Finely Stratified Rerandomization Designs

Max Cytrynbaum*

July 25, 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 covariates used for stratification, and linear control over the rerandomization covariates. We also introduce new rerandomization criteria, allowing for nonlinear imbalance metrics and proposing a minimax scheme that optimizes the balance criterion using pilot data or prior information provided by the researcher. While the asymptotic distribution of generalized method of moments (GMM) estimators under stratified rerandomization is generically non-Gaussian, we show how to restore asymptotic normality using optimal ex-post linear adjustment. This allows us to provide simple asymptotically exact inference methods for superpopulation parameters, as well as efficient conservative inference methods for 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 estimation and inference on 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. Modulo this residual truncated term, we show that 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 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.

Our third contribution is to study optimization of the balance criterion itself. We

[1] For example, Cytrynbaum (2023) 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 (2024), 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 (2024).

suggest 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 consistent with this prior information is small. If the prior information set contains the truth, this design provides strong control over the variance of the truncated normal term in the asymptotic distribution. Building on this, we show that if the prior information set is a confidence region estimated from pilot data, then this minimax design bounds the truncated normal variance with high probability.

Our fourth contribution is to provide simple t-statistic 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 methods are asymptotically conservative, but still exploit 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 (2024), 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 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 essen-

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

tially 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 (2024) 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 fixed dimension.

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 (2023), 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 (2024). 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 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.

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

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 $\{1, \ldots, n\} = \bigcup_s s$ disjointly and $|s| = k$. Let $\psi = \psi(R)$ with $\psi \in \mathbb{R}^{d_\psi}$ denote a vector of stratification variables, which may be continuous or discrete. Suppose the groups satisfy the 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 (2024) 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 $n(\bar{X}_1 - \bar{X}_0)'\Sigma_n(\bar{X}_1 - \bar{X}_0) \leq \epsilon^2$ 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 $\mathcal{I}_n \in A$ 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 (2024), and Bai et al. (2024). Definition 2.1 allows for

[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 (2024) 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 (2024). 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.

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 $T \in \{1, \dots, m\}$ 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 $D_{1:n} \sim \text{CR}(p)$. Cytrynbaum (2024) shows that $\text{CR}(p)$ 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 $D|W \sim \text{Bernoulli}(p)$ define $\phi(W, \theta) = E[g(D, R, S, \theta)|W]$, so that $E[\phi(W, \theta)] = 0 \iff E[g(D, R, S, \theta)] = 0$.

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 $\theta_0 = E[Y(1) - Y(0)] =$ ATE, the average treatment effect, and $\theta_n = E_n[Y_i(1) - Y_i(0)] =$ 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 $Y =$

[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).

$Y(D(Z))$. 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:

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

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

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

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 $\theta_n = \operatorname{argmin}_\theta E_n[(\tau_i - X_i'\theta)^2|C_i = 1]$ is the within-sample best linear predictor. In the case of perfect compliance $C_i = 1$ for all i, this is $\theta_n = \operatorname{argmin}_\theta E_n[(\tau_i - X_i'\theta)^2]$, 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

[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.

method of moments (GMM) estimation of the causal parameters θ_0 and θ_n under stratified rerandomization.

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* $\text{Leb}(\partial A) = 0$,[12] *and require $E[\text{Var}(h|\psi)] \succ 0$ 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]

(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 $a(W,\theta) \equiv \text{Var}(D)(g_1(W,\theta) - g_0(W,\theta))$. For Horvitz-Thompson weights $H = (D-p)/(p-p^2)$, a simple calculation shows that we can expand

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

[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.

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 $\widehat{\theta} = E_n[H_i Y_i]$ 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 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 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 chance 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 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 $R_{B(0,\epsilon)} \sim \{\gamma_0' Z_h \,|\, |Z_h|_2 \leq \epsilon\} \xrightarrow{p} 0$ 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 $V = v_D^{-1} E[\mathrm{Var}(\Pi a(W, \theta_0)|\psi)] \leq v_D^{-1} \mathrm{Var}(\Pi a(W, \theta_0))$, 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

[14] See Proposition 8.16 in the appendix for a formal statement.

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 intuitive, closed form variance expressions and 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 may be possible 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 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

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

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

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

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 $V_a = \min_{\gamma \in \mathbb{R}^{d_h}} v_D^{-1} E[\text{Var}(\bar{Y} - \gamma' h | \psi)]$. This can be rewritten in terms of the residuals of a partially linear regression of $\bar{Y}$ on ψ and h:

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

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:

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

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.

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.

[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).

4.1 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

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

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 $A = \{x : |x|_2 \leq \epsilon\}$.

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*

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

[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.

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 rerandomization schemes, as shown in the next example.

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, one can show that $t(\beta)$ is strictly convex, so $\beta \to \log f(x, \beta)$ is strictly concave for all x.[17] Then the score $m(X, \beta) = \nabla_\beta \log f(X, \beta)$ has a unique solution $E[m(X, \beta^*)] = 0$, and quasi-MLE estimation 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 $E_n[H_i t(\widehat{\beta})] = t(\widehat{\beta}) E_n[H_i] = 0$, 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 rerandomization based on 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 $r(x) = ((x_j)_j, (x_j x_l)_{j<l})$. The parameters β_{jl} model correlation between the binary variables x_j and x_l. Categorical variables with more than two lev-

[17] In particular, this holds for β s.t. $t(\beta) < \infty$. See e.g. Wainwright and Jordan (2008) Chapter 3 for an introduction to the properties of exponential families and the log partition function $t(\beta)$.

[18] This is known as the Ising model in statistical physics. See Wainwright and Jordan (2008) chapter 6 for efficient MLE algorithms in this family.

els and and higher order interactions can easily be accommodated. By the discussion above, a design that rerandomizes based on quasi-MLE density estimates in this family 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." 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 $p(X, \beta) = L(X'\beta)$ 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).

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 8.5.

Theorem 4.7 (Propensity Rerandomization)**.** *Suppose $D_{1:n}$ is as in Definition 4.6. Require Assumptions 3.2, 8.12. Then $\sqrt{n}(\widehat{\theta} - \theta_n)|W_{1:n} \Rightarrow \mathcal{N}(0, V_a) + R$.*

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

The residual $R \sim \gamma_0' Z_h \,|\, Z_h' \operatorname{Var}(h)^{-1} Z_h \leq \epsilon$ for $Z_h \sim \mathcal{N}(0, v_D^{-1} E[\operatorname{Var}(h|\psi)])$ 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 $n(\bar{h}_1 - \bar{h}_0)' \operatorname{Var}_n(h_i)^{-1}(\bar{h}_1 - \bar{h}_0) \leq \epsilon$. Equivalently, propensity rerandomization behaves like linear rerandomization with $\mathcal{I}_n = \sqrt{n}(\bar{h}_1 - \bar{h}_0)$ and ellipsoidal acceptance region $A = \operatorname{Var}(h)^{1/2} B(0, \epsilon)$.[19]

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 $Z_h \sim \mathcal{N}(0, v_D^{-1} E[\operatorname{Var}(h|\psi)])$, 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 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 rerandomization acceptance region $A \subseteq \mathbb{R}^{d_h}$. For simplicity and intuition, we first restrict to the case of estimating $\theta_n = \text{SATE}$, generalizing in what follows.

Imbalance Decomposition. The difference of means estimator $\widehat{\theta} = E_n[Y_i(1) - Y_i(0)] + E_n[H_i \bar{Y}_i] = \theta_n + E_n[H_i \bar{Y}_i]$ for $H_i = (D_i - p)/(p - p^2)$. Intuitively, the scaled errors $\sqrt{n}(\widehat{\theta} - \theta_n) = \sqrt{n} E_n[H_i \bar{Y}_i]$ are driven by imbalances in the outcome levels $\bar{Y}$ between treatment arms. The previous section showed that if covariates h are predictive of $\bar{Y}$, we can reduce these imbalances by rerandomizing until $\sqrt{n} E_n[H_i h_i] = \sqrt{n}(\bar{h}_1 - \bar{h}_0) \in A$. To study the role of the acceptance region A, let (γ_0, t_0) be solutions to the partially linear prediction problem in Equation 3.6 and consider the expansion[20]

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

We use this to decompose the imbalance in outcome levels $\bar{Y}$ into imbalances in the covariates ψ and h and residuals e. In particular, we can now write the imbalance decomposition

$$E_n[H_i \bar{Y}_i] = E_n[H_i t_0(\psi_i)] + \gamma_0' E_n[H_i h_i] + E_n[H_i e_i] \equiv I_1 + I_2 + I_3.$$

[19] 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.

[20] This is without loss of generality. Note that we do not impose well-specification $E[e|\psi, h] = 0$.

The analysis in Section 3 showed that for *any* acceptance region $A \subseteq \mathbb{R}^{d_h}$:

(1) The ψ imbalance component $I_1 = \sqrt{n}E_n[H_i t_0(\psi_i)] \xrightarrow{p} 0$ due to stratification.

(2) The components $I_2 + I_3 = \gamma_0'\sqrt{n}E_n[H_i h_i] + \sqrt{n}E_n[H_i e_i] \Rightarrow R_A + v_D^{-1/2}\mathcal{N}(0, \text{Var}(e))$ are asymptotically independent, with $\text{Var}(e)$ not depending on A.

In particular, it suffices to choose A to minimize the component $I_2 = \gamma_0'\sqrt{n}E_n[H_i h_i]$. This suggests an oracle acceptance criterion, rerandomizing until $|\gamma_0'\sqrt{n}E_n[H_i h_i]| \leq \epsilon$, with acceptance region $A = \{a : |\gamma_0' a| \leq \epsilon\}$. However, this acceptance region is infeasible since γ_0 is unknown at design-time. Instead, we take a minimax approach, allowing the researcher to incorporate prior information about γ_0.

5.1 Minimax Rerandomization

Suppose that we know $\gamma_0 \in B$ for some prior information set $B \subseteq \mathbb{R}^{d_h}$. Fix $\epsilon > 0$ and consider a "minimax" style acceptance criterion, rerandomizing the treatments $D_{1:n}$ until

$$\sup_{\gamma \in B} |\gamma'\sqrt{n}E_n[H_i h_i]| \leq \epsilon. \tag{5.1}$$

Note that the function $f_B(x) = \sup_{\gamma \in B} |\gamma' x|$ is convex, so we can also interpret this as a convex imbalance penalty, rerandomizing until $f_B(\mathcal{I}_n) \leq \epsilon$ for imbalance metric $\mathcal{I}_n = \sqrt{n}E_n[H_i h_i]$, generalizing the quadratic penalty in Example 2.2. Our first result shows that this minimax design is of the form studied in the Section 3, characterizing the acceptance region induced by this convex penalty.

Theorem 5.1. *The following hold:*

(a) (Rerandomization). The acceptance criterion $\sup_{\gamma \in B} |\gamma'\sqrt{n}E_n[H_i h_i]| \leq \epsilon \iff \sqrt{n}E_n[H_i h_i] \in A$ *for* $A = \epsilon B^\circ$ *with* $B^\circ = \{a : \sup_{\gamma \in B} |\gamma' a| \leq 1\} \subseteq \mathbb{R}^{d_h}$.[21]

(b) (Acceptance Region). $A = \epsilon B^\circ$ *is symmetric and convex. If* B *is bounded, then* A *is closed and has non-empty interior. If* B *is open, then* A *is bounded.*

(c) (Well-specification). If $\gamma_0 \in B$*, then* $\text{Var}(R_A) \leq \epsilon^2$.

Part (a) of Theorem 5.1 shows that the rerandomization criterion is of the form studied in Definition 2.1, with acceptance region $A = \epsilon B^\circ$. Part (b) shows that A is always symmetric and convex. In particular, the asymptotic distribution of $\widehat{\theta}$ is centered at zero. The set B° is known as the absolute polar of B, e.g. see Aliprantis and Border (2006). Part (c) of the theorem shows that if the prior information set B contains the true coefficient γ_0, then $\text{Var}(R_A) \leq \epsilon^2$. Then by independence, the asymptotic variance is within ϵ^2 of the optimal partially linear variance. If $\gamma_0 \notin B$ (misspecification), then

[21] Note that for a set $S \subseteq \mathbb{R}^d$, we have $\epsilon S = \{\epsilon s : s \in S\}$.

possibly $\text{Var}(R_A) > \epsilon^2$. However, note that misspecification does not affect our inference methods, which allow for general acceptance regions A.

Remark 5.2 (Acceptance Probability)**.** Note that the asymptotic acceptance probability $a(\epsilon) = P(Z_h \in \epsilon B^\circ)$ has $a(\epsilon) \to 0$ as $\epsilon \to 0$ and is monotonically increasing. For B bounded, the theorem shows that B° has non-empty interior. In this case, as $\epsilon \to \infty$ we have $\epsilon B^\circ \uparrow \mathbb{R}^{d_h}$ so $a(\epsilon) \to 1$. This shows that, at least in large samples, we can choose ϵ to achieve any desired acceptance probability $P(Z_h \in A) \in (0, 1)$. Under well-specification, any such choice of ϵ comes with a variance guarantee provided by the theorem.

5.2 Specifying Prior Information

Without pilot data, we are left to introspection to choose the prior information set B. Recall that γ_0 is the coefficient from the partially linear projection $\bar{Y} = \gamma_0' h + t_0(\psi) + e$. Intuitively, γ_0 parameterizes how much we expect the average outcome level to change given a unit change in h, holding ψ fixed. If the partially linear model happens to be well-specified, then $\gamma_0 = \nabla_b E[\bar{Y}|\psi, h = b]$. If $t_0(\psi) = t'\psi$ happens to be linear, then $\bar{Y} = c + \gamma_0' h + t'\psi + e$ and γ_0 is just an OLS coefficient. The following examples provide some reasonable prior information specifications and their associated acceptance regions. These examples rely on a general characterization of acceptance regions in Lemma 5.5 below.

Example 5.3 (Rectangle)**.** One natural way to specify prior information is to assume $\gamma_{0j} \in [l_j, u_j]$ for each $1 \le j \le d_h$, equivalent to setting $B = \prod_{j=1}^{d_h} [l_j, u_j]$. This allows sign constraints, e.g. $0 \le \gamma_{0j} \le m$ for some j and $-m \le \gamma_{0j} \le 0$ for others. Lemma 5.5 below shows that if $B = \prod_{j=1}^{d_h} [l_j, u_j]$, then $A = \epsilon B^\circ = \{a : |a'l + a'u| + \sum_j |a_j| u_j - |a_j| l_j \le 2\epsilon\}$, where $l = (l_j)_j$ and $u = (u_j)_j$. An example is shown in Figure 1. Note that the acceptance region A is conservative in directions aligned with the prior information set $B = [1, 2] \times [1, 3/2]$, guarding against covariate imbalances that are aligned with adverse coefficient values $\gamma_0 \in B$. A is more lenient in directions approximately orthogonal to B.

Example 5.4 (Ellipse)**.** Another natural specification is to guess $\gamma_0 \approx \bar{\gamma}$, setting $B = \bar{\gamma} + B_2(0, m)$, for an uncertainty parameter m. By the characterization in Lemma 5.5 below, $A = \epsilon B^\circ = \{a : |a'\bar{\gamma}| + m|a|_2 \le \epsilon\}$. More generally, if $B = \bar{\gamma} + \Sigma B_2(0, 1)$ for a positive-definite matrix Σ, the lemma shows that $A = \epsilon B^\circ = \{a : |a'\bar{\gamma}| + |\Sigma a|_2 \le \epsilon\}$. One natural application of this specification is when B is a Wald confidence region constructed using pilot data, as discussed below. An example is shown in Figure 1.

More generally, the following lemma provides a useful characterization of the acceptance region $A = \epsilon B^\circ$ from Theorem 5.1 for a large family of prior information set specifications. 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 $B_p(0, 1) = \{a : |a|_p \le 1\}$.

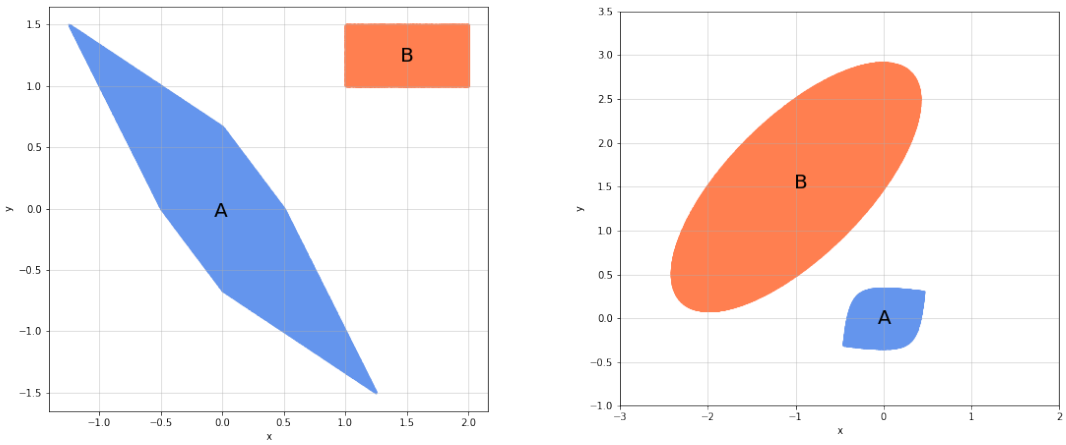

Figure 1: Prior information B and $A = \epsilon B^\circ$ for Examples 5.3 and 5.4.

Lemma 5.5 (Acceptance Regions)**.** *For $p \in [1, \infty]$, let $1/p + 1/q = 1$, setting $q = 1$ if $p = \infty$ and vice-versa. Suppose $B = x + \Sigma B_p(0, 1)$, for $x \in \mathbb{R}^{d_h}$ and Σ invertible. Then $A = \epsilon B^\circ = \{a : |a'x| + |\Sigma' a|_q \leq \epsilon\}$.*

5.3 Using Pilot Data

Next, we discuss an alternative strategy that uses pilot data to specify the set B. 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 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 the Wald region as

$$\widehat{B}_{pilot} = \widehat{\gamma} + c_\alpha^{1/2} m^{-1/2} \cdot \widehat{\Sigma}_{pilot}^{1/2} B_2(0, 1). \tag{5.2}$$

Using $\widehat{B}_{pilot}$ as a prior information set, by Example 5.4 we have acceptance region

$$\widehat{A}_{pilot} = \epsilon \widehat{B}^\circ_{pilot} = \{a : |a'\widehat{\gamma}_{pilot}| + m^{-1/2} c_\alpha^{1/2} |\widehat{\Sigma}^{1/2} a|_2 \leq \epsilon\}. \tag{5.3}$$

Note that the acceptance region $\widehat{A}_{pilot}$ grows with the pilot size m. This reflects smaller uncertainty about the true parameter γ_0, and thus less adversarial worst case imbalance $\sup_{\gamma \in \widehat{B}_{pilot}} |\gamma' E_n[H_i h_i]|$. Conversely, $\widehat{A}_{pilot}$ shrinks as the confidence parameter α and the scale of the variance estimate $\widehat{\Sigma}_{pilot}$ increases, reflecting greater uncertainty and a more conservative approach to covariate balances. Our next result shows that rerandomization with acceptance region $\widehat{A}_{pilot}$ controls the variance of the imbalance $R_A = \gamma_0' Z | Z \in \widehat{A}_{pilot}$ with high probability marginally over the realizations of the pilot data. The result is an

immediate consequence of Theorem 3.5 and Theorem 5.1.

Corollary 5.6 (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* $A = \widehat{A}_{pilot} = \epsilon \widehat{B}^{\circ}_{pilot}$*, then if Assumptions 3.1, 3.2 hold, then* $\sqrt{n}(\widehat{\theta} - \theta_n)|\mathcal{D}_{pilot} \Rightarrow v_D^{-1}\mathcal{N}(0, \mathrm{Var}(e)) + R_A$ *and* $\mathrm{Var}(R_A|\mathcal{D}_{pilot}) \leq \epsilon^2$ *with probability* $\geq 1-\alpha$*.*

Formally, the pilot estimate of γ_0 and Wald region could be constructed as in Robinson (1988). In practice, a simple 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 $E[Y_T|W] = (1-p)Y(1) + pY(0) = \bar{Y}$.

General Parameters. For completeness, we extend the preceding work to general parameters θ_n as in Definition 2.5. Let $\Pi a(W, \theta_0)$ be the assignment influence function. As in Equation 3.7, consider the partially linear decomposition

$$\Pi a(W, \theta_0) = \gamma_0' h + t_0(\psi) + e, \qquad E[e|\psi] = 0,\ E[eh] = 0.$$

Note that $e \in \mathbb{R}^{d_\theta}$ and $E[e|\psi] = 0$ is interpreted componentwise. Consider prior information sets B_j for each γ_0^j with $1 \leq j \leq d_\theta$, where $\gamma_0^j \in \mathbb{R}^{d_h}$ is the jth column of γ_0. The final result of this section bounds the asymptotic imbalance term R_A if all these prior information sets are well specified.

Theorem 5.7. *Let* $D_{1:n}$ *as in Definition 2.1 with* $A = \cap_{j=1}^{d_\theta} \epsilon B_j^{\circ}$*. Then* $\sqrt{n}(\widehat{\theta} - \theta_n)|W_{1:n} \Rightarrow \mathcal{N}(0, V_a) + R_A$*, as defined in Theorem 3.5. If* $\gamma_0^j \in B_j\ \forall j$*, then* $\max_{j=1}^{d_\theta} \mathrm{Var}((R_A)_{jj}) \leq \epsilon^2$*.*

Note that by construction the conservative acceptance region $A = \cap_{j=1}^{d_\theta} \epsilon B_j^{\circ}$ is symmetric and convex.

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, provided in Section 7 below.

Let w denote the covariates used for ex-post adjustment and suppose $E[|w|_2^2] < \infty$.

Definition 6.1 (Adjusted GMM)**.** Suppose that $\widehat{\alpha} \xrightarrow{p} \alpha \in \mathbb{R}^{d_w \times d_g}$. 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 (2023) for the special case $\theta_0 = \mathrm{ATE}$. Motivated by Proposition 6.2, we define the optimal linear adjustment coefficient as the minimizer of the asymptotic variance $V(\alpha)$, in the positive semidefinite sense.

Optimal Adjustment Coefficient. Define the coefficient

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

Note that if $w = h$ then $\alpha_0 = \gamma_0$, as 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 $\alpha_0 = E[\mathrm{Var}(w|\psi)]^{-1} E[\mathrm{Cov}(w, \Pi a(W, \theta_0)|\psi)]$. Observe that the optimal adjustment coefficient α_0 varies with the stratification variables ψ, as observed in Cytrynbaum (2024) 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 as in Definition 2.1. Require Assumption 3.1, 3.2. Let $h \subseteq w$ and suppose that $\widehat{\alpha} \xrightarrow{p} \alpha_0$. 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. For nonlinear models, the 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 $\widehat{\theta}_{adj} = \widehat{\theta} - E_n[H_i \widehat{\alpha}' w_i]$.

Similarly to two-step efficient GMM, this process could be iterated until convergence to improve finite sample properties. One feasible estimator of the optimal coefficient α_0 is given in the following theorem. To state the result, define the within-group partialled covariates $\check{w}_i = w_i - \sum_{j \in s(i)} w_j$, where group $s(i)$ contains unit i in Definition 2.1. Let $\widehat{\Pi} \xrightarrow{p} \Pi$ consistently estimate the linearization matrix and denote the score evaluation $\widehat{g}_i \equiv g(D_i, R_i, S_i, \widehat{\theta})$.

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

In some cases, α_0 is independent of θ_0. For example, if $a(W,\theta) = a_1(\psi,\theta) + a_2(W)$ then $\alpha_0 = E[\text{Var}(w|\psi)]^{-1}E[\text{Cov}(w, \Pi a_2(W)|\psi)]$ does not depend on θ_0. In such cases, one-step optimal adjustment is possible.

Corollary 6.5 (One-step Adjustment)**.** *Suppose $a(W,\theta) = a_1(\psi,\theta) + a_2(W)$. Then for any $\theta \in \Theta$, substituting $g_i = g(D_i, R_i, S_i, \theta)$ for $\widehat{g}_i$ in $\widehat{\alpha}$ above, 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 with score $g(Y,D,X,\theta) = (HY - X'\theta)X$ and $\theta_n = \text{argmin}_\theta\, E_n[(\tau_i - X_i'\theta)^2]$, recall that $a(W,\theta_0) = \bar{Y}X$ and $\Pi = E[XX']^{-1}$. Letting $\theta = 0$ gives $g(Y,D,X,0) = HYX$. By Corollary 6.5, we have $\widehat{\alpha} = \alpha_0 + o_p(1)$ for the adjustment coefficient matrix

$$\widehat{\alpha} = v_D E_n[\check{w}_i \check{w}_i']^{-1} E_n[H_i^2 Y_i \check{w}_i X_i'] E_n[X_i X_i']^{-1}. \tag{6.2}$$

This allows us to estimate treatment effect heterogeneity relative to a low-dimensional vector of important covariates X, adjusting optimally for a larger set of covariates w ex-post in order to improve precision, as well as restore asymptotic normality when $A \neq \mathbb{R}^{d_h}$. In the case $X = 1$ (SATE estimation), $\widehat{\alpha}$ is equivalent to the "tyranny-of-the-minority" style estimator proposed in Cytrynbaum (2023).

7 Inference

In this section, we provide novel methods for inference on general causal parameters under stratified rerandomization designs. We make crucial use of asymptotic normality of the optimally adjusted estimator $\widehat{\theta}_{adj}$, shown in Theorem 6.3. For the superpopulation parameter θ_0, we provide asymptotically exact inference methods. The asymptotic variance for estimating the finite population parameter θ_n is generally not identified. In this case, we provide conservative variance estimation that still reflects the precision gains due to stratification and rerandomization.

7.1 Asymptotically Exact Inference

To define our variance estimator, we begin with some definitions. Let $\mathcal{S}_n$ denote the set of groups constructed in Definition 2.1. For each $s \in \mathcal{S}_n$ define the centroid $\bar{\psi}_s = |s|^{-1}\sum_{i\in s}\psi_i$. Let $\nu : \mathcal{S}_n \to \mathcal{S}_n$ be a bijective matching between groups satisfying $\nu(s) \neq s$, $\nu^2 = \text{Id}$, and the homogeneity condition

$$\frac{1}{n}\sum_{s\in\mathcal{S}_n} |\bar{\psi}_s - \bar{\psi}_{\nu(s)}|_2^2 = o_p(1). \tag{7.1}$$

In practice, ν is obtained by simply matching the group centroids $\bar{\psi}_s$ into pairs using the Derigs (1988) non-bipartite matching algorithm. Let $\mathcal{S}_n^\nu = \{s \cup \nu(s) : s \in \mathcal{S}_n\}$ be the unions of paired groups formed by this matching. Denote $a(s) = \sum_{i \in s} D_i$ and $k(s) = |s|$. Define the adjusted moment $\widehat{m}_i \equiv \widehat{\Pi}\widehat{g}_i - H_i\widehat{\alpha}'w_i$, where $\widehat{g}_i \equiv g(D_i, X_i, Y_i, \widehat{\theta}_{adj})$. Suppose that $\widehat{\Pi} \xrightarrow{p} \Pi$ and $\widehat{\alpha} \xrightarrow{p} \alpha_0$ for the optimal adjustment coefficient in Equation 6.1. For instance, we can use the consistent estimator provided by Theorem 6.4. Finally, define the variance estimator components

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

Using these terms, construct the variance estimator

$$\widehat{V} = \mathrm{Var}_n(\widehat{m}_i) - v_D(\widehat{v}_1 + \widehat{v}_0 - \widehat{v}_{10} - \widehat{v}_{10}'). \tag{7.2}$$

We require a slight strengthening of our GMM assumptions 3.2.

Assumption 7.1. *There exists $\theta_0 \in U \subseteq \Theta$ open s.t. $E[\sup_{\theta \in U} |\partial/\partial\theta' g_d(W, \theta)|_F^2] < \infty$.*

Under this condition, we can state our first inference result, showing consistent estimation of the asymptotic variance matrix in Theorem 6.3.

Theorem 7.2 (Inference)**.** *Suppose $D_{1:n}$ is as in Definition 2.1, and impose Assumptions 3.1, 3.2, 7.1. Then $\widehat{V} \xrightarrow{p} V_\phi + V_a^{adj}$.*

By Theorem 6.3, $\sqrt{n}(\widehat{\theta}_{adj} - \theta_0) \Rightarrow N(0, V_\phi + V_a^{adj})$. Then the variance estimation result above allows for joint inference on θ_0 using e.g. standard Wald-test or t-statistic based confidence regions.

7.2 Inference on the Finite Population Parameter

In this section, we provide asymptotically conservative inference on linear contrasts of the finite population parameter $c'\theta_n$.

As noted above, the asymptotic variance V_a^{adj} in Theorem 6.3 for estimating the finite population parameter θ_n is generically not identified. This happens because it depends on terms of the form $\mathrm{Var}(a|\psi) \propto \mathrm{Var}(g_1|\psi) + \mathrm{Var}(g_0|\psi) - 2\,\mathrm{Cov}(g_1, g_0|\psi)$, with $g_d = g(d, X, S(d), \theta_0)$. However, $S(1)$ and $S(0)$ are never simultaneously observed (Neyman (1990)), so $\mathrm{Cov}(g_1, g_0|\psi)$ is generically not identified. We work with linear contrasts $c'\theta_n$

since this allows us to tighten our upper bounds on the (non-identified) variance. To do so, let $\widehat{u}_1 = E_n[\frac{D_i}{p}\widehat{m}_i\widehat{m}_i'] - \widehat{v}_1$ and $\widehat{u}_0 = E_n[\frac{1-D_i}{1-p}\widehat{m}_i\widehat{m}_i'] - \widehat{v}_0$ using the estimator components above and consider the variance estimator

$$\widehat{V}_a(c) = v_D([c'\widehat{u}_1 c]^{1/2} + [c'\widehat{u}_0 c]^{1/2})^2. \tag{7.3}$$

By Theorem 6.3, we have $\sqrt{n}(c'\widehat{\theta}_{adj} - c'\theta_n)|W_{1:n} \Rightarrow \mathcal{N}(0, c'V_a^{adj}c)$. Our next result shows how to consistently estimate an upper bound on this asymptotic variance.

Theorem 7.3 (Inference)**.** *Suppose $D_{1:n}$ as in Definition 2.1 and impose Assumptions 3.1, 3.2, 7.1. Then $\widehat{V}_a(c) \xrightarrow{p} \bar{V}_a(c) \geq c'V_a^{adj}c$.*

The variance upper bound $\bar{V}_a(c) \geq c'(V_\phi + V_a^{adj})c$, so the confidence intervals derived from this approach are always weakly shorter than those using the variance estimator in Equation 7.2. See Section 8.8 in the appendix for an explicit comparison. The upper bound $\bar{V}_a(c)$ incorporates the efficiency gains from stratification, rerandomization, and adjustment. However, this upper bound is generally not sharp (Aronow et al. (2014)). We leave sharp upper bounds on the asymptotic variance matrix V_a^{adj} to future work.

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.

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*.

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

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

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

Cytrynbaum, M. (2024). 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*.

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.

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

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. (2024). 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*.

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

8 Proofs

8.1 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 8.1 (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 (2024).

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 8.2 (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 8.1, define the rerandomization distribution

$$Q(B|\mathcal{F}_n) = P(B|\mathcal{F}_n, \mathcal{I}_n \in T_n), \qquad Q(B) = E[Q(B|\mathcal{F}_n)]. \tag{8.1}$$

(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 8.2 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 8.9 below.

Lemma 8.3 (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 8.2.*

(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 8.3 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 8.3. 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.

8.2 Rerandomization Asymptotics

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

Theorem 8.4 (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 8.1. Then* $X_n \equiv \sqrt{n}E_n[H_i a(W_i)]$ *has* $X_n|\mathcal{F}_n \Rightarrow \mathcal{N}(0, V)$. *In particular, for each* $t \in \mathbb{R}^{d_a}$ *we have* $E[e^{it'X_n}|\mathcal{F}_n] = \phi(t) + o_p(1)$ *with* $\phi(t) = e^{-t'Vt/2}$ *and* $V = v_D^{-1}E[\mathrm{Var}(a|\psi)]$.

Proof. First consider the case $d_g = 1$. Define $u_i = a_i - E[a_i|\psi_i]$. By Lemma A.3 in Cytrynbaum (2024), 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 $l = 1, \ldots, n/k$ of $s(l) \in \mathcal{S}_n$, noting that $|\mathcal{S}_n| \leq n/k$. Define $D_{s(l)} = (D_i)_{i \in s(l)}$. Define $\mathcal{H}_{0,n} = \mathcal{F}_n$ and $\mathcal{H}_{j,n} = \sigma(\mathcal{F}_n, D_{s(l)}, l \in [j])$ for $j \geq 1$. Define $D_{l,n} = n^{-1/2}\sum_{i \in s(l)}(D_i - p)u_i$ and $S_{j,n} = \sum_{i=1}^{j} D_{i,n}$.

(1) We claim that $(S_{j,n}, \mathcal{H}_{j,n})_{j \geq 1}$ is an MDS. Adaptation is clear from our definitions.

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

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

(2). Next, we compute the variance process. By the same argument in (1), we have

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

By Lemma C.10 of Cytrynbaum (2024), we have $\mathrm{Cov}(D_s, D_t|\mathcal{F}_n)\mathbb{1}(s, t \in s(l)) = -l(k -$

$l)/k^2(k-1) \equiv c$ and $\text{Var}(D_i|\mathcal{F}_n) = p - p^2$. Then we may expand σ_n^2 as

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

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

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

Then we have

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

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

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

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

Then using the inequality above we compute

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

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 (2024), we have shown that $E[e^{it\sqrt{n}E_n[(D_i-p)a_i]}|\mathcal{F}_n] = \phi(t) + o_p(1)$ for $\phi(t) = e^{-t^2V/2}$ with $V = (p-p^2)E[\mathrm{Var}(a|\psi)]$.

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

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

Lemma 8.5. *Let Definition 8.2 hold. Let $\widehat{\Delta}_a = E_n[H_i a_i]$ and $\rho = (a,h)$. Fix $t \in \mathbb{R}^{d_a}$. Let $(Z_a, Z_h) \sim \mathcal{N}(0,\Sigma)$ for $\Sigma = v_D^{-1}E[\mathrm{Var}(\rho|\psi)]$. Then under P in Definition 8.1*

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

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

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

For the second equality, note that $e^{it_3'\tau_n} \xrightarrow{p} e^{it_3'\tau}$ by continuous mapping. Then $R_n = e^{it_1'\sqrt{n}\widehat{\Delta}_a + it_2'\sqrt{n}\widehat{\Delta}_h}(e^{it_3'\tau_n} - e^{it_3'\tau}) = o_p(1)$. Clearly $|R_n| \leq 2$, so $E[|R_n||\mathcal{F}_n] = o_p(1)$ by Lemma 8.19. 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 8.4 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 8.16. Fix $t \in \mathbb{R}$ and define $G(z_1,z_2,x) = e^{it'z_1}\mathbb{1}(b(z_2,x) \leq 0)$ and note that

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

Define $E_G = \{w : G(\cdot) \text{ not continuous at } w\}$. By Proposition 8.16, if $P(B \in E_G) = 0$ then $E[G(B_n)|\mathcal{F}_n] = E[G(B)] + o_p(1) = E[G(Z_a, Z_h, \tau)] + o_p(1)$, which is the required claim.

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

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

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

Theorem 8.6 (Asymptotic Distribution)**.** *Let Definition 8.2 hold. Suppose that* $(Z_a, Z_h) \sim v_D^{-1} E[\mathrm{Var}((a, h)|\psi)]$. *Then under* Q *in Definition 8.2 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 8.5, $a_n \xrightarrow{p} a_\infty$ and $b_n \xrightarrow{p} b_\infty$, with $b_\infty > 0$ by assumption in Definition 8.2. 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 8.5. 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 8.17. Letting $\tilde{Z}_a = Z_a - \gamma_0' Z_h$, by Lemma 8.17 $\tilde{Z}_a \perp\!\!\!\perp Z_h$ 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 8.16, we have shown $Z_a\,|\,Z_h \in T \stackrel{d}{=} \tilde{Z}_a + [\gamma_0' Z_h\,|\,Z_h \in T]$, 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

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

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 8.19. 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). □

Lemma 8.7 (Linearization). *Suppose Definition 8.2 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 8.3 below for the proof of this lemma.

Proof of Theorem 3.5. We claim that the conditions of Definition 8.2 hold. This will allow us to apply our general rerandomization asymptotics in Theorem 8.6 and linearization in Lemma 8.7. 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 8.2. For part (b), $P(b(Z_h) = 0) = P(Z_h \in \partial A) = 0$ since $\text{Leb}(\partial A) = 0$ and by absolute continuity of Z_h relative to Lebesgue measure Leb. We also have $P(Z_h \in T) = P(Z_h \in A) > 0$ since Z_h is full measure by $E[\text{Var}(h|\psi)] \succ 0$ and since A has non-empty interior.

This proves the claim. Then by Lemma 8.7, $\sqrt{n}(\widehat{\theta}-\theta_n) = \sqrt{n}E_n[H_i \Pi a(W_i,\theta_0)]+o_p(1)$. The result now follows immediately by Slutsky and Theorem 8.6(a), letting $a \to \Pi a$. Likewise, Corollary 3.8 follows from Theorem 8.6(b), letting $\phi \to \Pi\phi$. □

Proof of Corollary 3.6. By Theorem 3.5, since $A = \mathbb{R}^{d_h}$ we have $\sqrt{n}(\widehat{\theta} - \theta_n)|W_{1:n} \Rightarrow \mathcal{N}(0,V_a) + R$, independent RV's with $V_a = v_D^{-1}E[\text{Var}(\Pi a(W,\theta_0) - \gamma_0' h|\psi)]$ and $R \sim \gamma_0' Z_h$ for $Z_h \sim \mathcal{N}(0, v_D^{-1}E[\text{Var}(h|\psi)])$. Then $\mathcal{N}(0,V_a) + R \sim \mathcal{N}(0,V)$ with $V = V_a + \text{Var}(\gamma_0' Z_h) = v_D^{-1}E[\text{Var}(\Pi a(W,\theta_0) - \gamma_0' h + \gamma_0' h|\psi)] - 2v_D^{-1}E[\text{Cov}(\Pi a(W,\theta_0) - \gamma_0' h, \gamma_0' h|\psi)] = v_D^{-1}E[\text{Var}(\Pi a(W,\theta_0)|\psi)]$. The covariance term is zero by Lemma 8.17. The second statement follows by setting $\psi = 1$. □

8.3 GMM Linearization

This section collects proofs needed for the key linearization result in Lemma 8.7. 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 8.8 (ULLN)**.** *Working under P in Definition 8.1:*

(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

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

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

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

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

Next we show $\|g_n - g_0\|_{\infty, \Theta} = o_p(1)$. We have $(g_n - g_0)(\theta) = E_n[\phi(W_i, \theta)] - E[\phi(W, \theta)]$. Under our assumptions, $|E_n[\phi(W_i, \theta)] - E[\phi(W, \theta)]|_{\infty, \Theta} = o_p(1)$ and $g_0(\theta) = E[\phi(W, \theta)]$ is continuous by Lemma 2.4 of Newey and McFadden (1994). This proves the second claim.

For the statement about objective functions, observe that

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

The first inequality by telescoping, then Cauchy-Schwarz, then using equivalence of finite-dimensional vector space norms and $\sup_\theta a(\theta) b(\theta) \leq \sup_\theta a(\theta) \sup_\theta b(\theta)$ 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 $|g_0|_{\infty,\Theta} \leq E[\sup_{\theta \in \Theta} \phi(W, \theta)] < \infty$. Also $|M_n|_2 = O_p(1)$ and $|M_n - M|_2 = o_p(1)$ by continuous mapping. Taking $\sup_{\theta \in \Theta}$ on both sides gives the result. The proof that $|H_n - H_0|_{\infty,K} = o_p(1)$ is identical. By triangle inequality, this proves the claim.

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 $m' \geq 1$. 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 $m \geq m'$. 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 8.9 (Consistency)**.** *Under the distribution P in Definition 8.1, 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 $\widehat{\theta} \xrightarrow{p} \theta_0$. □

Proof of Lemma 8.7. By Lemma 8.3, it suffices to show the result under P in Definition 8.1. 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

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

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 8.8. 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 8.8. 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 8.4, 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 8.7. 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. □

8.4 Linearization for M-Estimation

In this section, we extend our key result to M-estimation $\widehat{\theta} = \text{argmax}_{\theta \in \Theta} E_n[m(D_i, R_i, S_i, \theta)]$. 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 8.10 (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 8.2, this suffices to show that the main results of Sections 3-7 also apply to M-estimators with multiple local maxima.

Lemma 8.11 (Linearization)**.** *Suppose Definition 8.2 and Assumption 8.10 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 8.3, 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 8.8 and also $|E_n[\varphi_m(W_i,\theta)] - E[\varphi_m(W,\theta)]|_{\infty,\Theta} = o_p(1)$, all by Lemma 8.8. Then by extremum consistency, we have $\theta_n \xrightarrow{p} \theta_0$ and $\widehat{\theta} \xrightarrow{p} \theta_0$. By Lemma 8.8 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 8.7, with a slightly simpler first order condition. □

8.5 Nonlinear Rerandomization

Proof of Theorem 4.3. We first prove a slightly more general result, allowing for overidentified GMM estimation with positive definite weighting matrix $\Delta_n \xrightarrow{p} \Delta$. For $|x|_{2,A}^2 = x'Ax$, define

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

Define $g^1(D,X,\beta) = Dm(X,\beta)$ and $g^0(D,X,\beta) = (1-D)m(X,\beta)$. Under the expansion in Equation 3.1, we have $\phi^1(X,\beta) = pg^1(1,X,\beta) = pm(X,\beta)$ and $a^1(X,\beta) = v_D g^1(1,X,\beta) = v_D m(X,\beta)$. Similarly, $\phi^0(X,\beta) = (1-p)g^0(0,X,\beta) = (1-p)m(X,\beta)$ and $a^0(X,\beta) = -v_D g^0(0,X,\beta) = -v_D m(X,\beta)$. Note that $E[g^1(D,X,\beta)] = pE[m(X,\beta)]$ and $E[g^0(D,X,\beta)] = (1-p)E[m(X,\beta)]$, so the GMM parameters $\beta_1 = \beta_0 = \beta^*$, where 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 8.7 to GMM estimation using g^1 and g^0, under the measure P in Definition 8.1 we have

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

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

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

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

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

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

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

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

For the corollary, consider letting $\widehat{\beta} \in \mathrm{argmin}_{\beta \in \mathbb{R}^{d_\beta}} |E_n[m(X_i, \beta)]|_{2,\Delta_n}^2$. Relative to the expansion in Equation 3.1, $a_m(X_i, \beta) = 0$ and $\phi_m(X_i, \beta) = m(X_i, \beta)$, with linearization matrix Π_m as above. Then by Lemma 8.7 $\sqrt{n}(\widehat{\beta} - \beta^*) = \Pi_m E_n[m(X_i, \beta^*)] + o_p(1) = O_p(1)$. 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 $|v^n_{ijk}| \leq \sup_{\beta\in U} |(\partial/\partial\beta')m(X_i,\beta)|_F \in L_1$ by definition of S_n and $\tilde{\beta}_i \in [\beta^*, \widehat{\beta}]$ for each n, so by domination $(v^n_{ijk})_n$ is uniformly integrable, so $E_n[H_i v^n_{ijk}] = o_p(1)$ by Lemma A.2 of Cytrynbaum (2024). 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 8.12 (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, $E[\mathrm{Var}(h|\psi)]$, $\mathrm{Var}(h)$ are full rank.

Proof of Theorem 4.7. By assumption, $\widehat{\beta}$ is a GMM estimator for $m(D_i, X_i, \beta) = D_i \frac{L'(X_i'\beta)X_i}{L(X_i'\beta)} - (1 - D_i)\frac{L'(X_i'\beta)X_i}{1-L(X_i'\beta)}$. Let c such that $L(c) = p$. Then $\beta^* = (c, 0)$ has $E[m(D, X, \beta^*)] = E[H_i L'(c) X_i] = 0$. Relative to the decomposition in Equation 3.1, we have $\phi(X, \beta) = p\frac{L'(X_i'\beta)X_i}{L(X_i'\beta)} - (1-p)\frac{L'(X_i'\beta)X_i}{1-L(X_i'\beta)}$ and $a(X,\beta) = v_D(\frac{L'(X_i'\beta)X_i}{L(X_i'\beta)} + \frac{L'(X_i'\beta)X_i}{1-L(X_i'\beta)})$. Since $L(X_i'\beta^*) = L(c) = p$, apparently we have $\phi(X, \beta^*) = 0$ and $a(X, \beta^*) = L'(c)X_i$. It's easy to see $\mathrm{Var}(h) \succ 0$ implies $E[XX'] \succ 0$ for $X = (1, h)$. A calculation shows that $G_m = E[\frac{\partial}{\partial\beta'}\phi(X, \beta^*)] = -L'(c)^2 E[X_i X_i']$, so $\Pi_m = -G_m^{-1} = \frac{1}{L'(c)^2}E[X_i X_i']^{-1}$. By Lemma 8.7, we have shown

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

Consider rerandomizing until $\mathcal{J}_n = nE_n[(p - L(X_i'\widehat{\beta}))^2] \leq \epsilon^2$. 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

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

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

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

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

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

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

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

Define the function $b(x, y) = x' \mathrm{Var}(h)^{-1} x + y - \epsilon$. Then $\mathcal{J}_n \leq \epsilon \iff b(\mathcal{I}_n, \xi_n) \leq 0$ for $\mathcal{I}_n = \sqrt{n}E_n[H_i h_i]$ and $\xi_n \xrightarrow{p} 0$. Clearly, $x \to b(x, 0)$ is continuous. Also note $E[|h|_2^2] < \infty$ by assumption. Finally, for $Z_h \sim \mathcal{N}(0, E[\mathrm{Var}(h|\psi)])$, have $P(b(Z_h, 0) = 0) = P(Z_h' \mathrm{Var}(h)^{-1} Z_h = \epsilon^2) = 0$ since $E[\mathrm{Var}(h|\psi)]$ is full rank. Then this rerandomization satisfies all the conditions in Definition 8.2. By Lemma 8.7, 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 8.6, have $\sqrt{n}E_n[H_i \Pi a(W_i)]|\mathcal{F}_n \Rightarrow \mathcal{N}(0, V_a) + R$ with $R \sim \gamma_0' Z_h | Z_h \in T \sim \gamma_0' Z_h | Z_h' \mathrm{Var}(h)^{-1} Z_h \leq \epsilon$ for acceptance region $T = \{x : b(x, 0) \leq 0\} = \{x : x' \mathrm{Var}(h)^{-1} x \leq \epsilon\}$ and

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

This finishes the proof. □

8.6 Covariate Adjustment

Proof of Theorem 3.12. By Lemma 8.7, $\sqrt{n}(\widehat{\theta^*} - \theta_n)$ 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 8.6, $\sqrt{n}E_n[H_i \beta(W_i, \theta_0)]|\mathcal{F}_n \Rightarrow \mathcal{N}(0, V)$ with $V = v_D^{-1} \mathrm{Var}(\beta(W, \theta_0))$. 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 8.4, 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 8.7. The first statement now follows from Slutsky and Theorem 8.4. 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 8.6, $\sqrt{n}(\widehat{\theta}_{adj} - \theta_n)|\mathcal{F}_n \Rightarrow \mathcal{N}(0,V) + R$, independent with

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

The residual term $R \sim \beta_0' Z_h \,|\, Z_h \in A$. Then it suffices to show that $\beta_0 = 0$. Define $a_{\Pi\alpha} = \Pi a(W,\theta_0) - \alpha_0' w$. By Lemma 8.17, it further suffices to show $\beta_0 = 0$ solves $E[\mathrm{Var}(h|\psi)]\beta_0 = E[\mathrm{Cov}(h, a_{\Pi\alpha}|\psi)]$, i.e. that $E[\mathrm{Cov}(h, a_{\Pi\alpha}|\psi)] = 0$. To do so, note that $E[\mathrm{Cov}(h, a_{\Pi\alpha}|\psi)] = E[\mathrm{Cov}(h, (\Pi a - \alpha_0' w)|\psi)] = E[\mathrm{Cov}(h, \Pi a|\psi)] - E[\mathrm{Cov}(h, w|\psi)]\alpha_0$. By assumption, $E[\mathrm{Var}(w|\psi)]\alpha_0 = E[\mathrm{Cov}(w, \Pi a|\psi)]$. Since $h \subseteq w$, we have

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

This shows that $[\mathrm{Cov}(h, a_{\Pi\alpha}|\psi)] = 0$, so $\beta_0 = 0$ is a solution, proving the claim. This finishes the proof of the statement for θ_n. The result for θ_0 follows trivially, as in Corollary 3.8. □

Proof of Theorem 6.4. By Lemma 8.3, it suffices to show the result under P in Definition 8.1. By Lemma 8.13, $E_n[\check{w}_i \check{w}_i'] = k^{-1}(k-1)E[\mathrm{Var}(w|\psi)] + o_p(1)$. 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. We have $\widehat{\Pi} \xrightarrow{p} \Pi$ by assumption. Then it suffices to show $E_n[\check{w}_i(D_i - p)\widehat{g}_i'] = k^{-1}(k-1)E[\mathrm{Cov}(w, a|\psi)] + o_p(1)$. First, claim $E_n[\check{w}_i(D_i - p)\widehat{g}_i'] = E_n[\check{w}_i(D_i - p)g_i(\theta_0)'] + o_p(1)$, for $g_i(\theta) \equiv g(D_i, X_i, S_i, \theta)$. 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 $|E_n[\check{w}_i(D_i - p)(g_i(\widehat{\theta}) - g_i(\theta_0))']|_2 \leq E_n[|\check{w}_i|_2 |g_i(\widehat{\theta}) - g_i(\theta_0)|_2] \leq |\widehat{\theta} - \theta_0|_2 E_n[|\check{w}_i|_2 |\frac{\partial g_i}{\partial \theta'}(\tilde{\theta}_i)|_2] \leq |\widehat{\theta} - \theta_0|_2 (E_n[|\check{w}_i|_2^2] + E_n[|\frac{\partial g_i}{\partial \theta'}(\tilde{\theta}_i)|_2^2])$ 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 8.15. Similarly, $E_n[|\check{w}_i|_2^2] \leq E_n[|w_i|_2^2] = O_p(1)$ by the bound in Lemma 8.13. Since $|\widehat{\theta} - \theta_0|_2 = o_p(1)$ by Theorem 3.5, this proves the claim.

Next, claim $E_n[\check{w}_i(D_i - p)g_i(\theta_0)'] = E_n[\check{w}_i a(W_i,\theta_0)'] + o_p(1)$. By definition, we have $E_n[\check{w}_i(D_i - p)g_i(\theta_0)'] = E_n[\check{w}_i(D_i - p)\phi(W_i,\theta_0)'] + \mathrm{Var}(D)^{-1}E_n[(D_i - p)^2 \check{w}_i a(W_i,\theta_0)'] \equiv A_n + B_n$. Expanding $(D_i - p)^2$, $B_n = \mathrm{Var}(D)^{-1}E_n[[\mathrm{Var}(D) + (D_i - p)(1-2p)]\check{w}_i a(W_i,\theta_0)'] = E_n[\check{w}_i a(W_i,\theta_0)'] + \frac{1-2p}{\mathrm{Var}(D)}E_n[(D_i - p)\check{w}_i a(W_i,\theta_0)']$. Since $\phi = pg_1 + (1-p)g_0$ and $a = \mathrm{Var}(D)(g_1 - g_0)$, apparently it suffices to show $E_n[(D_i - p)\check{w}_i g_d(W_i,\theta_0)'] = o_p(1)$ for

each $d = 0, 1$. Since $E[|g_d(W_i, \theta_0)|_2^2] < \infty$, this follows from Lemma 8.13. Finally, $E_n[\check{w}_i a(W_i, \theta_0)'] = k^{-1}(k-1)E[\text{Cov}(w_i, a(W_i, \theta_0)|\psi_i)] + o_p(1)$ since $E[|w|_2^2 + |g_d|_2^2] < \infty$ and by applying Lemma 8.13 componentwise. This finishes the proof. □

Lemma 8.13. *Let $E[w_i^2 + v_i^2] < \infty$ with $w_i, v_i \in \sigma(W_i)$. Then under P in Definition 8.1, $E_n[(D_i - p)\check{w}_i\check{v}_i] = o_p(1)$ and $E_n[(D_i - p)\check{w}_i v_i] = o_p(1)$. Also $E_n[\check{w}_i\check{v}_i] = \frac{k-1}{k}E[\text{Cov}(w, v|\psi)] + o_p(1)$.*

Proof. First, note $|s|^{-1}\sum_{i\in s}\check{w}_i^2 = |s|^{-1}\sum_{i\in s}(w_i - |s|^{-1}\sum_{j\in s}w_j)^2 = \text{Var}_s(w_i) \leq E_s[w_i^2] = |s|^{-1}\sum_{i\in s}w_i^2$. Then in particular $\sum_{i\in s}\check{w}_i^2 \leq \sum_{i\in s}w_i^2$ and $E_n[\check{w}_i^2] \leq E_n[w_i^2]$. 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 (2024). By Lemma C.7 of Cytrynbaum (2024), 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 $|u_s| \leq \sum_{i\in s}|\check{w}_i\check{v}_i| \leq \sum_{i\in s}\check{w}_i^2 + \sum_{i\in s}\check{v}_i^2 \leq \sum_{i\in s}w_i^2 + \sum_{i\in s}v_i^2$ by Young's inequality and the bound above. Note that for any positive constants $(a_k)_{k=1}^m$ we have $\sum_k a_k\mathbb{1}(\sum_k a_k > c) \leq m\sum_k a_k\mathbb{1}(a_k > c/m)$. Applying this fact and the upper bounds gives

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

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, $E_n[\check{w}_i\check{v}_i] = k^{-1}(k-1)(E[w_iv_i] - E[E[w_i|\psi_i]E[v_i|\psi_i]]) + o_p(1) = k^{-1}(k-1)E[\text{Cov}(w_i, v_i|\psi_i)] + o_p(1)$ as claimed. The analysis of the term $\widehat{v}_{10}$ in Lemma A.6 of Cytrynbaum (2024) shows

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

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

8.7 Acceptance Region Optimization

Proof of Theorem 5.1. First we prove part (a). Define the function $f(a) = \sup_{b\in B} |b'a|$. As the sup of linear functions, f is convex (e.g. Rockafellar (1996)). Then the sublevel set $A \equiv \{a : f(a) \leq 1\}$ is convex. Note that $f(a) = f(-a)$, so A is symmetric. For the main statement of the theorem, let $a_n = \sqrt{n}E_n[H_i h_i]$. Clearly, f is positive homogeneous, i.e. $f(\lambda a) = \lambda f(a)$ for $\lambda \geq 0$. Then note that the LHS event occurs iff $f(a_n) \leq \epsilon \iff f(a_n/\epsilon) \leq 1 \iff a_n/\epsilon \in A \iff a_n \in \epsilon \cdot A$. This proves the main statement.

Next, we prove (b). Symmetry and convexity were already shown. Suppose B is bounded. Then by Cauchy-Schwarz $f(a) \leq |a|_2 \sup_{b\in B} |b|_2 < \infty$ for any $a \in \mathbb{R}^{d_h}$. Then f is a proper function, so f is continuous by Corollary 10.1.1. of Rockafellar (1996). Then $A = f^{-1}([0,1])$ is closed. Moreover, the open set $f^{-1}((1/3, 2/3)) \subseteq f^{-1}([0,1]) = A$, so A has non-empty interior. Suppose that B is open. Then B contains an open ball $B(x,\delta)$ for some $x \in \mathbb{R}^{d_h}$ and $\delta > 0$. Fix $a \in \mathbb{R}^{d_h}$ and define $b(a) = x + \text{sgn}(a'x)\frac{\delta}{2|a|}a$. By assumption, $b(a) \in B$. Then $f(a) = \sup_{b\in B} |b'a| \geq |b(a)'a| = |a'x + \text{sgn}(a'x)(\delta/2)|a|| = |a'x| + (\delta/2)|a| \geq (\delta/2)|a|$. Then $f(a) = \sup_{b\in B} |a'b| \geq (\delta/2)|a|$, so $A \subseteq B(0, 2/\delta)$.

Finally, we prove (c). Note that $R_A = \gamma_0' Z | Z \in \epsilon B^\circ$. By symmetry of ϵB°, we have $E[Z|Z \in \epsilon B^\circ] = 0$. Denote $W = Z/\epsilon$. Then we calculate

$$\begin{aligned}\text{Var}(R_A|Z \in \epsilon B^\circ) &= E[(\gamma_0' Z)^2 | Z \in \epsilon B^\circ] \leq E[\sup_{\gamma\in B} |\gamma' Z|^2 | Z \in \epsilon B^\circ] \\ &= \epsilon^2 E[\sup_{\gamma\in B} |\gamma' W|^2 | W \in B^\circ] \leq \epsilon^2 \cdot 1 = \epsilon^2.\end{aligned}$$

The first inequality is by well-specification The final inequality follows since $W \in B^\circ = \{a : \sup_{b\in B} |b'a| \leq 1\}$. This finishes the proof. □

Proof of Lemma 5.5. For $B = x + \Sigma B_p$ we compute the upper bound.

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

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

8.8 Inference

Proof of Theorem 7.2. By Lemma 8.3, it suffices to show the result under P in Definition 8.1. Define $m_i = \Pi g_i(\theta_0) - H_i\alpha_0' w_i$, the population version of $\widehat{m}_i$. Also define $m_{1i} = \Pi g_{1i}(\theta_0) - \alpha_0' w_i/p$ and $m_{0i} = \Pi g_{0i}(\theta_0) + \alpha_0' w_i/(1-p)$. We may expand

$$\begin{aligned}\phi_b &\equiv pm_{1i} + (1-p)m_{0i} = pg_{1i} + (1-p)g_{0i} = \Pi\phi(W,\theta_0),\\ a_b &\equiv v_D(m_{1i} - m_{0i}) = \Pi a(W,\theta_0) - \alpha_0' w_i.\end{aligned}$$

By Theorem 6.3, we need to estimate $V = \mathrm{Var}(\phi_b) + v_D^{-1}E[\mathrm{Var}(a_b|\psi)] = \mathrm{Var}(\phi_b) + v_D^{-1}E[a_b a_b'] - v_D^{-1}E[E[a_b|\psi]E[a_b|\psi]'] \equiv V_1 - V_2$. We expand $V_1 = \mathrm{Var}(\phi_b) + v_D^{-1}E[a_b a_b']$ as

$$\begin{aligned}V_1 &= \mathrm{Var}(pm_{1i} + (1-p)m_{0i}) + v_D E[(m_{1i} - m_{0i})(m_{1i} - m_{0i})']\\ &= E[(pm_{1i} + (1-p)m_{0i})(pm_{1i} + (1-p)m_{0i})'] + v_D E[(m_{1i} - m_{0i})(m_{1i} - m_{0i})']\\ &= (p^2 + v_D)E[m_{1i}m_{1i}'] + ((1-p)^2 + v_D)E[m_{0i}m_{0i}']\\ &= pE[m_{1i}m_{1i}'] + (1-p)E[m_{0i}m_{0i}'] = \mathrm{Var}_n(\widehat{m}_i) + o_p(1).\end{aligned}$$

The second equality since $E[\phi_b] = 0$, and the final equality by Lemma 8.14. By Lemma 8.15, we also have

$$\begin{aligned}V_2 &= v_D^{-1}E[E[a_b|\psi]E[a_b|\psi]'] = v_D(E[E[m_{1i}|\psi]E[m_{1i}|\psi]'] + E[E[m_{0i}|\psi]E[m_{0i}|\psi]'])\\ &\quad - v_D(E[E[m_{1i}|\psi]E[m_{0i}|\psi]'] + E[E[m_{0i}|\psi]E[m_{1i}|\psi]'])\\ &= v_D(\widehat{v}_1 + \widehat{v}_0 - \widehat{v}_{10} - \widehat{v}_{10}') + o_p(1).\end{aligned}$$

This finishes the proof. □

Proof of Theorem 7.3. By Lemma 8.3, it suffices to show the result under P in Definition 8.1. With notation as in the proof of Theorem 7.2, by Theorem 6.3, $c'(\widehat{\theta} - \theta_0) \Rightarrow \mathcal{N}(0, V_a(c))$ with variance $V_a(c) = v_D^{-1}c'E[\mathrm{Var}(a_b|\psi)]c$ for $a_b = v_D(m_{1i} - m_{0i})$. Then $V_a(c) = v_D c'E[\mathrm{Var}(m_{1i} - m_{0i}|\psi)]c$ may be expanded as

$$v_D \cdot c'(E[\mathrm{Var}(m_{1i}|\psi)] + E[\mathrm{Var}(m_{0i}|\psi)] - 2E[\mathrm{Cov}(m_{1i}, m_{0i}|\psi)])c.$$

Note that by Cauchy-Schwarz and Jensen we have the bound

$$\begin{aligned}&-2c'E[\mathrm{Cov}(m_{1i}, m_{0i}|\psi)]c \le 2|E[\mathrm{Cov}(c'm_{1i}, c'm_{0i}|\psi)]|\\ &\le 2E[\mathrm{Var}(c'm_{1i}|\psi)^{1/2}\,\mathrm{Var}(c'm_{0i}|\psi)^{1/2}] \le 2(E[\mathrm{Var}(c'm_{1i}|\psi)]E[\mathrm{Var}(c'm_{0i}|\psi)])^{1/2}\\ &= 2(c'E[\mathrm{Var}(m_{1i}|\psi)]c \cdot c'E[\mathrm{Var}(m_{0i}|\psi)]c)^{1/2}.\end{aligned}$$

Then we bound

$$V_a(c) \leq \bar{V}_a(c) \equiv v_D[(c'E[\mathrm{Var}(m_{1i}|\psi)]c)^{1/2} + (c'E[\mathrm{Var}(m_{0i}|\psi)]c)^{1/2}]^2.$$

Note $E[\mathrm{Var}(m_{1i}|\psi)] = E[m_{1i}m_{1i}'] - E[E[m_{1i}|\psi_i]E[m_{1i}|\psi_i]'] = E_n[\frac{D_i}{p}\widehat{m}_i\widehat{m}_i'] - \widehat{v}_1 + o_p(1)$ by Lemma 8.14 and Lemma 8.15. Similarly, $E[\mathrm{Var}(m_{0i}|\psi)] = E_n[\frac{1-D_i}{1-p}\widehat{m}_i\widehat{m}_i'] - \widehat{v}_0 + o_p(1)$. Then for $\widehat{u}_1 = E_n[\frac{D_i}{p}\widehat{m}_i\widehat{m}_i'] - \widehat{v}_1$ and $\widehat{u}_0 = E_n[\frac{1-D_i}{1-p}\widehat{m}_i\widehat{m}_i'] - \widehat{v}_0$ by continuous mapping

$$\widehat{V}_a(c) = v_D([c'\widehat{u}_1c]^{1/2} + [c'\widehat{u}_0c]^{1/2})^2 \xrightarrow{p} \bar{V}_a(c) \geq V_a(c).$$

This finishes the proof. □

Comparison of Variances. The superpopulation variance is

$$\begin{aligned} V(c) &= \mathrm{Var}(c'\phi_b) + v_D \cdot (E[\mathrm{Var}(c'm_{1i}|\psi)] + E[\mathrm{Var}(c'm_{0i}|\psi)] - 2E[\mathrm{Cov}(c'm_{1i}, c'm_{0i}|\psi)]) \\ &= p^2\,\mathrm{Var}(c'm_{1i}) + (1-p)^2\,\mathrm{Var}(c'm_{0i}) + v_D \cdot (E[\mathrm{Var}(c'm_{1i}|\psi)] + E[\mathrm{Var}(c'm_{0i}|\psi)]). \end{aligned}$$

Then the variance gap $V(c) - V_a(c)$ is

$$\begin{aligned} &p^2\,\mathrm{Var}(c'm_{1i}) + (1-p)^2\,\mathrm{Var}(c'm_{0i}) - 2v_D(E[\mathrm{Var}(m_{1i}|\psi)] \cdot E[\mathrm{Var}(m_{0i}|\psi)])^{1/2} \\ &= p^2\,\mathrm{Var}(E[c'm_{1i}|\psi_i]) + (1-p)^2\,\mathrm{Var}(E[c'm_{0i}|\psi_i]) \\ &+ (pE[\mathrm{Var}(c'm_{1i}|\psi)]^{1/2} - (1-p)E[\mathrm{Var}(c'm_{0i}|\psi)]^{1/2})^2 \geq 0. \end{aligned}$$

Lemma 8.14. *Impose Assumptions 3.1, 3.2, 7.1. Then under P in Definition 8.1, the following hold:*

(a) $E_n[\frac{D_i}{p}\widehat{m}_i\widehat{m}_i'] = E[m_{1i}m_{1i}'] + o_p(1)$ *and* $E_n[\frac{1-D_i}{1-p}\widehat{m}_i\widehat{m}_i'] = E[m_{0i}m_{0i}'] + o_p(1)$.

(b) $\mathrm{Var}_n(\widehat{m}_i) = pE[m_{1i}m_{1i}'] + (1-p)E[m_{0i}m_{0i}'] + o_p(1)$.

Proof. For (a), consider the first statement. We may expand this as

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

For $g_i = g_i(\theta_0)$, we have $|\widehat{m}_i - m_i|_2 = |\widehat{\Pi}\widehat{g}_i - \Pi g_i - H_i(\widehat{\alpha} - \alpha)'w_i|_2 \lesssim |\widehat{\Pi} - \Pi|_2|\widehat{g}_i|_2 + |\Pi|_2|\widehat{g}_i - g_i|_2 + |\widehat{\alpha} - \alpha_0|_2|w_i|_2$. Then the first term above has

$$\begin{aligned} |E_n[(D_i/p)\widehat{m}_i(\widehat{m}_i - m_i)']| \leq |\widehat{\Pi} - \Pi|_2E_n[|\widehat{m}_i|_2|\widehat{g}_i|_2] + |\Pi|_2E_n[|\widehat{m}_i|_2|\widehat{g}_i - g_i|_2] \\ + |\widehat{\alpha} - \alpha_0|_2E_n[|\widehat{m}_i|_2|w_i|_2]. \end{aligned}$$

We claim this term is $o_p(1)$. Note that $|\widehat{\Pi} - \Pi|_2 = o_p(1)$ and $|\widehat{\alpha} - \alpha_0|_2 = o_p(1)$ by assumption. Then applying Cauchy-Schwarz, it suffices to show $E_n[|\widehat{m}_i|_2^2 + |\widehat{g}_i|_2^2 + |w_i|_2^2] = O_p(1)$ 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 $E_n[|\widehat{m}_i|_2^2] = E_n[|\widehat{\Pi}\widehat{g}_i - H_i\widehat{\alpha}'w_i|_2^2] \leq 2E_n[|\widehat{\Pi}\widehat{g}_i|_2^2] + 2E_n[|\widehat{\alpha}'w_i|_2^2] \leq 2|\widehat{\Pi}|_2^2 E_n[|\widehat{g}_i|_2^2] + 2|\widehat{\alpha}|_2^2 E_n[|w_i|_2^2]$, so clearly it suffices to show $E_n[|\widehat{g}_i|_2^2] = O_p(1)$ to handle this term.

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

$$|\frac{\partial g_{1i}}{\partial \theta'}(\tilde{\theta}_i)|_2^2 + |\frac{\partial g_{0i}}{\partial \theta'}(\tilde{\theta}_i)|_2^2 \leq |\frac{\partial g_{1i}}{\partial \theta'}(\tilde{\theta}_i)|_F^2 + |\frac{\partial g_{0i}}{\partial \theta'}(\tilde{\theta}_i)|_F^2 = \sum_{d=0,1}\sum_{k=1}^{d_g} |\nabla g_{di}^k(\tilde{\theta}_{ik})|_2^2$$

$$\leq \sum_{d=0,1}\sum_{k=1}^{d_g} \sup_{\theta \in U} |\nabla g_{di}^k(\theta)|_2^2 \equiv \bar{U}_i.$$

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

Then we have shown $|E_n[(D_i/p)\widehat{m}_i(\widehat{m}_i - m_i)']| = o_p(1)$ and $E_n[(D_i/p)(\widehat{m}_i - m_i)m_i'] = o_p(1)$ by an identical argument. This shows that $E_n[(D_i/p)\widehat{m}_i\widehat{m}_i'] = E_n[(D_i/p)m_i m_i'] + o_p(1)$. Next, we have $E_n[(D_i/p)m_i m_i'] = E_n[(D_i/p)m_{1i}m_{1i}'] = E_n[m_{1i}m_{1i}'] + o_p(1) = E[m_{1i}m_{1i}'] + o_p(1)$. The first equality is by definition of m_i, m_{1i}. The second equality by Lemma A.2 of Cytrynbaum (2024) and the third equality by vanilla WLLN, using $E[|m_i|_2^2] < \infty$. This finishes our proof of the first statement of (a), and the second statement follows by symmetry.

For (b), note $E_n[\widehat{m}_i\widehat{m}_i'] = pE_n[\frac{D_i}{p}\widehat{m}_i\widehat{m}_i'] + (1-p)E_n[\frac{1-D_i}{1-p}\widehat{m}_i\widehat{m}_i'] = pE[\frac{D_i}{p}m_{1i}m_{1i}'] + (1-p)E[m_{0i}m_{0i}'] + o_p(1)$ by part (a) of the lemma. Moreover, $E_n[\widehat{m}_i] = E_n[\widehat{\Pi}\widehat{g}_i - H_i\widehat{\alpha}'w_i] = \widehat{\Pi}E_n[\widehat{g}_i] + o_p(1)$. 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 8.8. Then $\mathrm{Var}_n(\widehat{m}_i) = pE[m_{1i}(\theta_0)m_{1i}(\theta_0)'] + (1-p)E[m_{0i}(\theta_0)m_{0i}(\theta_0)'] + o_p(1)$, finishing the proof. □

Lemma 8.15. *Require Assumptions 3.1, 3.2, 7.1. Then under P in Definition 8.1, the variables in the statement of Theorem 7.2 have $\widehat{v}_{10} \xrightarrow{p} E[E[m_{1i}(\theta_0)|\psi]E[m_{0i}(\theta_0)|\psi]']$, $\widehat{v}_1 \xrightarrow{p} E[E[m_{1i}(\theta_0)|\psi]E[m_{1i}(\theta_0)|\psi]']$, and $\widehat{v}_0 \xrightarrow{p} E[E[m_{0i}(\theta_0)|\psi]E[m_{0i}(\theta_0)|\psi]']$.*

Proof. Let $\widehat{v}_1^o$ denote the oracle version of $\widehat{v}_1$, substituting $m_i = \Pi g_i(\theta_0) - H_i\alpha_0'w_i$ for $\widehat{m}_i$, and similarly for $\widehat{v}_0^o$, $\widehat{v}_{10}^o$. In Lemma A.6 of Cytrynbaum (2024), set $A_i = m_{1i}$ and $B_i = m_{1i}$. Applying the lemma componentwise, $\widehat{v}_1^o \xrightarrow{p} E[E[m_{1i}(\theta_0)|\psi]E[m_{1i}(\theta_0)|\psi]']$, $\widehat{v}_0^o \xrightarrow{p} E[E[m_{0i}(\theta_0)|\psi]E[m_{0i}(\theta_0)|\psi]']$, and $\widehat{v}_{10}^o \xrightarrow{p} E[E[m_{1i}(\theta_0)|\psi]E[m_{0i}(\theta_0)|\psi]']$. Then it

suffices to show that $\widehat{v}_1 - \widehat{v}_1^o = o_p(1)$, $\widehat{v}_0 - \widehat{v}_0^o = o_p(1)$, 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

$$\begin{aligned} |B_{ij}|_2 &\leq |\widehat{m}_i - m_i|_2 |m_j|_2 = |\widehat{\Pi}\widehat{g}_i - \Pi g_i - H_i(\widehat{\alpha} - \alpha)' w_i|_2 |m_j|_2 \\ &\leq |\widehat{\Pi} - \Pi|_2 |\widehat{g}_i|_2 |m_j|_2 + |\Pi|_2 |\widehat{g}_i - g_i|_2 |m_j|_2 + |\widehat{\alpha} - \alpha_0|_2 |w_i|_2 |m_j|_2. \end{aligned}$$

Then $B_n = n^{-1} \sum_{s \in \mathcal{S}_n^\nu} \sum_{i,j \in s} |\widehat{\Pi} - \Pi|_2 |\widehat{g}_i|_2 |m_j|_2 + |\Pi|_2 |\widehat{g}_i - g_i|_2 |m_j|_2 + |\widehat{\alpha} - \alpha_0|_2 |w_i|_2 |m_j|_2 \equiv B_{n1} + B_{n2} + B_{n3}$. 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 $B_{n3} \lesssim |\widehat{\alpha} - \alpha_0|_2 E_n[|w_i|_2^2 + |m_i|_2^2]$. 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 8.14. Note $E_n[|m_i|_2^2] = E_n[|\Pi g_i - H_i \alpha_0' w_i|_2^2] \leq 2E_n[|\Pi g_i|_2^2] + 2E_n[|\alpha_0' w_i|_2^2] \leq 2|\Pi|_2^2 E_n[|g_i|_2^2] + 2|\alpha_0|_2^2 E_n[|w_i|_2^2] = O_p(1)$ 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

$$\begin{aligned} B_{n3} &= n^{-1} \sum_{s \in \mathcal{S}_n^\nu} \sum_{i,j \in s} |\Pi|_2 |\widehat{g}_i - g_i|_2 |m_j|_2 \leq |\widehat{\theta} - \theta_0|_2 |\Pi|_2 \cdot n^{-1} \sum_{s \in \mathcal{S}_n^\nu} \sum_{i,j \in s} |\frac{\partial g_i}{\partial \theta'}(\tilde{\theta}_i)|_2 |m_j|_2 \\ &\lesssim |\widehat{\theta} - \theta_0|_2 |\Pi|_2 E_n[|\frac{\partial g_i}{\partial \theta'}(\tilde{\theta}_i)|_2^2 + |m_i|_2^2] = o_p(1). \end{aligned}$$

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 8.14. 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. □

8.9 Lemmas

Proposition 8.16 (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*

$$E[g(A_n)|\mathcal{F}_n] = E[g(A)] + o_p(1). \tag{8.2}$$

See Cytrynbaum (2021) for the proof.

Lemma 8.17. *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 8.18 (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 8.19. *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 8.20. *$A_n = O_p(1) \iff A_n = o_p(c_n)$ 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 $n = 1, \dots N$. 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 8.3. 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 8.5, $P(\mathcal{I}_n \in T_n|\mathcal{F}_n) \xrightarrow{p} P(Z_h \in T) = c$, so $P(S_n) \to 1$. We have the upper bound

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

The first equality by definition of Q. The first inequality by the definition of S_n. The final inequality by additivity of measures. Then for $r_n \equiv (1 - \mathbb{1}(S_n))Q(B_n|\mathcal{F}_n)$, we have

$Q(B_n|\mathcal{F}_n) = \mathbb{1}(S_n)Q(B_n|\mathcal{F}_n) + r_n$. Note that $|r_n| \leq 1$ and $r_n \xrightarrow{p} 0$, so $E_Q[r_n] = o(1)$ by modes of convergence. Then expand $Q(B_n)$ as

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

The second equality follows from part (a), and the final equality by tower law. The $o_p(1)$ results follow by setting $B_n = \{R_n > \epsilon\}$. The $O_p(1)$ results follow by the $o_p(1)$ statement and Lemma 8.20. □