterize the law of $Z_a\,|Z_h \in T$. Define $\phi(t) \equiv E\left[e^{it'Z_a}|Z_h \in T\right]$. Let $\gamma_0 \in \mathbb{R}^{d_h \times d_g}$ satisfy the normal equations $E[\mathrm{Var}(h|\psi)]\gamma_0 = E[\mathrm{Cov}(h, a|\psi)]$. Such a $\gamma_0$ exists and satisfies the stated inequality by Lemma C.10. Letting $\tilde{Z}_a = Z_a - \gamma_0' Z_h$, by Lemma C.10 $\tilde{Z}_a \perp\!\!\!\perp Z_h$ and $\tilde{Z}_a$ is Gaussian. Then $\tilde{Z}_a \perp\!\!\!\perp (Z_h, \mathbb{1}(Z_h \in T))$. Recall that $A \perp\!\!\!\perp (S, T) \implies A \perp\!\!\!\perp S\,|\,T$. Using this fact, we have $\tilde{Z}_a \perp\!\!\!\perp Z_h\,|\,Z_h \in T$. Then for any $t \in \mathbb{R}^{d_g}$

$$\begin{aligned}\phi(t) &= E[e^{it'Z_a}|Z_h \in T] = E[e^{it'\tilde{Z}_a}e^{it'\gamma_0'Z_h}|Z_h \in T] \\ &= E[e^{it'\tilde{Z}_a}|Z_h \in T]E[e^{it'\gamma_0'Z_h}|Z_h \in T] = E[e^{it'\tilde{Z}_a}]E[e^{it'\gamma_0'Z_h}|Z_h \in T].\end{aligned}$$

By Proposition C.8, we have shown $Z_a\,|\,Z_h \in T \overset{d}{=} \tilde{Z}_a + [\gamma_0' Z_h\,|\,Z_h \in T]$, where the RHS is a sum of independent random variables with the given distributions. Clearly $E[\tilde{Z}_a] = 0$ and $\mathrm{Var}(\tilde{Z}_a) = v_D^{-1}E[\mathrm{Var}(a - \gamma_0' h|\psi)]$. This finishes (a).

Next we prove (b). We may expand $\sqrt{n}(X_n - E[\phi(W)]) = \sqrt{n}(E_n[\phi(W_i)] - E[\phi(W)]) + \sqrt{n}\widehat{\Delta}_a \equiv A_n + B_n$. We have $A_n \Rightarrow \mathcal{N}(0, V_\phi)$ with $V_\phi = \mathrm{Var}(\phi(W))$ by vanilla CLT. Then let $t \in \mathbb{R}^{d_a}$ and calculate

$$E_Q\left[e^{it'X_n}\right] = E_Q\left[e^{it'A_n}E_Q\left[e^{it'B_n}|\mathcal{F}_n\right]\right] = \phi(t)E_Q\left[e^{it'A_n}\right] + o(1) = \phi(t)e^{-t'V_\phi t/2} + o(1).$$

The first equality since $A_n \in \mathcal{F}_n$. The second equality since

$$\left|E_Q\left[e^{it'A_n}(E_Q\left[e^{it'B_n}|\mathcal{F}_n\right] - \phi(t))\right]\right| \leq E_Q\left[|E_Q\left[e^{it'B_n}|\mathcal{F}_n\right] - \phi(t)|\right] = o(1).$$

To see this, note that the integrand is $o_p(1)$ by our work above. It is also bounded so it converges to zero in $L_1(Q)$ by Lemma C.9. The final equality since $A_n \in$

$\mathcal{F}_n = \sigma(W_{1:n}, \pi_n)$ and the marginal distribution of $(W_{1:n}, \pi_n)$ is identical under $P$ and $Q$ by definition. Then $E_Q\left[e^{it'A_n}\right] = E_P\left[e^{it'A_n}\right] = e^{-t'V_\phi t/2} + o(1)$ by vanilla CLT. Then we have shown

$$E_Q\left[e^{it'X_n}\right] = e^{-t'(V_\phi+V_a)t/2}E[e^{it'\gamma_0'Z_h}|Z_h \in B] + o(1).$$

This finishes the proof of (b). □

## A.2 Proof of Main Results

Next, we prove Theorem 3.5 and Corollary 3.7. See Section C.2 below for the proof of the following lemma, which uses a novel ULLN for GMM estimation under fine stratification.

**Lemma A.7** (Linearization)**.** *Suppose Definition A.2 and Assumption 3.2 hold. Let* $\Pi = -(G'MG)^{-1}G'M$. *Then* $\sqrt{n}(\widehat{\theta} - \theta_n) = \sqrt{n}E_n[H_i\Pi a(W_i, \theta_0)] + o_p(1)$ *and* $\sqrt{n}(\widehat{\theta} - \theta_0) = \sqrt{n}E_n[\Pi\phi(W_i, \theta_0) + H_i\Pi a(W_i, \theta_0)] + o_p(1)$.

*Proof of Theorem 3.5.* We claim that the conditions of Definition A.2 hold. This will allow us to apply our general rerandomization asymptotics in Theorem A.6 and linearization in Lemma A.7. To check part (a), define $b(x, y) = b(x) = d(x, A) - d(x, A^c)$, where $d(x, A) = \inf_{s\in\mathbb{R}^{d_h}} |x - s|_2$. It's well known that $x \to d(x, S)$ is continuous for any set $S$, so $b$ is continuous. The sample and population regions $T_n = T = \{x : b(x) \leq 0\}$. If $b(x) \leq 0$ then $d(x, A) = 0$, so $x \in A \cup \partial A \subseteq A$ by closedness. If $b(x) > 0$ then $x \notin A$. This shows $T_n = A$, so $\{\mathcal{I}_n \in T_n\} = \{\mathcal{I}_n \in A\}$. Then our criterion is of the form in Definition A.2. For part (b), $P(b(Z_h) = 0) = P(Z_h \in \partial A) = 0$ since $\text{Leb}(\partial A) = 0$ and by absolute continuity of $Z_h$ relative to Lebesgue measure. We also have $P(Z_h \in T) = P(Z_h \in A) > 0$ since $Z_h$ is full measure by $E[\text{Var}(h|\psi)] \succ 0$ and since $A$ has non-empty interior This proves the claim. Then by Lemma A.7, $\sqrt{n}(\widehat{\theta} - \theta_n) = \sqrt{n}E_n[H_i\Pi a(W_i, \theta_0)] + o_p(1)$. The result now follows immediately by Slutsky and Theorem A.6(a), letting $a \to \Pi a$. Likewise, Corollary 3.7 follows from Theorem A.6(b), letting $\phi \to \Pi\phi$. □

*Proof of Corollary 3.6.* By Theorem 3.5, since $A = \mathbb{R}^{d_h}$ we have $\sqrt{n}(\widehat{\theta}-\theta_n)|W_{1:n} \Rightarrow \mathcal{N}(0, V_a) + R$, independent RV's with $V_a = v_D^{-1}E[\text{Var}(\Pi a(W, \theta_0) - \gamma_0'h|\psi)]$ and $R \sim \gamma_0'Z_h$ for $Z_h \sim \mathcal{N}(0, v_D^{-1}E[\text{Var}(h|\psi)])$. Then $\mathcal{N}(0, V_a)+R \sim \mathcal{N}(0, V)$ with $V = V_a + \text{Var}(\gamma_0'Z_h) = v_D^{-1}E[\text{Var}(\Pi a(W, \theta_0) - \gamma_0'h + \gamma_0'h|\psi)] - 2v_D^{-1}E[\text{Cov}(\Pi a(W, \theta_0) - \gamma_0'h, \gamma_0'h|\psi)] = v_D^{-1}E[\text{Var}(\Pi a(W, \theta_0)|\psi)]$. The covariance term is zero by Lemma C.10. The second statement follows by setting $\psi = 1$. □

# References


Abadie, A., Athey, S., Imbens, G. W., and Wooldridge, J. M. (2020). Sampling-based versus design-based uncertainty in regression analysis. *Econometrica*.

Abadie, A. and Imbens, G. W. (2008). Estimation of the conditional variance in paired experiments. *Annales d'Economie et de Statistique*, pages 175–187.

Abadie, A., Imbens, G. W., and Zheng, F. (2014). Inference for misspecified models with fixed inference for misspecified models with fixed regressors. *Journal of the American Statistical Association*, 109(508).

Aliprantis, C. D. and Border, K. C. (2006). *Infinite Dimensional Analysis: A Hitchhiker's Guide*. Springer.

Angrist, J. D., Oreopoulos, P., and Williams, T. (2013). New evidence on college achievement awards. *Journal of Human Resources*.

Armstrong, T. (2022). Asymptotic efficiency bounds for a class of experimental designs.

Aronow, P., Green, D. P., and Lee, D. K. K. (2014). Sharp bounds on the variance in randomized experiments. *Annals of Statistics*.

Bai, Y. (2022). Optimality of matched-pair designs in randomized controlled trials. *American Economic Review*.

Bai, Y., Jiang, L., Romano, J. P., Shaikh, A. M., and Zhang, Y. (2024a). Covariate adjustment in experiments with matched pairs. *Journal of Econometrics*.

Bai, Y., Romano, J. P., and Shaikh, A. M. (2021). Inference in experiments with matched pairs. *Journal of the American Statistical Association*.

Bai, Y., Shaikh, A. M., and Tabord-Meehan, M. (2024b). On the efficiency of finely stratified experiments.

Bugni, F. A., Canay, I. A., and Shaikh, A. M. (2018). Inference under covariate-adaptive randomization. *Journal of the American Statistical Association*.

Cochran, W. G. (1977). *Sampling Techniques*. John Wiley and Sons, 3 edition.

Cytrynbaum, M. (2021). Essays on experimental design. Dissertation.

Cytrynbaum, M. (2024a). Covariate adjustment in stratified experiments. *Quantitative Economics*.

Cytrynbaum, M. (2024b). Optimal stratification of survey experiments.

Derigs, U. (1988). Solving non-bipartite matching problems via shortest path techniques. *Annals of Operations Research*, 13:225–261.

Ding, P., Feller, A., and Miratrix, L. W. (2019). Decomposing treatment effect variation. *Journal of the American Statistical Association*.

Ding, P. and Zhao, A. (2024). No star is good news: A unified look at rerandomization based on -values from covariate balance tests. *Journal of Econometrics*.

Fogarty, C. B. (2018). Regression-assisted inference for the average treatment effect in paired experiments. *Biometrika*, 105(4).

Graham, B. S. (2011). Efficiency bounds for missing data models with semiparametric efficiency bounds for missing data models with semiparametric restrictions. *Econometrica*.

Graham, B. S. (2012). Inverse probability tilting for moment condition models inverse probability tilting for moment condition models with missing data. *Review of Economics Studies*.

Hansen, M. H., Hurwitz, W. N., and Madow, W. G. (1953). *Sample Survey Methods and Theory*. Wiley.

Imbens, G. and Angrist, J. D. (1994). Identification and estimation of local average treatment effects. *Econometrica*, 62.

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

Li, Y., Kang, L., and Huang, X. (2021). Covariate balancing based on kernel density estimates for controlled experiments. *Statistical Theory and Related Fields*.

Lin, W. (2013). Agnostic notes on regression adjustments to experimental data: Reexamining freedman's critique. *The Annals of Applied Statistics*, 7(1):295–318.

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

Liu, Z., Han, T., Rubin, D. B., and Deng, K. (2023). Bayesian criterion for rerandomization.

Morgan, K. L. and Rubin, D. B. (2012). Rerandomization to improve covariate balance in experiments. *Annals of Statistics*, 40(2).

Newey, W. K. and McFadden, D. (1994). Large sample estimation and hypothesis testing. *Handbook of Econometrics*, IV.

Neyman, J. S. (1990). On the application of probability theory to agricultural experiments. essay on principles. *Statistical Science*.

Niehaus, P. and Muralidharan, K. (2017). Experimentation at scale. *Journal of Economic Perspectives*.

Pollard, D. (1984). *Convergence of Stochastic Processes*. Springer-Verlag.

Ren, J. (2023). Model-assisted complier average treatment effect estimates in randomized experiments with non-compliance and a binary outcome. *Journal of Business and Economic Statistics*.

Robinson, P. M. (1988). Root-n-consistent semiparametric regression. *Econometrica*, 56(4).

Rockafellar, T. R. (1996). *Convex Analysis*. Princeton University Press.

Schindl, K. and Branson, Z. (2024). A unified framework for rerandomization using quadratic forms.

Tauchen, G. (1985). Diagnostic testing and evaluation of maximum likelihood models. *Journal of Econometrics*.

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes*. Springer.

Wainwright, M. J. and Jordan, M. I. (2008). Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*.

Wang, B. and Li, F. (2024a). Asymptotic inference with flexible covariate adjustment under rerandomization and stratified rerandomization.

Wang, X., Wang, T., and Liu, H. (2021). Rerandomization in stratified randomized experiments. *Journal of the American Statistical Association*.

Wang, Y. and Li, X. (2022). Rerandomization with diminishing covariate imbalance and diverging number of covariates. *Annals of Statistics*.

Wang, Y. and Li, X. (2024b). Asymptotic theory of best-choice rerandomization using the mahalanobis distance. Working Paper.

Xu, R. (2021). Potential outcomes and finite population inference for m-estimators. *Econometrics Journal*.

Zhao, Q. (2019). Covariate balancing propensity score by tailored loss functions. *Annals of Statistics*.

# Online Appendix

## B Appendix

### B.1 Beliefs From Pilot Data

Here, we discuss an alternative to the a priori beliefs in Section 5 that uses pilot data to specify the set $B$ in a data-driven way. Suppose we have access to a dataset $\mathcal{D}_{pilot} \perp\!\!\!\perp (W_{1:n}, D_{1:n})$ of size $m$. Suppose $\sqrt{m}(\widehat{\gamma}_{pilot} - \gamma_0) \approx \mathcal{N}(0, \widehat{\Sigma}_{pilot})$ for some pilot estimator $\widehat{\gamma}_{pilot}$, discussed below. Consider forming the Wald region $\widehat{B}_{pilot} = \{\gamma : m(\widehat{\gamma}_{pilot} - \gamma)'\widehat{\Sigma}_{pilot}^{-1}(\widehat{\gamma}_{pilot} - \gamma) \leq c_\alpha\}$ using critical value $P(\chi^2_{d_h} \leq c_\alpha) = 1 - \alpha$ for $\alpha \in (0, 1)$. Equivalently, one can write this Wald region as

$$\widehat{B}_{pilot} = \widehat{\gamma} + c_\alpha^{1/2} m^{-1/2} \cdot \widehat{\Sigma}_{pilot}^{1/2} B_2(0, 1). \tag{B.1}$$

Viewing this $1 - \alpha$ confidence region as a belief set, Lemma 5.3 implies that the corresponding minimax acceptance region is

$$\widehat{A}_{pilot} = \epsilon \widehat{B}^\circ_{pilot} = \{x : |x'\widehat{\gamma}_{pilot}| + \frac{c_\alpha^{1/2} |\widehat{\Sigma}^{1/2} x|_2}{m^{1/2}} \leq \epsilon\}. \tag{B.2}$$

Note that the acceptance region $\widehat{A}_{pilot}$ expands as the pilot size $m$ is larger. This reflects smaller uncertainty about the true parameter $\gamma_0$, and thus less adversarial worst case imbalance $\sup_{\gamma \in \widehat{B}_{pilot}} |\gamma' \sqrt{n}(\bar{h}_1 - \bar{h}_0)|$. Conversely, $\widehat{A}_{pilot}$ shrinks as the confidence parameter $\alpha$ and variance estimate $\widehat{\Sigma}_{pilot}$ increase, reflecting greater uncertainty and a more conservative approach to covariate imbalances. Our next result shows that rerandomization with acceptance region $\widehat{A}_{pilot}$ controls the variance of the residual imbalance $R_A = \gamma_0' Z_h | Z_h \in \widehat{A}_{pilot}$ with high probability, marginally over the realizations of the pilot data. The result is an immediate consequence of Theorem 3.5 and Theorem 5.5.

**Corollary B.1** (Pilot Data)**.** *Suppose* $P(\gamma_0 \in \widehat{B}_{pilot}) \geq 1 - \alpha$*, for* $\mathcal{D}_{pilot} \perp\!\!\!\perp (W_{1:n}, D_{1:n})$*. Let* $D_{1:n}$ *as in Definition 2.1 with* $A = \widehat{A}_{pilot} = \epsilon \widehat{B}^\circ_{pilot}$*. If Assumptions 3.1, 3.2 hold, then* $\sqrt{n}(\widehat{\theta} - \theta_n) | \mathcal{D}_{pilot} \Rightarrow v_D^{-1} \mathcal{N}(0, E[\mathrm{Var}(\bar{Y} - \gamma_0' h | \psi)]) + R_A$*, where* $\mathrm{Var}(R_A | \mathcal{D}_{pilot}) \leq \epsilon^2$ *with probability* $\geq 1 - \alpha$*.*

Formally, the pilot estimate of $\gamma_0$ and Wald region could be constructed as in Robinson (1988). A simpler practical approach suggested by the theory is to let $\widehat{\gamma}_{pilot}, \widehat{\Sigma}_{pilot}$ be point and variance estimators from the regression $Y_T \sim 1 + h + \psi$,

for the "tyranny of the minority" (Lin (2013)) outcomes $Y_T = (1-p)DY/p + p(1-D)Y/(1-p)$, noting that $E[Y_T|W] = (1-p)Y(1) + pY(0) = \bar{Y}$.

## B.2 Comparing Rerandomization Types

| | | | MSE | | Cover | | CI Width | |
|---|---|---|---|---|---|---|---|---|
| $\theta_n$ | Mod. | **SR** Type | $\widehat{\theta}$ | $\widehat{\theta}_{adj}$ | Pop. | Fin. | Pop. | Fin. |
| SATE | 2 | MH | 1.00 | 1.03 | 0.96 | 0.99 | 1.00 | 0.82 |
| | | Prop | 1.04 | 1.05 | 0.95 | 0.98 | 1.00 | 0.81 |
| | | Best1 | 0.99 | 1.02 | 0.96 | 0.98 | 1.00 | 0.81 |
| | | Best2 | 0.99 | 1.07 | 0.95 | 0.98 | 1.00 | 0.82 |
| | | Opt1 | 1.00 | 1.08 | 0.95 | 0.98 | 1.00 | 0.82 |
| | | Opt2 | 1.01 | 1.02 | 0.95 | 0.98 | 1.00 | 0.82 |
| | 3 | MH | 1.00 | 1.06 | 0.96 | 0.98 | 1.00 | 0.77 |
| | | Prop | 1.02 | 1.06 | 0.95 | 0.98 | 1.00 | 0.77 |
| | | Best1 | 0.99 | 1.04 | 0.96 | 0.99 | 1.00 | 0.77 |
| | | Best2 | 1.01 | 1.11 | 0.95 | 0.99 | 1.00 | 0.77 |
| | | Opt1 | 1.00 | 1.08 | 0.95 | 0.99 | 1.00 | 0.77 |
| | | Opt2 | 0.99 | 1.03 | 0.96 | 0.99 | 1.00 | 0.77 |
| CATE | 2 | MH | 1.00 | 1.03 | 0.97 | 0.98 | 1.00 | 1.00 |
| | | Prop | 0.99 | 1.01 | 0.97 | 0.99 | 1.00 | 1.01 |
| | | Best1 | 1.00 | 1.03 | 0.97 | 0.98 | 1.00 | 1.01 |
| | | Best2 | 1.04 | 1.06 | 0.97 | 0.97 | 1.00 | 1.01 |
| | | Opt1 | 1.00 | 1.03 | 0.98 | 0.98 | 1.00 | 1.01 |
| | | Opt2 | 0.97 | 1.00 | 0.98 | 0.98 | 1.00 | 1.01 |
| | 3 | MH | 1.00 | 1.09 | 0.97 | 0.99 | 1.00 | 0.81 |
| | | Prop | 0.96 | 1.03 | 0.97 | 0.99 | 0.99 | 0.81 |
| | | Best1 | 1.00 | 1.08 | 0.96 | 0.99 | 1.00 | 0.81 |
| | | Best2 | 1.02 | 1.09 | 0.96 | 0.99 | 1.01 | 0.82 |
| | | Opt1 | 1.00 | 1.08 | 0.96 | 0.99 | 1.00 | 0.81 |
| | | Opt2 | 0.99 | 1.09 | 0.97 | 0.99 | 1.01 | 0.82 |

Table 3: Stratified Rerandomization Types

In Table 3 we compare different types of stratified rerandomization acceptance criteria. **MH** is Mahalanobis rerandomization, as in Table 1. **Prop** is the propensity-based rerandomization in Definition 4.6, using Logit $L(x) = (1+e^{-x})^{-1}$ and $X = (1, w)$. Designs **Opt1** and **Opt2** refer to the optimal acceptance regions in Section 5. The belief sets are both well-specified, with either high uncertainty $B_1 = \{x : |x - \gamma_0|_2 \leq 1\}$ or low uncertainty $B_2 = \{x : |x - \gamma_0|_2 \leq 1/10\}$, respectively. In all designs, we set the balance threshold $\epsilon(\alpha)$ so $P(Z_h \in A) = 1/500$. Finally, in **Best1** and **Best2** we rerandomize by implementing the best allocation out of either $k = 500$ or $k = 2500$ stratified draws, according to the minimal Mahalanobis imbalance metric. Note that such "best-of-k" stratified rerandomization designs are not formally covered by our theory.[37] In ad-

[37] Recent work by Wang and Li (2024b) provided the first formal results for "best-of-k" designs in the case without stratification.

dition to $\theta_n$ = SATE, we also provide efficiency and inference results for the treatment effect heterogeneity parameter from Example 3.10. In particular, let $\alpha_n = \text{argmin}_\alpha E_n[(Y_i(1) - Y_i(0) - \alpha'(1, r_{1i}))^2]$. We define $\theta_n$ to be the coefficient on $r_1$, denoting $\theta_n = \text{CATE}$ in the table. Cover Pop. and CI Width Pop. refer to inference on the corresponding superpopulation parameter $\theta_0$. Next, we summarize a few findings from Table 3. Theorem 4.7 showed that **Prop** was first-order equivalent to **MH**, and this is supported by finite-sample evidence in the table. We find that best of $k$ style rerandomization and Mahalnobis rerandomization with acceptance probability $\alpha \approx 1/k$ are indistinguishable in practice. In particular, our inference methods also work well for this design. We don't find major finite sample efficiency improvements from using the optimal acceptance regions in Section 5 in this experiment.

### B.3 Empirical Application Details

The full set of covariates from the baseline survey in Angrist et al. (2013) used in our imputation procedure is HS GPA, sex, year in college, mother and father's education, whether survey question 1 was answered correctly, age, native language, attempted credits, and financial stress. The vector $X$ consists of these basic covariates and all of their pairwise interactions. As noted in Section 9, for the ITT potential outcomes we set $\widehat{T}(z) = T = Y$ if $Z = z$ and impute $\widehat{T}(z) = \widehat{m}_z^T(X) + \widehat{\sigma}_z^T(X)\epsilon_z$ if $Z = 1 - z$. The function $\widehat{m}_z^T(X)$ is estimated using LASSO, regressing $TZ/p$ on $X$ for $z = 1$ and $T(1-Z)/(1-p)$ on $X$ for $z = 0$, with regularization parameter chosen by cross-validation. The variance function $\widehat{\sigma}_z^T(X)$ is estimated by random forests to preserve positivity, regressing $(T_i - \widehat{m}_1^T(X))^2 Z_i/p$ on $X_i$ for $z = 1$ and $(T_i - \widehat{m}_0^T(X))^2(1 - Z_i)/(1 - p)$ on $X_i$ for $z = 0$. The potential treatments $\widehat{D}(z) \in \{0, 1\}$ are imputed similarly, with $\widehat{D}(z) = D$ if $Z = z$ and $\widehat{D}(z) = \mathbb{1}(\widehat{m}_z^D(X) + \widehat{\sigma}_z^D(X)u_z \geq 1/2)$ with $u_z \sim \mathcal{N}(0, 1)$ and both $\widehat{m}_z^D(X), \widehat{\sigma}_z^D(X)$ estimated by cross-validated random forests, with estimation procedure identical to the ITT outcomes above.

## C Additional Proofs

### C.1 Proof of Conditional CLT

First, we provide a proof of the conditional CLT under $P$, Theorem A.4 above.

*Proof of Theorem A.4.* First consider the case $d_g = 1$. Define $u_i = a_i - E[a_i|\psi_i]$.

By Lemma A.3 in Cytrynbaum (2024b), since $E[a_i^2] < \infty$ we have $\sqrt{n}E_n[(D_i - p)E[a_i|\psi_i]] = o_p(1)$. Then it suffices to study $\sqrt{n}E_n[(D_i - p)u_i]$. To do so, we will use a martingale difference sequence (MDS) CLT. Fix an ordering $l = 1, \dots, n/k$ of $s(l) \in \mathcal{S}_n$, noting that $|\mathcal{S}_n| \leq n/k$. Define $D_{s(l)} = (D_i)_{i \in s(l)}$. Define $\mathcal{H}_{0,n} = \mathcal{F}_n$ and $\mathcal{H}_{j,n} = \sigma(\mathcal{F}_n, D_{s(l)}, l \in [j])$ for $j \geq 1$. Define $D_{l,n} = n^{-1/2} \sum_{i \in s(l)} (D_i - p)u_i$ and $S_{j,n} = \sum_{i=1}^{j} D_{i,n}$. (1) Claim $(S_{j,n}, \mathcal{H}_{j,n})_{j \geq 1}$ is an MDS. Adaptation is clear.

$$E[(D_i - p)\mathbb{1}(i \in s(j))|\mathcal{H}_{j-1,n}] = E[(D_i - p)\mathbb{1}(i \in s(j))|\mathcal{F}_n, (D_{s(l)})_{l=1}^{j-1}]$$
$$= E[(D_i - p)\mathbb{1}(i \in s(j))|\mathcal{F}_n] = E[(D_i - p)|\mathcal{F}_n]\mathbb{1}(i \in s(j)) = 0.$$

The second equality since $D_{s(j)} \perp\!\!\!\perp (D_{s(l)})_{l \neq j} | \mathcal{F}_n$. Then we compute $E[Z_{j,n}|\mathcal{H}_{j-1,n}] = n^{-1/2} \sum_{i \in s(l)} u_i E[(D_i - p)|\mathcal{H}_{j-1,n}] = 0$. This shows the MDS property. (2). Next, we compute the variance process. By the same argument in (1),

$$\sigma_n^2 \equiv \sum_{j=1}^{n/k} E[Z_{j,n}^2|\mathcal{H}_{j-1,n}] = n^{-1} \sum_{j=1}^{n/k} \left( \sum_{r \neq t \in s(j)} u_r u_t \operatorname{Cov}(D_s, D_t|\mathcal{F}_n) + \sum_{i \in s(j)} u_i^2 \operatorname{Var}(D_i|\mathcal{F}_n) \right)$$

By Lemma C.10 of Cytrynbaum (2024b), we have $\operatorname{Cov}(D_s, D_t|\mathcal{F}_n)\mathbb{1}(s, t \in s(l)) = -l(k-l)/k^2(k-1) \equiv c$ and $\operatorname{Var}(D_i|\mathcal{F}_n) = p - p^2$. Then we may expand $\sigma_n^2$ as

$$cn^{-1} \sum_{j=1}^{n/k} \sum_{r \neq t \in s(j)} u_r u_t + (p - p^2)E_n[u_i^2] \equiv cn^{-1} \sum_{j=1}^{n/k} v_j + (p - p^2)E_n[u_i^2] \equiv T_{n1} + T_{n2}.$$

First consider $T_{n1}$. Our plan is to apply the WLLN in Lemma C.7 of Cytrynbaum (2024b) to show $T_{n1} = o_p(1)$. Define $\mathcal{F}_n^\psi = \sigma(\psi_{1:n}, \pi_n)$ so that $\mathcal{S}_n \in \mathcal{F}_n^\psi$. For $r \neq t$ we have $E[u_r u_t|\psi_{1:n}, \pi_n] = E[u_r E[u_t|\psi_{1:n}, u_r, \pi_n]|\psi_{1:n}, \pi_n] = E[u_r E[u_t|\psi_t]|\psi_{1:n}, \pi_n] = 0$. The second equality follows by applying $(A, B) \perp\!\!\!\perp C \implies A \perp\!\!\!\perp C|B$ with $A = u_t$, $B = \psi_t$ and $C = (\psi_{-t}, u_r, \pi_n)$. Then $E[v_j|\mathcal{F}_n^\psi] = 0$ for $j \in [n/k]$. Next, observe that for any positive constants $(a_k)_{k=1}^m$ we have $\sum_k a_k \mathbb{1}(\sum_k a_k > c) \leq m \sum_k a_k \mathbb{1}(a_k > c/m)$ and $ab\mathbb{1}(ab > c) \leq a^2\mathbb{1}(a^2 > c) + b^2\mathbb{1}(b^2 > c)$. Then for $c_n \to \infty$ with $c_n = o(\sqrt{n})$ we have

$$|v_j|\mathbb{1}(|v_j| > c_n) \leq \sum_{r \neq t \in s(j)} |u_r u_t| \mathbb{1}\left( \sum_{r \neq t \in s(j)} |u_r u_t| > c_n \right)$$
$$\leq k^2 \sum_{r \neq t \in s(j)} |u_r u_t| \mathbb{1}(|u_r u_t| > c_n/k^2) \leq 2k^3 \sum_{r \in s(j)} u_r^2 \mathbb{1}(u_r^2 > c_n/k^2).$$

Then we have

$$n^{-1}E\left[\sum_{j=1}^{n/k} E[|v_j|\mathbb{1}(|v_j|>c_n)|\mathcal{F}_n^{\psi}\right] \leq 2k^3 E_n\left[E\left[u_i^2\mathbb{1}(u_i^2>c_n/k^2)|\psi_{1:n},\pi_n\right]\right] \equiv A_n.$$

Then $E[A_n] = 2k^3E[E_n[E[u_i^2\mathbb{1}(u_i^2 > c_n/k^2)|\psi_i]]] = 2k^3E[u_i^2\mathbb{1}(u_i^2 > c_n/k^2)] \to 0$ as $n \to \infty$. The first equality is by the conditional independence argument above, the second equality is tower law, and the limit by dominated convergence since $E[u_i^2] \leq E[a_i^2] < \infty$ by the contraction property of conditional expectation. Then $A_n = o_p(1)$ by Markov inequality. The conclusion $cn^{-1}\sum_{j=1}^{n/k} v_j = o_p(1)$ now follows by Lemma C.7 of Cytrynbaum (2024b). For $T_{n2}$, we have $E_n[u_i^2] \xrightarrow{p} E[u_i^2] = E[\mathrm{Var}(a|\psi)]$ by vanilla WLLN. Then we have shown $\sigma_n^2 \xrightarrow{p} (p-p^2)E[\mathrm{Var}(a|\psi)]$.

(3) Finally, we show Lindberg $\sum_{j=1}^{n/k} E[Z_{j,n}^2\mathbb{1}(|Z_{j,n}|>\epsilon)|\mathcal{H}_{0,n}] = o_p(1)$.

$$Z_{j,n}^2\mathbb{1}(|Z_{j,n}|>\epsilon) = Z_{j,n}^2\mathbb{1}(Z_{j,n}^2>\epsilon^2) \leq n^{-1}\sum_{r,t\in s(j)}|u_ru_t|\mathbb{1}\left(n^{-1}\sum_{r,t\in s(j)}|u_ru_t|>\epsilon^2\right)$$
$$\leq k^2n^{-1}\sum_{r,t\in s(j)}|u_ru_t|\mathbb{1}\left(|u_ru_t|>n\epsilon^2/k^2\right) \leq k^3n^{-1}\sum_{r\in s(j)}u_r^2\mathbb{1}\left(u_r^2>n\epsilon^2/k^2\right).$$

Then using the inequality above we compute

$$E\left[\sum_{j=1}^{n/k}E[Z_{j,n}^2\mathbb{1}(|Z_{j,n}|>\epsilon)|\mathcal{H}_{0,n}]\right] \leq k^3E\left[n^{-1}\sum_{j=1}^{n/k}\sum_{r\in s(j)}E[u_r^2\mathbb{1}\left(u_r^2>n\epsilon^2/k^2\right)|\mathcal{F}_n^{\psi}]\right]$$
$$= k^3E\left[E_n\left[E[u_i^2\mathbb{1}\left(u_i^2>n\epsilon^2/k^2\right)|\psi_i]\right]\right] = k^3E\left[u_i^2\mathbb{1}\left(u_i^2>n\epsilon^2/k^2\right)\right] = o(1).$$

The first equality by the conditional independence argument above. The second equality by dominated convergence. Then $\sum_{j=1}^{n/k}E[Z_{j,n}^2\mathbb{1}(|Z_{j,n}|>\epsilon)|\mathcal{H}_{0,n}] = o_p(1)$ by Markov. This finishes the proof of the Lindberg condition. Since $\mathcal{H}_{0,n} = \mathcal{F}_n$, by Theorem C.4 in Cytrynbaum (2024b), we have shown that $E[e^{it\sqrt{n}E_n[(D_i-p)a_i]}|\mathcal{F}_n] = \phi(t) + o_p(1)$ for $\phi(t) = e^{-t^2V/2}$ with $V = (p-p^2)E[\mathrm{Var}(a|\psi)]$.

Finally, consider $\dim(a) \geq 1$. Fix $t \in \mathbb{R}^{d_g}$ and let $\bar{a}(W_i) = t'a(W_i) \in \mathbb{R}$. Then we have $X_n(t) \equiv X_n't = E_n[(D_i-p)a(W_i)]'t = E_n[(D_i-p)a(W_i)'t] = E_n[(D_i-p)\bar{a}(W_i)]$. By the previous result $E[e^{iX_n(t)}|\mathcal{F}_n] \xrightarrow{p} e^{-v(t)/2}$ with variance $v(t) = E[\mathrm{Var}(\bar{a}|\psi)] = E[\mathrm{Var}(t'a|\psi)] = t'E[\mathrm{Var}(a|\psi)]t = t'Vt$. Then we have shown $E[e^{it'X_n}|\mathcal{F}_n] = e^{-t'Vt/2} + o_p(1)$ as claimed. □

## C.2 GMM Linearization

This section collects proofs needed for the key linearization result in Lemma A.7. First, define the following curves and objective functions

$$g_0(\theta) = E[\phi(W_i, \theta)], \quad g_n(\theta) = E_n[\phi(W_i, \theta)], \quad \widehat{g}(\theta) = E_n[\phi(W_i, \theta)] + E_n[H_i a(W_i, \theta)].$$
$$H_0(\theta) = g_0(\theta)' M g_0(\theta), \quad H_n(\theta) = g_n(\theta)' M g_n(\theta), \quad \widehat{H}(\theta) = \widehat{g}(\theta)' M_n \widehat{g}(\theta)$$

Define $\widehat{G}(\theta) = (\partial/\partial\theta')\widehat{g}(\theta)$ and $G_n(\theta) = (\partial/\partial\theta')g_n(\theta)$ and $G_0(\theta) = (\partial/\partial\theta')g_0(\theta)$. Define $G = G_0(\theta_0)$. For each $d \in \{0, 1\}$, define $g_d(W, \theta) = g(d, X, S(d), \theta)$.

**Lemma C.1** (ULLN)**.** *Working under $P$ in Definition A.1:*

*(a) If Assumption 3.2(b) holds, $\|\widehat{g} - g_0\|_{\infty,\Theta} = o_p(1)$, $\|g_n - g_0\|_{\infty,\Theta} = o_p(1)$, and $g_0(\theta)$ is continuous. If also $M_n \xrightarrow{p} M$ then $|H_n - H_0|_{\infty,\Theta} = o_p(1)$ and $|\widehat{H} - H_0|_{\infty,\Theta} = o_p(1)$.*

*(b) If Assumption 3.2(c) holds, then there is an open ball $U \subseteq \Theta$ with $\theta_0 \in U$ and $\|\widehat{G}_n - G_0\|_{\infty,U} = o_p(1)$ and $\|G_n - G_0\|_{\infty,U} = o_p(1)$. Also, $G_0(\theta)$ is continuous on $U$ for $G_0(\theta) = \partial/\partial\theta' E[\phi(W, \theta)]$.*

*Proof.* Consider (a). First we show $\|\widehat{g} - g_0\|_{\infty,\Theta} = o_p(1)$, modifying the approach used in the iid setting in Tauchen (1985). It suffices to prove the statement componentwise. Then without loss assume $d_g = 1$ and fix $\epsilon > 0$. Note also that $\phi, a$ are linear combinations of $g_d$ for $d \in \{0, 1\}$, so $\phi$ and $a$ inherit the properties in Assumption 3.2. We have $(\widehat{g} - g_n)(\theta) = E_n[H_i a(W_i, \theta)]$. For each $\theta \in K$ define $U_{\theta m} = B(\theta, m^{-1})$ and $\bar{v}_{\theta m}(D_i, W_i) = \sup_{\bar{\theta} \in U_{\theta m}} H_i a(W_i, \theta)$. Then $\bar{v}_{\theta m}(D_i, W_i)$ may be expanded

$$\begin{aligned}
&\sup_{\bar{\theta} \in U_{\theta m}} H_i a(W_i, \bar{\theta}) = \frac{D_i}{p} \sup_{\bar{\theta} \in U_{\theta m}} a(W_i, \bar{\theta}) + \frac{1 - D_i}{1 - p} \sup_{\bar{\theta} \in U_{\theta m}} -a(W_i, \bar{\theta}) \\
&= \sup_{\bar{\theta} \in U_{\theta m}} a(W_i, \bar{\theta}) + \sup_{\bar{\theta} \in U_{\theta m}} -a(W_i, \bar{\theta}) \\
&+ H_i((1 - p) \sup_{\bar{\theta} \in U_{\theta m}} a(W_i, \bar{\theta}) + p \inf_{\bar{\theta} \in U_{\theta m}} a(W_i, \bar{\theta})) \equiv f_{\theta m}(W_i) + H_i r_{\theta m}(W_i).
\end{aligned}$$

In particular, $E[\bar{v}_{\theta m}(X_i)] = E[f_{\theta m}(W_i)]$. Note both expectations exist by the envelope condition in Assumption 3.2. By continuity at $\theta$, $f_{\theta m}(W_i) \to a(W_i, \theta) - a(W_i, \theta) = 0$ as $m \to \infty$. Also $|f_{m\theta}(W_i)| \lesssim \sup_{\bar{\theta} \in U_{\theta m}} |a(W_i, \bar{\theta})| \leq \sup_{\theta \in \Theta} |a(W_i, \theta)|$. Then by our envelope assumption $\sup_m f_{\theta m}(W_i) \in L_1(P)$, so $\lim_m E[\bar{v}_{\theta m}(D_i, W_i)] = \lim_m E[f_{\theta m}(W_i)] = 0$ by dominated convergence. For each $\theta$, let $m(\theta)$ be such that $E[f_{\theta m(\theta)}(W_i)] \leq \epsilon$. Then $\{U_{\theta m(\theta)} : \theta \in \Theta\}$ is an open cover of $\Theta$, so by compactness

it admits a finite subcover $\{U_{\theta_l,m(\theta_l)}\}_{l=1}^{L(\epsilon)} \equiv \{U_l\}_{l=1}^{L(\epsilon)}$. Next, for each $(\theta, m)$ we claim $E_n[\bar{v}_{\theta m}(D_i, W_i)] = E[f_{\theta m}(W_i)] + o_p(1)$. We have $E_n[f_{\theta m}(W_i)] = E[f_{\theta m}(W_i)] + o_p(1)$ by WLLN since $E[f_{\theta m}(W_i)] < \infty$ as just shown. Similarly, we have

$$|r_{\theta m}(W_i)| = |(1-p) \sup_{\bar{\theta} \in U_{\theta m}} a(W_i, \bar{\theta}) + p \inf_{\bar{\theta} \in U_{\theta m}} a(W_i, \bar{\theta})| \leq \sup_{\bar{\theta} \in U_{\theta m}} |a(W_i, \bar{\theta})| \in L_1(P).$$

Then $E_n[H_i r_{\theta m}(W_i)] = o_p(1)$ by Lemma A.2 in Cytrynbaum (2024b). This proves the claim. Define $f_l(W)$ and $r_l(W)$ to be the functions above evaluated at $(\theta_l, m(\theta_l))$. Putting this all together, we have

$$\begin{aligned}\sup_{\theta \in K} E_n[H_i a(W_i, \theta)] &\leq \max_{l=1}^{L(\epsilon)} \sup_{\theta \in U_l} E_n[H_i a(W_i, \theta)] \leq \max_{l=1}^{L(\epsilon)} E_n[v_{\theta_l m(\theta_l)}(D_i, W_i)] \\ &= \max_{l=1}^{L(\epsilon)} (E[f_{\theta_l m(\theta_l)}(W_i)] + T_{nl}) \leq \epsilon + \max_{l=1}^{L(\epsilon)} T_{nl} = \epsilon + o_p(1).\end{aligned}$$

By symmetry, also $\sup_{\theta \in K} -E_n[H_i a(W_i, \theta)] \leq \epsilon + o_p(1)$. Then $\sup_{\theta \in K} |E_n[H_i a(W_i, \theta)]| \leq 2\epsilon + o_p(1)$. Since $\epsilon > 0$ was arbitrary, this finishes the proof of (1).

Next we show $\|g_n - g_0\|_{\infty,\Theta} = o_p(1)$. We have $(g_n - g_0)(\theta) = E_n[\phi(W_i, \theta)] - E[\phi(W, \theta)]$. Under our assumptions, $|E_n[\phi(W_i, \theta)] - E[\phi(W, \theta)]|_{\infty,\Theta} = o_p(1)$ and $g_0(\theta) = E[\phi(W, \theta)]$ is continuous by Lemma 2.4 of Newey and McFadden (1994). This proves the second claim. The statement about objective functions now follows by algebra, since $|\widehat{H}(\theta) - H_n(\theta)| \lesssim |\widehat{g} - g_n|_{\infty,\Theta}||M_n|_2|\widehat{g}|_{\infty,\Theta} + |g_n|_{\infty,\Theta}|M_n - M|_2|\widehat{g}|_{\infty,\Theta} + |g_n|_{\infty,\Theta}|M|_2|\widehat{g} - g_n|_{\infty,\Theta}$. We have $|g_n|_{\infty,\Theta}, |\widehat{g}|_{\infty,\Theta} = o_p(1) + |g_0|_{\infty,\Theta} = O_p(1)$ since $|g_0|_{\infty,\Theta} \leq E[\sup_{\theta \in \Theta} \phi(W, \theta)] < \infty$. Also $|M_n|_2 = O_p(1)$ and $|M_n - M|_2 = o_p(1)$ by continuous mapping. Taking $\sup_{\theta \in \Theta}$ on both sides gives the result. The proof that $|H_n - H_0|_{\infty,K} = o_p(1)$ is identical. By triangle inequality, this proves the claim. The proof of (2) is similar. □

**Lemma C.2** (Consistency)**.** *Under the distribution $P$ in Definition A.1, if Assumption 3.2 holds then $\widehat{\theta} - \theta_0 = o_p(1)$ and $\theta_n - \theta_0 = o_p(1)$.*

*Proof.* By definition, $\widehat{\theta} = \text{argmin}_{\theta \in \Theta} \widehat{H}(\theta)$. Moreover, $g_n(\theta_n) = 0$ so $H_n(\theta_n) = 0$ and $\theta_n \in \text{argmin}_{\theta \in \Theta} H_n(\theta)$. For (2), since $g_0(\theta_0) = 0$ uniquely and $\text{rank}(M) = d_g$, then $H_0(\theta)$ is uniquely minimized at $\theta_0$. Then by uniform convergence of $\widehat{H}, H_n$ to $H_0$, extremum consistency (e.g. Theorem 2.1 in Newey and McFadden (1994)) implies that $\theta_n \xrightarrow{p} \theta_0$ and $\widehat{\theta} \xrightarrow{p} \theta_0$. □

*Proof of Lemma A.7.* By Lemma A.3, it suffices to show the result under $P$ in Definition A.1. Since $\widehat{\theta} = \text{argmin}_{\theta \in \Theta} \widehat{H}(\theta)$, we have $\nabla_\theta \widehat{H}(\widehat{\theta}) = 0$, which is

$\widehat{G}(\widehat{\theta})' M_n \widehat{g}(\widehat{\theta}) = 0$. By differentiability in Assumption 3.2 and applying Taylor's Theorem componentwise, for each $k \in [d_g]$ and some $\tilde{\theta}_k \in [\theta_0, \widehat{\theta}]$ we have $\widehat{g}(\widehat{\theta}) = \widehat{g}(\theta_0) + \frac{\partial \widehat{g}_k}{\partial \theta'}(\tilde{\theta}_k)_{k=1}^{d_g}(\widehat{\theta} - \theta_0)$. Arguing exactly as in Newey and McFadden (1994), we find $\sqrt{n}(\widehat{\theta} - \theta_0) = -(G'MG)^{-1}G'M\sqrt{n}\widehat{g}(\theta_0) + o_p(1) = \Pi\sqrt{n}\widehat{g}(\theta_0) + o_p(1)$. This proves the second statement of Lemma A.7. For the first statement, we substitute $\theta_n, H_n, G_n$ for $\widehat{\theta}, \widehat{H}, \widehat{G}$ in the Newey and McFadden (1994) argument, obtaining $\sqrt{n}(\theta_n - \theta_0) = \Pi\sqrt{n}g_n(\theta_0) + o_p(1)$. Then we have $\sqrt{n}(\widehat{\theta} - \theta_n) = \sqrt{n}(\widehat{\theta} - \theta_0 + \theta_0 - \theta_n) = \Pi\sqrt{n}(\widehat{g}(\theta_0) - g_n(\theta_0)) + o_p(1) = \Pi\sqrt{n}E_n[H_i a(W_i, \theta_0)] + o_p(1)$. This finishes the proof. □

## C.3 Nonlinear Rerandomization

*Proof of Theorem 4.3.* We first prove a slightly more general result, allowing for over-identified GMM estimation with positive definite weighting matrix $\Delta_n \xrightarrow{p} \Delta$. For $|x|_{2,A}^2 = x'Ax$, define $\widehat{\beta}_d \in \operatorname{argmin}_{\beta \in \mathbb{R}^{d_\beta}} |E_n[\mathbb{1}(D_i = d)m(X_i, \beta)]|_{2,\Delta_n}^2$. Define $g^1(D, X, \beta) = Dm(X, \beta)$ and $g^0(D, X, \beta) = (1 - D)m(X, \beta)$. Under the expansion in Equation 3.1, we have $\phi^1(X, \beta) = pg^1(1, X, \beta) = pm(X, \beta)$ and $a^1(X, \beta) = v_D g^1(1, X, \beta) = v_D m(X, \beta)$. Similarly, $\phi^0(X, \beta) = (1 - p)g^0(0, X, \beta) = (1 - p)m(X, \beta)$ and $a^0(X, \beta) = -v_D g^0(0, X, \beta) = -v_D m(X, \beta)$. Note that $E[g^1(D, X, \beta)] = pE[m(X, \beta)]$ and $E[g^0(D, X, \beta)] = (1 - p)E[m(X, \beta)]$, so the GMM parameters $\beta_1 = \beta_0 = \beta^*$, where $\beta^*$ uniquely solves $E[m(X, \beta^*)] = 0$. Let $G_m = E[(\partial/\partial\beta')m(X, \beta^*)]$, full rank by assumption. Then $G^1 = E[(\partial/\partial\beta')g^1(D, X, \beta^*)] = pE[(\partial/\partial\beta')m(X, \beta^*)] = pG_m$ and $\Pi^1 = -((G^1)'\Delta G^1)^{-1}(G^1)'\Delta = -p^{-1}(G_m'\Delta G_m)^{-1}G_m'\Delta \equiv p^{-1}\Pi_m$. By symmetry, we have $\Pi^0 = (1 - p)^{-1}\Pi_m$. Observe that

$$
\begin{aligned}
(\Pi^1\phi^1 - \Pi^0\phi^0)(X, \beta) &= p^{-1}\Pi_m pm(X, \beta) - (1 - p)^{-1}\Pi_m(1 - p)m(X, \beta) = 0, \\
(\Pi^1 a^1 - \Pi^0 a^0)(X, \beta) &= p^{-1}\Pi_m v_D m(X, \beta) - (1 - p)^{-1}\Pi_m v_D(-m(X, \beta)) \\
&= (1 - p)\Pi_m m(X, \beta) + p\Pi_m m(X, \beta) = \Pi_m m(X, \beta).
\end{aligned}
$$

Then applying Lemma A.7 to GMM estimation using $g^1$ and $g^0$, under the measure $P$ in Definition A.1 we have

$$
\begin{aligned}
&\sqrt{n}(\widehat{\beta}_1 - \widehat{\beta}_0) = \sqrt{n}(\widehat{\beta}_1 - \beta^* - (\widehat{\beta}_0 - \beta^*)) = \sqrt{n}\Pi^1 E_n[\phi^1(X_i, \beta^*) + H_i a^1(X_i, \beta^*)] \\
&- \sqrt{n}\Pi^0 E_n[\phi^0(X_i, \beta^*) + H_i a^0(X_i, \beta^*)] + o_p(1) = \sqrt{n}\Pi_m E_n[H_i m(X, \beta^*)] + o_p(1).
\end{aligned}
$$

Then Definition 4.1 is an example of Definition 2.1 with $\mathcal{I}_n = \sqrt{n}E_n[H_i h_i] + o_p(1)$ for $h_i = \Pi_m m(X_i, \beta^*)$. Then Theorem 3.5 holds with $h_i = \Pi_m m(X_i, \beta^*)$. Consider

the exactly identified case, so $\Pi_m = -G_m^{-1}$ and $h_i = -G_m^{-1}m(X_i, \beta^*)$. Then by Theorem 3.5, $\sqrt{n}(\widehat{\theta} - \theta_n)|W_{1:n} \Rightarrow \mathcal{N}(0, V_a) + R_A$. Denote $\Pi a = \Pi a(W, \theta_0)$ and $m = m(X, \beta^*)$. Then the rerandomization coefficient $\gamma_0$ is

$$\begin{aligned}
&\gamma_0 = E[\mathrm{Var}(h|\psi)]^{-1}E[\mathrm{Cov}(h, \Pi a|\psi)] = -E[\mathrm{Var}(G_m^{-1}m|\psi)]^{-1}E[\mathrm{Cov}(G_m^{-1}m, \Pi a|\psi)] \\
&= -E[G_m^{-1}\,\mathrm{Var}(m|\psi)(G_m^{-1})']^{-1}E[G_m^{-1}\,\mathrm{Cov}(m, \Pi a|\psi)] \\
&= -G_m' E[\mathrm{Var}(m|\psi)]^{-1}E[\mathrm{Cov}(m, \Pi a|\psi)].
\end{aligned}$$

Then $V_a = v_D^{-1}E[\mathrm{Var}(\Pi a - \gamma_0'(-G_m^{-1}m)|\psi)] = v_D^{-1}E[\mathrm{Var}(\Pi a - \gamma_0' m)|\psi)]$, where

$$\gamma_0 = \underset{\gamma \in \mathbb{R}^{d_\beta \times d_\theta}}{\mathrm{argmin}}\, v_D^{-1}E[\mathrm{Var}(\Pi a - \gamma' m|\psi)].$$

From above, we have $\gamma_0 = -G_m'\gamma_0$. Then the residual term

$$\begin{aligned}
R_A &\sim \gamma_0' Z_h \,|\, Z_h \in A \sim -\gamma_0' G_m Z_h \,|\, Z_h \in A \sim -\gamma_0' G_m Z_h \,|\, (-G_m^{-1})(-G_m)Z_h \in A \\
&\sim \gamma_0' Z_m \,|\, -G_m^{-1} Z_m \in A \sim \gamma_0' Z_m \,|\, Z_m \in -G_m A.
\end{aligned}$$

$Z_h \sim \mathcal{N}(0, v_D^{-1}E[\mathrm{Var}(h|\psi)])$, so $Z_m = G_m Z_h \sim \mathcal{N}(0, v_D^{-1}G_m E[\mathrm{Var}(h|\psi)]G_m') \sim \mathcal{N}(0, v_D^{-1}E[\mathrm{Var}(G_m h|\psi)]) \sim \mathcal{N}(0, v_D^{-1}E[\mathrm{Var}(m|\psi)])$ since $G_m h = G_m G_m^{-1} m = m(X, \beta^*)$. Summarizing, we have shown $V_a = v_D^{-1}E[\mathrm{Var}(\Pi a - \gamma_0' m|\psi)]$ and $R_A \sim \gamma_0' Z_m \,|\, Z_m \in G_m A$ for $Z_m \sim \mathcal{N}(0, v_D^{-1}E[\mathrm{Var}(m|\psi)])$.

For the corollary, consider letting $\widehat{\beta} \in \mathrm{argmin}_{\beta \in \mathbb{R}^{d_\beta}} |E_n[m(X_i, \beta)]|_{2,\Delta_n}^2$. Relative to the expansion in Equation 3.1, $a_m(X_i, \beta) = 0$ and $\phi_m(X_i, \beta) = m(X_i, \beta)$, with linearization matrix $\Pi_m$ as above. Then by Lemma A.7 $\sqrt{n}(\widehat{\beta} - \beta^*) = \Pi_m E_n[m(X_i, \beta^*)] + o_p(1) = O_p(1)$. Consider setting $h_i = m(X_i, \widehat{\beta})$. By the mean value theorem, $m(X_i, \widehat{\beta}) - m(X_i, \beta^*) = \frac{\partial m(X_i, \tilde{\beta}_i)}{\partial \beta}(\widehat{\beta} - \beta^*)$, where the $\tilde{\beta}_i \in [\beta^*, \widehat{\beta}]$ may change by row. Then we have

$$\sqrt{n}E_n[H_i m(X_i, \widehat{\beta})] - \sqrt{n}E_n[H_i m(X_i, \beta^*)] = E_n[H_i(\partial/\partial\beta')m(X_i, \tilde{\beta}_i)]\sqrt{n}(\widehat{\beta} - \beta^*).$$

We claim that $E_n[H_i(\partial/\partial\beta')m(X_i, \tilde{\beta}_i)] = o_p(1)$. Let $U$ open s.t. $E[\sup_{\beta \in U} |m(X_i, \beta)|_F] < \infty$ and define $S_n = \{\widehat{\beta} \in U\}$. Then by consistency $E_n[H_i(\partial/\partial\beta')m(X_i, \tilde{\beta}_i)]\mathbb{1}(S_n^c) = o_p(1)$. Define $v_{ijk}^n = \mathbb{1}(S_n)((\partial/\partial\beta')m(X_i, \tilde{\beta}_i))_{jk}$. By the definition of $\widehat{\beta}$, clearly $v_{ijk}^n \in \mathcal{F}_n = \sigma(W_{1:n}, \pi_n)$. Moreover, we have $|v_{ijk}^n| \leq \sup_{\beta \in U} |(\partial/\partial\beta')m(X_i, \beta)|_F \in L_1$ by definition of $S_n$ and $\tilde{\beta}_i \in [\beta^*, \widehat{\beta}]$ for each $n$, so by domination $(v_{ijk}^n)_n$ is uniformly integrable, so $E_n[H_i v_{ijk}^n] = o_p(1)$ by Lemma A.2 of Cytrynbaum (2024b). This proves the claim, showing that $\mathcal{I}_n = \sqrt{n}E_n[H_i m(X_i, \widehat{\beta})] = \sqrt{n}E_n[H_i m(X_i, \beta^*)] +$

$o_p(1)$. Note that $\mathcal{I}_n \in \widehat{G}_m A \iff \widehat{G}_m^{-1}\mathcal{I}_n \in A$ and $\widehat{G}_m^{-1}\mathcal{I}_n = \widehat{G}_m^{-1}\sqrt{n}E_n[H_i m(X_i,\beta^*)] + o_p(1) = G_m^{-1}\sqrt{n}E_n[H_i m(X_i,\beta^*)] + o_p(1)$. The result follows from Theorem 3.5. □

**Assumption C.3** (Propensity Rerandomization)**.** *Impose the following:*

*(a) Let $L$ be twice differentiable, with $|L'|_\infty, |L''|_\infty < \infty$. For each $p \in (0,1)$, there is a unique $c$ with $L(c) = p$. Also, $|L'(c)| > 0$.*

*(b) The score $m(D_i, X_i, \beta) = D_i \frac{L'(X_i'\beta)X_i}{L(X_i'\beta)} - (1-D_i)\frac{L'(X_i'\beta)X_i}{1-L(X_i'\beta)}$ satisfies condition 3.2. The solution to Equation 4.3 exists.*

*(c) $X = (1, h)$ for $E[|h|_2^2] < \infty$. Also, $E[\mathrm{Var}(h|\psi)]$, $\mathrm{Var}(h)$ are full rank.*

*Proof of Theorem 4.7.* By assumption, $\widehat{\beta}$ is a GMM estimator for $m(D_i, X_i, \beta) = D_i \frac{L'(X_i'\beta)X_i}{L(X_i'\beta)} - (1-D_i)\frac{L'(X_i'\beta)X_i}{1-L(X_i'\beta)}$. Let $c$ such that $L(c) = p$. Then $\beta^* = (c, 0)$ has $E[m(D, X, \beta^*)] = E[H_i L'(c) X_i] = 0$. Relative to the decomposition in Equation 3.1, we have $\phi(X,\beta) = p\frac{L'(X_i'\beta)X_i}{L(X_i'\beta)} - (1-p)\frac{L'(X_i'\beta)X_i}{1-L(X_i'\beta)}$ and $a(X,\beta) = v_D(\frac{L'(X_i'\beta)X_i}{L(X_i'\beta)} + \frac{L'(X_i'\beta)X_i}{1-L(X_i'\beta)})$. Since $L(X_i'\beta^*) = L(c) = p$, apparently we have $\phi(X,\beta^*) = 0$ and $a(X,\beta^*) = L'(c)X_i$. It's easy to see $\mathrm{Var}(h) \succ 0$ implies $E[XX'] \succ 0$ for $X = (1,h)$. A calculation shows that $G_m = E[\frac{\partial}{\partial\beta'}\phi(X,\beta^*)] = -v_D^{-1}L'(c)^2 E[X_i X_i']$, so $\Pi_m = -G_m^{-1} = \frac{v_D}{L'(c)^2}E[X_iX_i']^{-1}$. By Lemma A.7, we have shown

$$\begin{aligned}\sqrt{n}(\widehat{\beta} - \beta^*) &= \sqrt{n}\Pi_m E_n[\phi(X_i,\beta^*) + H_i a(X_i,\beta^*)] + o_p(1) \\ &= v_D \frac{\sqrt{n}}{L'(c)} E[X_iX_i']^{-1} E_n[H_iX_i] + o_p(1).\end{aligned}$$

Consider rerandomizing until $\mathcal{J}_n = nE_n[(p - L(X_i'\widehat{\beta}))^2] \le \epsilon^2$. Then for $\beta^*$ s.t. $L(x'\beta^*) = p$, the above quantity is $nE_n[(L(X_i'\widehat{\beta}) - L(X_i'\beta^*))^2]$. By Taylor's Theorem, $L(X_i'\widehat{\beta}) - L(X_i'\beta^*) = L'(\xi_i)(X_i'\widehat{\beta} - X_i'\beta^*) = L'(\xi_i)X_i'(\widehat{\beta} - \beta^*)$ for some $\xi_i \in [X_i'\beta^*, X_i'\widehat{\beta}]$. Then we have

$$\mathcal{J}_n = n(\widehat{\beta} - \beta^*)' E_n[X_iX_i'L'(\xi_i)^2](\widehat{\beta} - \beta^*).$$

Claim that $E_n[X_iX_i'L'(\xi_i)^2] = E_n[X_iX_i'L'(X_i'\beta^*)^2] + o_p(1)$. If so, then $E_n[X_iX_i'L'(\xi_i)^2] = L'(c)^2E_n[X_iX_i'] + o_p(1) = L'(c)^2E[X_iX_i'] + o_p(1)$. To see this, note that $|L'(X_i'\beta^*)^2 - L'(\xi_i)^2| = |L'(X_i'\beta^*) - L'(\xi_i)||L'(X_i'\beta^*) + L'(\xi_i)| \le 2|L'|_\infty|L''|_\infty|X_i'\beta^* - \xi_i|_2 \lesssim |X_i'\beta^* - X_i'\widehat{\beta}|_2 \le |X_i|_2|\beta^* - \widehat{\beta}|_2$. Then we have

$$\begin{aligned}&|E_n[X_iX_i'L'(\xi_i)^2] - E_n[X_iX_i'L'(X_i'\beta^*)^2]|_2 \le E_n[|X_i|_2^2|L'(X_i'\beta^*)^2 - L'(\xi_i)^2|] \\ &\lesssim E_n[|X_i|_2^3]|\beta^* - \widehat{\beta}|_2 = o_p(1)\end{aligned}$$

The last equality if $E_n[|X_i|_2^3] = o_p(n^{1/2})$. Note that $E_n[|X_i|_2^3] \leq E_n[|X_i|_2^2] \max_{i=1}^n |X_i|_2 = O_p(1)o_p(n^{1/2})$ since $E[|X_i|_2^2] < \infty$ by assumption, using Lemma C.8 of Cytrynbaum (2024b). Using the claim, $\sqrt{n}(\widehat{\beta} - \beta^*) = O_p(1)$, and the linear expansion of $\sqrt{n}(\widehat{\beta} - \beta^*)$, $\mathcal{J}_n = L'(c)^2 n(\widehat{\beta} - \beta^*)' E[X_i X_i'](\widehat{\beta} - \beta^*) + o_p(1)$, which is

$$\begin{aligned}
&= v_D^2 L'(c)^2 (L'(c)^{-1} E[X_i X_i']^{-1} \sqrt{n} E_n[H_i X_i])' E[X_i X_i'] \\
&\times (L'(c)^{-1} E[X_i X_i']^{-1} \sqrt{n} E_n[H_i X_i]) + o_p(1) \\
&= v_D^2 \sqrt{n} E_n[H_i X_i]' E[X_i X_i']^{-1} \sqrt{n} E_n[H_i X_i] + o_p(1).
\end{aligned}$$

Note $E_n[H_i] = O_p(n^{-1})$ by stratification. Since $X = (1, h)$, $\sqrt{n}E_n[H_i X_i]' = (0, \sqrt{n}E_n[H_i h_i]') + O_p(n^{-1/2})$. Also, by block inversion $(E[X_i X_i']^{-1})_{hh} = \mathrm{Var}(h_i)^{-1}$. For some $\xi_n = o_p(1)$

$$\begin{aligned}
\mathcal{J}_n &= v_D^2 (0, \sqrt{n}E_n[H_i h_i]') E[X_i X_i']^{-1} (0, \sqrt{n}E_n[H_i h_i]')' + o_p(1) \\
&= v_D^2 \sqrt{n} E_n[H_i h_i]' (E[X_i X_i']^{-1})_{hh} \sqrt{n} E_n[H_i h_i] + o_p(1) \\
&= v_D^2 \sqrt{n} E_n[H_i h_i]' \mathrm{Var}(h_i)^{-1} \sqrt{n} E_n[H_i h_i] + \xi_n.
\end{aligned}$$

Define the function $b(x, y) = v_D^2 x' \mathrm{Var}(h)^{-1} x + y - \epsilon$. Then $\mathcal{J}_n \leq \epsilon \iff b(\mathcal{I}_n, \xi_n) \leq 0$ for $\mathcal{I}_n = \sqrt{n}E_n[H_i h_i]$ and $\xi_n \xrightarrow{p} 0$. Clearly, $x \to b(x, 0)$ is continuous. Also note $E[|h|_2^2] < \infty$ by assumption. Finally, for $Z_h \sim \mathcal{N}(0, E[\mathrm{Var}(h|\psi)])$, have $P(b(Z_h, 0) = 0) = P(Z_h' \mathrm{Var}(h)^{-1} Z_h = \epsilon^2) = 0$ since $E[\mathrm{Var}(h|\psi)]$ is full rank. Then this rerandomization satisfies all the conditions in Definition A.2. By Lemma A.7, the GMM estimator $\sqrt{n}(\widehat{\theta} - \theta_0) = \sqrt{n}E_n[H_i \Pi a(W_i, \theta_0)] + o_p(1)$ under this rerandomization. By Theorem A.6, have $\sqrt{n}E_n[H_i \Pi a(W_i)]|\mathcal{F}_n \Rightarrow \mathcal{N}(0, V_a) + R$ with residual variable

$$R \sim \gamma_0' Z_h | Z_h \in T \sim \gamma_0' Z_h | v_D^2 \cdot Z_h' \mathrm{Var}(h)^{-1} Z_h \leq \epsilon$$

for acceptance region $T = \{x : b(x, 0) \leq 0\} = \{x : v_D^2 \cdot x' \mathrm{Var}(h)^{-1} x \leq \epsilon\}$ and

$$V_a = \min_{\gamma \in \mathbb{R}^{d_h \times d_\theta}} v_D^{-1} E[\mathrm{Var}(\Pi a(W) - \gamma' h | \psi)].$$

This finishes the proof. □

## C.4 Covariate Adjustment

*Proof of Proposition 6.2.* Since $\widehat{\theta}_{adj} = \widehat{\theta} - E_n[H_i \widehat{\alpha}' w_i]$ for $\widehat{\alpha} \xrightarrow{p} \alpha$ and $E_n[H_i w_i] = O_p(n^{-1/2})$ by Theorem A.4, then $\widehat{\theta}_{adj} = \widehat{\theta} - E_n[H_i \alpha' w_i] + o_p(n^{-1/2}) = E_n[H_i(\Pi a(W_i, \theta_0) -$

$\alpha' w_i)] + o_p(n^{-1/2})$, the final equality by Lemma A.7. The first statement now follows from Slutsky and Theorem A.4. The second statement follows by the same argument used in the proof of Corollary 3.7. □

*Proof of Theorem 6.3.* By the same argument in the proof of Proposition 6.2, we have $\widehat{\theta}_{adj} = E_n[H_i(\Pi a(W_i, \theta_0) - \alpha_0' w_i)] + o_p(n^{-1/2})$. Then by Theorem A.6, $\sqrt{n}(\widehat{\theta}_{adj} - \theta_n)|\mathcal{F}_n \Rightarrow \mathcal{N}(0, V) + R$, independent with

$$V = v_D^{-1} E[\mathrm{Var}(\Pi a(W) - \alpha_0' w - \beta_0' h|\psi)] = \min_{\beta \in \mathbb{R}^{d_h \times d_\theta}} v_D^{-1} E[\mathrm{Var}(\Pi a(W) - \alpha_0' w - \beta' h|\psi)].$$

The residual term $R \sim \beta_0' Z_h \,|\, Z_h \in A$. Then it suffices to show that $\beta_0 = 0$. Define $a_{\Pi\alpha} = \Pi a(W, \theta_0) - \alpha_0' w$. By Lemma C.10, it further suffices to show $\beta_0 = 0$ solves $E[\mathrm{Var}(h|\psi)]\beta_0 = E[\mathrm{Cov}(h, a_{\Pi\alpha}|\psi)]$, i.e. that $E[\mathrm{Cov}(h, a_{\Pi\alpha}|\psi)] = 0$. To do so, note that $E[\mathrm{Cov}(h, a_{\Pi\alpha}|\psi)] = E[\mathrm{Cov}(h, (\Pi a - \alpha_0' w)|\psi)] = E[\mathrm{Cov}(h, \Pi a|\psi)] - E[\mathrm{Cov}(h, w|\psi)]\alpha_0$. By assumption, $E[\mathrm{Var}(w|\psi)]\alpha_0 = E[\mathrm{Cov}(w, \Pi a|\psi)]$. Since $h \subseteq w$, we have

$$\begin{aligned} E[\mathrm{Cov}(h, w|\psi)]\alpha_0 &= (E[\mathrm{Var}(w|\psi)])_{hw}\alpha_0 = (E[\mathrm{Var}(w|\psi)]\alpha_0)_{h\theta} \\ &= (E[\mathrm{Cov}(w, \Pi a|\psi)])_{h\theta} = E[\mathrm{Cov}(h, \Pi a|\psi)] \end{aligned}$$

This shows that $[\mathrm{Cov}(h, a_{\Pi\alpha}|\psi)] = 0$, so $\beta_0 = 0$ is a solution, proving the claim. This finishes the proof of the statement for $\theta_n$. The result for $\theta_0$ follows as in Corollary 3.7. □

In Section 7, we defined $\beta_d = E[\mathrm{Var}(w|\psi)]^{-1} E[\mathrm{Cov}(w, v_D \Pi g_d(W, \theta_0)|\psi)]$ and estimator $\widehat{\beta}_d = v_D E_n[\check{w}_i \check{w}_i']^{-1} \mathrm{Cov}_n(\check{w}_i, \widehat{\Pi}\widehat{g}_i | D_i = d)$. By definition $\alpha_0 = \beta_1 - \beta_0$ and $\widehat{\alpha} = \widehat{\beta}_1 - \widehat{\beta}_0$. Then for Theorem 6.4, apparently it suffices to show $\widehat{\beta}_d = \beta_d + o_p(1)$.

**Theorem C.4** (Adjustment Coefficients)**.** *Suppose $D_{1:n}$ is as in Definition 2.1. Require 3.1, 3.2. Assume that $E[\mathrm{Var}(w|\psi)] \succ 0$. Then $\widehat{\beta}_d = \beta_d + o_p(1)$ for $d = 0, 1$.*

*Proof of Theorem C.4.* By Lemma A.3, it suffices to show the result under $P$ in Definition A.1. First consider estimating $\beta_1$. By Lemma C.5, $E_n[\check{w}_i \check{w}_i'] = k^{-1}(k-1)E[\mathrm{Var}(w|\psi)] + o_p(1)$. Then if $E[\mathrm{Var}(w|\psi)] \succ 0$, $E_n[\check{w}_i \check{w}_i']^{-1} \xrightarrow{p} k(k-1)^{-1} E[\mathrm{Var}(w|\psi)]^{-1}$ by continuous mapping. $\widehat{\Pi} \xrightarrow{p} \Pi$ by assumption. Then it suffices to show

$$\begin{aligned} \mathrm{Cov}_n(\check{w}_i, \widehat{g}_i | D_i = 1) &= E_n[(D_i/p)\check{w}_i \widehat{g}_i'] - E_n[\check{w}_i | D_i = 1] E_n[\widehat{g}_i | D_i = 1] \\ &= \frac{k-1}{k} E[\mathrm{Cov}(w, g_1(\theta_0)|\psi)] + o_p(1). \end{aligned}$$

That $E_n[\check{w}_i|D_i=1]=o_p(1)$ can be shown similar to Lemma C.5 below. Then consider the first term. First, claim that $E_n[(D_i/p)\check{w}_i\widehat{g}_i']=E_n[(D_i/p)\check{w}_i g_i']+o_p(1)$. By Taylor's theorem, $|g_i(\widehat{\theta})-g_i(\theta_0)|_2 \leq |\frac{\partial g_i}{\partial \theta'}(\tilde{\theta}_i)|_2|\widehat{\theta}-\theta_0|_2$, where $\tilde{\theta}_i \in [\theta_0,\widehat{\theta}]$ may change by row. Then using $|xy'|_2 \leq |x|_2|y|_2$, we have $|E_n[(D_i/p)\check{w}_i(g_i(\widehat{\theta})-g_i(\theta_0))']|_2 \leq E_n[|\check{w}_i|_2|g_i(\widehat{\theta})-g_i(\theta_0)|_2] \leq |\widehat{\theta}-\theta_0|_2 E_n[|\check{w}_i|_2|\frac{\partial g_i}{\partial \theta'}(\tilde{\theta}_i)|_2] \leq |\widehat{\theta}-\theta_0|_2(E_n[|\check{w}_i|_2^2]+E_n[|\frac{\partial g_i}{\partial \theta'}(\tilde{\theta}_i)|_2^2])$ by Young's inequality. We showed $E_n[|\frac{\partial g_i}{\partial \theta'}(\tilde{\theta}_i)|_2^2]=O_p(1)$ in the proof of Lemma C.7. Similarly, $E_n[|\check{w}_i|_2^2] \leq E_n[|w_i|_2^2]=O_p(1)$ by the bound in Lemma C.5. Since $|\widehat{\theta}-\theta_0|_2=o_p(1)$ by Theorem 3.5, this proves the claim. Next, we calculate

$$\begin{aligned} E_n[(D_i/p)\check{w}_i g_i'] &= E_n[(D_i/p)\check{w}_i g_{1i}'] = p^{-1}E_n[(D_i-p)\check{w}_i g_{1i}'] + E_n[\check{w}_i g_{1i}'] \\ &= E_n[\check{w}_i g_{1i}'] + o_p(1) = k^{-1}(k-1)E[\mathrm{Cov}(w, g_{1i}|\psi)] + o_p(1). \end{aligned}$$

The first equality since $g_{1i}=g(1,R_i,S_i(1),\theta_0)$. The third and fourth equalities by Lemma C.5, since $E[|w|_2^2+|g_1|_2^2]<\infty$ Then we have shown $\widehat{\beta}_1 \xrightarrow{p} \beta_1$, and $\widehat{\beta}_0 \xrightarrow{p} \beta_0$ by symmetry. □

**Lemma C.5.** *Let $E[w_i^2+v_i^2]<\infty$ with $w_i, v_i \in \sigma(W_i)$. Then under $P$ in Definition A.1, $E_n[(D_i-p)\check{w}_i\check{v}_i]=o_p(1)$ and $E_n[(D_i-p)\check{w}_i v_i]=o_p(1)$. Also $E_n[\check{w}_i\check{v}_i]=\frac{k-1}{k}E[\mathrm{Cov}(w,v|\psi)]+o_p(1)$.*

We omit the proof, since this is a slight restatement of Lemma A.8 in Cytrynbaum (2024a).

## C.5 Acceptance Region Optimization

*Proof of Proposition 5.1.* First we prove part (a). Define the function $f(a)=\sup_{b\in B}|b'a|$. As the sup of linear functions, $f$ is convex (e.g. Rockafellar (1996)). Then the sublevel set $A \equiv \{a : f(a) \leq 1\}$ is convex. Note that $f(a)=f(-a)$, so $A$ is symmetric. For the main statement of the theorem, let $a_n=\sqrt{n}E_n[H_i h_i]$. Clearly, $f$ is positive homogeneous, i.e. $f(\lambda a)=\lambda f(a)$ for $\lambda \geq 0$. Then note that the LHS event occurs iff $f(a_n) \leq \epsilon \iff f(a_n/\epsilon) \leq 1 \iff a_n/\epsilon \in A \iff a_n \in \epsilon \cdot A$. This proves the main statement. Suppose $B$ is bounded. Then by Cauchy-Schwarz $f(a) \leq |a|_2 \sup_{b\in B}|b|_2 < \infty$ for any $a \in \mathbb{R}^{d_h}$. Then $f$ is a proper function, so $f$ is continuous by Corollary 10.1.1. of Rockafellar (1996). Then $A=f^{-1}([0,1])$ is closed. Moreover, the open set $f^{-1}((1/3,2/3)) \subseteq f^{-1}([0,1])=A$, so $A$ has non-empty interior. Suppose that $B$ is open. Then $B$ contains an open ball $B(x,\delta)$ for some $x \in \mathbb{R}^{d_h}$ and $\delta>0$. Fix $a \in \mathbb{R}^{d_h}$ and define $b(a)=x+\mathrm{sgn}(a'x)\frac{\delta}{2|a|}a$. By assumption, $b(a) \in B$. Then $f(a)=$

$\sup_{b\in B}|b'a| \geq |b(a)'a| = |a'x + \operatorname{sgn}(a'x)(\delta/2)|a|| = |a'x| + (\delta/2)|a| \geq (\delta/2)|a|$. Then $f(a) = \sup_{b\in B}|a'b| \geq (\delta/2)|a|$, so $A \subseteq B(0, 2/\delta)$. □

*Proof of Theorem 5.5.* First we show the set $A_0$ is feasible in Equation 5.3. We have $L_{\gamma A} = T_\gamma + \gamma' Z_{hA}$, where $T_\gamma \sim \mathcal{N}(0, V(\gamma))$ and $T_\gamma \perp\!\!\!\perp Z_{hA}$. Then $\operatorname{bias}(L_{\gamma A}|Z_h) = E[L_{\gamma A}|Z_{hA}] = E[T_\gamma|Z_{hA}] + \gamma' Z_{hA} = \gamma' Z_{hA}$. For $A_0 = \epsilon B^\circ$, $\sup_{\gamma\in B}|\operatorname{bias}(L_{\gamma A}|Z_h)| = \sup_{\gamma\in B}|\gamma' Z_{hA}|$. Note $Z_{hA} \in \epsilon B^\circ$, so $Z_{hA}/\epsilon \in B^\circ$. Then we have

$$\sup_{\gamma\in B}|\gamma' Z_{hA}| \leq \epsilon \cdot \sup_{b\in B^\circ}\sup_{\gamma\in B}|\gamma' b| \leq \epsilon \cdot 1.$$

The final inequality by definition of $B^\circ$. This shows that $A_0$ is feasible. We claim $A_0$ is optimal. Suppose for contradiction that there exists $A \subseteq \mathbb{R}^{d_h}$ with $\operatorname{Leb}(A \triangle A_0) \neq 0$ and $P(Z_h \in A) > P(Z_h \in A_0)$. Clearly $A \not\subseteq A_0$. Then $\operatorname{Leb}(A \setminus A_0) > 0$, so $P(Z_h \in A \setminus A_0) > 0$ by absolute continuity. For any $x \in A \setminus A_0 \subseteq (\epsilon B^\circ)^c$, we must have $\sup_{\gamma\in B}|\gamma' x| > \epsilon$. Then $\{\sup_{\gamma\in B}\operatorname{bias}(L_{\gamma A}|Z_h) > \epsilon\} = \{\sup_{\gamma\in B}|Z_{hA}| > \epsilon\} \supseteq \{Z_{hA} \in A \setminus A_0\}$. $B$ is totally bounded by assumption, so as in the proof of Proposition 5.1, we have $\sup_{\gamma\in B}|Z_{hA}| = p_B(Z_{hA})$ for $p_B$ continuous. Then the event $\{\sup_{\gamma\in B}|Z_{hA}| > \epsilon\} = \{p_B(Z_{hA}) > \epsilon\}$ is measurable. Then note $P(\sup_{\gamma\in B}\operatorname{bias}(L_{\gamma A}|Z_h) > \epsilon) \geq P(Z_h \in A \setminus A_0) > 0$, which contradicts feasibility of $A$, proving the claim. □

*Proof of Lemma 5.3.* For $B = x + \Sigma B_p$ we compute the upper bound.

$$\begin{aligned}\sup_{b\in B}|a'b| &= \sup_{u\in\Sigma B_p}|a'x + a'u| \leq |a'x| + \sup_{u\in\Sigma B_p}|a'\Sigma\Sigma^{-1}u| \\ &= |a'x| + \sup_{v\in B_p}|(\Sigma' a)'v| = |a'x| + |\Sigma' a|_q.\end{aligned}$$

Before proceeding, we claim that for any $z \in \mathbb{R}^{d_h}$, we have $\max_{v\in B_p} v'z = \max_{v\in B_p}|v'z|$. Clearly $\max_{v\in B_p} v'z \leq \max_{v\in B_p}|v'z|$. Since $B_p$ is compact and $v \to v'z$ continuous, $v^* \in \operatorname{argmax}_{v\in B_p}|v'z|$ exists. Then $\max_{v\in B_p}|v'z| = |z'v^*| = z'v^* \operatorname{sgn}(z'v^*) = z'w$ for $w = v^* \operatorname{sgn}(z'v^*) \in B_p$ since $v^* \in B_p$. Then $\max_{v\in B_p}|v'z| = z'w \leq \max_{w\in B_p} z'w$. This proves the claim. Next, define $b(a) = x + \operatorname{sgn}(a'x)\Sigma v(a)$ with $v(a) \in \operatorname{argmax}_{v\in B_p} v'\Sigma' a$, which exists by compactness and continuity. Note $b(a) \in B$ by construction. We may calculate $|a'b(a)| = |a'x + \operatorname{sgn}(a'x)a'\Sigma v(a)|$. By the claim, $a'\Sigma v(a) \geq 0$. Then by matching signs, $|a'x + \operatorname{sgn}(a'x)a'\Sigma v(a)| = |a'x| + |\operatorname{sgn}(a'x)a'\Sigma v(a)| = |a'x| + |a'\Sigma v(a)|$. By the claim again, this is $|a'x| + a'\Sigma v(a) = |a'x| + \max_{v\in B_p}|a'\Sigma v| = |a'x| + |\Sigma' a|_q$. Combining with the upper bound above, we have shown that $\sup_{b\in B}|a'b| = |a'x| + |\Sigma' a|_q$. □

## C.6 Inference

In what follows, recall the within-arm influence functions $m_1 = v_D \Pi g_1 - \beta_1' w$ and $m_0 = v_D \Pi g_0 - \beta_0' w$ defined in Section 7.

*Proof of Theorem 7.1.* By two applications of Cauchy-Schwarz, we can upper bound

$$\begin{aligned} |E[\mathrm{Cov}(c'm_1, c'm_0|\psi)]| &\leq E[\mathrm{Var}(c'm_1|\psi)^{1/2}\,\mathrm{Var}(c'm_0|\psi)^{1/2}] \\ &\leq E[\mathrm{Var}(c'm_1|\psi)]^{1/2} E[\mathrm{Var}(c'm_0|\psi)]^{1/2}. \end{aligned}$$

This gives $V_a^{adj}(c) \leq v_D^{-1}(\tilde{\sigma}_1^2(c) + \tilde{\sigma}_0^2(c) + 2\tilde{\sigma}_1(c)\tilde{\sigma}_0(c)) = v_D^{-1}(\tilde{\sigma}_1(c) + \tilde{\sigma}_0(c))^2$. The second equality in the theorem is an algebraic identity. □

*Proof of Theorem 7.6.* By Lemma A.3, it suffices to show the result under $P$ in Definition A.1. Note that by Lemma C.6 and Lemma C.7, we have $\widehat{u}_1 = E_n[\frac{D_i}{p}\widehat{m}_i\widehat{m}_i'] - \widehat{v}_1 = E[m_1 m_1'] - E[E[m_1|\psi]E[m_1|\psi]'] + o_p(1) = E[\mathrm{Var}(m_1|\psi)] + o_p(1)$, and similarly $\widehat{u}_0 = E[\mathrm{Var}(m_0|\psi)] + o_p(1)$. Then $v_D^{-1}([c'\widehat{u}_1 c]^{1/2} + [c'\widehat{u}_0 c]^{1/2})^2 = \bar{V}_a(c) + o_p(1)$ by continuous mapping. This finishes the proof. □

*Proof of Theorem 7.8.* By Lemma A.3, it suffices to show the result under $P$ in Definition A.1. Denoting $\phi = \phi(W, \theta_0)$, $a = a(W, \theta_0)$, we have $\kappa_i(\theta_0) = \Pi g_i(\theta_0) - H_i\alpha_0' w_i = \Pi(\phi + H(a - \alpha_0' w_i))$. Then we may calculate

$$\begin{aligned} \mathrm{Var}(\kappa_i) = \mathrm{Var}(\Pi\phi) + v_D^{-1}E[(\Pi a - \alpha_0' w)^2] &= \mathrm{Var}(\Pi\phi) + v_D^{-1}E[\mathrm{Var}(\Pi a - \alpha_0' w|\psi)] \\ &+ v_D^{-1}E[E[\Pi a - \alpha_0' w|\psi]E[\Pi a - \alpha_0' w|\psi]']. \end{aligned}$$

This shows that $V_a = \mathrm{Var}(\kappa_i) - v_D^{-1}E[E[\Pi a - \alpha_0' w|\psi]E[\Pi a - \alpha_0' w|\psi]']$. The proof of Theorem 7.6 showed that $\alpha_0 = \beta_1 - \beta_0$. Also $\Pi a(W, \theta_0) = v_D \Pi(g_1 - g_0)(W, \theta_0)$ by definition. Then $\Pi a(W, \theta_0) - \alpha_0' w = v_D \Pi g_1 - \beta_1' w - (v_D \Pi g_0 - \beta_0' w) = m_1 - m_0$. Apparently,

$$\begin{aligned} V_a &= \mathrm{Var}(\kappa_i) - v_D^{-1}E[E[m_1 - m_0|\psi]E[m_1 - m_0|\psi]'] \\ &= \mathrm{Var}_n(\widehat{\kappa}_i) - v_D^{-1}(\widehat{v}_1 + \widehat{v}_0 - \widehat{v}_{10} - \widehat{v}_{10}') + o_p(1). \end{aligned}$$

This finishes the proof. □

**Lemma C.6.** *Impose Assumptions 3.1, 3.2, 7.5. Then under $P$ in Definition A.1, $E_n[\frac{D_i}{p}\widehat{m}_i\widehat{m}_i'] = E[m_1 m_1'] + o_p(1)$ and $E_n[\frac{1-D_i}{1-p}\widehat{m}_i\widehat{m}_i'] = E[m_0 m_0'] + o_p(1)$. Also, we have $\mathrm{Var}_n(\widehat{\kappa}_i) = \mathrm{Var}(\kappa_i) + o_p(1)$.*

*Proof.* For (a), consider the first statement. Note that $D_i\widehat{m}_i = v_D\widehat{\Pi}D_i\widehat{g}_i - D_i\widehat{\beta}_1'w_i$ and $D_i m_i = v_D\Pi D_i g_i - D_i\beta_1'w_i = D_i m_{1i}$. We can expand $E_n[(D_i/p)\widehat{m}_i\widehat{m}_i']$ as

$$E_n[(D_i/p)\widehat{m}_i(\widehat{m}_i - m_i)'] + E_n[(D_i/p)(\widehat{m}_i - m_i)m_i'] + E_n[(D_i/p)m_i m_i'].$$

For the first term, we have $E_n[(D_i/p)\widehat{m}_i(\widehat{m}_i - m_i)'] = p^{-1}E_n[D_i\widehat{m}_i(D_i\widehat{m}_i - D_i m_i)']$.

$$\begin{aligned}|D_i\widehat{m}_i - D_i m_i|_2 &= |D_i v_D\widehat{\Pi}\widehat{g}_i - D_i v_D\Pi g_i - D_i(\widehat{\beta}_1 - \beta_1)'w_i|_2 \\ &\lesssim |\widehat{\Pi} - \Pi|_2|\widehat{g}_i|_2 + |\Pi|_2|\widehat{g}_i - g_i|_2 + |\widehat{\beta}_1 - \beta_1|_2|w_i|_2.\end{aligned}$$

Then using $|xy'|_2 \leq |x|_2|y|_2$ and triangle inequality, the first term above has

$$\begin{aligned}|E_n[D_i\widehat{m}_i(D_i\widehat{m}_i - D_i m_i)']| \leq\ & |\widehat{\Pi} - \Pi|_2 E_n[|D_i\widehat{m}_i|_2|\widehat{g}_i|_2] + |\Pi|_2 E_n[|D_i\widehat{m}_i|_2|\widehat{g}_i - g_i|_2] \\ &+ |\widehat{\beta}_1 - \beta_1|_2 E_n[|D_i\widehat{m}_i|_2|w_i|_2].\end{aligned}$$

We claim this term is $o_p(1)$. Note that $|\widehat{\Pi} - \Pi|_2 = o_p(1)$ and $|\widehat{\beta}_1 - \beta_1|_2 = o_p(1)$ by assumption. Then applying Cauchy-Schwarz, it suffices to show $E_n[|D_i\widehat{m}_i|_2^2 + |\widehat{g}_i|_2^2 + |w_i|_2^2] = O_p(1)$ and $E_n[|\widehat{g}_i - g_i|_2^2] = o_p(1)$. First, note $E_n[|w_i|_2^2] = O_p(1)$ since $E[|w|_2^2] < \infty$. Next, note $E_n[|D_i\widehat{m}_i|_2^2] = E_n[|v_D D_i\widehat{\Pi}\widehat{g}_i - D_i\widehat{\beta}_1'w_i|_2^2] \leq 2E_n[|\widehat{\Pi}\widehat{g}_i|_2^2] + 2E_n[|\widehat{\beta}_1'w_i|_2^2] \leq 2|\widehat{\Pi}|_2^2 E_n[|\widehat{g}_i|_2^2] + 2|\widehat{\beta}_1|_2^2 E_n[|w_i|_2^2]$, so suffices to show $E_n[|\widehat{g}_i|_2^2] = O_p(1)$.

We start by showing that $E_n[|\widehat{g}_i - g_i|_2^2] = o_p(1)$. By the mean value theorem $g_i(\widehat{\theta}) - g_i(\theta_0) = \frac{\partial g_i}{\partial\theta'}(\tilde{\theta}_i)(\widehat{\theta} - \theta_0)$, where $\tilde{\theta}_i \in [\theta_0, \widehat{\theta}]$ may change by row. Then we have $E_n[|g_i(\widehat{\theta}) - g_i(\theta_0)|_2^2] \leq |\widehat{\theta} - \theta_0|_2^2 E_n[|\frac{\partial g_i}{\partial\theta'}(\tilde{\theta}_i)|_2^2]$, so it suffices to show $E_n[|\frac{\partial g_i}{\partial\theta'}(\tilde{\theta}_i)|_2^2] = O_p(1)$. Since $g_i(\theta) = D_i g_{1i}(\theta) + (1 - D_i)g_{0i}(\theta)$ for all $\theta$, $|\frac{\partial g_i}{\partial\theta'}(\tilde{\theta}_i)|_2^2 \leq 2|\frac{\partial g_{1i}}{\partial\theta'}(\tilde{\theta}_i)|_2^2 + 2|\frac{\partial g_{0i}}{\partial\theta'}(\tilde{\theta}_i)|_2^2$. Define the event $S_n = \{\widehat{\theta} \in U\}$. Then on $S_n$ we have

$$\begin{aligned}&|\frac{\partial g_{1i}}{\partial\theta'}(\tilde{\theta}_i)|_2^2 + |\frac{\partial g_{0i}}{\partial\theta'}(\tilde{\theta}_i)|_2^2 \leq |\frac{\partial g_{1i}}{\partial\theta'}(\tilde{\theta}_i)|_F^2 + |\frac{\partial g_{0i}}{\partial\theta'}(\tilde{\theta}_i)|_F^2 = \sum_{d=0,1}\sum_{k=1}^{d_g}|\nabla g_{di}^k(\tilde{\theta}_{ik})|_2^2 \\ &\leq \sum_{d=0,1}\sum_{k=1}^{d_g}\sup_{\theta\in U}|\nabla g_{di}^k(\theta)|_2^2 \equiv \bar{U}_i.\end{aligned}$$

Then $E_n[|\frac{\partial g_i}{\partial\theta'}(\tilde{\theta}_i)|_2^2]\mathbb{1}(S_n) \leq E_n[\bar{U}_i]\mathbb{1}(S_n) = O_p(1)$ since $E[\sup_{\theta\in U}|\nabla g_{di}^k(\theta)|_2^2] < \infty$ by assumption. Then $E_n[|\frac{\partial g_i}{\partial\theta'}(\tilde{\theta}_i)|_2^2] = O_p(1)$ since $P(S_n^c) \to 0$. This finishes the proof of $E_n[|\widehat{g}_i - g_i|_2^2] = o_p(1)$. Finally, the claim $E_n[|\widehat{g}_i|_2^2] = O_p(1)$ is clear since $E_n[|\widehat{g}_i|_2^2] \leq 2E_n[|\widehat{g}_i - g_i|_2^2] + 2E_n[|g_i|_2^2] = o_p(1) + O_p(1)$ by the preceding claim.

Then we have shown $|E_n[(D_i/p)\widehat{m}_i(\widehat{m}_i - m_i)']| = o_p(1)$ and $E_n[(D_i/p)(\widehat{m}_i - m_i)m_i'] = o_p(1)$ by an identical argument. This shows that $E_n[(D_i/p)\widehat{m}_i\widehat{m}_i'] =$

$E_n[(D_i/p)m_i m_i'] + o_p(1)$. Next, we claim $E_n[(D_i/p)m_i m_i'] = E_n[(D_i/p)m_{1i}m_{1i}'] = E_n[m_{1i}m_{1i}'] + o_p(1) = E[m_{1i}m_{1i}'] + o_p(1)$. The first equality is by definition of $m_i(D_i, W_i, \theta_0)$ and $m_{1i}(W_i, \theta_0)$. The second equality by Lemma A.2 of Cytrynbaum (2024b) and the third equality by vanilla WLLN, both using $E[|m_i|_2^2] < \infty$. This finishes our proof of the first statement of (a), and the second statement follows by symmetry.

Next consider the final statement. Note that $\widehat{\kappa}_i = \widehat{\Pi}\widehat{g}_i - H_i\widehat{\alpha}'w_i$ and $\kappa_i = \Pi g_i(\theta_0) - H_i\alpha_0'w_i$. Then $D_i\widehat{\kappa}_i = D_i\widehat{\Pi}\widehat{g}_i - D_i(1/p)\widehat{\alpha}'w_i$, which is of the form studied above. Then $E_n[\frac{D_i}{p}\widehat{\kappa}_i\widehat{\kappa}_i'] = E[\kappa_{1i}\kappa_{1i}'] + o_p(1)$ for score $\kappa_{1i} = \Pi g_{1i} - (1/p)\alpha_0'w_i$ with $D_i\kappa_i = D_i\kappa_{1i}$. Arguing similarly for $D_i = 0$, we have $E_n[\widehat{\kappa}_i\widehat{\kappa}_i'] = pE_n[\frac{D_i}{p}\widehat{\kappa}_i\widehat{\kappa}_i'] + (1-p)E_n[\frac{1-D_i}{1-p}\widehat{\kappa}_i\widehat{\kappa}_i'] = pE[\kappa_{1i}\kappa_{1i}'] + (1-p)E[\kappa_{0i}\kappa_{0i}'] + o_p(1) = E[D_i\kappa_{1i}\kappa_{1i}'] + E[(1-D_i)\kappa_{0i}\kappa_{0i}'] + o_p(1) = E[\kappa_i\kappa_i'] + o_p(1)$. Moreover, $E_n[\widehat{\kappa}_i] = E_n[\widehat{\Pi}\widehat{g}_i - H_i\widehat{\alpha}'w_i] = \widehat{\Pi}E_n[\widehat{g}_i] + o_p(1)$. Note that $E_n[\widehat{g}_i] = \widehat{g}(\widehat{\theta})$ and $\widehat{g}(\widehat{\theta}) - \widehat{g}(\theta_0) = g_0(\widehat{\theta}) - g_0(\theta_0) + o_p(1) = o_p(1)$. The first equality since $|\widehat{g} - g_0|_{\Theta,\infty} = o_p(1)$ and the second by continuous mapping, using Lemma C.1. Then $\mathrm{Var}_n(\widehat{\kappa}_i) = E[\kappa_i\kappa_i'] + o_p(1)$. □

**Lemma C.7.** *Require Assumptions 3.1, 3.2, 7.5. Then under $P$ in Definition A.1, the estimators in the statement of Theorem 7.8 have $\widehat{v}_{10} \xrightarrow{p} E[E[m_{1i}|\psi]E[m_{0i}|\psi]']$, and $\widehat{v}_1 \xrightarrow{p} E[E[m_{1i}|\psi]E[m_{1i}|\psi]']$, and $\widehat{v}_0 \xrightarrow{p} E[E[m_{0i}|\psi]E[m_{0i}|\psi]']$.*

*Proof.* Let $\widehat{v}_1^o$ the oracle version of $\widehat{v}_1$ with $m_i = v_D\Pi g_i(\theta_0) - D_i\beta_1'w_i - (1-D_i)\beta_0'w_i$ substituted for $\widehat{m}_i$, and similarly define oracle versions $\widehat{v}_0^o$, $\widehat{v}_{10}^o$ of $\widehat{v}_0, \widehat{v}_{10}$. Note $D_i m_i = D_i m_{1i} = D_i(v_D\Pi g_{1i}(\theta_0) - \beta_1'w_i)$. In Lemma A.6 of Cytrynbaum (2024b), set $A_i = m_{1i}$ and $B_i = m_{1i}$. Applying the lemma componentwise gives $\widehat{v}_1^o \xrightarrow{p} E[E[m_{1i}|\psi]E[m_{1i}|\psi]']$. Similarly, we have $\widehat{v}_0^o \xrightarrow{p} E[E[m_{0i}|\psi]E[m_{0i}|\psi]']$, and $\widehat{v}_{10}^o \xrightarrow{p} E[E[m_{1i}|\psi]E[m_{0i}|\psi]']$. Then it suffices to show $\widehat{v}_1 - \widehat{v}_1^o = o_p(1)$, $\widehat{v}_0 - \widehat{v}_0^o = o_p(1)$, and $\widehat{v}_{10} - \widehat{v}_{10}^o = o_p(1)$. For the first statement, expand

$$\widehat{v}_1 - \widehat{v}_1^o = (np)^{-1}\sum_{s\in\mathcal{S}_n^\nu}\frac{1}{a(s)-1}\sum_{i\neq j\in s} D_iD_j(\widehat{m}_i\widehat{m}_j' - m_im_j')$$

Expand $\widehat{m}_i\widehat{m}_j' - m_im_j' = \widehat{m}_i(\widehat{m}_j' - m_j') + (\widehat{m}_i - m_i)m_j' \equiv A_{ij} + B_{ij}$. Using triangle inequality, $a(s)-1 \geq 1$ and $p > 0$, we calculate $\widehat{v}_1^o - \widehat{v}_1 \lesssim n^{-1}\sum_{s\in\mathcal{S}_n^\nu}\sum_{i,j\in s}|A_{ij}|_2 + |B_{ij}|_2 \equiv A_n + B_n$. First consider $B_n$. Using that $|xy'|_2 \leq |x|_2|y|_2$, we have

$$\begin{aligned}|B_{ij}|_2 \leq |\widehat{m}_i - m_i|_2|m_j|_2 &= |v_D\widehat{\Pi}\widehat{g}_i - v_D\Pi g_i - D_i(\widehat{\beta}_1 - \beta_1)'w_i - (1-D_i)(\widehat{\beta}_0 - \beta_0)'w_i|_2|m_j|_2 \\ &\leq |\widehat{\Pi} - \Pi|_2|\widehat{g}_i|_2|m_j|_2 + |\Pi|_2|\widehat{g}_i - g_i|_2|m_j|_2 + 2\max_{d=0,1}|\widehat{\beta}_d - \beta_d|_2|w_i|_2|m_j|_2.\end{aligned}$$

Then $B_n = n^{-1}\sum_{s\in\mathcal{S}_n^\nu}\sum_{i,j\in s}|\widehat{\Pi}-\Pi|_2|\widehat{g}_i|_2|m_j|_2 + |\Pi|_2|\widehat{g}_i - g_i|_2|m_j|_2 + 2\max_{d=0,1}|\widehat{\beta}_d -$

$\beta_d|_2|w_i|_2|m_j|_2 \equiv B_{n1} + B_{n2} + B_{n3}$. Consider $B_{n1}$. This is

$$B_{n1} = |\widehat{\Pi} - \Pi|_2 \cdot n^{-1} \sum_{s \in \mathcal{S}_n^\nu} \sum_{i,j \in s} |\widehat{g}_i|_2 |m_j|_2 \leq |\widehat{\Pi} - \Pi|_2 \cdot (2n)^{-1} \sum_{s \in \mathcal{S}_n^\nu} \sum_{i,j \in s} |\widehat{g}_i|_2^2 + |m_j|_2^2$$
$$\leq |\widehat{\Pi} - \Pi|_2 \cdot (2n)^{-1} \sum_{s \in \mathcal{S}_n^\nu} |s| \sum_{i \in s} |\widehat{g}_i|_2^2 + |m_i|_2^2 \lesssim |\widehat{\Pi} - \Pi|_2 E_n[|\widehat{g}_i|_2^2 + |m_i|_2^2].$$

By an identical argument $B_{n3} \lesssim \max_{d=0,1} |\widehat{\beta}_d - \beta_d|_2 E_n[|w_i|_2^2 + |m_i|_2^2]$. Then to show $B_{n1} + B_{n3} = o_p(1)$, suffices to show $E_n[|w_i|_2^2 + |m_i|_2^2 + |\widehat{g}_i|_2^2] = O_p(1)$. That $E_n[|w_i|_2^2 + |\widehat{g}_i|_2^2] = O_p(1)$ was shown in the proof of Lemma C.6. Note $E_n[|m_i|_2^2] = E_n[|v_D \Pi g_i(\theta_0) - D_i \beta_1' w_i - (1 - D_i)\beta_0' w_i|_2^2] \leq 2E_n[|\Pi g_i|_2^2] + 2E_n[|D_i \beta_1' w_i + (1 - D_i)\beta_0' w_i|_2^2] \leq 2|\Pi|_2^2 E_n[|g_i|_2^2] + 2 \max_{d=0,1} |\beta_d|_2^2 E_n[|w_i|_2^2] = O_p(1)$ since $E[|g_i|_2^2] < \infty$ by assumption. Then $B_{n1} + B_{n3} = o_p(1)$. Finally, consider $B_{n2}$. By the mean value theorem $g_i(\widehat{\theta}) - g_i(\theta_0) = \frac{\partial g_i}{\partial \theta'}(\tilde{\theta}_i)(\widehat{\theta} - \theta_0)$, where $\tilde{\theta}_i \in [\theta_0, \widehat{\theta}]$ may change by row. Then we have

$$B_{n2} = n^{-1} \sum_{s \in \mathcal{S}_n^\nu} \sum_{i,j \in s} |\Pi|_2 |\widehat{g}_i - g_i|_2 |m_j|_2 \leq |\widehat{\theta} - \theta_0|_2 |\Pi|_2 \cdot n^{-1} \sum_{s \in \mathcal{S}_n^\nu} \sum_{i,j \in s} |\frac{\partial g_i}{\partial \theta'}(\tilde{\theta}_i)|_2 |m_j|_2$$
$$\lesssim |\widehat{\theta} - \theta_0|_2 |\Pi|_2 E_n[|\frac{\partial g_i}{\partial \theta'}(\tilde{\theta}_i)|_2^2 + |m_i|_2^2] = o_p(1).$$

The final equality follows since $E_n[|\frac{\partial g_i}{\partial \theta'}(\tilde{\theta}_i)|_2^2 = O_p(1)$, as shown in the proof of Lemma C.6. Then we have shown $B_n = o_p(1)$, and $A_n = o_p(1)$ is identical. This completes the proof that $\widehat{v}_1 - \widehat{v}_1^o = o_p(1)$, and the proof of $\widehat{v}_0 - \widehat{v}_0^o = o_p(1)$, and $\widehat{v}_{10} - \widehat{v}_{10}^o = o_p(1)$ are identical. □

## C.7 Lemmas

**Proposition C.8** (Lévy). *Consider probability spaces $(\Omega_n, \mathcal{G}_n, P_n)$ and $\sigma$-algebras $\mathcal{F}_n \subseteq \mathcal{G}_n$. We say $A_n \in \mathbb{R}^d$ has $A_n|\mathcal{F}_n \Rightarrow A$ if $\phi_n(t) \equiv E[e^{it'A_n}|\mathcal{F}_n] = E[e^{it'A}|\mathcal{F}_n] + o_p(1)$ for each $t \in \mathbb{R}^d$. If $g : \mathbb{R}^d \to \mathbb{C}$ is bounded, measurable, and $P(A \in \{a : g(\cdot)$ discontinuous at $a\}) = 0$ then we have*

$$E[g(A_n)|\mathcal{F}_n] = E[g(A)] + o_p(1). \tag{C.1}$$

**Lemma C.9.** *Consider probability spaces $(\Omega_n, \mathcal{G}_n, P_n)$ and $\sigma$-algebras $\mathcal{F}_n \subseteq \mathcal{G}_n$. Suppose $0 \leq A_n \leq B < \infty$ and $A_n = o_p(1)$. Then $E[A_n|\mathcal{F}_n] = o_p(1)$.*

See Cytrynbaum (2021) for the proofs.

**Lemma C.10.** *The following statements hold*

(a) *There exists* $\gamma_0 \in \mathbb{R}^{d_h \times d_a}$ *solving* $E[\mathrm{Var}(h|\psi)]\gamma_0 = E[\mathrm{Cov}(h, a|\psi)]$. *For any solution, we have* $E[\mathrm{Var}(a - \gamma_0' h|\psi)] \preceq E[\mathrm{Var}(a - \gamma' h|\psi)]$ *for all* $\gamma \in \mathbb{R}^{d_h \times d_a}$.

(b) *Let* $Z = (Z_a, Z_h)$ *a random variable with* $\mathrm{Var}(Z) = E[\mathrm{Var}((a, h)|\psi)] \equiv \Sigma$ *and define* $\tilde{Z}_a = Z_a - \gamma_0' Z_h$. *Then* $\mathrm{Cov}(\tilde{Z}_a, Z_h) = 0$. *In particular, if* $(Z_a, Z_h)$ *are jointly Gaussian, then* $\tilde{Z}_a$ *is Gaussian with* $\tilde{Z}_a \perp\!\!\!\perp Z_h$.

*Proof.* In the notation of (b), it suffices to show $\Sigma_{hh}\gamma_0 = \Sigma_{ha}$. If $\mathrm{rank}(\Sigma_{hh}) = 0$ then $Z_h = c_h$ a.s. for constant $c_h$ and $\Sigma_{ha} = \mathrm{Cov}(Z_h, Z_a) = 0$. Then any $\gamma \in \mathbb{R}^{d_h \times d_a}$ is a solution. Then suppose $\mathrm{rank}(\Sigma_{hh}) = r \geq 1$. Let $\Sigma_{hh} = U\Lambda U'$ be the compact SVD with $U \in \mathbb{R}^{d_h \times r}$ and $\mathrm{rank}(\Lambda) = r$, and $U'U = I_r$. We claim $Z_h = UU'Z_h$ a.s. Calculate $\mathrm{Var}((UU' - I)Z_h) = (UU' - I)U\Lambda U'(UU' - I) = 0$. Note that $\Sigma_{hh}\gamma = \Sigma_{ha} \iff \mathrm{Var}(Z_h)\gamma = \mathrm{Cov}(Z_h, Z_a) \iff \mathrm{Var}(UU'Z_h)\gamma = \mathrm{Cov}(UU'Z_h, Z_a) \iff U[\mathrm{Var}(U'Z_h)U'\gamma - \mathrm{Cov}(U'Z_h, Z_a)] = 0$. Define $\bar{Z}_h = U'Z_h$ and note $\mathrm{Var}(\bar{Z}_h) = U'U\Lambda U'U = \Lambda \succ 0$. Then let $\bar{\gamma} = \mathrm{Var}(\bar{Z}_h)^{-1}\mathrm{Cov}(\bar{Z}_h, a)$ so that $\mathrm{Var}(\bar{Z}_h)\bar{\gamma} - \mathrm{Cov}(\bar{Z}_h, Z_a) = 0$. Then it suffices to find $\gamma$ such that $U'\gamma = \bar{\gamma}$. Since $U' : \mathbb{R}^{d_h} \to \mathbb{R}^r$ is onto, there exists $\gamma^k$ with $U'\gamma^k = \bar{\gamma}^k$. Then let $\gamma_0^k \in [\gamma^k + \ker(U')]$ and set $\gamma_0 = (\gamma_0^k : k = 1, \ldots, d_a)$, so that $U'\gamma_0 = \bar{\gamma}$. Then $\Sigma_{hh}\gamma_0 = \Sigma_{ha}$ by work above. For the optimality statement, calculate

$$
\begin{aligned}
&E[\mathrm{Var}(a - \gamma' h|\psi)] = \Sigma_{aa} - \Sigma_{ah}\gamma - \gamma'\Sigma_{ha} + \gamma'\Sigma_{hh}\gamma = \Sigma_{aa} - \Sigma_{ah}(\gamma - \gamma_0 + \gamma_0) \\
&- (\gamma - \gamma_0 + \gamma_0)'\Sigma_{ha} + \gamma'\Sigma_{hh}\gamma = \Sigma_{aa} - 2\gamma_0'\Sigma_{hh}\gamma_0 - (\gamma - \gamma_0)'\Sigma_{ha} - \Sigma_{ah}(\gamma - \gamma_0) \\
&+ \gamma'\Sigma_{hh}\gamma \propto -(\gamma - \gamma_0)'\Sigma_{hh}\gamma_0 - \gamma_0'\Sigma_{hh}(\gamma - \gamma_0) + \gamma'\Sigma_{hh}\gamma = -(\gamma - \gamma_0)'\Sigma_{hh}\gamma_0 \\
&- \gamma_0'\Sigma_{hh}(\gamma - \gamma_0) + \gamma'\Sigma_{hh}\gamma + (\gamma - \gamma_0 + \gamma_0)'\Sigma_{hh}(\gamma - \gamma_0 + \gamma_0) \\
&= \gamma_0'\Sigma_{hh}\gamma_0 + (\gamma - \gamma_0)'\Sigma_{hh}(\gamma - \gamma_0).
\end{aligned}
$$

Then $E[\mathrm{Var}(a - \gamma' h|\psi)] - E[\mathrm{Var}(a - \gamma_0' h|\psi)]) = (\gamma - \gamma_0)'\Sigma_{hh}(\gamma - \gamma_0)$ and for any $a \in \mathbb{R}^{d_a}$ we have $a'(\gamma - \gamma_0)'\Sigma_{hh}(\gamma - \gamma_0)a \geq 0$ since $\Sigma_{hh} \succeq 0$. This proves the claim. Finally, we have $\mathrm{Cov}(\tilde{Z}_a, Z_h) = \mathrm{Cov}(Z_a - \gamma_0' Z_h, Z_h) = \Sigma_{ah} - \gamma_0'\Sigma_{hh} = 0$. The final statement follows from well-known facts about the normal distribution. □

**Lemma C.11.** $A_n = O_p(1) \iff A_n = o_p(c_n)$ *for every sequence* $c_n \to \infty$.

*Proof.* It suffices to consider $A_n \geq 0$. The forward direction is clear. For the backward direction, suppose for contradiction that there exists $\epsilon > 0$ such that $\sup_{n \geq 1} P(A_n > M) > \epsilon$ for all $M$. Then find $n_k$ such that $P(A_{n_k} > k) > \epsilon$ for each $k \geq 1$. We claim $n_k \to \infty$. Suppose not and $\liminf_k n_k \leq N < \infty$. Then let $k(j) \to \infty$ such that $n_{k(j)} \leq N$ for all $j$. Choose $M' < \infty$ such that $P(A_n > M') < \epsilon$ for all $n = 1, \ldots N$. Then for $k(j) > M'$ we have $P(A_{n_{k(j)}} > k(j)) \leq$

$P(A_{n_{k(j)}} > M') < \epsilon$, which is a contradiction. Then apparently $\lim_k n_k = +\infty$. Define $Z_j = \{i : i \geq j\}$. Regard the sequence $n_k$ as map $n : \mathbb{N} \to \mathbb{N}$. For $m \in \text{Image}(n)$, define $n^\dagger(m) = \min n^{-1}(m)$. It's easy to see that $n^\dagger(m_k) \to \infty$ for $\{m_k\}_k \subseteq \text{Image}(n)$ with $m_k \to \infty$. Then write

$$\sup_{k \geq j} P(A_{n_k} > k) = \sup_{m \in n(Z_j)} \sup_{a \in n^{-1}(m)} P(A_m > a) \leq \sup_{m \in n(Z_j)} P(A_m > n^\dagger(m))$$

Note $A_{m_k}/n^\dagger(m_k) = o_p(1)$ by assumption for any $\{m_k\}_k \subseteq \text{Image}(n)$ with $m_k \to \infty$. Then we have

$$\limsup_k P(A_{n_k} > k) = \lim_j \sup_{k \geq j} P(A_{n_k} > k) = \lim_j \sup_{m \in n(Z_j)} P(A_m > n^\dagger(m)) = o(1).$$

This is a contradiction, which completes the proof. □