Finely Stratified Rerandomization Designs

Max Cytrynbaum*

December 31, 2024

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 rerandomization designs, allowing for imbalance metrics based on nonlinear estimators. We also propose a novel minimax scheme that uses pilot data or prior information to minimize the computational cost of rerandomization, subject to a strict bound on statistical efficiency. 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. This enables simple asymptotically exact inference on super-population parameters, as well as efficient conservative inference on finite population parameters.

*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 within small groups of units tightly matched on baseline covariate information, makes unadjusted estimators like difference of means semiparametrically efficient.[2] In finite samples, however, the performance of such designs can deteriorate rapidly with the dimension of the stratification variables 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-groups 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 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 residual truncated term, the asymptotic variance of unadjusted estimation under stratified rerandomization is the same as that of semiparametrically adjusted GMM (e.g. Graham (2011)) under an iid design. Intuitively, stratified rerandomization implements partially linear regression adjustment "by design."

Our second contribution is to introduce several novel forms of rerandomization based on nonlinear balance criteria. For example, we allow acceptance or rejection of an allocation based on the difference of covariate density estimates within each treatment arm, attempting to balance nonlinear features of the covariate distribution. Similarly, we propose a design that rerandomizes until a nonlinear estimate of the propensity score is approximately constant, effectively forcing the covariates to have no predictive power for treatment assignments. In both cases, these nonlinear rerandomization schemes are asymptotically equivalent to standard rerandomization based on a difference of covariate means, but with an implicit choice of covariates and acceptance region, which we characterize.

[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. (2024) for more detailed 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).

Our third contribution is to study optimization of the balance criterion itself. We propose a novel minimax approach that allows the researcher to specify prior information about the relationship between covariates and outcomes, then rerandomizes until the worst case correlation between treatments and covariates consistent with this prior is small. We prove that this design minimizes the (asymptotic) computational cost of rerandomization, subject to a strict bound on statistical efficiency over the set of DGP's consistent with the prior. If the prior information set contains the truth, this design strictly bounds the asymptotic variance within a small additive factor of the optimal semiparametrically adjusted variance. Extending this result, we show that if the prior information set is a confidence region estimated from pilot data, then this minimax design bounds the asymptotic variance in the main experiment with high probability.

Our fourth contribution is to provide simple t-statistic and Wald based inference methods for general causal parameters under stratified rerandomization designs. To do this, we first characterize and provide a feasible implementation of the optimal ex-post linear adjustment for GMM estimation under stratified rerandomization.[4] Crucially, optimal ex-post adjustment makes the asymptotic distribution insensitive to the rerandomization acceptance criterion, removing the truncated normal term from the limiting distribution and restoring asymptotic normality. For superpopulation parameters, our inference methods are asymptotically exact. For finite population parameters, our inference is conservative due to non-identification of the asymptotic variance, but still exploits the efficiency gains from both stratified rerandomization and ex-post optimal adjustment.

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. (2024). Some important theoretical contributions to the literature on rerandomization include 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. (2024). Wang et al. (2021) study estimation of the

[4]This extends recent work on optimal adjustment under pure stratified randomization for ATE estimation, e.g. see Cytrynbaum (2024a), Bai et al. (2023), or Liu and Yang (2020).

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. (2024) study estimation of superpopulation parameters defined by moment equalities under pure stratified randomization. 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. (2024).[5] In concurrent 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. 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 a specific quadratic rerandomization using a Bayesian criterion, in both cases for rerandomization without stratification. We provide 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. 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. (2023), to stratified rerandomization and GMM parameters. Finally our inference methods (Section 7) build on previous work by Abadie and Imbens (2008), Bai et al. (2021), and Cytrynbaum (2024b). To the best of our knowledge we provide the first asymptotically exact inference for causal GMM parameters under stratified rerandomization, as well as conservative inference for their finite population analogues.

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

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

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 $\bar{a}_1 = E_n[a_i D_i]/E_n[D_i]$ and $\bar{a}_0 = E_n[a_i(1-D_i)]/E_n[(1-D_i)]$ 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 s with $\{1,\dots,n\} = \bigcup_s s$ 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 homogeneity 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 π_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-adaptive "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 ψ space. Cytrynbaum (2024b) provides algorithms to match units into groups that satisfy this condition for any fixed k.

Example 2.2 (Matched Pairs Rerandomization)**.** For $k = 2$, the optimal matched pairs in Equation 2.1 can be found by Derigs (1988) algorithm. Suppose we have done so, and consider rerandomizing until the imbalance criterion $n(\bar{X}_1 - \bar{X}_0)'\Sigma_n(\bar{X}_1 - \bar{X}_0) \le \epsilon^2$ is satisfied for positive-definite $\Sigma_n \xrightarrow{p} \Sigma$.[8] Let $\mathcal{I}_n \equiv \Sigma_n^{1/2}\sqrt{n}(\bar{X}_1 - \bar{X}_0) = \sqrt{n}(\bar{h}_1 - \bar{h}_0) + o_p(1)$ for modified covariates $h = \Sigma^{1/2}X$. This quadratic acceptance criterion is equivalent to

[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 $\mathcal{I}_n = \sqrt{n}(\bar{h}_1 - \bar{h}_0) + o_p(1)$ under "pure" stratified randomization, the design in steps (1) and (2) only, studied e.g. in Cytrynbaum (2024b). We give several examples below.

[8] Several recent papers in the statistics literature have considered such criteria. See e.g. Morgan and Rubin (2012), Li et al. (2018), Wang et al. (2021) among others.

$\mathcal{I}_n \in A$ for acceptance region $A = \{x : |x|_2 \leq \epsilon\}$. We study the efficiency consequences of different covariates and acceptance regions in detail in Sections 3 and 5 below.

Example 2.3 (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. (2024). 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 fixed strata $T \in \{1, \ldots, m\}$ and fixed m, as in Bugni et al. (2018), which can be obtained in this framework by setting $\psi = T$ and matching units into groups s at random within each $\{i : T_i = k\}$.

Example 2.4 (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 $\sum_i d_i = np$.[9] If so, we denote $D_{1:n} \sim \mathrm{CR}(p)$. Cytrynbaum (2024b) shows that $\mathrm{CR}(p)$ randomization can be obtained by setting $\psi = 1$ and $A = \mathbb{R}^{d_h}$ in Definition 2.1, matching units into groups at random. Intuitively, random matched k-tuples is equivalent to complete randomization.

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 $D|W \sim \mathrm{Bernoulli}(p)$ 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 introduce our causal estimands.

Definition 2.5 (Causal Estimands)**.** The *superpopulation* estimand θ_0 is the unique solution to $E[\phi(W, \theta)] = 0$. The *finite population* estimand θ_n is the unique solution to $E_n[\phi(W_i, \theta)] = 0$.

In what follows, we study GMM estimation of both θ_0 and θ_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 θ_n that do not appear to have been considered previously in the literature. Note that GMM estimation of θ_0 under pure stratification was studied in Bai et al. (2024) for the exactly identified case. Our finite population parameter θ_n can be viewed as a causal version of the finite population estimand defined in Abadie et al. (2014).[10]

Example 2.6 (ATE)**.** 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 $\theta_0 = E[Y(1) - Y(0)] =$

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

[10] See also the related finite population estimands studied under iid sampling and assignment in Xu (2021) and Kakehi and Otsu (2024).

ATE, the average treatment effect, and $\theta_n = E_n[Y_i(1) - Y_i(0)] = \text{SATE}$, the sample average treatment effect.

For a more interesting example, consider the best parametric predictor of treatment effect heterogeneity in experiments with noncompliance.

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 local average treatment effect $\text{LATE} = E[Y(1) - Y(0)|C = 1]$. Let $H = (Z - p)/(p - p^2)$ and consider the score function $g(Z, D, Y, X, \theta) = (HY - HD \cdot f(X, \theta))\nabla_\theta f(X, \theta)$. Using standard LATE manipulations,

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

Then $E[\phi(W, \theta)] = 0$ is the first order condition of a treatment effect prediction problem in the complier population. In particular, for $\tau \equiv Y(1) - Y(0)$, the parameter θ_0 is the best parametric predictor of treatment effects for compliers:

$$\theta_0 = \underset{\theta}{\text{argmin}}\, E[(\tau - f(X, \theta))^2 | C = 1].$$

For example, if Y is binary then $Y(1) - Y(0) \in \{-1, 0, 1\}$, so a scaled link function model $f(X, \theta) = 2L(X'\theta) - 1$ may be appropriate. We can easily estimate marginal effects by adding $m(X_i, \theta, \beta) = \beta - (\partial/\partial\theta')f(X_i, \theta)$ to the score function.

Example 2.8 (Finite Population Heterogeneity)**.** Continuing Example 2.7, note that for $\tau_i = Y_i(1) - Y_i(0)$ the corresponding finite population parameter is

$$\theta_n = \underset{\theta}{\text{argmin}}\, E_n[(\tau_i - f(X_i, \theta))^2 | C_i = 1]. \tag{2.2}$$

We can view θ_n as a "SATE-like" version of θ_0, the best parametric predictor of treatment effects in the *within-sample* complier population. θ_n may be a more appropriate target for experiments run in a convenience sample. If $f(X, \theta) = X'\theta$ linear, then $\theta_n = \text{argmin}_\theta\, E_n[(\tau_i - X_i'\theta)^2 | C_i = 1]$ is the within-sample best linear predictor. In the case of perfect compliance $C_i = 1$ for all i, this is $\theta_n = \text{argmin}_\theta\, E_n[(\tau_i - X_i'\theta)^2]$, a finite-sample version of the best linear predictor of the conditional average treatment effect (CATE). The case $X = 1$ recovers $\theta_n = E_n[Y_i(1) - Y_i(0)|C_i = 1]$, the finite-population LATE, studied e.g. in Ren (2023). Our inference methods in Section 7 produce tighter confidence intervals for these finite population parameters than θ_0, since we only need to account for the uncertainty due to random assignment, with no sampling uncertainty.

GMM Estimation. Let positive-definite weighting matrix $M_n \in \mathbb{R}^{d_g \times d_g}$ with $M_n \xrightarrow{p}$

$M \succ 0$. For sample moment $\widehat{g}(\theta) \equiv E_n[g(D_i, R_i, S_i, \theta)]$, the GMM estimator[11] is

$$\widehat{\theta} = \operatorname*{argmin}_{\theta \in \Theta} \widehat{g}(\theta)' M_n' \widehat{g}(\theta). \tag{2.3}$$

In the exactly identified case, $\widehat{\theta}$ solves $\widehat{g}(\widehat{\theta}) = 0$. In the next section, we study generalized method of moments (GMM) estimation of the causal parameters θ_0 and θ_n under stratified rerandomization.

Remark 2.9 (M-estimation)**.** Our results below also extend to M-estimators of the form $\widehat{\theta} = \operatorname{argmax}_\theta E_n[m(D_i, R_i, S_i, \theta)]$, even when $\widehat{\theta}$ cannot be formulated as a GMM estimator e.g. due to the existence of local maxima. For example, this happens in some nonconvex problems in density estimation (Newey and McFadden (1994)). We briefly discuss this extension in Section 11.4.

3 Asymptotics for GMM Estimation

In this section, we characterize the asymptotic distribution of the GMM estimator $\widehat{\theta}$ under stratified rerandomization designs, as in Definition 2.1. We show that the variance under stratified rerandomization is proportional to the residuals of a partially linear regression model, up to an extra term that reflects 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* $\mathrm{Leb}(\partial A) = 0$,[12] *and require $E[\mathrm{Var}(h|\psi)] \succ 0$ and $E[|\psi|_2^2 + |h|_2^2] < \infty$.*

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 Θ is compact.[13]

[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 ∂A denotes the boundary of A, the limit points of both A and A^c.

[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.

(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.

Orthogonal Expansion. Recall $\phi(W,\theta) = E[g(D,R,S,\theta)|W]$ for $W = (R, S(1), S(0))$. Define the assignment influence component $a(W,\theta) \equiv \mathrm{Var}(D)(g_1(W,\theta) - g_0(W,\theta))$. For Horvitz-Thompson weights $H = (D-p)/(p-p^2)$, a simple calculation shows that we can expand

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

Our work below shows that $a(W,\theta)$ parameterizes estimator variance due to random assignment, while $\phi(W,\theta)$ parameterizes the variance due to random sampling. We work directly with this expansion in what follows.

Example 3.3 (SATE)**.** Continuing Example 2.6 above, let $\bar{Y} = (1-p)Y(1) + pY(0)$, a convex combination that summarizes each unit's potential outcome level. Then for the score $g(D,Y,\theta) = HY - \theta$, we have $a(W,\theta) = \bar{Y}$. A simple calculation shows that for $\widehat{\theta} = E_n[H_iY_i]$ and $\theta_n = E_n[Y_i(1) - Y_i(0)]$, we have

$$\widehat{\theta} - \theta_n = E_n[H_i a(W_i)] = \frac{\mathrm{Cov}_n(D_i, \bar{Y}_i)}{\mathrm{Var}_n(D_i)}. \tag{3.2}$$

Intuitively, the term $E_n[H_i a(W_i)]$ from Equation 3.1 isolates the estimator variance due to chance in-sample correlations between the assignments D_i and outcome levels $\bar{Y}_i$. By contrast, $\phi(W,\theta) = Y(1) - Y(0) - \theta$ does not depend on assignments D_i, and $\mathrm{Var}(\phi(W,\theta)) = \mathrm{Var}(Y(1) - Y(0))$ isolates the estimator variance due to random sampling of units and heterogeneity of individual treatment effects.

3.1 Finite Population Estimand

Our first theorem studies GMM estimation of the finite population estimand θ_n, which solves $E_n[\phi(W_i,\theta_n)] = 0$. We extend these results to θ_0 in Corollary 3.8 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 $\Pi = -G^{-1}$. For brevity, we also denote $v_D = \mathrm{Var}(D) = p - p^2$.

Before stating the main result, we first derive the influence function for GMM estimation of θ_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 Equation 3.2 above, showing that

$$\widehat{\theta} - \theta_n = \frac{\mathrm{Cov}_n(D_i, \Pi a(W_i, \theta_0))}{\mathrm{Var}_n(D_i)} + o_p(n^{-1/2}).$$

This implies that to first order, the errors in estimating θ_n are driven by the random in-sample correlations between treatment assignments D_i and the assignment influence function $\Pi a(W_i, \theta_0)$. Our main theorem shows that, by balancing ψ and h, stratified rerandomization reduces these correlations, 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*

$$V_a = \min_{\gamma \in \mathbb{R}^{d_h \times d_\theta}} v_D^{-1} E[\mathrm{Var}(\Pi a(W, \theta_0) - \gamma' h|\psi)]. \tag{3.3}$$

Let γ_0 be optimal in Equation 3.3. The term R_A is a truncated Gaussian

$$R_A \sim \gamma_0' Z_h \,|\, Z_h \in A, \qquad Z_h \sim \mathcal{N}(0, v_D^{-1} E[\mathrm{Var}(h|\psi)]). \tag{3.4}$$

Note that the variance matrix $V_a \in \mathbb{R}^{d_\theta \times d_\theta}$, so the minimum should be interpreted in the positive semidefinite sense. In particular, we say $V(\gamma_0) = \min_\gamma V(\gamma)$ 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 R_A. The normal term $\mathcal{N}(0, V_a)$ only depends on the "treatment assignment" component of the influence function, $\Pi a(W, \theta_0)$. The variance is attenuated nonparametrically by the stratification variables ψ and linearly by rerandomization covariates h.

Residual Imbalance. The truncated Gaussian term $R_A \sim \gamma_0' Z_h \,|\, Z_h \in A$ arises from leftover covariate imbalances due to slackness in the rerandomization acceptance criterion, $\sqrt{n}(\bar{h}_1 - \bar{h}_0) \in A$, since $A \neq \{0\}$. If the acceptance region 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 asymptotically unbiased. In principle, R_A could be made negligible relative to $\mathcal{N}(0, V_a)$ in large samples by choosing a small enough acceptance region A. For example, if $A = B(0, \epsilon)$ then $R_{B(0,\epsilon)} \sim \{\gamma_0' Z_h \,|\, |Z_h|_2 \le \epsilon\} \xrightarrow{p} 0$ as $\epsilon \to 0$. However, in finite samples and for small enough ϵ, this acceptance region may be infeasible. We study a minimax style criterion to choose an efficient acceptance 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, as defined in Examples 2.3 and 2.4.

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 $\sqrt{n}(\widehat{\theta} - \theta_n)|W_{1:n} \Rightarrow \mathcal{N}(0, V)$ with $V = v_D^{-1} E[\mathrm{Var}(\Pi a(W, \theta_0)|\psi)]$.*

In particular, if $D_{1:n} \sim \mathrm{CR}(p)$ *then* $V = v_D^{-1}\,\mathrm{Var}(\Pi a(W,\theta_0))$.

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

Remark 3.7 (Design-Based Asymptotics)**.** Our results above show that $\sqrt{n}(\widehat{\theta}-\theta_n)|W_{1:n} \Rightarrow \mathcal{N}(0,V_a)+R_A$, conditional on the sampled data $W_{1:n} = (R_i, S_i(1), S_i(0))_{i=1}^n$.[14] This result is "design-based" in the sense that the variance in the limiting distribution arises solely due to randomness of the treatment assignments $D_{1:n}$. However, we impose structure on the sequence of populations $W_{1:n}$ ex-ante, assuming each population is drawn from a fixed measure, $W_i \sim F$. This allows us to provide simple variance expressions that connect our results with the superpopulation-based literature on GMM and partially linear adjustment in econometrics. By contrast, the "sequence of finite populations model" often used in the statistics literature (e.g. Li et al. (2018)) begins with an arbitrary sequence of finite populations $(W_{i,n})_{i=1}^n$, imposing the minimal structure needed for certain moments to converge ex-post. It would be interesting to extend our results to this setting, but we leave this to future work.

3.2 Superpopulation Estimand

The next result extends Theorem 3.5 to the superpopulation estimand θ_0, which uniquely solves $E[\phi(W,\theta_0)] = 0$.

Corollary 3.8 (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* $V_\phi = \mathrm{Var}(\Pi\phi(W,\theta_0))$ *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*

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

Comparing Corollary 3.8 with the results above, we see that targeting θ_0 instead of θ_n adds an extra independent Gaussian term $\mathcal{N}(0,V_\phi)$ to the asymptotic distribution. Intuitively, V_ϕ arises due to iid random sampling of $\Pi\phi(W,\theta_0)$. Notice that stratification and rerandomization only affect the assignment influence function component $\Pi a(W,\theta_0)$, while the sampling influence component $\Pi\phi(W,\theta_0)$ is irreducible. In this sense, the statistical consequences of different designs and adjustment strategies all happen at the level

[14] See Proposition 11.17 in the appendix for a formal statement.

of the finite population estimand θ_n, while targeting the superpopulation estimand θ_0 just adds extra irreducible noise. For pure stratification, Bai et al. (2024) were the first to derive an analogue of part (b) of Corollary 3.8 in the exactly identified case, under different GMM regularity conditions than we use here.

Example 3.9 (SATE)**.** Continuing Example 2.6, we had $\phi(W, \theta) = Y(1) - Y(0) - \theta$, so $G = 1$ and $\Pi = 1$. As above, $a(W, \theta) = (1-p)Y(1) + pY(0) \equiv \bar{Y}$. The GMM estimator $\widehat{\theta} = \bar{Y}_1 - \bar{Y}_0$ is just difference of means. Then by Theorem 3.5 and Corollary 3.8, we have $\sqrt{n}(\widehat{\theta} - \text{SATE})|W_{1:n} \Rightarrow \mathcal{N}(0, V_a) + R_A$ 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_ϕ, which only appears when estimating the superpopulation estimand θ_0, reflects sampling variance due to treatment effect heterogeneity. The term V_a is the variance due to random assignment, caused by random in-sample correlations between treatments D and outcome levels $\bar{Y}$. Covariate-adaptive randomization and adjustment can be used to reduce V_a, while V_ϕ is an irreducible sampling variance.

Remark 3.10. Wang et al. (2021) study SATE estimation under stratified rerandomization in the sequence of finite populations framework. Relative to Wang et al. (2021), by imposing the tight-matching condition 2.1 we are able to derive a simple closed form for the asymptotic variance in terms of the measure $W \sim F$, showing an equivalence with partially linear regression adjustment.

Example 3.11 (Treatment Effect Heterogeneity)**.** Continuing Example 2.7, consider the case with perfect compliance $D = Z$ and $f(X, \theta) = X'\theta$. Then we can use the slightly modified score $g(D, X, Y, \theta) = (HY - X'\theta)X$. Then for $\tau = Y(1) - Y(0)$ we have $\phi(W, \theta_0) = (\tau - X'\theta_0)X$, and the parameters θ_n, θ_0 are the best linear predictors of treatment effect heterogeneity

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

It's also easy to see that $a(W, \theta_0) = \bar{Y}X$ and $\Pi = E[XX']^{-1}$. Then for $e = \tau - X'\theta_0$, the variance matrices in Corollary 3.8 are

$$V_\phi = E[XX']^{-1}E[e^2XX']E[XX']^{-1} \qquad V_a = \min_{\gamma \in \mathbb{R}^{d_h \times d_x}} v_D^{-1} E[\text{Var}(\bar{Y}\Pi X - \gamma' h|\psi)].$$

The expression for V_a shows that if we want to precisely estimate treatment effect heterogeneity, it is important to stratify and rerandomize not only the variables that predict outcome levels $\bar{Y}$, but also their interactions with the heterogeneity variable X.

3.3 Equivalence with Partially Linear Adjustment

Example 3.9 showed that, up to the rerandomization imbalance R_A, the unadjusted estimator $\widehat{\theta} = \bar{Y}_1 - \bar{Y}_0$ has asymptotic variance $V_a = \min_{\gamma \in \mathbb{R}^{d_h}} v_D^{-1} E[\mathrm{Var}(\bar{Y} - \gamma' h | \psi)]$. This can be rewritten in terms of the residuals of a partially linear regression of $\bar{Y}$ on ψ and h:

$$V_a = \min_{\substack{\gamma \in \mathbb{R}^{d_h} \\ t \in L_2(\psi)}} v_D^{-1} \mathrm{Var}(\bar{Y} - \gamma' h - t(\psi)). \tag{3.6}$$

More generally, Theorem 3.5 shows that under stratified rerandomization designs, the usual GMM estimator $\widehat{\theta}$ behaves like semiparametrically adjusted GMM in the iid setting. Formally, let $\mathcal{L}(\psi) = L_2^{d_\theta}(\psi)$ be the d_θ-fold Cartesian product of $L_2(\psi)$, the space of square-integrable functions. Then the variance due to random assignment V_a in Theorem 3.5 is can be written in terms of the residuals of the influence function $\Pi a(W, \theta_0)$ in a partially linear regression on ψ and h:

$$V_a = \min_{\substack{\gamma \in \mathbb{R}^{d_h \times d_\theta} \\ t \in \mathcal{L}(\psi)}} v_D^{-1} \mathrm{Var}\left(\Pi a(W, \theta_0) - \gamma' h - t(\psi)\right). \tag{3.7}$$

Intuitively, stratified rerandomization does partially linear regression adjustment "by design," providing nonparametric control over ψ and linear control over h. For a more explicit equivalence statement, define $m(\psi, h) = \gamma_0' h + t_0(\psi)$ to be the partially linear function achieving the optimum in Equation 3.7. Define the oracle semiparametrically adjusted GMM estimator

$$\widehat{\theta}^* = \widehat{\theta} - E_n[H_i m(\psi_i, h_i)]. \tag{3.8}$$

For example, for the SATE estimation problem one can show that $\widehat{\theta}^*$ is just an oracle version of the usual augmented inverse propensity weighting (AIPW) estimator (Robins and Rotnitzky (1995)), with partially linear regression models in each arm.[15]

Theorem 3.12 (Partially Linear Adjustment)**.** *Suppose that $D_{1:n} \sim \mathrm{CR}(p)$. The oracle partially linearly adjusted GMM estimator $\sqrt{n}(\widehat{\theta}^* - \theta_n)|W_{1:n} \Rightarrow \mathcal{N}(0, V_a)$, with variance V_a as defined in Theorem 3.5.*

Under a completely randomized design, we require ex-post semiparametric adjustment to achieve V_a. Under stratified rerandomization, however, the simple GMM estimator $\widehat{\theta}$ automatically achieves V_a, up to the leftover imbalance term R_A.

[15] Feasible partially linear adjustment in an iid mean estimation problem with missing data was studied in Wang et al. (2004). See also the related semiparametric adjustment for GMM parameters in Graham (2011).

4 Nonlinear Rerandomization

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

4.1 GMM Rerandomization

First, we generalize the imbalance metric $\mathcal{I}_n$ introduced in Definition 2.1, allowing rejection of a treatment allocation $D_{1:n}$ based on potentially nonlinear features of the in-sample distribution of treatments and covariates $(D_i, X_i)_{i=1}^n$. We can define a large class of nonlinear imbalance metrics by letting $m(X_i, \beta)$ be a score function and considering within-arm GMM estimators $\widehat{\beta}_1$ and $\widehat{\beta}_0$ defined by

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

We propose to rerandomize until the within-arm parameter estimates are approximately equal, $\sqrt{n}(\widehat{\beta}_1 - \widehat{\beta}_0) \approx 0$.

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. (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).

Intuitively, the generalized imbalance metric $\mathcal{I}_n^m$ allows us to randomize until possibly nonlinear features of the covariates are balanced between the treatment and control groups. Observe that 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.

Example 4.2 (Density Rerandomization)**.** Let $f(X, \beta)$ be a parametric density model for covariates X, which may be misspecified. After drawing $D_{1:n}$ by stratified randomization, consider forming (quasi) maximum likelihood estimators $\widehat{\beta}_1 \in \operatorname{argmax}_\beta E_n[D_i \log f(X_i, \beta)]$ and $\widehat{\beta}_0 \in \operatorname{argmax}_\beta E_n[(1 - D_i) \log f(X_i, \beta)]$ for the density of covariates assigned to each treatment arm, rerandomizing until the estimated parameters $\sqrt{n}|\widehat{\beta}_1 - \widehat{\beta}_0|_2 \leq \epsilon$. Under regularity conditions,[16] $\widehat{\beta}_d$ are GMM estimators as in Equation 4.1 with score function $m(X_i, \beta) = \nabla_\beta \log f(X_i, \beta)$, so this procedure is a GMM rerandomization with acceptance region $A = \{x : |x|_2 \leq \epsilon\}$.

[16] 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.

Let β^* be the unique solution to $E[m(X, \beta^*)] = 0$ and define $G_m = E[(\partial/\partial\beta')m(X_i, \beta^*)]$. 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 $h_i = m(X_i, \beta^*)$ 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*

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

The residual $R \sim \gamma_0' Z_m \,|\, Z_m \in G_m A$ for $Z_m \sim \mathcal{N}(0, v_D^{-1} E[\mathrm{Var}(m(X_i, \beta^*)|\psi)])$, where γ_0 is optimal in Equation 4.2.

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 in Equation 4.1. This suggests an equivalent, but computationally much simpler design with only one round of nonlinear estimation. In particular, let $\widehat{\beta}$ solve $E_n[m(X_i, \widehat{\beta})] = 0$ be the pooled GMM estimator and set rerandomization covariates $\widehat{h}_i = m(X_i, \widehat{\beta})$, rerandomizing until $\sqrt{n} E_n[H_i \widehat{h}_i] \in G_m A$. The next result shows that this design, which generalizes Definition 2.1 to allow for estimated covariates, is asymptotically equivalent to the GMM rerandomization in Definition 4.1.

Corollary 4.4. *Suppose Assumption 3.1, 3.2 hold and let $m(X, \beta)$ be 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 $G_m A$. 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.*

Corollary 4.4 is a useful tool for showing the equivalence of computationally intensive designs based on nonlinear estimation with simpler linear schemes. For instance, Example 4.5 shows that density rerandomization over an exponential family with sufficient statistic $r(X_i)$ is asymptotically equivalent to linear rerandomization with $h_i = r(X_i)$.

Example 4.5 (Density Rerandomization)**.** Continuing Example 4.2, for $x \in \mathcal{X}$ and a sufficient statistic $r(x) \in \mathbb{R}^{d_r}$, define the exponential family $f(x, \beta) = \exp(\beta' r(x) - t(\beta))$, with $t(\beta) = \log \int_{\mathcal{X}} \exp(\beta' r(x)) d\nu(x)$ for some measure ν on $\mathcal{X}$. If the sufficient statistics $(r_j(x))_{j=1}^k$ are ν-a.s. linearly independent, then $\beta \to \log f(x, \beta)$ is strictly concave for all x.[17] Then $E[m(X, \beta)] = 0$ has a unique solution for score $m(X, \beta) = \nabla_\beta \log f(X, \beta)$, showing that quasi-MLE in this family can be formulated as a GMM problem. By Corollary 4.4, density rerandomization using $f(x, \beta)$ is asymptotically equivalent to linear rerandomization with $\widehat{h}_i = \nabla_\beta \log f(X_i, \widehat{\beta}) = r(X_i) - t(\widehat{\beta})$. Since $E_n[H_i t(\widehat{\beta})] =$

[17] This holds since the log partition function $t(\beta)$ is strictly convex for β 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 $t(\beta)$.

$t(\widehat{\beta})E_n[H_i] = 0$, this is in turn equivalent to linear rerandomization with $h_i = r(X_i)$, directly balancing the sufficient statistics for the family. For example, if $x \in \{\pm 1\}^k$ are binary variables, consider density estimation in the graphical model[18]

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

The sufficient statistic is $r(x) = ((x_j)_j, (x_j x_l)_{j<l})$, and the parameters β_{jl} model correlation between the binary variables x_j and x_l. By the discussion above, a design that rerandomizes based on the difference of quasi-MLE density estimates in this family[19] is asymptotically equivalent to the much simpler linear rerandomization in Definition 2.1 with covariates $h_i = ((x_j)_j, (x_j x_l)_{j<l})$.

4.2 Propensity Score Rerandomization

To motivate a propensity score based rerandomization procedure, note that under stratified randomization we have $E[D_i|X_i] = p$ for all units. In finite samples, however, 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 are predictive of treatment assignments post-randomization, a form of "in-sample confounding," which of vanishes as $n \to \infty$ but affects precision. To prevent this, we could, for instance, reject allocations where $|\widehat{p}(B) - p| > \epsilon$ for some collection of sets B. To make this idea tractable without fully discretizing, consider a parametric propensity model $p(X, \beta) = L(X'\beta)$ for smooth link function L and define the MLE estimator

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

We can measure the average gap between the estimated and true propensity score using

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

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 $D_{1:n}$ and estimate the propensity model in Equation 4.3. (3) If imbalance $\mathcal{J}_n \leq \epsilon$, accept. Otherwise, repeat from (2).

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

[19]This is well-motivated when ψ 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 ψ.

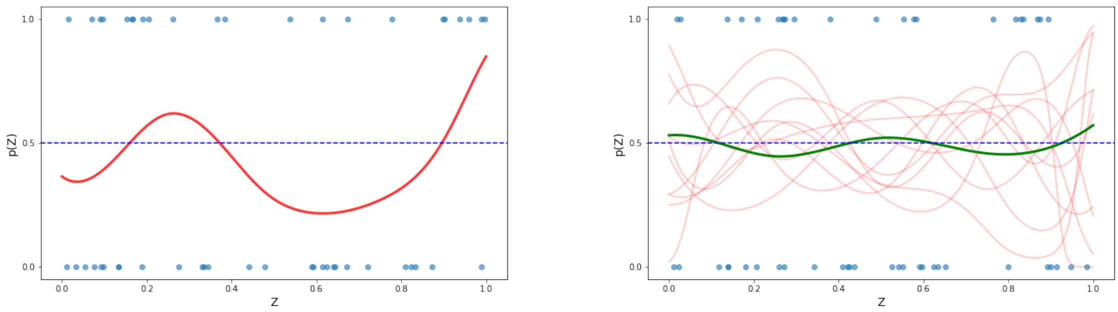

Figure 1: Propensity rerandomization (Definition 4.6) with $p = 1/2$ for $Z \sim \text{Unif}[0, 1]$ and $X = B(Z)$ a B-spline basis. LHS: $D_{1:n}$ and estimated propensity with $\widehat{p}(Z) \ll 1/2$, for $Z \in [0.4, 0.9]$. RHS: Accepted allocation $D_{1:n}$ with $\mathcal{J}_n \leq \epsilon$

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 acceptance region. We require some extra regularity conditions on the link function L, which for brevity we state in Appendix 11.5.

Theorem 4.7 (Propensity Rerandomization)**.** *Suppose $D_{1:n}$ is as in Definition 4.6. Require Assumptions 3.2, 11.12. 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 γ_0 optimal in the equation above.

Theorem 4.7 shows that for any sufficiently regular link function, propensity rerandomization is asymptotically equivalent to the quadratic rerandomization design in Example 2.2, 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 $A = \text{Var}(h)^{1/2} B(0, \epsilon v_D^{-2})$.[20]

Implicit Acceptance Regions. Both nonlinear designs in this section turned out to be equivalent to the standard rerandomization scheme in Definition 2.1, with a specific, implicit choice of rerandomization moments and acceptance region determined by the choice of score m and marginal covariate distribution. However, this implicit choice is not likely to be optimal, since the residual term in the asymptotic error distribution $R_A \sim \gamma_0' Z_h | Z_h \in A$ depends on both the covariates $Z_h \sim \mathcal{N}(0, v_D^{-1} E[\text{Var}(h|\psi)])$, and the partially linear coefficient γ_0. This coefficient is determined by the *joint* distribution of the assignment influence function $\Pi a(W, \theta_0)$ and covariates (ψ, h). In the next section, we

[20] 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.

show how to use prior information about this joint distribution to optimize the acceptance region and bound the variance of R_A.

5 Optimizing Acceptance Regions

In this section, we study efficient choice of the acceptance region $A \subseteq \mathbb{R}^{d_h}$. We propose a novel 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 first restrict to the case of estimating $\theta_n = \text{SATE}$. Example 3.9 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 γ_0 comes from the partially linear regression[21]

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

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 γ_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 acceptance region $A = \{x : |\gamma_0' x| \leq \epsilon\}$, reducing the problem to one dimension from arbitrary $\dim(h)$. Of course, this oracle design is infeasible since γ_0 is generally unknown when designing the experiment.

5.1 Minimax Rerandomization

Since γ_0 is unknown at design-time, we instead take a minimax approach that incorporates prior information about the coefficient γ_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. Our next result shows that Equation 5.2 is a linear rerandomization design, characterizing the implicit acceptance region A.

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

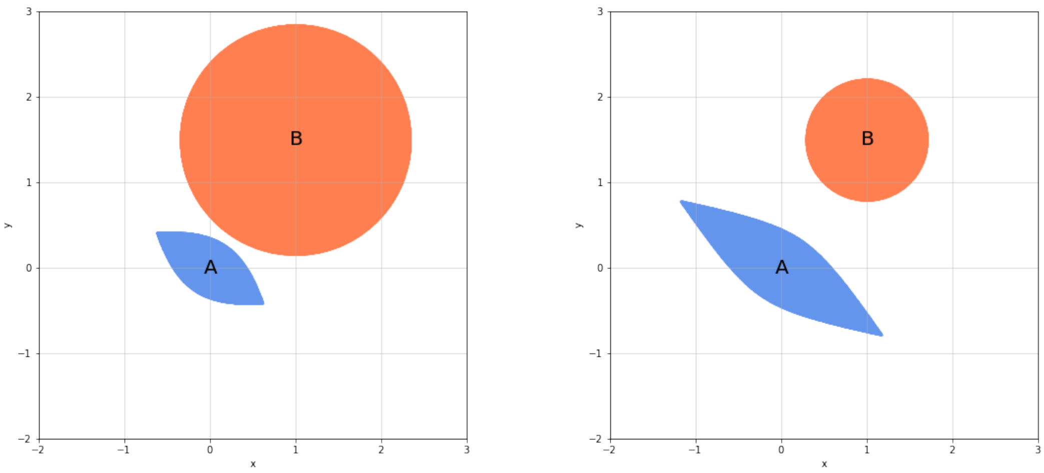

Figure 2: Prior information B and $A_0 = \epsilon B^\circ$ for Example 5.2.

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.*[22]

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 shows that the corresponding acceptance region is $A_0 = \{x : |x'\bar{\gamma}| + u|x|_2 \leq \epsilon\}$. For small u, acceptance region A_0 mimics the oracle, allowing very large imbalances $\sqrt{n}(\bar{h}_1 - \bar{h}_0)$ 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

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

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 $A_0 = \{x : |x'\bar{\gamma}| + |U'x|_q \leq \epsilon\}$.*

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. $0 \leq \gamma_{0j} \leq m$ for some j and $-m \leq \gamma_{0j} \leq 0$ for others. 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 γ_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*[23]

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

[23]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 $\{A \in \mathcal{B}(\mathbb{R}^{d_h}) : \text{Leb}(A \triangle A_0) = 0\}$, where $\triangle$ denotes symmetric difference.

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 3.6.*

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 ϵ. 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 ϵ^2 of the optimal partially linear variance V_a in Equation 3.7.

Remark 5.6 (Integral Probability Metric)**.** We briefly note another interesting interpretation of the design in Equation 5.2. For distributions P, Q and a function class $\mathcal{F}$, the integral probability metric is a pseudo-distance between distributions defined by $\rho(P, Q; \mathcal{F}) \equiv \sup_{f \in \mathcal{F}} |E_P[f(X)] - E_Q[f(X)]|$.[24] Let function class $\mathcal{F}_B = \{\gamma' h : \gamma \in B\}$ and let $\widehat{P}_d$ denote the empirical distribution of $h_i | D_i = d$ for $d = 0, 1$. Then we have

$$\sup_{\gamma \in B} |\gamma' \sqrt{n}(\bar{h}_1 - \bar{h}_0)| \leq \epsilon \iff \sqrt{n} \rho(\widehat{P}_1, \widehat{P}_0; \mathcal{F}_B) \leq \epsilon.$$

This shows that the minimax design rerandomizes until within-arm empirical distribution of covariates h are balanced according to $\rho(\widehat{P}_1, \widehat{P}_0; \mathcal{F}_B)$, a distance between $h_i|D_i = 1$ and $h_i|D_i = 0$ that is only sensitive to the projections $\gamma' h$ with $\gamma \in B$ that may actually matter for estimating $\theta_n = \text{SATE}$. By doing so, we maximize the size of the acceptance region (and acceptance probability) subject to the statistical guarantee in Theorem 5.5.

5.3 Beliefs From Pilot Data

Next, we discuss an alternative strategy that uses pilot data to specify the set B in a data-driven way. Suppose we have access to $\mathcal{D}_{pilot} \perp\!\!\!\perp (W_{1:n}, D_{1:n})$ 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

$$\widehat{B}_{pilot} = \widehat{\gamma} + c_\alpha^{1/2} m^{-1/2} \cdot \widehat{\Sigma}_{pilot}^{1/2} B_2(0, 1). \tag{5.4}$$

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

$$\widehat{A}_{pilot} = \epsilon \widehat{B}_{pilot}^\circ = \{x : |x' \widehat{\gamma}_{pilot}| + \frac{c_\alpha^{1/2} |\widehat{\Sigma}^{1/2} x|_2}{m^{1/2}} \leq \epsilon\}. \tag{5.5}$$

Note that the acceptance region $\widehat{A}_{pilot}$ expands as the pilot size m is larger. This

[24] The pseudometric ρ is also referred to as the maximum mean discrepancy. This is a commonly used statistic in two-sample testing, see e.g. Gretton et al. (2008).

reflects smaller uncertainty about the true parameter γ_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 α 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 5.7 (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 $\sqrt{n}(\widehat{\theta} - \theta_n) | \mathcal{D}_{pilot} \Rightarrow v_D^{-1} \mathcal{N}(0, \mathrm{Var}(e)) + R_A$, where $\mathrm{Var}(R_A | \mathcal{D}_{pilot}) \leq \epsilon^2$ with probability $\geq 1 - \alpha$.*

Formally, the pilot estimate of γ_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 $E[Y_T | W] = (1-p)Y(1) + pY(0) = \bar{Y}$.

6 Restoring Normality

In this section, we study optimal linearly adjusted GMM estimation under stratified rerandomization. We show that, to first order, optimal ex-post linear adjustment completely removes the impact of the acceptance region A and imbalance term R_A, restoring asymptotic normality. This enables standard t-statistic and Wald-test based inference on the parameters θ_n and θ_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 $E[|w|_2^2] < \infty$.

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 $\sqrt{n}(\widehat{\theta}_{adj} - \theta_n) | W_{1:n} \Rightarrow \mathcal{N}(0, V(\alpha))$ with $V(\alpha) = v_D^{-1} E[\mathrm{Var}(\Pi a(W, \theta_0) - \alpha' w | \psi)]$ and $\sqrt{n}(\widehat{\theta}_{adj} - \theta_0) \Rightarrow \mathcal{N}(0, V_\phi + V(\alpha))$.*

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

Optimal Adjustment Coefficient. Define the coefficient

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

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 the optimal adjustment coefficient α_0 varies with the stratification variables ψ, as observed in Cytrynbaum (2024b) and Bai et al. (2023) for ATE estimation. The main result of this section shows that adjustment by a consistent estimate of α_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. Equation 6.1 shows that for nonlinear models the optimal coefficient α_0 may depend on the unknown parameter θ_0. This suggests a two-step adjustment strategy, where we

(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 $\breve{w}_i = w_i - \sum_{j \in s(i)} w_j$, 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

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

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, α_0 may not depend on θ_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.

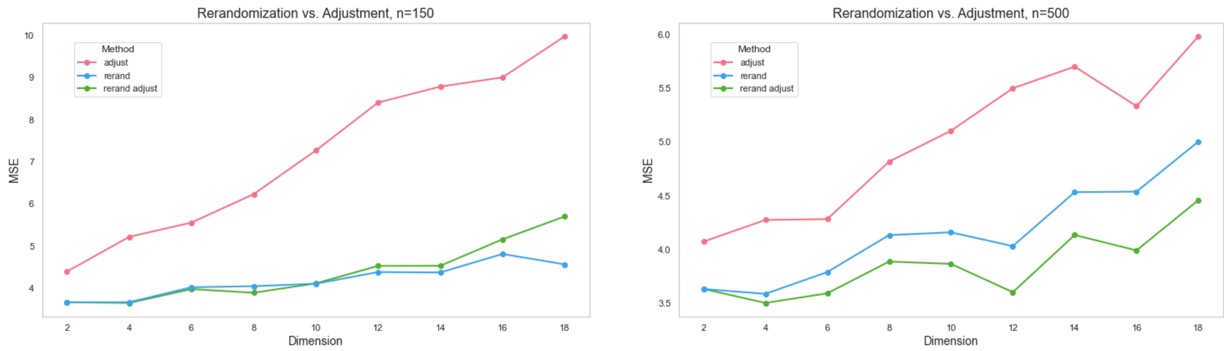

Figure 3: MSE for Adjustment vs. Rerandomization, $n = 150$ and $n = 500$.

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, we have* $\widehat{\alpha} = \alpha_0 + o_p(1)$*.*

One-step adjustment is possible in many linear GMM problems, including the best linear predictor of treatment effects parameter in Example 3.11.

Example 6.6 (Treatment Effect Heterogeneity)**.** Continuing Example 3.11, suppose we want to estimate treatment effect heterogeneity relative to an important covariate X, while adjusting optimally ex-post for larger set of covariates w to improve precision and restore asymptotic normality under rerandomization. We use GMM score $g(Y, D, X, \theta) = (HY - X'\theta)X$ for estimating the heterogeneity parameter $\theta_n = \text{argmin}_\theta \, E_n[(Y_i(1) - Y_i(0) - X_i'\theta)^2]$. 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[25]

$$\widehat{\alpha} = E_n[\check{w}_i \check{w}_i']^{-1} \left[(1-p)\,\mathrm{Cov}_n(\check{w}_i, Y_i X_i | D_i = 1) + p\,\mathrm{Cov}_n(\check{w}_i, Y_i X_i | D_i = 0)\right] E_n[X_i X_i']^{-1}.$$

We apply this adjustment in our simulations and empirical application to Angrist et al. (2013) in Section 9 below.

6.1 Double Robustness from Rerandomization

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

[25] In the case $X = 1$ (SATE estimation), $\widehat{\alpha}$ is equivalent to the stratified "tyranny-of-the-minority" style estimator in Cytrynbaum (2024a).

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

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

For simplicity, consider the case of SATE estimation with $\psi = 1$. For the difference of means estimator $\widehat{\theta}$, by a simple calculation $\widehat{\theta} - \text{SATE} = E_n[\bar{Y}_i|D_i = 1] - E_n[\bar{Y}_i|D_i = 0]$. Let $\bar{Y} = c + \gamma_0' h + e$ with $e \perp (1, h)$ be the decomposition from Equation 5.1. Then we can decompose the estimation error $\widehat{\theta} - \text{SATE}$ into imbalances in h and e:

$$\sqrt{n}(\widehat{\theta} - \text{SATE}) = \sqrt{n}\gamma_0'(\bar{h}_1 - \bar{h}_0) + \sqrt{n}(\bar{e}_1 - \bar{e}_0). \tag{6.3}$$

Rerandomizing until $\sqrt{n}(\bar{h}_1 - \bar{h}_0)$ is small shrinks the first imbalance term, ideally making its variance negligible relative to $\text{Var}(e) = \text{Var}(\bar{Y} - \gamma_0' h)$. Suppse $h = w$ so the adjustment coefficient $\alpha_0 = \gamma_0$. Then writing $\widehat{\alpha} = \widehat{\gamma}$, the adjusted estimator becomes $\widehat{\theta}_{adj} = \widehat{\theta} - \widehat{\gamma}' E_n[H_i h_i] = \widehat{\theta} - \widehat{\gamma}'(\bar{h}_1 - \bar{h}_0)$, so that

$$\sqrt{n}(\widehat{\theta}_{adj} - \text{SATE}) = \sqrt{n}(\gamma_0 - \widehat{\gamma})'(\bar{h}_1 - \bar{h}_0) + \sqrt{n}(\bar{e}_1 - \bar{e}_0). \tag{6.4}$$

This decomposition shows a novel 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. Consider the LHS of Figure 3, where γ_0 is not estimated well for small n and large $\dim(w)$, and adjustment performs poorly. This effect is exacerbated by stratification, since the partialling operation $\breve{w}_i = w_i - \sum_{j \in s(i)} w_j$ tends to decrease the variance of the regressors w_i, making estimation of α_0 more difficult.[27] However, adjustment + rerandomization still performs well due to double robustness.

The first imbalance term has a "product of errors" structure, similar to the product of nuisance estimation errors for doubly-robust estimators in the literature on Neyman orthogonal estimating equations (e.g. Chernozhukov et al. (2017)). This shows that even when both $\widehat{\gamma} - \gamma_0$ and $\bar{h}_1 - \bar{h}_0$ are small, we can get an extra benefit from combining the two methods. This is shown in the RHS of Figure 3 with $n = 500$, where adjustment is competitive, but rerandomization still performs better, and stratification + rerandomization is even more efficient due to this product structure. Such double robustness also holds for more general casual parameters and GMM estimators. Let $h = w$ and consider the partially linear decomposition $\Pi a(W, \theta_0) = \gamma_0' h + t(\psi) + e$ with $e \perp h$ and $E[e|\psi] = 0$ and $\bar{t}_d = E_n[t(\psi_i)|D_i = d]$. Our work shows that

$$\begin{aligned} \widehat{\theta}_{adj} - \theta_n &= (\gamma_0 - \widehat{\gamma})'(\bar{h}_1 - \bar{h}_0) + (\bar{t}_1 - \bar{t}_0) + o_p(n^{-1/2}) \\ &= (\gamma_0 - \widehat{\gamma})'(\bar{h}_1 - \bar{h}_0) + o_p(n^{-1/2}). \end{aligned}$$

[27]For example, in our empirical application the condition number of the design matrix $E_n[\breve{w}_i \breve{w}_i]$ increases as we stratify more finely.

The second equality is a consequence of fine stratification on ψ.

Summarizing, this discussion highlights a double robustness property that explains the additoinal finite-sample precision gains from combining rerandomization with ex-post adjustment. This effect is likely to be especially important in regimes where the optimal adjustment coefficient α_0 is poorly estimated, such as for small n, large $\dim(w)$, ill-conditioned design matrix $E_n[\check{w}_i\check{w}_i'] \approx E[\mathrm{Var}(w|\psi)]$. A full theory of high-dimensional stratification, rerandomization, and ex-post adjustment is beyond the scope of the current work, but this is an interesting area for future research.[28]

7 Variance Bounds and Inference Methods

In this section, we provide methods for inference on causal GMM parameters under stratified rerandomization designs. We make crucial use of asymptotic normality of the optimally adjusted estimator $\widehat{\theta}_{adj}$ in the previous section. The asymptotic variance for estimating the finite population parameter θ_n is generally not identified, so we provide novel variance upper bounds allowing for conservative inference that still reflects the precision gains from stratified rerandomization. The asymptotic variance for estimating the superpopulation parameter θ_0 is identified, and in this case we provide asymptotically exact inference methods.

7.1 Variance Bounds

For intuition, we briefly review the classical variance bounds for θ_n = SATE estimation under completely randomized assignment. In this case, we have $\sqrt{n}(\widehat{\theta} - \mathrm{SATE}) \Rightarrow \mathcal{N}(0, V_a)$ with $V_a = \mathrm{Var}(D)^{-1}\,\mathrm{Var}(\bar{Y})$ for $\bar{Y} = (1-p)Y(1)+pY(0)$. The variance $\mathrm{Var}(\bar{Y}) \propto \mathrm{Cov}(Y(1), Y(0))$, so V_a is not identified. Let $\sigma_d^2 = \mathrm{Var}(Y(d))$ and $\tau = Y(1) - Y(0)$. The simple Cauchy-Schwarz bound $|\,\mathrm{Cov}(Y(1), Y(0))| \leq \sigma_1\sigma_0$ and algebra gives

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

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

To define the bounds, recall that Theorem 6.3 showed $\sqrt{n}(\widehat{\theta}_{adj} - \theta_n) \Rightarrow N(0, V_a^{adj})$ with $V_a^{adj} = v_D^{-1} E[\mathrm{Var}(\Pi a(W, \theta_0) - \alpha_0' w|\psi)]$, where $\alpha_0 = E[\mathrm{Var}(w|\psi)]^{-1} E[\mathrm{Cov}(w, \Pi a(W, \theta_0)|\psi)]$ was the optimal adjustment coefficient. By definition, $\Pi a(W, \theta_0) = v_D \Pi(g_1(W, \theta_0) -$

[28] High-dimensional rerandomization creates analytical complications from conditioning on $\sqrt{n}(\bar{h}_1 - \bar{h}_0) \in A$ with $\dim(h)$ is growing. A recent theoretical breakthrough on this question was achieved by Wang and Li (2024b) for the case of complete rerandomization, without stratification or adjustment.

$g_0(W,\theta_0))$. Note the adjustment coefficient may be decomposed $\alpha_0 = \beta_1 - \beta_0$ for coefficients $\beta_d = E[\text{Var}(w|\psi)]^{-1}E[\text{Cov}(w, v_D \Pi g_d(W,\theta_0)|\psi)]$. Denote $g_d = g_d(W,\theta_0)$ 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

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

Similarly to above, the cross-term $V_a^{adj}(c) \propto E[\text{Cov}(c'm_1, c'm_0|\psi)]$ is generically not identified. The next result provides a sequence of Neyman-style identified upper bounds

Theorem 7.1 (Variance Bounds)**.** *Define* $\tilde{\sigma}_d^2(c) = E[\text{Var}(c'm_d|\psi)]$ *for* $d \in \{0,1\}$.

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

Dividing by v_D*, denote bounds above as* $V_a^{adj}(c) \le \bar{V}_a^{adj}(c) \le \check{V}_a^{adj}(c)$.

We derive a consistent estimator of the sharper 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 $\theta_n = \text{SATE}$ under pure stratified randomization ($A = \mathbb{R}^{d_h}$) and optimal ex-post adjustment.

Example 7.2 (Pure Stratification)**.** Let $\theta_n = \text{SATE}$ 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 $\delta_1 = \text{argmin}_\delta E[\text{Var}(Y(1) - \delta' w|\psi)]$ and $m_1 = (1-p)(Y(1) - \delta_1' w)$. Similarly, $m_0 = p(Y(0) - \delta_0' w)$ with $\delta_0 = \text{argmin}_\delta E[\text{Var}(Y(0) - \delta' w|\psi)]$. After algebra, the bound $\bar{V}_a^{adj}$ on the optimally adjusted variance $V_a^{adj} = \min_\gamma v_D^{-1} E[\text{Var}(\bar{Y} - \gamma' w|\psi)]$ is

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

For $\psi = 1$ (complete randomization) and $w = 0$ (no adjustment), we recover the sharper Neyman bound in Equation 7.1. If $\psi \neq 0$, we get a "stratified" Neyman bound:

$$\bar{V}_a^{adj} = \frac{E[\text{Var}(Y(1)|\psi)]}{p} + \frac{E[\text{Var}(Y(0)|\psi)]}{1-p} - (E[\text{Var}(Y(1)|\psi)]^{1/2} - E[\text{Var}(Y(0)|\psi)]^{1/2})^2.$$

Remark 7.3 (Sharp Bounds)**.** For $\theta_n = \text{SATE}$ estimation and 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 GMM estimators and stratified rerandomization designs in our current setting. However, this construction and the associated variance estimators are quite involved, so we leave this significant extension to future work.

7.2 Inference on the Finite Population Parameter

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

Variance Estimator. We begin with some definitions. Let $\mathcal{S}_n$ denote the set of groups 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 β_d above by $\widehat{\beta}_d = v_D E_n[\breve{w}_i\breve{w}_i']^{-1} \operatorname{Cov}_n(\breve{w}_i, \widehat{g}_i|D_i = d)\widehat{\Pi}'$. For $\widehat{g}_i \equiv g(D_i, X_i, S_i, \widehat{\theta}_{adj})$, define the estimated within-arm influence functions $\widehat{m}_{1i} \equiv \widehat{\Pi}\widehat{g}_i - \widehat{\beta}_1' w_i$ and $\widehat{m}_{0i} \equiv \widehat{\Pi}\widehat{g}_i - \widehat{\beta}_0' w_i$. First, assume each group has at least two treated and control units, $2 \le a(s) \le k(s) - 2$ for all $s \in \mathcal{S}$. We relax this in a moment. Define

$$
\begin{aligned}
\widehat{v}_1 &= n^{-1} \sum_{s\in\mathcal{S}_n} \frac{1}{a(s)-1} \sum_{i\neq j\in s} \widehat{m}_{1i}\widehat{m}_{1j}' D_i D_j / p \\
\widehat{v}_0 &= n^{-1} \sum_{s\in\mathcal{S}_n} \frac{1}{(k-a)(s)-1} \sum_{i\neq j\in s} \widehat{m}_{0i}\widehat{m}_{0j}' (1-D_i)(1-D_j)/(1-p) \\
\widehat{v}_{10} &= n^{-1} \sum_{s\in\mathcal{S}_n} \frac{k}{a(k-a)}(s) \sum_{i,j\in s} \widehat{m}_{1i}\widehat{m}_{0j}' D_i (1-D_j).
\end{aligned}
$$

Collapsed Strata. If $a(s) = 1$ or $a(s) = k(s) - 1$, as in the case of matched pairs and matched triples designs, the inner sums in $\widehat{v}_1$ and $\widehat{v}_0$ are empty. In this case, we follow[29] the method of collapsed strata (Hansen et al. (1953)), agglomerating the original groups $s \in \mathcal{S}_n$ into larger "group patches." For instance, in a matched triples design with $p = 1/3$, we agglomerate two triples into a larger group s' of 6 units with two treated units, $a(s') = 2$. 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 $\nu(s) \neq s$, $\nu^2 = \mathrm{Id}$, 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, ν 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 group patches. If $a(s) = 1$ or $a(s) = k(s) - 1$, we replace $\mathcal{S}_n$ with $\mathcal{S}_n^\nu$ in the definitions of $\widehat{v}_1$ and $\widehat{v}_0$ above.

[29] Also see Abadie and Imbens (2008), Bai et al. (2021), and Cytrynbaum (2024b) more recent results. In particular, Bai et al. (2021) showed asymptotic exactness of the collapsed strata method below for matched pairs designs.

Finally, let $\widehat{u}_1 = E_n[\frac{D_i}{p}\widehat{m}_{1i}\widehat{m}_{1i}'] - \widehat{v}_1$ and $\widehat{u}_0 = E_n[\frac{1-D_i}{1-p}\widehat{m}_{0i}\widehat{m}_{0i}'] - \widehat{v}_0$ and define

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

The theorem requires a slight strengthening of GMM Assumption 3.2.

Assumption 7.4. *There 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 7.5 (Inference)**.** *Suppose $D_{1:n}$ as in Definition 2.1 and impose Assumptions 3.1, 3.2, 7.4. Then $\widehat{V}_a(c) \xrightarrow{p} \bar{V}_a^{adj}(c) \geq V_a^{adj}(c)$.*

Combining Theorem 6.3 and Theorem 7.5, the interval $\widehat{C}_{fin} \equiv [c'\widehat{\theta}_{adj} \pm z_{1-\alpha/2}\widehat{V}_a(c)^{1/2}/\sqrt{n}]$ has $P(c'\theta_n \in \widehat{C}_{fin}) \geq 1 - \alpha - o(1)$.

7.3 Inference on Superpopulation Parameter

The asymptotic variance $V = V_\phi + V_a^{adj}$ for adjusted estimation of the superpopulation parameter θ_0 under stratified rerandomization (Theorem 6.3) is identified. In this case, we can modify the approach above to provide asymptotically exact inference methods. Let $\widehat{m}_i = \widehat{\Pi}\widehat{g}_i - D_i\widehat{\beta}_1' w_i - (1-D_i)\widehat{\beta}_0' w_i$ and define variance estimator

$$\widehat{V} = \mathrm{Var}_n(\widehat{m}_i) - v_D(\widehat{v}_1 + \widehat{v}_0 - \widehat{v}_{10} - \widehat{v}_{10}'). \tag{7.3}$$

Theorem 7.6 (Inference)**.** *Suppose $D_{1:n}$ is as in Definition 2.1, and impose Assumptions 3.1, 3.2, 7.4. 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 θ_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 $P(c'\theta_0 \in \widehat{C}_{pop}) = 1 - \alpha - o(1)$.

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 ψ, 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 β_d as follows:

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

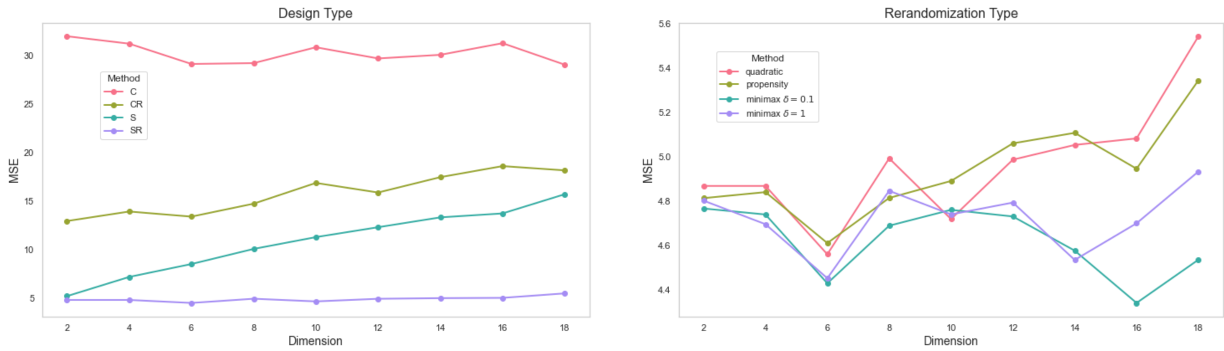

Figure 4: Designs and rerandomization types for $n = 150$, varying $\dim(r)$.

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 Model 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 Model 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 $r \sim \mathcal{N}(0, \Sigma)$. For Tables 1 and 2, we let $\Sigma = I_m$. For Table 3 below, we set $\Sigma_{ii} = 1$ and $\Sigma_{ij} = (1/2)(m-1)^{-1}$ for $i \neq j$. The residuals $(e_1, e_0) \sim \mathcal{N}(0, \tilde{\Sigma})$ with $\mathrm{Var}(e_d) = 4$, $\mathrm{Corr}(e_1, e_0) = 0.8$, and $(e_1, e_0) \perp\!\!\!\perp r$. We set $p = 1/2$ in all simulations, corresponding to matched pairs rerandomization for ψ, h non-constant.

In Table 1, we compare the efficiency and inference properties of various designs for estimating $\theta_n = \mathrm{SATE}$. 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.[30] 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), with acceptance probability $\alpha = 1/500$. $\widehat{\theta}$ is the unadjusted GMM estimator of Definition 2.3, 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 θ_n using the (conservative) finite population variance bound estimator $\widehat{V}_a(c)$ in Section 7.2 and confidence interval $\widehat{C}_{fin} = [\widehat{\theta}_{adj} \pm 1.96 \cdot \widehat{V}_a(c)^{1/2}/\sqrt{n}]$. Cover Pop. presents coverage of θ_0 for $\widehat{C}_{pop} = [\widehat{\theta}_{adj} \pm 1.96 \cdot \widehat{V}^{1/2}/\sqrt{n}]$, using asymptotically exact variance estimator $\widehat{V}$ from Section 7.3. CI Width Fin. and Pop. report the width confidence

[30]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

intervals, normalized so that the width of $\widehat{C}_{pop}$ is 1 for $\widehat{\theta}_{adj}$ and design **C**. We provide additional results for Model 4 in Figure 4, letting $n = 150$ and finely varying $\dim(r)$. **CR** refers to complete rerandomization, without stratification.

Next, 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 $\widehat{\theta}_{adj}$), 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 γ_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.[31] By contrast, coverage is approximately nominal for designs **S** and **SR**. Note that $\widehat{C}_{fin}$ is often much smaller than $\widehat{C}_{pop}$, showing that experimenters only interested in covering θ_n can potentially report smaller confidence intervals.

In Table 2 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.[32] In addition to θ_n = SATE, we also provide efficiency and inference results for the treatment effect heterogeneity parameter from Example 3.11. In particular, let $\alpha_n = \operatorname{argmin}_\alpha E_n[(Y_i(1) - Y_i(0) - \alpha'(1, r_{1i}))^2]$. We define θ_n to be the coefficient on r_1, denoting θ_n = CATE in the table. Cover Pop. and CI Width Pop. refer to inference on the corresponding superpopulation parameter θ_0.

Next, we summarize a few findings from Table 2. 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,

[31]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.

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

			MSE		Cover		CI Width	
θ_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 2: Stratified Rerandomization Types

				MSE		Cover		CI Width	
M.	Dim	Inter.	Design	$\widehat{\theta}$	$\widehat{\theta}_{adj}$	Pop.	Fin.	Pop.	Fin.
1	15	No	C	1.00	0.95	0.94	0.97	1.00	0.83
			SR	0.96	0.96	0.94	0.98	1.01	0.87
	30	Yes	C	1.00	0.91	0.93	0.95	0.93	0.70
			SR	0.90	0.90	0.94	0.98	0.98	0.80
2	15	No	C	1.00	0.88	0.94	0.95	1.00	0.87
			SR	0.62	0.62	0.95	0.97	0.85	0.74
	30	Yes	C	1.00	0.60	0.92	0.96	0.70	0.67
			SR	0.61	0.60	0.98	0.99	0.90	0.82
3	15	No	C	1.00	0.87	0.94	0.97	1.00	0.85
			SR	0.52	0.53	0.96	0.98	0.87	0.64
	30	Yes	C	1.00	0.76	0.89	0.93	0.79	0.66
			SR	0.51	0.60	0.96	0.99	0.95	0.78
4	15	No	C	1.00	0.92	0.94	0.96	1.00	0.93
			SR	0.87	0.87	0.94	0.97	0.98	0.91
	30	Yes	C	1.00	0.89	0.92	0.94	0.90	0.83
			SR	0.86	0.87	0.95	0.97	0.97	0.91

Table 3: Interacted Design for $\theta_n = \text{CATE}$

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. We provide additional results for Model 4 in Figure 4, showing that **Opt1** and **Opt2** reduce the curse of dimensionality for rerandomization, since we are able to downweight less important dimensions of h. Finally, in Table 3, we provide additional simulation results for estimating the heterogeneity parameter θ_n = CATE. In particular, Example 3.11 showed that if the experimenter is interested in treatment effect heterogeneity along dimension r_1, then they should balance variables ψ, h and w predictive of the interaction $\bar{Y} r_1$, not just the outcome level $\bar{Y}$. The designs in Table 3 are as above for no interactions (Inter. = No). In the "Yes" case, we add interactions so that rerandomization and ex-post adjustment covariates $h = (r, r \cdot r_1)$, and $w = (r, r \cdot r_1)$, keeping $\psi = r_1$. This significantly increases efficiency for $\widehat{\theta}_{adj}$ under design **C**, with smaller efficiency gains for design **SR**.

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 high grades to first and second year students at a large Canadian university. They estimated the effect of the program on future student GPA, successful graduation, and other outcomes. They measured several 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 $D = 1$ 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 D and estimate both intention-to-treat (ITT) effects and effects on compliers (LATE). Let $A \in \{0, 1\}$ denote endogeneous decision to engage with the program, with $A(d)$ the potential treatments, $Y(a)$ the potential outcomes, and $T(d) = Y(A(d))$ the ITT potential outcomes with realized outcome $T = Y(A(D)) = Y$. Angrist et al. (2013) 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, including both rerandomized and standard 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 $\widehat{T}(d) = T = Y$ if $D = d$ in the observed data, and

		MSE		Cover		CI Width	
θ_n (ITT)	Design	$\widehat{\theta}$	$\widehat{\theta}_{adj}$	Pop.	Fin.	Pop.	Fin.
SATE	C	1.21	1.00	0.94	0.98	1.00	0.96
	CR	1.05	1.00	0.94	0.98	1.00	0.96
	S	1.01	1.00	0.94	0.98	0.99	0.95
	SR	1.00	1.01	0.94	0.97	0.99	0.95
	F	1.05	0.98	0.94	0.97	0.98	0.93
	FR	0.98	0.97	0.94	0.98	0.98	0.93
	F+	0.96	0.98	0.96	0.98	1.00	0.95
	FR+	0.97	0.97	0.95	0.98	1.00	0.95
CATE (Fin.)	C	3.22	1.00	0.93	0.97	1.00	1.02
	CR	1.80	0.96	0.95	0.98	0.99	1.00
	S	3.09	1.01	0.94	0.97	0.99	1.01
	SR	1.73	0.94	0.95	0.98	0.99	1.01
	F	3.13	1.04	0.94	0.97	1.01	1.02
	FR	1.72	0.99	0.95	0.98	1.01	1.01
	F+	2.91	0.99	0.95	0.97	1.02	1.03
	FR+	1.48	0.97	0.95	0.98	1.02	1.02
CATE (GPA)	C	3.01	1.00	0.93	0.98	1.00	1.02
	CR	1.63	0.93	0.94	0.98	0.99	1.00
	S	1.37	0.91	0.94	0.98	0.95	0.98
	SR	1.02	0.87	0.96	0.99	0.94	0.97
	F	0.86	0.91	0.96	0.98	1.03	0.95
	FR	0.80	0.81	0.97	0.99	1.01	0.93
	F+	1.34	1.36	0.95	0.96	1.33	1.28
	FR+	1.27	1.27	0.96	0.97	1.32	1.27

Table 4: ITT Parameters

impute $\widehat{T}(d) = \widehat{m}_d^T(X) + \widehat{\sigma}_d^T(X)\epsilon_d$ if $D = 1-d$, where $\widehat{m}_d(X)$, $\widehat{\sigma}_d(X)$ are estimated using cross-validated LASSO and random forests applied to 11 baseline covariates their full pairwise interactions. The residual $\epsilon_d \sim \mathcal{N}(0,1)$. We similarly impute missing potential treatments $\widehat{A}(d)$ for all units with $\widehat{A}(d) = A$ if $D = d$. See Section 10.1 for more details on this procedure.

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

		MSE		Cover		CI Width	
θ_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
C-LATE (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
C-LATE (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 5: LATE Parameters

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 3.11. 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+**.[33] We let $p = 3/10$ and $n = 1200$ for all.

We present empirical results for several causal estimands. Table 4 presents results on the ITT estimands SATE $= E_n[T_i(1) - T_i(0)]$ and "CATE," the coefficient on x_i in

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

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

$$\text{CATE} = E_n[T_i(1) - T_i(0)|F_i = 1] - E_n[T_i(1) - T_i(0)|F_i = 0].$$

Table 5 presents efficiency and inference results for LATE-style treatment effects on compliers. In particular, if $C_i = \mathbb{1}(A_i(1) - A_i(0) > 0)$ is a compliance indicator then LATE $= E_n[Y_i(1) - Y_i(0)|C_i = 1]$ and CLATE (Example 2.8) is the coefficient on x_i in

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

In both tables, Cover Pop. and CI Width Pop. refer to inference on the corresponding superpopulation estimands θ_0, e.g. $\theta_0 = \operatorname{argmin}_\theta E[(Y(1) - Y(0) - \theta'(1, x))^2 | C = 1]$ for $\theta_n =$ CLATE and $\theta_0 = E[T(1) - T(0)] =$ ATE for $\theta_n =$ SATE. 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 heterogeneity variables CATE and CLATE than for average effects SATE and LATE. Finely stratified rerandomization **FR** is the efficient for the majority of estimands, while **SR** is slightly more efficient for estimating treatment effect heterogeneity CATE and CLATE along the financial need variable $F \in \{0, 1\}$. Confidence intervals broadly have correct coverage. The width of $\widehat{C}_{fin}$ for inference on θ_n is slightly smaller than $\widehat{C}_{pop}$ for inference on θ_0 on average, with the largest improvements for estimating CATE (GPA) and CLATE (GPA).

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

References

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. (2023). Covariate adjustment in experiments with matched pairs.

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. (2024). 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*.

Chernozhukov, V., Chetverikov, D., Duflo, E., Hansen, C., and Newey, W. (2017). Double / debiased / neyman machine learning of treatment effects. *American Economics Review Papers and Proceedings*, 107(5):261–265.

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

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

Gretton, A., Borgwardt, K. M., Rasch, M. J., Scholkopf, B., and Smola, A. (2008). A kernel method for the two-sample problem. *Journal of Machine Learning Research.*

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.

Kakehi, H. and Otsu, T. (2024). Finite-population inference via gmm estimator.

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.

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.*

Robins, J. M. and Rotnitzky, A. (1995). Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association*, 90:122–129.

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, Q., Linton, O., and Hardle, W. (2004). Semiparametric regression analysis with missing response at random. *Journal of the American Statistical Association*.

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*.

10 Appendix

10.1 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 $\widehat{T}(d) = T = Y$ if $D = d$ and impute $\widehat{T}(d) = \widehat{m}_d^T(X) + \widehat{\sigma}_d^T(X)\epsilon_d$ if $D = 1 - d$. The function $\widehat{m}_d^T(X)$ is estimated using LASSO, regressing TD/p on X for $d = 1$ and $T(1-D)/(1-p)$ on X for $d = 0$, with regularization parameter chosen by cross-validation. The variance function $\widehat{\sigma}_d^T(X)$ is estimated by random forests to preserve positivity, regressing $(T_i - \widehat{m}_1^T(X))^2 D_i/p$ on X_i for $d = 1$ and $(T_i - \widehat{m}_0^T(X))^2(1 - D_i)/(1-p)$ on X_i for $d = 0$. The potential treatments $\widehat{A}(d) \in \{0, 1\}$ are imputed similarly, with $\widehat{A}(d) = A$ if $D = d$ and $\widehat{A}(d) = \mathbb{1}(\widehat{m}_d^A(X) + \widehat{\sigma}_d^A(X)u_d \geq 1/2)$ with $u_d \sim \mathcal{N}(0, 1)$ and both $\widehat{m}_d^A(X), \widehat{\sigma}_d^A(X)$ estimated by cross-validated random forests, with estimation procedure identical to the ITT outcomes above.

11 Proofs

11.1 Rerandomization Distribution

In what follows, we carefully distinguish between the the law of the data $(W_{1:n}, D_{1:n})$ under "pure" stratified randomization, which we denote by P, and the law under rerandomized stratification, which we denote by Q. First, we formally define pure stratification.

Definition 11.1 (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, as studied in Cytrynbaum (2024b).

Next, we slightly generalize the rerandomization designs introduced in Definition 2.1, which will be useful for our study of nonlinear rerandomization in Section 4. We let Q denote the law of $(W_{1:n}, D_{1:n})$ under this design.

Definition 11.2 (Rerandomization)**.** Consider the following:

(a) (Acceptance Regions). Suppose $\mathcal{I}_n = \sqrt{n}\widehat{\Delta}_h + o_p(1)$ for $\widehat{\Delta}_h = E_n[H_i h_i]$ with $H_i = (D_i - p)/(p - p^2)$ and $\tau_n = \tau + o_p(1)$ 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. We accept $D_{1:n}$ if $\mathcal{I}_n \in T_n$.

(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 11.1, define the rerandomization distribution

$$Q(B|\mathcal{F}_n) = P(B|\mathcal{F}_n, \mathcal{I}_n \in T_n), \quad Q(B) = E[Q(B|\mathcal{F}_n)]. \tag{11.1}$$

(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 $E[|\phi|_2^2 + |h|_2^2] < \infty$.

Our work below shows that rerandomization as in Definition 2.1 of the main text specializes Definition 11.2 to $b(x, y) = b(x) = d(x, A) - d(x, A^c)$ for distance function $d(x, A) = \inf_{z \in \mathbb{R}^{d_h}} |x - z|_2$.

The following essential lemma shows that the high level properties (e.g. convergence in probability) of P are inherited by the rerandomized version Q. The proof is given in Section 11.9 below.

Lemma 11.3 (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 11.2.*

(a) If $B_n \in \mathcal{F}_n$ then $P(B_n) = Q(B_n)$. In particular, if a random variable 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.

Equipped with this lemma, we will take the following approach: (1) show linearization of the GMM estimator $\widehat{\theta}$ about θ_n and θ_0 under P, (2) invoke Lemma 11.3 to show these properties still hold under Q, then (3) prove distributional convergence of the simpler linearized quantities directly under Q. GMM linearization (1) is discussed in Section 11.3. For (3), the next section derives the conditional asymptotic distribution of quantities of the form $\sqrt{n}E_n[H_i a(W_i)]$ under the rerandomization measure Q.

11.2 Rerandomization Asymptotics

Before studying rerandomization, we first establish a CLT for pure stratified designs, conditional on the data $W_{1:n}$.

Theorem 11.4 (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 11.1. 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 $V = v_D^{-1}E[\text{Var}(a|\psi)]$.*

Proof. 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 $|\mathcal{S}_n| \le n/k$. 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 $j \ge 1$. 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) We claim that $(S_{j,n}, \mathcal{H}_{j,n})_{j \ge 1}$ is an MDS. Adaptation is clear from our definitions.

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

The second equality since $D_{s(j)} \perp\!\!\!\perp (D_{s(l)})_{l \ne j}|\mathcal{F}_n$. 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), we have

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

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 σ_n^2 as

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

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 $r \ne t$ 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 $\sum_k a_k \mathbb{1}(\sum_k a_k > c) \le m \sum_k a_k \mathbb{1}(a_k > c/m)$ and $ab\mathbb{1}(ab > c) \le a^2\mathbb{1}(a^2 > c) + b^2\mathbb{1}(b^2 > c)$. Then for $c_n \to \infty$ with $c_n = o(\sqrt{n})$ we have

$$
\begin{aligned}
&|v_j|\mathbb{1}(|v_j| > c_n) \le \sum_{r \ne t \in s(j)} |u_r u_t| \mathbb{1}\left(\sum_{r \ne t \in s(j)} |u_r u_t| > c_n \right) \\
&\le k^2 \sum_{r \ne t \in s(j)} |u_r u_t| \mathbb{1}(|u_r u_t| > c_n/k^2) \le 2k^3 \sum_{r \in s(j)} u_r^2 \mathbb{1}(u_r^2 > c_n/k^2).
\end{aligned}
$$

Then we have

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

Then $E[A_n] = 2k^3E[E_n[E\left[u_i^2\mathbb{1}(u_i^2>c_n/k^2)|\psi_i\right]]] = 2k^3E[u_i^2\mathbb{1}(u_i^2 > c_n/k^2)] \to 0$ 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 $E[u_i^2]\le E[a_i^2]<\infty$ 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 the Lindberg condition $\sum_{j=1}^{n/k} E[Z_{j,n}^2\mathbb{1}(|Z_{j,n}|>\epsilon)|\mathcal{H}_{0,n}] = o_p(1)$.

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

Then using the inequality above we compute

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

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 $\dim(a)\ge 1$. 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. □

Next, we provide asymptotic theory for stratified rerandomization. The following definition generalizes Definition 2.1 in Section 1.

Lemma 11.5. *Let Definition 11.2 hold. Let $\widehat{\Delta}_a = E_n[H_ia_i]$ 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 11.1*

$$E\left[e^{it'\sqrt{n}\widehat{\Delta}_a} \mathbb{1}\left(\mathcal{I}_n \in T_n\right) | \mathcal{F}_n\right] = E\left[e^{it' Z_a} \mathbb{1}\left(Z_h \in T\right)\right] + o_p(1).$$

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

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

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 $|R_n| \le 2$, so $E[|R_n||\mathcal{F}_n] = o_p(1)$ by Lemma 11.20. 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 11.4 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 11.17. Fix $t \in \mathbb{R}$ and define $G(z_1, z_2, x) = e^{it'z_1}\mathbb{1}(b(z_2, x) \le 0)$ and note that

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

Define $E_G = \{w : G(\cdot) \text{ not continuous at } w\}$. By Proposition 11.17, 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 $g(z_2, x) = \mathbb{1}(b(z_2, x) \le 0)$ 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 discontinuity points $E_g^1 \subseteq \{z_h : b(z_h, \tau) = 0\}$. $\square$

Theorem 11.6 (Asymptotic Distribution)**.** *Let Definition 11.2 hold. Suppose that* $(Z_a, Z_h) \sim v_D^{-1} E[\mathrm{Var}((a, h)|\psi)]$. *Then under* Q *in Definition 11.2 the following hold:*

(a) We have $\sqrt{n}E_n[H_i a(W_i)]|\mathcal{F}_n \Rightarrow Z_a|Z_h \in T = \mathcal{N}(0, V_a) + R$*, 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*

$$\sqrt{n}(X_n - E[\phi(W)]) \Rightarrow Z_\phi + Z_a|Z_h \in T = \mathcal{N}(0, V_\phi) + \mathcal{N}(0, V_a) + R.$$

The RV's are independent with $V_\phi = \mathrm{Var}(\phi(W))$.

Proof. First, we prove (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 11.5, $a_n \xrightarrow{p} a_\infty$ and $b_n \xrightarrow{p} b_\infty$, with $b_\infty > 0$ by assumption in Definition 11.2. 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 11.5. 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 γ_0 exists and satisfies the stated inequality by Lemma 11.18. Letting $\tilde{Z}_a = Z_a - \gamma_0' Z_h$, by Lemma 11.18 $\tilde{Z}_a \perp\!\!\!\perp Z_h$ 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 11.17, 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 the proof of (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 11.20. 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). $\square$

Lemma 11.7 (Linearization). *Suppose Definition 11.2 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* $\sqrt{n}(\widehat{\theta}-\theta_0) = \sqrt{n}E_n[\Pi\phi(W_i,\theta_0) + H_i\Pi a(W_i,\theta_0)] + o_p(1)$.

See Section 11.3 below for the proof of this lemma.

Proof of Theorem 3.5. We claim that the conditions of Definition 11.2 hold. This will allow us to apply our general rerandomization asymptotics in Theorem 11.6 and linearization in Lemma 11.7. 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 11.2. For part (b), $P(b(Z_h) = 0) = P(Z_h \in \partial A) = 0$ since $\mathrm{Leb}(\partial A) = 0$ and by absolute continuity of Z_h relative to Lebesgue measure Leb. We also have $P(Z_h \in T) = P(Z_h \in A) > 0$ since Z_h is full measure by $E[\mathrm{Var}(h|\psi)] \succ 0$ and since A has non-empty interior.

This proves the claim. Then by Lemma 11.7, $\sqrt{n}(\widehat{\theta}-\theta_n) = \sqrt{n}E_n[H_i\Pi a(W_i,\theta_0)] + o_p(1)$. The result now follows immediately by Slutsky and Theorem 11.6(a), letting $a \to \Pi a$. Likewise, Corollary 3.8 follows from Theorem 11.6(b), letting $\phi \to \Pi\phi$. $\square$

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 $V_a = v_D^{-1}E[\mathrm{Var}(\Pi a(W,\theta_0) - \gamma_0'h|\psi)]$ and $R \sim \gamma_0'Z_h$ for $Z_h \sim \mathcal{N}(0, v_D^{-1}E[\mathrm{Var}(h|\psi)])$. Then $\mathcal{N}(0,V_a) + R \sim \mathcal{N}(0,V)$ with $V = V_a + \mathrm{Var}(\gamma_0'Z_h) = v_D^{-1}E[\mathrm{Var}(\Pi a(W,\theta_0)-\gamma_0'h+\gamma_0'h|\psi)] - 2v_D^{-1}E[\mathrm{Cov}(\Pi a(W,\theta_0)-\gamma_0'h, \gamma_0'h|\psi)] = v_D^{-1}E[\mathrm{Var}(\Pi a(W,\theta_0)|\psi)]$. The covariance term is zero by Lemma 11.18. The second statement follows by setting $\psi = 1$. $\square$

11.3 GMM Linearization

This section collects proofs needed for the key linearization result in Lemma 11.7. 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 11.8 (ULLN)**.** *Working under P in Definition 11.1:*

(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 ϕ, a are linear combinations of g_d for $d \in \{0, 1\}$, so ϕ 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 θ, $f_{\theta m}(W_i) \to a(W_i, \theta) - a(W_i, \theta) = 0$ 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 θ, let $m(\theta)$ s.t. $E[f_{\theta m(\theta)}(W_i)] \leq \epsilon$. Then $\{U_{\theta m(\theta)} : \theta \in \Theta\}$ is an open cover of Θ, 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 (θ, 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

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

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

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

By symmetry, also $\sup_{\theta\in K} -E_n[H_i a(W_i,\theta)] \le \epsilon + o_p(1)$. Then $\sup_{\theta\in K}|E_n[H_i a(W_i,\theta)]| \le 2\epsilon + o_p(1)$. 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.

For the statement about objective functions, observe that

$$\begin{aligned}&|\widehat{H}(\theta) - H_n(\theta)| = |\widehat{g}(\theta)'M_n\widehat{g}(\theta) - g_n(\theta)'Mg_n(\theta)| \le |(\widehat{g} - g_n)(\theta)'M_n\widehat{g}(\theta)| \\ &+ |g_n(\theta)'(M_n - M)\widehat{g}(\theta)| + |g_n(\theta)'M(\widehat{g} - g_n)(\theta)| \le |\widehat{g} - g_n|_2(\theta)||M_n|_2|\widehat{g}(\theta)|_2 \\ &+ |g_n(\theta)|_2|M_n - M|_2|\widehat{g}(\theta)|_2 + |g_n(\theta)|_2|M|_2|\widehat{g} - g_n|_2(\theta) \lesssim |\widehat{g} - g_n|_{\infty,\Theta}||M_n|_2|\widehat{g}|_{\infty,\Theta} \\ &+ |g_n|_{\infty,\Theta}|M_n - M|_2|\widehat{g}|_{\infty,\Theta} + |g_n|_{\infty,\Theta}|M|_2|\widehat{g} - g_n|_{\infty,\Theta}.\end{aligned}$$

The first inequality by telescoping, then Cauchy-Schwarz, then using equivalence of finite-dimensional vector space norms and $\sup_\theta a(\theta)b(\theta) \le \sup_\theta a(\theta)\sup_\theta b(\theta)$ for positive a, b. We have $|g_n|_{\infty,\Theta}, |\widehat{g}|_{\infty,\Theta} = o_p(1) + |g_0|_{\infty,\Theta} = O_p(1)$ since $|g_0|_{\infty,\Theta} \le E[\sup_{\theta\in\Theta}\phi(W,\theta)] < \infty$. 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.

Next consider (2). Let $U_1 \subseteq \tilde{U}$ an open set $\theta_0 \in U_1$ such that the closed $1/m'$ enlargement $\tilde{U}_1^{1/m'} \subseteq \tilde{U}$ for some $m' \ge 1$. Set $\tilde{\Theta} = \tilde{U}_1^{1/m'}$, which is compact. As in the proof of (1), let $U_{\theta m} = B(\theta, m^{-1})$ for $m \ge m'$. The conclusion now follows from the exact argument in (1), applied to the alternate moment functions $\tilde{g}_z(W_i,\theta) \equiv \partial/\partial\theta' g_z(W_i,\theta)$. In particular, uniform convergence holds on any open set $U \subseteq \tilde{\Theta} \subseteq \tilde{U}$. The final statement about $G_0(\theta)$ follows by dominated convergence. □

Lemma 11.9 (Consistency). *Under the distribution P in Definition 11.1, if Assumption 3.2 holds then $\widehat{\theta} - \theta_0 = o_p(1)$ and $\theta_n - \theta_0 = o_p(1)$.*

Proof. By definition, $\widehat{\theta} = \operatorname{argmin}_{\theta\in\Theta} \widehat{H}(\theta)$. Moreover, $g_n(\theta_n) = 0$ so $H_n(\theta_n) = 0$ and $\theta_n \in \operatorname{argmin}_{\theta\in\Theta} H_n(\theta)$. For (2), since $g_0(\theta_0) = 0$ uniquely and $\operatorname{rank}(M) = d_g$, then $H_0(\theta)$ is uniquely minimized at θ_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 $\widehat{\theta} \xrightarrow{p} \theta_0$. $\square$

Proof of Lemma 11.7. By Lemma 11.3, it suffices to show the result under P in Definition 11.1. Since $\widehat{\theta} = \operatorname{argmin}_{\theta\in\Theta} \widehat{H}(\theta)$, 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

$$\widehat{g}(\widehat{\theta}) = \widehat{g}(\theta_0) + \frac{\partial \widehat{g}_k}{\partial \theta'}(\tilde{\theta}_k)_{k=1}^{d_g}(\widehat{\theta} - \theta_0).$$

Then we may expand

$$0 = \widehat{G}(\widehat{\theta})' M_n[\widehat{g}(\theta_0) + \frac{\partial \widehat{g}_k}{\partial \theta'}(\tilde{\theta}_k)_{k=1}^{d_g}(\widehat{\theta} - \theta_0)]$$
$$\widehat{\theta} - \theta_0 = -(\widehat{G}(\widehat{\theta})' M_n \frac{\partial \widehat{g}_k}{\partial \theta'}(\tilde{\theta}_k)_{k=1}^{d_g})^{-1} \widehat{G}(\widehat{\theta})' M_n \widehat{g}(\theta_0).$$

On the event $S_n = \{\widehat{\theta} \in U\}$, $\tilde{\theta}_k \in U$ for each k. Then $\mathbb{1}(S_n)|\frac{\partial \widehat{g}_k}{\partial \theta'}(\tilde{\theta}_k)_{k=1}^{d_g} - \frac{\partial g_{0k}}{\partial \theta'}(\tilde{\theta}_k)_{k=1}^{d_g}|_F^2 \leq \sum_{k=1}^{d_g} \sup_{\theta\in U} |\frac{\partial \widehat{g}_k}{\partial \theta'}(\theta) - \frac{\partial g_{0k}}{\partial \theta'}(\theta)|_2^2 \leq d_g \sup_{\theta\in U} |\widehat{G}(\theta) - G_0(\theta)|_F^2 = o_p(1)$ by Lemma 11.8. Similarly, $\mathbb{1}(S_n)|\widehat{G}(\widehat{\theta}) - G_0(\widehat{\theta})|_F^2 \leq \sup_{\theta\in U} |\widehat{G}(\theta) - G_0(\theta)|_F^2 = o_p(1)$. Moreover, since $\widehat{\theta} \xrightarrow{p} \theta_0$ and $\tilde{\theta}_k \in [\theta_0, \widehat{\theta}]\ \forall k$, we have $\mathbb{1}(S_n)|G_0(\widehat{\theta}) - G(\theta_0)|_F^2 = o_p(1)$ and $\mathbb{1}(S_n)|\frac{\partial g_{0k}}{\partial \theta'}(\tilde{\theta}_k)_{k=1}^{d_g} - G(\theta_0)|_F^2 = o_p(1)$, using continuous mapping and continuity of $\theta \to G_0(\theta)$ on U, shown in Lemma 11.8. Since $P(S_n) \to 1$, we have shown $|\widehat{G}(\widehat{\theta}) - G(\theta_0)|_F^2 = o_p(1)$ and $|\frac{\partial \widehat{g}_k}{\partial \theta'}(\tilde{\theta}_k)_{k=1}^{d_g} - G(\theta_0)|_F^2 = o_p(1)$. Since $\widehat{g}(\theta_0) = O_p(n^{-1/2})$ by Theorem 11.4, by the work above and continuous mapping theorem we have

$$\begin{aligned}\sqrt{n}(\widehat{\theta} - \theta_0) &= -(\widehat{G}(\widehat{\theta})' M_n \frac{\partial \widehat{g}_k}{\partial \theta'}(\tilde{\theta}_k)_{k=1}^{d_g})^{-1} \widehat{G}(\widehat{\theta})' M_n \sqrt{n}\widehat{g}(\theta_0) \\ &= -(G'MG)^{-1} G'M\sqrt{n}\widehat{g}(\theta_0) + o_p(1) = \Pi\sqrt{n}\widehat{g}(\theta_0) + o_p(1).\end{aligned}$$

This proves the second statement of Lemma 11.7. For the first statement, substituting θ_n, H_n, G_n for $\widehat{\theta}, \widehat{H}, \widehat{G}$ in the above argument, we have $\sqrt{n}(\theta_n - \theta_0) = \Pi\sqrt{n} g_n(\theta_0) + o_p(1)$. 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. $\square$

11.4 Linearization for M-Estimation

In this section, we extend our key result to M-estimation $\widehat{\theta} = \operatorname{argmax}_{\theta\in\Theta} E_n[m(D_i, R_i, S_i, \theta)]$. M-estimation is often equivalent to GMM with score $\nabla_\theta m(D, R, S, \theta)$, e.g. if $\theta \to m(\cdot, \theta)$ is strictly concave. However, this equivalence fails when $E[m(D,R,S,\theta)]$ has local maxima, violating GMM identification (Assumption 3.2). E.g. see Newey and McFadden (1994) for examples. To handle such cases, in this section we analyze M-estimation under weaker conditions. Let $m_d(W,\theta) = m(d,R,S(d),\theta)$ and define $\varphi_m(W,\theta) = E[m(D,R,S,\theta)|W] = pm_1(W,\theta) + (1-p)m_0(W,\theta)$ and $\theta_n = \text{argmax}_{\theta\in\Theta} E_n[\varphi_m(W_i,\theta)]$. Define $g(D,R,S,\theta) = \nabla_\theta m(D,R,S,\theta)$ and let ϕ, a as in the main text, e.g. $\phi(W,\theta) = \nabla_\theta E[m(D,R,S,\theta)|W]$.

Assumption 11.10 (M-estimation)**.** *The following conditions hold for $d \in \{0,1\}$:*

(a) (Consistency). $\theta_0 = \text{argmax}_{\theta\in\Theta} E[\varphi_m(W,\theta)]$ uniquely and $E[\sup_{\theta\in\Theta}|m_d(W,\theta)|_2] < \infty$. Also $\theta \to m_d(W,\theta)$ is continuous almost surely and Θ is compact.

(b) (CLT). Let $g_d(W,\theta) = \nabla_\theta m_d(W,\theta)$. We have $E[g_d(W,\theta_0)^2] < \infty$. There exists a neighborhood $\theta_0 \in U \subseteq \Theta$ such that $G_d(W,\theta) \equiv \partial/\partial\theta' g_d(W,\theta) = (\partial^2/\partial\theta\partial\theta')m_d(W,\theta)$ exists and is continuous. Also $E[\sup_{\theta\in U}|G_d(W,\theta)|_F] < \infty$.

The next result extends our key lemma to this setting. Combined with the results of Section 11.2, this suffices to show that the main results of Sections 3-7 also apply to M-estimators with multiple local maxima.

Lemma 11.11 (Linearization)**.** *Suppose Definition 11.2 and Assumption 11.10 hold for the M-estimator $\widehat{\theta}$. Let $G = E[(\partial^2/\partial\theta\partial\theta')m(W,\theta_0)]$ and set $\Pi = -G^{-1}$. 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. By Lemma 11.3, it suffices to show the result under the distribution P. We have $|E_n[m(D_i,R_i,S_i,\theta)] - E[\varphi_m(W,\theta)]|_{\infty,\Theta} = o_p(1)$, $\theta \to E[\varphi_m(W,\theta)]$ continuous, and 11.8 and also $|E_n[\varphi_m(W_i,\theta)] - E[\varphi_m(W,\theta)]|_{\infty,\Theta} = o_p(1)$, all by Lemma 11.8. Then by extremum consistency, we have $\theta_n \xrightarrow{p} \theta_0$ and $\widehat{\theta} \xrightarrow{p} \theta_0$. By Lemma 11.8 again, 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 $\|G_n - G_0\|_{\infty,U} = o_p(1)$ for $\widehat{G}_n(\theta) = (\partial^2/\partial\theta\partial\theta')E_n[m(D_i,R_i,S_i,\theta)]$, $G_n(\theta) = (\partial^2/\partial\theta\partial\theta')E_n[\varphi_m(W_i,\theta)]$, and $G_0(\theta) = (\partial^2/\partial\theta\partial\theta')E[\varphi_m(W,\theta)]$. Also, $G_0(\theta)$ is continuous on U. Defining $\widehat{g}(\theta) = E_n[(\partial/\partial\theta)m(D_i,R_i,S_i,\theta)]$ and $g_n(\theta) = E_n[\varphi_m(W_i,\theta)]$, by optimality we have $\widehat{g}(\widehat{\theta}) = 0$ and $g_n(\theta_n) = 0$. Then result now follows exactly by the proof of Lemma 11.7, with a slightly simpler first order condition. □

11.5 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 $|x|^2_{2,A} = x'Ax$, define

$$\widehat{\beta}_d \in \underset{\beta\in\mathbb{R}^{d_\beta}}{\text{argmin}}\, |E_n[\mathbb{1}(D_i = d)m(X_i,\beta)]|^2_{2,\Delta_n}.$$

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 β^* uniquely solves $E[m(X,\beta^*)] = 0$. Let $G_m = E[(\partial/\partial\beta')m(X,\beta^*)]$, which is 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 11.7 to GMM estimation using g^1 and g^0, under the measure P in Definition 11.1 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 γ_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

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

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

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

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

For the corollary, consider letting $\widehat{\beta} \in \text{argmin}_{\beta \in \mathbb{R}^{d_\beta}} |E_n[m(X_i, \beta)]|_{2,\Delta_n}^2$. 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 Π_m as above. Then by Lemma 11.7 $\sqrt{n}(\widehat{\beta} - \beta^*) = \Pi_m E_n[m(X_i, \beta^*)] + o_p(1) = O_p(1)$. 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 $|v_{ijk}^n| \le \sup_{\beta \in U} |(\partial/\partial\beta')m(X_i, \beta)|_F \in L_1$ 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 $\mathcal{I}_n = \sqrt{n}E_n[H_i m(X_i, \widehat{\beta})] = \sqrt{n}E_n[H_i m(X_i, \beta^*)] + o_p(1)$. The result now follows from Theorem 3.5. □

Assumption 11.12 (Propensity Rerandomization)**.** *Impose the following conditions.*

(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) Covariates $X = (1, h)$ for $E[|h|_2^2] < \infty$. Also, $E[\text{Var}(h|\psi)]$, $\text{Var}(h)$ 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 $\text{Var}(h) \succ 0$ 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 $\Pi_m = -G_m^{-1} = \frac{v_D}{L'(c)^2}E[X_i X_i']^{-1}$. By Lemma

11.7, 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 β^* 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 $E_n[|X_i|_2^3] \le E_n[|X_i|_2^2]\max_{i=1}^n |X_i|_2 = O_p(1)o_p(n^{1/2})$ since $E[|X_i|_2^2] < \infty$ by assumption, using Lemma C.8 of Cytrynbaum (2024b). Then using the claim, $\sqrt{n}(\widehat{\beta}-\beta^*) = O_p(1)$, and the linear expansion of $\sqrt{n}(\widehat{\beta}-\beta^*)$ above, we have shown $\mathcal{J}_n = L'(c)^2n(\widehat{\beta}-\beta^*)'E[X_iX_i'](\widehat{\beta}-\beta^*) + o_p(1)$, which is

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

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

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

Define the function $b(x,y) = v_D^2x'\,\mathrm{Var}(h)^{-1}x + y - \epsilon$. Then $\mathcal{J}_n \le \epsilon \iff b(\mathcal{I}_n, \xi_n) \le 0$ for $\mathcal{I}_n = \sqrt{n}E_n[H_ih_i]$ 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 11.2. By Lemma 11.7, 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 11.6, have $\sqrt{n}E_n[H_i \Pi a(W_i)] | \mathcal{F}_n \Rightarrow \mathcal{N}(0, V_a) + R$ with residual variable

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

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

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

This finishes the proof. □

11.6 Covariate Adjustment

Proof of Theorem 3.12. By Lemma 11.7, $\sqrt{n}(\widehat{\theta}^* - \theta_n)$ may be expanded as

$$\begin{aligned}\sqrt{n}(\widehat{\theta} - \theta_n - E_n[H_i m(\psi_i, h_i)]) &= \sqrt{n}E_n[H_i(\Pi a(W_i, \theta_0) - m(\psi_i, h_i))] + o_p(1) \\ &\equiv \sqrt{n}E_n[H_i \beta(W_i, \theta_0)] + o_p(1).\end{aligned}$$

By Theorem 11.6, $\sqrt{n}E_n[H_i \beta(W_i, \theta_0)] | \mathcal{F}_n \Rightarrow \mathcal{N}(0, V)$ with $V = v_D^{-1} \operatorname{Var}(\beta(W, \theta_0))$. Since $\beta(W, \theta_0) = \Pi a(W, \theta_0) - \gamma_0' h - t_0(\psi)$ for (γ_0, t_0) solving Equation 3.7, this completes the proof. □

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 11.4, 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 11.7. The first statement now follows from Slutsky and Theorem 11.4. The second statement follows by the same argument used in the proof of Corollary 3.8. □

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 11.6, $\sqrt{n}(\widehat{\theta}_{adj} - \theta_n) | \mathcal{F}_n \Rightarrow \mathcal{N}(0, V) + R$, independent with

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

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 11.18, it further suffices to show $\beta_0 = 0$ solves $E[\operatorname{Var}(h|\psi)]\beta_0 = E[\operatorname{Cov}(h, a_{\Pi\alpha}|\psi)]$, i.e. that $E[\operatorname{Cov}(h, a_{\Pi\alpha}|\psi)] = 0$. To do so, note that $E[\operatorname{Cov}(h, a_{\Pi\alpha}|\psi)] = E[\operatorname{Cov}(h, (\Pi a - \alpha_0' w)|\psi)] = E[\operatorname{Cov}(h, \Pi a|\psi)] - E[\operatorname{Cov}(h, w|\psi)]\alpha_0$. 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 θ_n. The result for θ_0 follows trivially, as in Corollary 3.8. □

We are required to estimate $\beta_1 = E[\mathrm{Var}(w|\psi)]^{-1}E[\mathrm{Cov}(w, v_D\Pi g_1(W, \theta_0)|\psi)]$. Define

$$\widehat{\beta}_1 = v_D E_n[\check{w}_i\check{w}_i']^{-1}E_n[(D_i/p)\check{w}_i\widehat{g}_i']\widehat{\Pi}' \qquad \widehat{\beta}_0 = v_D E_n[\check{w}_i\check{w}_i']^{-1}E_n[(1-D_i)/(1-p)\check{w}_i\widehat{g}_i']\widehat{\Pi}'.$$

Theorem 11.13. *Suppose $D_{1:n}$ is as in Definition 2.1. Require Assumption 3.1, 3.2. Assume that $E[\mathrm{Var}(w|\psi)] \succ 0$. Define $\widehat{\alpha} = v_D E_n[\check{w}_i\check{w}_i']^{-1}E_n[H_i\check{w}_i\widehat{g}_i']\widehat{\Pi}'$. Then $\widehat{\alpha} = \alpha_0 + o_p(1)$ and $\widehat{\beta}_d = \beta_d + o_p(1)$ for $d = 0, 1$.*

Proof of Theorem 11.13. By Lemma 11.3, it suffices to show the result under P in Definition 11.1. First consider estimating β_1. By Lemma 11.14, $E_n[\check{w}_i\check{w}_i'] = k^{-1}(k-1)E[\mathrm{Var}(w|\psi)]+o_p(1)$. Then if $E[\mathrm{Var}(w|\psi)] \succ 0$, we have $E_n[\check{w}_i\check{w}_i']^{-1} \xrightarrow{p} k(k-1)^{-1}E[\mathrm{Var}(w|\psi)]^{-1}$ by continuous mapping. $\widehat{\Pi} \xrightarrow{p} \Pi$ by assumption. Then it suffices to show that $E_n[(D_i/p)\check{w}_i\widehat{g}_i'] = k^{-1}(k-1)E[\mathrm{Cov}(w, g_1(\theta_0)|\psi)]$. 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 $\tilde{\theta}_i$ 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 11.16. Similarly, $E_n[|\check{w}_i|_2^2] \leq E_n[|w_i|_2^2] = O_p(1)$ by the bound in Lemma 11.14. 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 11.14, 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. Finally note that since $H_i = \frac{D_i}{p} - \frac{1-D_i}{1-p}$, we have

$$\begin{aligned} \alpha_0 &= E[\mathrm{Var}(w|\psi)]^{-1}E[\mathrm{Cov}(w, \Pi a(W, \theta_0)|\psi)] \\ &= E[\mathrm{Var}(w|\psi)]^{-1}E[\mathrm{Cov}(w, v_D\Pi(g_1 - g_0)(W, \theta_0)|\psi)] = \beta_1 - \beta_0 \\ &= \widehat{\beta}_1 - \widehat{\beta}_0 + o_p(1) = v_D E_n[\check{w}_i\check{w}_i']^{-1}E_n[H_i\check{w}_i\widehat{g}_i']\widehat{\Pi}' + o_p(1) = \widehat{\alpha} + o_p(1). \end{aligned}$$

This finishes the proof. □

Lemma 11.14. *Let $E[w_i^2+v_i^2]<\infty$ with $w_i, v_i \in \sigma(W_i)$. Then under P in Definition 11.1, $E_n[(D_i-p)\check{w}_i\check{v}_i] = o_p(1)$ and $E_n[(D_i-p)\check{w}_i v_i] = o_p(1)$. Also $E_n[\check{w}_i\check{v}_i] = \frac{k-1}{k}E[\mathrm{Cov}(w,v|\psi)]+o_p(1)$.*

Proof. First, note $|s|^{-1}\sum_{i\in s}\check{w}_i^2 = |s|^{-1}\sum_{i\in s}(w_i-|s|^{-1}\sum_{j\in s}w_j)^2 = \mathrm{Var}_s(w_i) \le E_s[w_i^2] = |s|^{-1}\sum_{i\in s}w_i^2$. Then in particular $\sum_{i\in s}\check{w}_i^2 \le \sum_{i\in s}w_i^2$ and $E_n[\check{w}_i^2]\le E_n[w_i^2]$. Write $E_n[(D_i-p)\check{w}_i\check{v}_i] = n^{-1}\sum_s u_s$ for $u_s = \sum_{i\in s}(D_i-p)\check{w}_i\check{v}_i$. Let $\mathcal{F}_n = \sigma(W_{1:n},\pi_n)$. Then $\mathcal{S}_n\in\mathcal{F}_n$, $E[u_s|\mathcal{F}_n]=0$ and $u_s \perp\!\!\!\perp u_{s'}|\mathcal{F}_n$ for $s\neq s'$ by Lemma C.10 and Lemma C.9 of Cytrynbaum (2024b). By Lemma C.7 of Cytrynbaum (2024b), it suffices to show $n^{-1}\sum_s E[|u_s|\mathbb{1}(|u_s|>c_n)|\mathcal{F}_n] = o_p(1)$ for some $c_n = o(\sqrt{n})$ with $c_n\to\infty$. Note that $|u_s|\le\sum_{i\in s}|\check{w}_i\check{v}_i| \le \sum_{i\in s}\check{w}_i^2+\sum_{i\in s}\check{v}_i^2 \le \sum_{i\in s}w_i^2+\sum_{i\in s}v_i^2$ by Young's inequality and the bound above. Note that for any positive constants $(a_k)_{k=1}^m$ we have $\sum_k a_k\mathbb{1}(\sum_k a_k > c)\le m\sum_k a_k\mathbb{1}(a_k>c/m)$. Applying this fact and the upper bounds gives

$$\begin{aligned}
&n^{-1}\sum_s E[|u_s|\mathbb{1}(|u_s|>c_n)|\mathcal{F}_n] \le n^{-1}\sum_s E\left[\sum_{i\in s}(w_i^2+v_i^2)\mathbb{1}(\sum_{i\in s}(w_i^2+v_i^2)>c_n)|\mathcal{F}_n\right] \\
&\le 2kn^{-1}\sum_s\sum_{i\in s}w_i^2\mathbb{1}(w_i^2>c_n/2k) + 2kn^{-1}\sum_s\sum_{i\in s}v_i^2\mathbb{1}(v_i^2>c_n/2k)
\end{aligned}$$

The final quantity is $2kE_n[w_i^2\mathbb{1}(w_i^2>c_n/2k)]+2kE_n[v_i^2\mathbb{1}(v_i^2>c_n/2k)] = o_p(1)$. This follows by Markov inequality since $E[E_n[w_i^2\mathbb{1}(w_i^2>c_n/2k)]] = E[w_i^2\mathbb{1}(w_i^2>c_n/2k)]\to 0$ for any $c_n\to\infty$ by dominated convergence. This proves the first statement, and the second statement follows by setting $\check{v}_i\to v_i$ above. For the final statement, calculate

$$\sum_{i\in s}\check{w}_i\check{v}_i = \sum_{i\in s}(w_i-k^{-1}\sum_{j\in s}w_j)(v_i-k^{-1}\sum_{j\in s}v_j) = k^{-1}(k-1)\sum_{i\in s}w_iv_i - k^{-1}\sum_{i\neq j\in s}v_iw_j$$

Clearly $n^{-1}k^{-1}(k-1)\sum_s\sum_{i\in s}w_iv_i = k^{-1}(k-1)E_n[w_iv_i] = k^{-1}(k-1)E[w_iv_i]+o_p(1)$. Then it suffices to show $(kn)^{-1}\sum_s\sum_{i\neq j\in s}v_iw_j = k^{-1}(k-1)E[E[w_i|\psi_i]E[v_i|\psi_i]]+o_p(1)$. If so, $E_n[\check{w}_i\check{v}_i] = k^{-1}(k-1)(E[w_iv_i]-E[E[w_i|\psi_i]E[v_i|\psi_i]])+o_p(1) = k^{-1}(k-1)E[\mathrm{Cov}(w_i,v_i|\psi_i)]+o_p(1)$ as claimed. The analysis of the term $\widehat{v}_{10}$ in Lemma A.6 of Cytrynbaum (2024b) shows

$$\begin{aligned}
&n^{-1}\sum_s\sum_{i\neq j\in s}v_iw_j = n^{-1}\sum_s\sum_{i\neq j\in s}E[v_i|\psi_i]E[w_j|\psi_j]+o_p(1) \\
&=(k-1)E_n[E[v_i|\psi_i]E[w_i|\psi_i]]+o_p(1) = (k-1)E[E[v_i|\psi_i]E[w_i|\psi_i]]+o_p(1).
\end{aligned}$$

By above work, this finishes our proof of the claim. □

11.7 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 $A \equiv \{a : f(a) \le 1\}$ 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 $\lambda \ge 0$. Then note that the LHS event occurs iff $f(a_n) \le \epsilon \iff f(a_n/\epsilon) \le 1 \iff a_n/\epsilon \in A \iff a_n \in \epsilon \cdot A$. This proves the main statement. Suppose B is bounded. Then by Cauchy-Schwarz $f(a) \le |a|_2 \sup_{b\in B}|b|_2 < \infty$ 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 $b(a) = x + \text{sgn}(a'x)\frac{\delta}{2|a|}a$. By assumption, $b(a) \in B$. Then $f(a) = \sup_{b\in B}|b'a| \ge |b(a)'a| = |a'x + \text{sgn}(a'x)(\delta/2)|a|| = |a'x| + (\delta/2)|a| \ge (\delta/2)|a|$. Then $f(a) = \sup_{b\in B}|a'b| \ge (\delta/2)|a|$, 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 $\text{bias}(L_{\gamma A}|Z_h) = E[L_{\gamma A}|Z_{hA}] = E[T_\gamma|Z_{hA}] + \gamma' Z_{hA} = \gamma' Z_{hA}$. For $A_0 = \epsilon B^\circ$, we have $\sup_{\gamma\in B}|\text{bias}(L_{\gamma A}|Z_h)| = \sup_{\gamma\in B}|\gamma' Z_{hA}|$. Note $Z_{hA} \in \epsilon B^\circ$, so $Z_{hA}/\epsilon \in B^\circ$. Then we have

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

The final inequality by definition of B°. 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 $\text{Leb}(A \triangle A_0) \ne 0$ and $P(Z_h \in A) > P(Z_h \in A_0)$. Clearly $A \not\subseteq A_0$. Then $\text{Leb}(A \backslash A_0) > 0$, so $P(Z_h \in A\backslash A_0) > 0$ by absolute continuity. For any $x \in A\backslash A_0 \subseteq (\epsilon B^\circ)^c$, we must have $\sup_{\gamma\in B}|\gamma' x| > \epsilon$. Then $\{\sup_{\gamma\in B}\text{bias}(L_{\gamma A}|Z_h) > \epsilon\} = \{\sup_{\gamma\in B}|Z_{hA}| > \epsilon\} \supseteq \{Z_{hA} \in A\backslash A_0\}$. 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 $P(\sup_{\gamma\in B}\text{bias}(L_{\gamma A}|Z_h) > \epsilon) \ge P(Z_h \in A \setminus A_0) > 0$, which contradicts feasibility of A, proving the claim. □

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

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

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 $\max_{v\in B_p} v'z \le \max_{v\in B_p}|v'z|$. 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 $\sup_{b \in B} |a'b| = |a'x| + |\Sigma' a|_q$. □

11.8 Inference

Proof of Theorem 7.6. By Lemma 11.3, it suffices to show the result under P in Definition 11.1. 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

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

This shows that $V_a = \operatorname{Var}(\kappa_i) - v_D^{-1} E[E[\Pi a - \alpha_0' w|\psi] E[\Pi a - \alpha_0' w|\psi]']$. The proof of Theorem 7.5 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 $\Pi a(W, \theta_0) - \alpha_0' w = v_D \Pi g_1 - \beta_1' w - (v_D \Pi g_0 - \beta_0' w) = \tilde{m}_1 - \tilde{m}_0$. Apparently,

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

This finishes the proof. □

Proof of Theorem 7.5. By Lemma 11.3, it suffices to show the result under P in Definition 11.1. With notation as in the proof of Theorem 7.6, by Theorem 6.3, $c'(\widehat{\theta} - \theta_0) \Rightarrow \mathcal{N}(0, V_a(c))$ with variance $V_a(c) = v_D^{-1} c' E[\operatorname{Var}(\Pi a(W, \theta_0) - \alpha_0' w|\psi)] c$. Recall that $\Pi a(W, \theta_0) = v_D \Pi(g_1 - g_0)(W, \theta_0)$. By optimality in Equation 3.7, for any choice of $b_0, b_1 \in \mathbb{R}^{d_w \times d_\theta}$ and $t_0, t_1 \in \mathcal{L}(\psi)$, we have

$$\begin{aligned}V_a(c) &\leq v_D^{-1} c' \operatorname{Var}(\Pi a(W, \theta_0) - b_1' w + b_0' w - t_1(\psi) + t_0(\psi)) c \\ &= v_D^{-1} c' \operatorname{Var}(v_D \Pi(g_1 - g_0) - b_1' w + b_0' w - t_1(\psi) + t_0(\psi)) c \\ &= v_D^{-1} c' \operatorname{Var}((v_D \Pi g_1 - b_1' w - t_1(\psi)) - (v_D \Pi g_1 - b_0' w - t_0(\psi))) c\end{aligned}$$

Define $\bar{m}_1(b_1, t_1) = v_D \Pi g_1 - b_1' w - t_1(\psi)$ and $\bar{m}_0(b_0, t_0) = v_D \Pi g_1 - b_0' w - t_0(\psi)$. Then by

Cauchy-Schwarz, we may bound the above by

$$v_D^{-1}(\mathrm{Var}(c'\bar{m}_1(b_1,t_1)-c'\bar{m}_0(b_0,t_0))\leq v_D^{-1}(\mathrm{Var}(c'\bar{m}_1(b_1,t_1))+\mathrm{Var}(c'\bar{m}_0(b_0,t_0))$$
$$+2|\operatorname{Cov}(c'\bar{m}_1(b_1,t_1),c'\bar{m}_0(b_0,t_0))|)\leq v_D^{-1}(\mathrm{Var}(c'\bar{m}_1(b_1,t_1))^{1/2}+\mathrm{Var}(c'\bar{m}_0(b_0,t_0))^{1/2})^2.$$

Next, we note that

$$\begin{aligned}\mathrm{Var}(c'\bar{m}_1(b_1,t_1)) &\leq \min_{b_1\in\mathbb{R}^{d_w\times d_\theta}}\min_{t_1\in\mathcal{L}(\psi)}\mathrm{Var}(c'(v_D\Pi g_1-b_1'w-t_1(\psi)))\\ &=\min_{b_1\in\mathbb{R}^{d_w\times d_\theta}}E[\mathrm{Var}(c'(v_D\Pi g_1-b_1'w)|\psi)]\end{aligned}$$

The minimum is achieved at $\beta_1=E[\mathrm{Var}(w|\psi)]^{-1}E[\mathrm{Cov}(w,v_D\Pi g_1(W,\theta_0)|\psi)]$, using the discussion after Equation 6.1. Define $\tilde{m}_1=v_D\Pi g_1-\beta_1'w$ and $\tilde{m}_0=v_D\Pi g_0-\beta_0'w$, with β_0 defined similarly. By symmetry, we have shown that for $d=0,1$

$$\min_{b_d\in\mathbb{R}^{d_w\times d_\theta}}\min_{t_d\in\mathcal{L}(\psi)}\mathrm{Var}(c'\bar{m}_d(b_d,t_d))=c'E[\mathrm{Var}(\tilde{m}_d|\psi)]c. \tag{11.2}$$

Putting everything together, by monotonicity of $(a,b)\to a+b+(ab)^{1/2}$ for $a,b\geq 0$,

$$\begin{aligned}V_a(c) &\leq \min_{b_0,b_1\in\mathbb{R}^{d_w\times d_\theta}}\min_{t_0,t_1\in\mathcal{L}(\psi)}v_D^{-1}(\mathrm{Var}(c'\bar{m}_1(b_1,t_1))^{1/2}+\mathrm{Var}(c'\bar{m}_0(b_0,t_0))^{1/2})^2\\ &=v_D^{-1}([c'E[\mathrm{Var}(\tilde{m}_1|\psi)]c]^{1/2}+[c'E[\mathrm{Var}(\tilde{m}_0|\psi)]c]^{1/2})^2\equiv\bar{V}_a(c).\end{aligned}$$

Note that by Lemma 11.15 and Lemma 11.16, we have $\widehat{u}_1=E_n[\frac{D_i}{p}\widehat{m}_i\widehat{m}_i']-\widehat{v}_1=E[\tilde{m}_{1i}\tilde{m}_{1i}']-E[E[\tilde{m}_{1i}|\psi]E[\tilde{m}_{1i}|\psi]']+o_p(1)=E[\mathrm{Var}(\tilde{m}_{1i}|\psi)]+o_p(1)$, and similarly $\widehat{u}_0=E[\mathrm{Var}(\tilde{m}_{0i}|\psi)]+o_p(1)$. Then $v_D^{-1}([c'\widehat{u}_1c]^{1/2}+[c'\widehat{u}_0c]^{1/2})^2=\bar{V}_a(c)+o_p(1)$ by continuous mapping. This finishes the proof. □

Let's try this again. The GMM variance is

$$\begin{aligned}v_DV_a(c)&=E[\mathrm{Var}(\Pi v_D(g_1-g_0)-\alpha_0'w|\psi)]=E[\mathrm{Var}(\Pi v_D(g_1-g_0)-(\beta_1-\beta_0)'w|\psi)]\\ &\equiv E[\mathrm{Var}(m_1-m_0|\psi)]=E[\mathrm{Var}(m_1|\psi)]+E[\mathrm{Var}(m_0|\psi)]-2E[\mathrm{Cov}(m_1,m_0|\psi)]\end{aligned}$$

For the moment, just consider the case $\Pi=1$ and $w=0$, so that $m_1=(p-p^2)g_1$. We want to make $v_D\operatorname{Var}(\phi)=v_D\operatorname{Var}(pY_1/p+(1-p)(-Y_0/(1-p)))=v_D\operatorname{Var}(Y_1-Y_0)$ appear at the end. Then we want to simplify $E[\mathrm{Var}(m_1-m_0|\psi)]+v_DE[\mathrm{Var}(\phi|\psi)]$. This is

$$E[\mathrm{Var}(m_1-m_0|\psi)]+v_DE[\mathrm{Var}(pg_1+(1-p)g_0|\psi)]$$

Note also that since $\Pi a(W,\theta_0) = v_D \Pi (g_1 - g_0)(W,\theta_0)$, the optimal adjustment coefficient

$$\alpha_0 = \underset{\alpha \in \mathbb{R}^{d_w \times d_\theta}}{\operatorname{argmin}} E[\operatorname{Var}(\Pi a(W,\theta_0) - \alpha' w | \psi)] = \beta_1 - \beta_0.$$

Define $\widehat{m}_i \equiv v_D \widehat{\Pi} \widehat{g}_i - D_i \widehat{\beta}_1' w_i - (1-D_i)\widehat{\beta}_0' w_i$. Define $m_i = v_D \Pi g_i - D_i \beta_1' w_i - (1-D_i)\beta_0' w_i$ for $g_i = g_i(\theta_0)$ and $\tilde{m}_{1i} = v_D \Pi g_{1i} - \beta_1' w_i$.

Lemma 11.15. *Impose Assumptions 3.1, 3.2, 7.4. Then under P in Definition 11.1, $E_n[\frac{D_i}{p}\widehat{m}_i \widehat{m}_i'] = E[\tilde{m}_1 \tilde{m}_1'] + o_p(1)$ and $E_n[\frac{1-D_i}{1-p}\widehat{m}_i \widehat{m}_i'] = E[\tilde{m}_0 \tilde{m}_0'] + o_p(1)$. Also, we have $\operatorname{Var}_n(\widehat{\kappa}_i) = \operatorname{Var}(\kappa_i) + o_p(1)$.*

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 $D_i m_i = v_D \Pi D_i g_i - D_i \beta_1' w_i = D_i \tilde{m}_{1i}$. Then we can expand $E_n[(D_i/p)\widehat{m}_i \widehat{m}_i']$ as

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

Consider 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 $|xy'|_2 \le |x|_2|y|_2$ and triangle inequality, the first term above has

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

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 $E_n[|D_i\widehat{m}_i|_2^2] = E_n[|v_D D_i \widehat{\Pi}\widehat{g}_i - D_i\widehat{\beta}_1' w_i|_2^2] \le 2E_n[|\widehat{\Pi}\widehat{g}_i|_2^2] + 2E_n[|\widehat{\beta}_1' w_i|_2^2] \le 2|\widehat{\Pi}|_2^2 E_n[|\widehat{g}_i|_2^2] + 2|\widehat{\beta}_1|_2^2 E_n[|w_i|_2^2]$, so clearly it 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 $E_n[|g_i(\widehat{\theta}) - g_i(\theta_0)|_2^2] \le |\widehat{\theta} - \theta_0|_2^2 E_n[|\frac{\partial g_i}{\partial \theta'}(\tilde{\theta}_i)|_2^2]$, 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 θ, $|\frac{\partial g_i}{\partial \theta'}(\tilde{\theta}_i)|_2^2 \le 2|\frac{\partial g_{1i}}{\partial \theta'}(\tilde{\theta}_i)|_2^2 + 2|\frac{\partial g_{0i}}{\partial \theta'}(\tilde{\theta}_i)|_2^2$. 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 $E_n[(D_i/p)\widehat{m}_i\widehat{m}_i'] = E_n[(D_i/p)m_im_i'] + o_p(1)$. Next, we claim $E_n[(D_i/p)m_im_i'] = E_n[(D_i/p)\tilde{m}_{1i}\tilde{m}_{1i}'] = E_n[\tilde{m}_{1i}\tilde{m}_{1i}'] + o_p(1) = E[\tilde{m}_{1i}\tilde{m}_{1i}'] + o_p(1)$. The first equality is by definition of $m_i(D_i, W_i, \theta_0)$ and $\tilde{m}_{1i}(W_i, \theta_0)$. 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 11.8. Then $\mathrm{Var}_n(\widehat{\kappa}_i) = E[\kappa_i\kappa_i'] + o_p(1)$. □

Lemma 11.16. *Require Assumptions 3.1, 3.2, 7.4. Then under P in Definition 11.1, the estimators in the statement of Theorem 7.6 have $\widehat{v}_{10} \xrightarrow{p} E[E[\tilde{m}_{1i}|\psi]E[\tilde{m}_{0i}|\psi]']$, and $\widehat{v}_1 \xrightarrow{p} E[E[\tilde{m}_{1i}|\psi]E[\tilde{m}_{1i}|\psi]']$, and $\widehat{v}_0 \xrightarrow{p} E[E[\tilde{m}_{0i}|\psi]E[\tilde{m}_{0i}|\psi]']$.*

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 $D_im_i = D_i\tilde{m}_{1i} = D_i(v_D\Pi g_{1i}(\theta_0) - \beta_1'w_i)$. In Lemma A.6 of Cytrynbaum (2024b), set $A_i = \tilde{m}_{1i}$ and $B_i = \tilde{m}_{1i}$. Applying the lemma componentwise gives $\widehat{v}_1^o \xrightarrow{p} E[E[\tilde{m}_{1i}|\psi]E[\tilde{m}_{1i}|\psi]']$. Similarly, we have $\widehat{v}_0^o \xrightarrow{p} E[E[\tilde{m}_{0i}|\psi]E[\tilde{m}_{0i}|\psi]']$, and $\widehat{v}_{10}^o \xrightarrow{p} E[E[\tilde{m}_{1i}|\psi]E[\tilde{m}_{0i}|\psi]']$. 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

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

Expand $\widehat{m}_i \widehat{m}_j' - m_i m_j' = \widehat{m}_i(\widehat{m}_j' - m_j') + (\widehat{m}_i - m_i)m_j' \equiv A_{ij} + B_{ij}$. 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

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

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 11.15. 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

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

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 11.15. 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. □

11.9 Lemmas

Proposition 11.17 (Lévy). *Consider probability spaces $(\Omega_n, \mathcal{G}_n, P_n)$ and σ-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*

$$E[g(A_n)|\mathcal{F}_n] = E[g(A)] + o_p(1). \tag{11.3}$$

See Cytrynbaum (2021) for the proof.

Lemma 11.18. *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 γ such that $U'\gamma = \bar{\gamma}$. Since $U' : \mathbb{R}^{d_h} \to \mathbb{R}^r$ is onto, there exists γ^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 11.19 (SVD)**.** *Suppose* $\Sigma \in \mathbb{R}^{m \times m}$ *is symmetric PSD with* $\mathrm{rank}(\Sigma) = r$. *Then* $\Sigma = U\Lambda U'$ *for* $U \in \mathbb{R}^{m \times r}$ *with* $U'U = I_r$ *and* Λ *diagonal.*

Proof. Since Σ is symmetric PSD, there exists $B'B = \Sigma$ for $\mathrm{rank}(B) = r$. Let VAU' be the compact SVD of B, with A diagonal. Then $\Sigma = B'B = UA^2U' \equiv U\Lambda U'$ with $U'U = I_r$. □

Lemma 11.20. *Consider probability spaces $(\Omega_n, \mathcal{G}_n, P_n)$ and σ-algebras $\mathcal{F}_n \subseteq \mathcal{G}_n$. Suppose $0 \leq A_n \leq B < \infty$ and $A_n = o_p(1)$. Then $E[A_n|\mathcal{F}_n] = o_p(1)$.*

Proof. For any $\epsilon > 0$, note that $E[A_n|\mathcal{F}_n] = E[A_n \mathbb{1}(A_n \leq \epsilon)|\mathcal{F}_n] + E[A_n \mathbb{1}(A_n > \epsilon)|\mathcal{F}_n] \leq \epsilon + BP(A_n > \epsilon|\mathcal{F}_n)$. We have $E[P(A_n > \epsilon|\mathcal{F}_n)] = P(A_n > \epsilon) = o(1)$ by tower law and assumption. Then $P(A_n > \epsilon|\mathcal{F}_n) = o_p(1)$ by Markov inequality. Then we have shown $E[A_n|\mathcal{F}_n] \leq \epsilon + T_n(\epsilon)$ with $T_n(\epsilon) = o_p(1)$. Fix $\delta > 0$ and let $\epsilon = \delta/2$. Then $P(E[A_n|\mathcal{F}_n] > \delta) \leq P(\delta/2 + T_n(\delta/2) > \delta) = P(T_n(\delta/2) > \delta/2) = o(1)$ since $T_n(\delta/2) = o_p(1)$. Since δ was arbitrary, we have shown that $E[A_n|\mathcal{F}_n] = o_p(1)$. □

Lemma 11.21. $A_n = O_p(1) \iff A_n = o_p(c_n)$ *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. □

Proof of Lemma 11.3. The first set of statements since $Q = P$ on $\mathcal{F}_n$ by definition. Let $c = P(Z_h \in T$, with $c > 0$ by assumption. Define $S_n = \{P(\mathcal{I}_n \in T_n|\mathcal{F}_n) \geq c/2\}$. Then by Lemma 11.5, $P(\mathcal{I}_n \in T_n|\mathcal{F}_n) \xrightarrow{p} P(Z_h \in T) = c$, so $P(S_n) \to 1$. We have the upper bound

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

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

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

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 11.21. $\square$