

Robust Inference for Weighted Estimands*

Vod Vilfort[†]

July 8, 2026

Abstract

Researchers often conduct inference on weighted estimands, defined as weighted averages of group-level effects. Example settings include event studies with cohort-level effects and experiments with site-level effects. Under heterogeneous effects, different weighting schemes yield estimands with distinct empirical and policy interpretations, leading to ambiguity and disagreement over the choice of weights. I establish bounds on differences between weighted estimands and confidence bounds on effect heterogeneity, which I use to construct estimators that minimize worst-case bias and confidence intervals that are uniformly valid over classes of weighted estimands. I apply these methods to an event study in Lakdawala, Nakasone, and Kho (2023), which studies the effects of school-based internet access on test scores. I find that results are robust to broad classes of weights. I then apply the methods to Tennessee’s Project STAR experiment and find that results are sensitive to small departures from baseline weights.

*I thank Isaiah Andrews, Anna Mikusheva, and Alberto Abadie for their guidance and support. I thank Kirill Borusyak, Patrick Kline, Ricky Li, Sarah Moon, Andreas Petrou-Zeniou, Jonathan Roth, Reza Sarfati, Liyang Sun, Whitney Zhang, and seminar participants from the MIT econometrics lunch for helpful comments and discussions. I gratefully acknowledge support from the Jerry A. Hausman Fellowship and the National Science Foundation Graduate Research Fellowship under Grant No. 1745302.

[†]Department of Economics, Massachusetts Institute of Technology, vod@mit.edu.

1 Introduction

In empirical research, many estimands can be expressed as weighted averages of group-level parameters.¹ These *weighted estimands* give concise summaries of the parameters of interest, making them frequent targets for statistical inference. However, conventional inference procedures can be sensitive to the choice of weights: conclusions reached under a researcher’s baseline weights may not hold under alternative weights entertained by readers.

For example, in event studies with multiple treatment cohorts, many common estimands recover weighted averages of cohort-level causal effects, including conventional two-way fixed effects (TWFE) regression estimands and recently proposed alternatives.² TWFE has a convenient and flexible structure, which can yield lower-variance estimators in practice (Armstrong, Kline and Sun, 2025). However, TWFE can place negative weight on some of the cohort-level effects, complicating the interpretation of the resulting estimands under heterogeneous effects (De Chaisemartin and d’Haultfoeuille, 2020; Goodman-Bacon, 2021; Sun and Abraham, 2021; Borusyak, Jaravel and Spiess, 2024). Recent alternatives allow a researcher to target weighted estimands that rule out negative weights by design, but this introduces an additional degree of freedom: the researcher must decide which positive weights to use.

This choice matters because different readers may prefer different weights, depending on the empirical or policy question that they have in mind (Callaway and Sant’Anna, 2021). One reader may wish to upweight early-treated cohorts, another later-treated cohorts; one may care about short-run effects, another long-run effects; one may prefer weights that represent the characteristics of a policy-relevant target population, another may favor weights that trade off representativeness against estimation precision. These alternatives may all be reasonable, but they need not lead to the same conclusion under heterogeneous effects. Given the potential for disagreement, researchers may wish to assess the robustness of conclusions to the choice of weights. Current practice seems to reflect this desire by reporting results for different weighting choices, such as default implementations of the recently proposed alternatives. But such exercises consider only a finite set of weighting choices—which may not cover the range of plausible

¹Under various assumptions, ordinary least squares (OLS) estimands can be expressed as weighted averages of conditional average treatment effects (CATEs), conditional mean derivatives, or other regression parameters that may vary across covariate cells and treatment intensities (Yitzhaki, 1996; Angrist, 1998; Aronow and Samii, 2016; Słoczyński, 2022; Goldsmith-Pinkham, Hull and Kolesár, 2024); two-stage least squares (TSLS) estimands can be expressed as weighted averages of local average treatment effects (LATEs), which may vary across compliance types and covariate cells (Imbens and Angrist, 1994; Heckman and Vytlačil, 2005; Angrist and Fernandez-Val, 2010; Kolesár, 2013; Słoczyński, 2020; Huntington-Klein, 2020; Coussens and Spiess, 2021; Blandhol et al., 2022; Abadie, Gu and Shen, 2024); and two-way fixed effects (TWFE) estimands can be expressed as weighted averages of average treatment effects on the treated (ATTs), which may vary across treatment cohorts and time periods (De Chaisemartin and d’Haultfoeuille, 2020; Callaway and Sant’Anna, 2021; Goodman-Bacon, 2021; Sun and Abraham, 2021; Athey and Imbens, 2022; Gardner, 2022; Borusyak, Jaravel and Spiess, 2024; Wing, Freedman and Hollingsworth, 2024).

²For a review, see Roth et al. (2023).

weighting schemes—and they generally do not provide formal inferential guarantees for claims that conclusions are robust to the choice of weights.

In this paper, I develop robust inference procedures that directly account for ambiguity and disagreement over the choice of weights. I consider a general framework in which a researcher observes asymptotically normal estimates for a vector of parameters. The researcher reports a conventional point estimator and confidence interval (CI) for a baseline estimand defined by a vector of baseline weights. At the same time, the researcher faces a broad class of alternative weights, giving rise to alternative estimands for which the baseline estimator may be biased and the baseline CI may undercover. To address these concerns, I develop robust estimators that mitigate worst-case bias and robust CIs that ensure valid coverage.

To proceed, I first establish a sharp bound on the difference between the baseline weighted estimand and any given alternative weighted estimand. This bound is the product of (i) the *heterogeneity in parameters*, defined as the square root of the residual sum of squares from a generalized least squares (GLS) regression of the parameters on a constant and (ii) the *distance between weights*, defined as the standard deviation of the difference in estimators under the baseline and alternative weights.

To develop my estimator, I study the problem of choosing weights to minimize the *maximum distance* between weights over a class of alternative weights. For a broad range of such classes, I show that this problem admits a unique solution. Moreover, I show that the corresponding estimator is optimal for minimizing the maximum bias over the class of alternative estimands under any bound on the heterogeneity in parameters. I therefore refer to this estimator as the *robust estimator*, and to the corresponding weights as the *minimax-bias weights*. I propose that researchers report the robust estimator alongside the baseline estimator. Intuitively, the robust estimator accounts for the bias concerns of readers who may disagree with the baseline weights. Moreover, since the minimax-bias weights depend only on the class of alternative weights, they provide a natural default for researchers facing ambiguity over their initial choice of baseline weights.

To develop my CI, I study the problem of constructing an upper confidence bound (UCB) for the heterogeneity in parameters. I show that a monotonic transformation of the heterogeneity in estimates produces an optimal UCB. This *heterogeneity UCB*—together with the maximum distance between weights—yields a simple adjustment to the critical values of a baseline CI. The resulting *robust CI* provides uniformly valid coverage over the class of alternative estimands. In particular, for each alternative estimand, the coverage probability is at least one minus the sum of (i) the significance level of the baseline CI and (ii) the significance level of the heterogeneity UCB. For example, suppose a 95% baseline CI excludes zero, leading the researcher to reject the null of no average effect at the 5% level. If zero remains excluded from a robust CI constructed using a 95% heterogeneity UCB, then readers with alternative weights can robustly reject the null of no average effect at the 10% level. To facilitate such inference procedures, I propose that

researchers report the robust CI alongside the baseline CI.

The robust estimator and CI require the researcher to specify a class of alternative weights. My framework accommodates any compact and convex class of alternatives. I highlight three such classes that cover a broad range of empirical contexts. Below I give a high-level overview of these classes, deferring precise definitions and details to Section 2 of the paper.

The first class is the *bounded variance class*, which restricts attention to weights that yield estimators with variance no larger than a given bound, ensuring that the alternative estimands can be estimated with a reasonable degree of precision. For example, when assessing robustness of results, a researcher may wish to consider alternative estimands that can be estimated at least as precisely as the baseline. Under the bounded variance class, my proposed procedures admit a convenient reduction to the GLS regression of the group-level estimates on a constant, where (i) the GLS estimator coincides with the robust estimator, (ii) the GLS variance determines the maximum distance between weights for any given choice of variance bound, and (iii) the GLS residual sum of squares determines the heterogeneity UCB. This reduction has two key implications. First, because the GLS estimator minimizes variance, the robust estimator is optimal for both bias and variance under the bounded variance class. Second, due to the structure in (ii) and (iii), readers can construct robust CIs for their own choices of variance bounds, provided the researcher reports the GLS variance and heterogeneity UCB. These convenient properties make the bounded variance class a natural benchmark.

The simplex is the set of nonnegative weights, which yields weighted estimands that do not extrapolate beyond the range of the parameters. However, the unrestricted simplex allows for some groups to receive zero weight, which can be undesirable in practice. To ensure that groups are represented, one can impose a floor on the simplex. The *truncated simplex class* is precisely the set of simplex weights that are bounded below by the given floor. Under positive baseline weights, the floor can be parameterized so that the truncated simplex yields the set of weights that are a given fraction between the baseline weights and the unrestricted simplex weights. In this form, the truncated simplex class allows one to investigate perturbations of a chosen magnitude from the baseline weights.

The simplex weights can be viewed as probability distributions over the groups, and thus as different target populations that readers may be interested in. For instance, when the groups index the sites of an experiment, baseline results may not generalize to future sites of interest to policymakers (Allcott, 2015). If the researcher observes site-level covariates, however, the differences in covariate means under the baseline and alternative weights can inform the extent of external validity. Based on this idea, the *covariate balance class* considers simplex weights whose covariate means differ from the baseline mean by at most a given constant. Under this class, readers can assess whether results are robust to alternative populations whose covariate means are not too far from the baseline population.

I illustrate my framework in two empirical applications. The first application is an event study in [Lakdawala, Nakasone and Kho \(2023\)](#), which uses the staggered rollout of internet access in Peruvian public schools to estimate the causal effects of school-based internet access on second grade test scores. The authors find a delayed achievement response: TWFE estimates are initially small but grow over time, indicating that schools require time to adapt to new internet access. The authors find similar patterns and magnitudes when using the [Sun and Abraham \(2021\)](#) estimator to account for negative weighting concerns, concluding that results are not sensitive to the use of TWFE. My robust estimator and CI give a stronger conclusion: results are robust to *classes* of nonnegative weights that yield comparable estimation precision to the [Sun and Abraham \(2021\)](#) weights. In particular, robustness only fails when attempting to cover weighted estimands that cannot be estimated precisely in the first place.

The second application is Tennessee’s Project STAR (Student/Teacher Achievement Ratio) experiment, which randomized students in seventy-nine Tennessee public elementary schools to classrooms of different sizes to estimate the causal effects of class size on test scores ([Achilles et al., 2008](#); [Krueger, 1999](#)). Project STAR has been used in many policy discussions, but the experiment’s selection of schools raises concerns about whether results are representative of Tennessee or U.S. schools more broadly ([Schanzenbach, 2006](#)). To investigate these issues, I use an equal weights baseline to model the distribution of STAR schools and the truncated simplex class to model departures from the STAR empirical distribution. I show that, while the baseline CI implies medium-sized positive effects, the robust CI includes small-to-zero effect sizes even under small departures from equal weighting. The robust CI shrinks when I intersect the truncated simplex with the covariate balance class based on school-level covariates, but results remain sensitive to small departures from the baseline weights.

While I focus on event studies and multisite experiments as the motivating examples, my framework applies broadly to settings where group-level parameters are averaged using weights that sum to one. My inference procedures are not designed for continuously indexed groups, but the bound on differences between weighted estimands has a natural analogue in that case.³ Finally, the inferential guarantees of the robust estimator and CI do not depend on how the underlying parameters are defined. For example, in event studies one can define the parameters as cohort-level difference-in-differences (DiD) estimands rather than cohort-level average causal effects ([Sun and Abraham, 2021](#)). This distinction affects the interpretation of the weighted estimands, but not the validity of the inference procedures.

A large literature highlights issues that arise when interpreting weighted estimands. In the context of my framework, these issues represent different sources of ambiguity and disagreement over the choice of weights. For example, regression-based estimands often recover weighted averages of treatment effects, but allow for negative weights ([De Chaisemartin and d’Haultfoeuille,](#)

³For a discussion, see footnote 12 under Proposition 1.

2020; Słoczyński, 2020; Goodman-Bacon, 2021; Mogstad, Torgovitsky and Walters, 2021; Sun and Abraham, 2021; Blandhol et al., 2022; Bhuller and Sigstad, 2024; Goldsmith-Pinkham, Hull and Kolesár, 2024). Negative weights can flip the signs of treatment effects, complicating the interpretation of an estimand. In such contexts, there may be (i) ambiguity over how much negative weighting matters and (ii) disagreement over what alternative weighting schemes to consider.⁴ Regarding the latter, a separate issue is that a positively weighted estimand can still lack empirical or policy relevance (Yitzhaki, 1996; Heckman and Vytlačil, 2005; Crump et al., 2006; Angrist and Fernandez-Val, 2010; Aronow and Samii, 2016; Li, Morgan and Zaslavsky, 2018; Słoczyński, 2022; Mogstad and Torgovitsky, 2024; Poirier and Słoczyński, 2024). In this case, disagreement may arise from policymakers interested in the effects of treatment on their own target populations, while ambiguity may arise from researchers who must navigate such disagreement when summarizing results.

Common approaches to these issues are to report the extent of negative weighting in one’s estimator or to examine the stability of results under alternative weighting schemes.⁵ Intuitively, such approaches convey information about how far one’s weights deviate from a benchmark, the degree of underlying heterogeneity, or some combination of both. My procedures build on this same intuition, but in a GLS-based geometry that yields a notion of heterogeneity amenable to inference with confidence bounds, drawing on statistical results on optimal quantile-unbiased estimation (Pfanzagl, 1994). This allows one to directly account for weighting issues by constructing robust CIs and estimators with explicit inferential guarantees, thereby facilitating robust inference for weighted estimands.

The remainder of this paper proceeds as follows. Section 2 develops the model setting and notation. Section 3 defines the heterogeneity and distance measures and establishes the bound on differences between weighted estimands. Section 4 develops the robust estimator and CI and establishes their properties. These results are developed under the assumption that group-level estimates are normally distributed with a known covariance matrix. Section 5 shows that, with asymptotically normal estimates and a consistent covariance matrix estimator, my proposed procedures are uniformly asymptotically valid over a broad class of data generating processes. Section 6 discusses the practical implementation of these procedures. Section 7 presents the empirical applications. Section 8 concludes. The supplemental appendices contain proofs and additional results.

2 Model Setting

Consider parameters $\theta_k \in \mathbb{R}$ indexed by groups $k \in \{1, \dots, K\}$, where $K \geq 2$. Let $\theta = (\theta_1, \dots, \theta_K)'$ denote the vector of parameters. Researchers often conduct inference on estimands that can

⁴Abadie et al. (2025) and Chiu et al. (2026) give different takeaways for issue (i) in the event study context.

⁵See Roth et al. (2023, Section 3.2.1) for a review of such approaches in the event study context.

be expressed as weighted averages of the group-level parameters:

$$\tau_w(\theta) = w'\theta = \sum_{k=1}^K w_k \theta_k, \quad w \in \mathcal{W}, \quad \mathcal{W} = \left\{ w \in \mathbb{R}^K : \mathbf{1}'w = 1 \right\},$$

where $w \in \mathcal{W}$ is a vector of weights, $\mathcal{W}$ is the set of all weights, and $\mathbf{1}$ is the vector of ones. I refer to $\tau_w(\theta)$ as a *weighted estimand*. I allow the weights w to be negative unless stated otherwise. For the case of nonnegative (i.e., convex) weights, I define the simplex $\mathcal{W}_+ = \{w \in \mathcal{W} : w \geq 0\}$.

Example (Event Studies). Given units i and time periods $t \in \{0, 1, \dots, T\}$, where $T \geq 2$, the researcher observes outcomes $Y_{it} \in \mathbb{R}$ and treatment indicators $D_{it} \in \{0, 1\}$. A unit i 's cohort group $G_i = \min\{t : D_{it} = 1\}$ is the time period when first treated—if never treated, then $G_i = \infty$. There are no treated units in the base period (i.e., $G_i > 0$). Moreover, treatment is “absorbing” in the sense that a treated unit remains treated (i.e., $t \geq G_i$ implies $D_{it} = 1$).⁶ Let $Y_{it} = Y_{it}(G_i)$, where $Y_{it}(g)$ denotes the potential outcome for unit i when assigned to cohort g . Consider the group-time average treatment effect on the treated (ATT):

$$\text{ATT}_{g,t} = E[Y_{it}(g) - Y_{it}(\infty) | G_i = g], \quad t \geq g. \quad (1)$$

This parameter gives the period t causal effect of being treated in cohort g versus being never-treated, averaged among the units i in cohort $G_i = g$. Under assumptions of parallel trends and no anticipation, $\text{ATT}_{g,t}$ is identified for different group-time pairs (Callaway and Sant’Anna, 2021; Roth et al., 2023). Let k index such pairs (g_k, t_k) . For parameters $\theta_k = \text{ATT}_k = \text{ATT}_{g_k, t_k}$ and weights $w \in \mathcal{W}$, the corresponding weighted estimand is

$$\tau_w(\theta) = \sum_{k=1}^K w_k \text{ATT}_k.$$

Many common event-study estimands can be expressed as weighted estimands, including conventional TWFE regression coefficients and recently proposed alternatives (De Chaisemartin and d’Haultfoeuille, 2020; Callaway and Sant’Anna, 2021; Goodman-Bacon, 2021; Sun and Abraham, 2021; Gardner, 2022; Roth and Sant’Anna, 2023; Borusyak, Jaravel and Spiess, 2024; Wing, Freedman and Hollingsworth, 2024). In this setting, the different weights $w \in \mathcal{W}$ represent different ways of summarizing treatment effect heterogeneity across cohort groups and time periods; see Callaway and Sant’Anna (2021, Table 1) for examples. $\triangle$

Example (Multisite Experiments). For units i , there are outcomes $Y_i \in \mathbb{R}$, treatment indicators $D_i \in \{0, 1\}$, and covariates $X_i \in \mathbb{R}^M$. Let $(Y_i(1), Y_i(0))$ denote potential outcomes under

⁶For any treatment that is not absorbing, one can define an indicator for ever having received the treatment, which will be an absorbing treatment by construction; see Sun and Abraham (2021) for an example.

treatment and control, so that $Y_i = D_i Y_i(1) + (1 - D_i) Y_i(0)$. The units come from different site populations k , represented by distributions P_k over the unit-level random variables. Letting $E_k = E_{P_k}$ denote expectations under P_k , the site-level average treatment effects (ATEs) are

$$\text{ATE}_k = E_k[Y_i(1) - Y_i(0)], \quad k = 1, \dots, K.$$

Under assumptions of random treatment assignment within each site, ATE_k is identified for each site k (Hotz, Imbens and Mortimer, 2005). For parameters $\theta_k = \text{ATE}_k$ and weights $w \in \mathcal{W}$, the corresponding weighted estimand is

$$\tau_w(\theta) = \sum_{k=1}^K w_k \text{ATE}_k.$$

For instance, the equal weights (EW) vector $w_{\text{EW}} = \mathbf{1}/K$ represents an empirical distribution over the set of sites—this has been considered in, for example, Allcott (2015, Table VI). More generally, each simplex vector $w \in \mathcal{W}_+$ represents a distribution over the set of sites. $\triangle$

2.1 Conventional Inference for Weighted Estimands

The researcher observes a vector of estimates $\hat{\theta}$ for the parameters θ . I model the relationship between $\hat{\theta}$ and θ as follows.

Assumption 1. $\hat{\theta} \sim N(\theta, \Sigma)$, where Σ is a known positive definite covariance matrix.

I denote probabilities, expectations, and variances under $\hat{\theta} \sim N(\theta, \Sigma)$ as $\mathbb{P}_\theta\{\cdot\}$, $\mathbb{E}_\theta[\cdot]$, and $\text{Var}_\theta(\cdot)$, respectively. Moreover, $\Phi(z)$ denotes the cumulative distribution function (CDF) of the standard normal distribution $N(0,1)$ and z_α denotes its α -quantile for $\alpha \in (0,1)$.

Assumption 1 is motivated by standard large-sample asymptotic results. For example, the central limit theorem yields asymptotically normal estimates, which motivates the normality condition. Likewise, the law of large numbers yields consistent covariance matrix estimation, which motivates the condition that Σ is known. I therefore develop my procedures under Assumption 1. In Section 5, I relax Assumption 1 and establish the asymptotic validity of my procedures.

For each vector of weights $w \in \mathcal{W}$, Assumption 1 implies $w'\hat{\theta} \sim N(w'\theta, \sigma_w^2)$, where $\sigma_w^2 = w'\Sigma w$. For a given target estimand $\tau_w(\theta) = w'\theta$, the conventional point estimator is defined as $\hat{\tau}_w = w'\hat{\theta}$ and the conventional CI at significance level α is defined as

$$CI_w = \begin{cases} [\hat{\tau}_w \pm z_{1-\alpha/2} \sigma_w], & \text{two-sided,} \\ (-\infty, \hat{\tau}_w + z_{1-\alpha} \sigma_w], & \text{one-sided (upper),} \\ [\hat{\tau}_w - z_{1-\alpha} \sigma_w, \infty), & \text{one-sided (lower).} \end{cases} \quad (2)$$

The conventional estimator $\hat{\tau}_w$ is unbiased for $\tau_w(\theta)$:

$$\mathbb{E}_\theta[\hat{\tau}_w] = \tau_w(\theta), \quad \forall \theta.$$

The conventional CI_w has exact coverage for $\tau_w(\theta)$:

$$\mathbb{P}_\theta\{\tau_w(\theta) \in CI_w\} = 1 - \alpha, \quad \forall \theta.$$

In this sense, the conventional statistics $(\hat{\tau}_w, CI_w)$ provide valid inference for $\tau_w(\theta)$.

Example (Event Studies, continued). Let $\hat{\theta}_k = \widehat{ATT}_k$ denote estimators of ATT_k , such as DiD-based estimators. Under standard conditions, such estimators are asymptotically normal with consistently estimable covariance matrices as the number of units goes to infinity (Callaway and Sant’Anna, 2021). Often one must also estimate the weights w . For example, the Sun and Abraham (2021) weights depend on the population distribution of treatment D_{it} , so that in practice one must use the sample distribution of D_{it} to estimate w . For ease of exposition, I develop my procedures assuming that w is known and later establish asymptotic validity under estimated weights in Section 5. $\triangle$

Example (Multisite Experiments, continued). Let $\hat{\theta}_k = \widehat{ATE}_k$ denote estimators of ATE_k , such as those based on sample differences in means between treatment and control groups. Under site-level CLTs, such estimators are asymptotically normal with consistently estimable covariance matrices as the number of units goes to infinity. Note that even in cases where the unit-level variables (Y_i, D_i, X_i) are unobserved or confidential, one may still have access to site-level ATE estimates $\hat{\theta} = (\widehat{ATE}_1, \dots, \widehat{ATE}_K)'$ and covariate means $\mathbf{X} = (E_1[X_i], \dots, E_K[X_i])'$, as in the case of the metadata in Allcott (2015). The application of my framework to multisite experiments requires only the site-level variables. I accommodate the case of estimated covariate means $\hat{E}_k[X_i]$ in my asymptotic results. $\triangle$

2.2 Classes of Alternative Weights

The researcher reports $(\hat{\tau}_w, \sigma_w)$ for some baseline choice of weights $w \in \mathcal{W}$. This report yields conventional statistics $(\hat{\tau}_w, CI_w)$ that provide valid inference for the baseline estimand $\tau_w(\theta)$. However, the researcher may also consider a class $\Lambda \subseteq \mathcal{W}$ of alternative weights $\lambda \in \Lambda$, yielding alternative estimands $\tau_\lambda(\theta)$ for which $(\hat{\tau}_w, CI_w)$ may be uninformative.

Example (Event Studies, continued). Callaway and Sant’Anna (2021) advocate choosing w to address well-posed empirical or policy questions.⁷ While this principle is ideal, two practical issues can arise, leading to consideration of a class of alternatives Λ .

⁷This perspective is also prevalent in other empirical settings, such as instrumental variables estimation under heterogeneous treatment effects (Heckman and Vytlacil, 2007; Mogstad and Torgovitsky, 2024).

Researcher Ambiguity. First, even with a well-posed question, the researcher may find it difficult to articulate corresponding weights w . This costly introspection presumably underlies the various default weighting schemes considered in the event studies literature.⁸ A researcher may choose a default specification for w and then assess robustness of results to other defaults. But to the extent that these defaults fail to capture the scope of plausible weights, such robustness exercises may not adequately account for researcher ambiguity over the choice of w . To do better, it seems useful to consider a broader class of alternatives for w , such as the example Λ developed below.

Reader Disagreement. Second, even absent introspection costs, the researcher may have to communicate results to readers who are interested in different questions and hence disagree over the choice of weights. For example, colleagues with different priors about the empirical setting may wish to highlight different types of treatment effect heterogeneity: e.g., across cohorts versus across time (Callaway and Sant’Anna, 2021). In such cases, Λ represents the range of questions that readers may be interested in, which can differ from the question addressed by w . $\triangle$

Example (Multisite Experiments, continued). Given the site-level populations $P_1, \dots, P_K$, a baseline vector of simplex weights $w \in \mathcal{W}_+$ represents the population $P_w = \sum_k w_k P_k$. The set of readers may include policymakers interested in the effect of treatment on their own target populations, leading to disagreement over the choice of weights. For concreteness, consider a policymaker with target population P_0 and a corresponding target parameter

$$\text{ATE}_0 = E_0[Y_i(1) - Y_i(0)].$$

This policymaker may find the baseline weights w to be palatable when P_0 resembles P_w , such as when the mean $\mu_0 = E_0[X_i]$ under P_0 is close to the mean $\mu_w(\mathbf{X}) = w' \mathbf{X}$ under P_w .⁹ In view of this, $\Lambda \subseteq \mathcal{W}_+$ can be specified as a class of populations that a range of policymakers may find to be palatable. Note that if there were a single known policymaker, one could employ canonical approaches for extrapolating from the site-level populations ($P_1, \dots, P_K$) to the target population ATE_0 , such as density reweighting under covariate shift assumptions.¹⁰ However, such avenues are less tractable in the face of multiple and potentially unknown policymakers. $\triangle$

Whether the issue at hand is researcher ambiguity, reader disagreement, or some combination of both, my framework assumes that it can be represented by a class of alternative weights

⁸See Callaway and Sant’Anna (2021, Table 1) and Roth et al. (2023, Table 2) for examples.

⁹This mirrors the logic advanced in Aronow and Samii (2016, page 255) for interpreting the representativeness of weighted estimands in regression contexts. In their language, $\mu_w(\mathbf{X})$ is the covariate mean for the *effective sample* of units spanned by the w -weighted sites.

¹⁰There are applications of this approach in economics (Hotz, Imbens and Mortimer, 2005; Stuart et al., 2011; Allcott, 2015; Dehejia, Pop-Eleches and Samii, 2021), health (Cole and Stuart, 2010; Hartman et al., 2015), and machine learning (Farahani et al., 2021; Zhou et al., 2022).

Λ . To make this modeling assumption a practical one, I restrict attention to classes Λ that have the following structure.

Assumption 2. $\Lambda \subseteq \mathcal{W}$ is nonempty, compact, and convex.

This structure ensures the existence and uniqueness of solutions to upcoming optimization problems. I now provide examples of classes Λ that satisfy Assumption 2 and map to various empirical contexts.

Example (Bounded Variance). To analyze the robustness of baseline inferences to alternative weights, one may wish to account for differences in how well the corresponding estimands can be estimated. Formally, one can consider weights λ for which the standard deviation of $\hat{\tau}_\lambda$ is at most r times that of the baseline $\hat{\tau}_w$, leading to the *bounded variance class*

$$\Lambda_\sigma(r) = \{\lambda \in \mathcal{W} : \sigma_\lambda \leq r \sigma_w\}, \quad r \geq \frac{\sigma_{\min}}{\sigma_w}, \quad \sigma_{\min}^2 = \min_{w \in \mathcal{W}} \sigma_w^2 = \frac{1}{\mathbf{1}' \Sigma^{-1} \mathbf{1}}. \quad (3)$$

This class reflects a preference for estimands that can be estimated with some reasonable degree of precision.¹¹ For instance, $r = 1$ represents the class of estimands that can be estimated at least as precisely as the baseline estimand. When convenient, I leave the dependence of $\Lambda_\sigma(r)$ on the *standard deviation ratio bound* r implicit, denoting $\Lambda_\sigma = \Lambda_\sigma(r)$. $\triangle$

Example (Truncated Simplex). The simplex $\mathcal{W}_+ = \{w \in \mathcal{W} : w \geq 0\}$ is the set of nonnegative weights, which yields estimands that do not extrapolate beyond the range of the parameters: $\min_k \theta_k \leq \tau_w(\theta) \leq \max_k \theta_k$. However, the simplex allows for some groups to receive zero weight, which can be undesirable in practice. To ensure that groups are represented, one can impose a floor on the simplex. In particular, given baseline simplex weights $w \in \mathcal{W}_+$, the floor $(1 - \epsilon)w$ yields the *truncated simplex class*

$$\Lambda_+(\epsilon) = \{\lambda \in \mathcal{W}_+ : \lambda \geq (1 - \epsilon)w\} = \{(1 - \epsilon)w + \epsilon \lambda : \lambda \in \mathcal{W}_+\} \quad \epsilon \in [0, 1]. \quad (4)$$

This class allows one to flexibly model departures from the baseline simplex weights w , where the *truncation parameter* ϵ is the fraction of a given alternative $\lambda \in \Lambda_+$ that is allowed to deviate from w . To ensure that a group k is represented in $\lambda \in \Lambda_+(\epsilon)$, one must use weights with $w_k > 0$. In the case of equal weighting, $w = w_{EW} = \mathbf{1}/K$, one can interpret ϵ as the maximum possible discrepancy in weights $|\lambda_k - \lambda_{k'}|$ between two groups k and k' . I give additional interpretations of ϵ when discussing practical implementations in Section 6. $\triangle$

Example (Covariate Balance). For $m \in \{1, \dots, M\}$, consider group-level covariates $\mathbf{X}_m \in \mathbb{R}^K$. A baseline simplex weight vector $w \in \mathcal{W}_+$ is a probability distribution over the set of groups, and

¹¹For example, Mogstad and Torgovitsky (2024, pages 13-14) note that “How interesting a target parameter is also cannot be divorced from the difficulty involved in estimating it.”

can thus be viewed as a population with covariate means $w' \mathbf{X}_m$. Likewise, alternative simplex weights $\lambda \in \mathcal{W}_+$ yield populations with covariate means $\lambda' \mathbf{X}_m$. Letting $|(\lambda - w)' \mathbf{X}_m|$ denote the covariate m balance gap and $\text{sd}(\mathbf{X}_m)$ the standard deviation of $\mathbf{X}_{m,k}$ across groups k , the set of populations with balance gaps no larger than $\bar{c}$ standard deviations for all m is given by the *covariate balance class*

$$\Lambda_X(\bar{c}) = \left\{ \lambda \in \mathcal{W}_+ : \max_{m \in \{1, \dots, M\}} \frac{|(\lambda - w)' \mathbf{X}_m|}{\text{sd}(\mathbf{X}_m)} \leq \bar{c} \right\}, \quad \bar{c} \geq 0, \quad \min_{m \in \{1, \dots, M\}} \text{sd}(\mathbf{X}_m) > 0. \quad (5)$$

This class allows one to interrogate a range of populations with different covariate profiles from the baseline. For ease of reference, I abbreviate the balance gap measure in Λ_X as

$$c_\lambda(\mathbf{X}) = \max_{m \in \{1, \dots, M\}} \frac{|(\lambda - w)' \mathbf{X}_m|}{\text{sd}(\mathbf{X}_m)}, \quad \mathbf{X} = (\mathbf{X}_1, \dots, \mathbf{X}_M).$$

Thus, I can write $\Lambda_X(\bar{c}) = \{\lambda \in \mathcal{W}_+ : c_\lambda(\mathbf{X}) \leq \bar{c}\}$ for *balance gap bound $\bar{c}$* . $\triangle$

Remark 1 (Intersections). One can also take intersections of the above classes. For example, intersecting the bounded variance class $\Lambda_\sigma(r)$ with the unrestricted simplex $\Lambda_+(1) = \mathcal{W}_+$ yields the *bounded variance simplex class*

$$\Lambda_\sigma^+(r) = \Lambda_\sigma(r) \cap \Lambda_+(1) = \{\lambda \in \mathcal{W}_+ : \sigma_\lambda \leq r \sigma_w\}, \quad r \geq \frac{\min_{w \in \mathcal{W}_+} \sigma_w}{\sigma_w}, \quad (6)$$

which considers convex weights that yield estimators with reasonable estimation precision. As a second example, one can intersect the covariate balance class $\Lambda_X(\bar{c})$ with the truncated simplex $\Lambda_+(\epsilon)$ to obtain the *truncated covariate balance class*

$$\Lambda_X^\epsilon(\bar{c}) = \Lambda_X(\bar{c}) \cap \Lambda_+(\epsilon) = \{(1 - \epsilon)w + \epsilon\lambda : \lambda \in \Lambda_X(\bar{c}/\epsilon)\}, \quad \epsilon > 0, \quad (7)$$

which considers ϵ -departures from the baseline w towards populations within a covariate mean radius $\bar{c}/\epsilon$ from the baseline. I will consider these intersections in the empirical applications, but I focus on the benchmark classes $(\Lambda_\sigma, \Lambda_+, \Lambda_X)$ when interpreting my theoretical results. $\triangle$

I conclude this section by formalizing the inference distortions that arise when one uses the baseline statistics $(\hat{\tau}_w, CI_w)$ to learn about alternative estimands $\tau_\lambda(\theta)$. In particular, I consider (i) the absolute bias of $\hat{\tau}_w$ for $\tau_\lambda(\theta)$, given by

$$|\mathbb{E}_\theta[\hat{\tau}_w] - \tau_\lambda(\theta)| = |\tau_\lambda(\theta) - \tau_w(\theta)| \geq 0,$$

and (ii) the noncoverage probability of CI_w for $\tau_\lambda(\theta)$, given (in the two-sided case) by

$$\mathbb{P}_\theta\{\tau_\lambda(\theta) \notin CI_w\} = \Phi\left(z_{\alpha/2} - \frac{\tau_\lambda(\theta) - \tau_w(\theta)}{\sigma_w}\right) + \Phi\left(z_{\alpha/2} + \frac{\tau_\lambda(\theta) - \tau_w(\theta)}{\sigma_w}\right) \geq \alpha.$$

The inference distortions are governed by the (absolute) difference in estimands $|\tau_\lambda(\theta) - \tau_w(\theta)|$: the larger this difference, the higher the bias and noncoverage. In particular, when $\lambda \neq w$, there exist parameter values θ where $|\tau_\lambda(\theta) - \tau_w(\theta)|$ is arbitrarily large so that bias is arbitrarily large and noncoverage is arbitrarily close to one. In such cases, $(\hat{\tau}_w, CI_w)$ is completely uninformative for $\tau_\lambda(\theta)$. Thus, to obtain useful inferences for $\tau_\lambda(\theta)$, one must account for potential differences $|\tau_\lambda(\theta) - \tau_w(\theta)|$ in the weighted estimands across $\lambda \in \Lambda$ and $\theta \in \mathbb{R}^K$.

3 Bounding the Difference in Estimands

In this section, I bound the difference in estimands in terms of the heterogeneity in parameters and the distance between weights. I then show how to infer these quantities, which will be the basis for constructing the robust estimator and CI in Section 4.

Heterogeneity in Parameters. Consider the generalized least squares (GLS) regression of the parameters θ on a constant $\mathbf{1}$ under weighting matrix Σ^{-1} . I define the *heterogeneity in parameters* as the square root of the GLS residual sum of squares:

$$H(\theta) = \sqrt{\min_{\gamma \in \mathbb{R}} (\theta - \mathbf{1}\gamma)' \Sigma^{-1} (\theta - \mathbf{1}\gamma)}.$$

By construction, the heterogeneity in θ is zero if and only if θ_k is constant across k . The above regression is uniquely minimized at the GLS estimand

$$\arg\min_{\gamma \in \mathbb{R}} (\theta - \mathbf{1}\gamma)' \Sigma^{-1} (\theta - \mathbf{1}\gamma) = w'_{\text{GLS}} \theta = \tau_{\text{GLS}}(\theta), \quad w_{\text{GLS}} = \frac{\Sigma^{-1} \mathbf{1}}{\mathbf{1}' \Sigma^{-1} \mathbf{1}}, \quad (8)$$

which yields the formula

$$H(\theta) = \sqrt{\theta' Q \theta}, \quad Q = \Sigma^{-1/2} A \Sigma^{-1/2}, \quad A = I - \frac{\Sigma^{-1/2} \mathbf{1} \mathbf{1}' \Sigma^{-1/2}}{\mathbf{1}' \Sigma^{-1} \mathbf{1}}, \quad (9)$$

where A is the annihilator matrix for $\Sigma^{-1/2} \mathbf{1}$ and I is the identity matrix. Thus, the squared heterogeneity can be represented as a quadratic form $\theta' Q \theta$ of the parameters. This particular quadratic form will facilitate quantile-unbiased inference in Section 3.2.

Distance Between Weights. The standard deviation $\|v\|_\Sigma = \sqrt{v' \Sigma v}$ of a given linear estimator $v' \hat{\theta}$ defines a norm on $v \in \mathbb{R}^K$. Given weights λ and w , I define the *distance between weights*

as the corresponding norm of their difference:

$$\|\lambda - w\|_{\Sigma} = \sqrt{(\lambda - w)' \Sigma (\lambda - w)} = \sqrt{\text{Var}_{\theta}(\hat{\tau}_{\lambda} - \hat{\tau}_w)}.$$

Intuitively, the distance $\|\lambda - w\|_{\Sigma}$ measures the disagreement between λ and w by taking the standard deviation of the corresponding difference in estimators.

Proposition 1. *For any $w \in \mathcal{W}$ and $\lambda \in \mathcal{W}$, the difference in estimands is bounded as*

$$|\tau_{\lambda}(\theta) - \tau_w(\theta)| \leq H(\theta) \|\lambda - w\|_{\Sigma}, \quad \forall \theta.$$

This bound is sharp in the sense that, given $\lambda \neq w$ and any $\eta \geq 0$, there exists θ with $H(\theta) = \eta$ for which the bound holds with equality.

Proof. See Appendix C.1. □

Proposition 1 shows that the difference in estimands is bounded by the product of (i) the heterogeneity $H(\theta)$, which is unknown due to the parameters θ being unobserved, and (ii) the distance $\|\lambda - w\|_{\Sigma}$, which is unknown due to ambiguity or disagreement over the alternative weights $\lambda \in \Lambda$.¹² In Section 3.1, I account for the latter by taking the maximum distance across alternative weights. In Section 3.2, I account for the former by constructing an upper confidence bound (UCB) on the heterogeneity.

3.1 Maximum Distance

Given baseline w and class of alternatives Λ , the *maximum distance* between weights is

$$\max_{\lambda \in \Lambda} \|\lambda - w\|_{\Sigma} = \sqrt{\max_{\lambda \in \Lambda} (\lambda - w)' \Sigma (\lambda - w)}.$$

Intuitively, the maximum distance measures the worst-case disagreement between the baseline and alternative weights across Λ . Under the structure on Λ from Assumption 2, the maximum is attained at some $\lambda^* \in \Lambda$, which can be viewed as the weights of a reader who disagrees with w the most. Below I analyze the structure of the maximum distance under the example classes from Section 2.2 and illustrate how this structure can facilitate communication between researchers and readers—I defer practical recommendations and implementation choices to Section 6.

¹²Proposition 1 is based on the Cauchy-Schwarz inequality, similar to Scheffé-style arguments for bounding data-dependent linear combinations to obtain uniformly valid inference (Scheffé, 1953; Lehmann and Romano, 2024). However, Proposition 1 considers nonrandom linear combinations and obtains bounds on the inference distortions themselves. Analogous bounds hold if one replaces Σ with another positive definite matrix when defining the heterogeneity and distance measures. It also has a natural analogue for continuously indexed groups, where one may define the weighted estimands as $\tau_w(\theta) = \int w(k) \theta(k) d\nu(k)$ and apply the same Cauchy-Schwarz argument in an $L^2(\nu)$ inner product.

Example (Bounded Variance, continued). One can show that

$$\max_{\lambda \in \Lambda_\sigma} \|\lambda - w\|_\Sigma = \sqrt{r^2 \sigma_w^2 - \sigma_{\min}^2} + \sqrt{\sigma_w^2 - \sigma_{\min}^2}, \quad \sigma_{\min}^2 = \min_{w \in \mathcal{W}} \sigma_w^2 = \frac{1}{\mathbf{1}' \Sigma^{-1} \mathbf{1}}, \quad (10)$$

which is a function of $(r, \sigma_w, \sigma_{\min})$. If the researcher reports the minimum variance $\sigma_{\min}^2$ alongside the baseline variance σ_w^2 , then a reader can compute the maximum distance for their own choice of r . Conveniently, $\sigma_{\min}^2$ coincides with the variance of the GLS estimator

$$\hat{\tau}_{\text{GLS}} = w'_{\text{GLS}} \hat{\theta} = \underset{\gamma \in \mathbb{R}}{\text{argmin}} (\hat{\theta} - \mathbf{1}\gamma)' \Sigma^{-1} (\hat{\theta} - \mathbf{1}\gamma), \quad \sigma_{\min}^2 = \sigma_{\text{GLS}}^2 = w'_{\text{GLS}} \Sigma w_{\text{GLS}}, \quad (11)$$

where w_{GLS} are the weights of the GLS estimand defined in (8). The above GLS regression is an input to my inference procedures for $H(\theta)$ in Section 3.2, so the GLS variance can be obtained at essentially no further cost. $\triangle$

Example (Truncated Simplex, continued). Letting $v_1, \dots, v_K$ denote the standard unit vectors, one can show that

$$\max_{\lambda \in \Lambda_+} \|\lambda - w\|_\Sigma = \epsilon \max_{\lambda \in \mathcal{W}_+} \|\lambda - w\|_\Sigma = \epsilon \max_{j \in \{1, \dots, K\}} \|v_j - w\|_\Sigma, \quad (12)$$

which is a function of $(\epsilon, \max_j \|v_j - w\|_\Sigma)$. If the researcher reports $\max_j \|v_j - w\|_\Sigma$, then a reader can compute the maximum distance for their own choice of ϵ . In the case of an equal weights baseline $w = w_{\text{EW}} = \mathbf{1}/K$ and a reader with weights $\lambda_0 \in \mathcal{W}_+$ in the simplex, $\epsilon = 1 - K \min_k \lambda_{0,k}$ is the smallest truncation parameter at which $\Lambda_+(\epsilon)$ contains the reader's weights λ_0 . $\triangle$

Example (Covariate Balance, continued). Compared to the above examples, there does not appear to be a transparent expression for $\max_{\lambda \in \Lambda_X} \|\lambda - w\|_\Sigma$. However, the researcher can plot it for a range of $\bar{c}$, which allows a reader to examine the maximum distance at their preferred value of $\bar{c}$ from the reported range. To support this exercise, the researcher can additionally report covariate statistics $(w' \mathbf{X}_m, \text{sd}(\mathbf{X}_m))_{m=1}^M$. This allows a given policymaker with covariate means $\mu_0 \in \mathbb{R}^M$ for a target population P_0 to compute the balance gap measure

$$c_0(\mathbf{X}) = \max_{m \in \{1, \dots, M\}} \frac{|\mu_{0,m} - w' \mathbf{X}_m|}{\text{sd}(\mathbf{X}_m)},$$

and check the maximum distance at $\bar{c}_0 = \min\{c : c_0(\mathbf{X}) \leq c\}$. Intuitively, $\Lambda_X(\bar{c}_0)$ is the smallest class for which there can exist weights $\lambda \in \Lambda_X(\bar{c}_0)$ that induce balance gaps consistent with the policymaker's target population. In this sense, $\Lambda_X(\bar{c}_0)$ is a minimal set of plausible candidates for P_0 and the maximum distance at $\bar{c}_0$ gives a corresponding measure of ambiguity. Note that the above discussion implicitly assumes that the number of covariates M is small enough for the policymaker to easily compute $c_0(\mathbf{X})$. $\triangle$

Remark 2 (Reporting Constraints). Λ represents the broad range of beliefs and objectives that researchers and readers may have, making it impractical for researchers to report $(\hat{\tau}_\lambda, \sigma_\lambda)$ for all $\lambda \in \Lambda$. Moreover, if there are confidentiality restrictions or communication costs, reporting the entire data $(\hat{\theta}, \Sigma)$ may not be a broadly applicable solution.¹³ With these issues in mind, my framework develops methods that facilitate inferences across $\lambda \in \Lambda$ based simply on (i) the baseline report $(\hat{\tau}_w, \sigma_w)$ and (ii) a supplementary report of low-dimensional statistics, where the latter loosely refers to statistics whose dimension (e.g., the number of columns and rows occupied in a table) does not grow with the number of groups K , thus precluding $(\hat{\theta}, \Sigma)$. $\triangle$

3.2 Heterogeneity UCB

I now show how to infer the unknown heterogeneity in parameters $H(\theta)$. In particular, I derive an *upper confidence bound* (UCB) for $H(\theta)$: given a significance level $\beta \in (0, 1)$, I construct an estimator $\hat{\eta}_{1-\beta}$ such that

$$\mathbb{P}_\theta\{H(\theta) \leq \hat{\eta}_{1-\beta}\} \geq 1 - \beta, \quad \forall \theta. \quad (13)$$

In words, $\hat{\eta}_{1-\beta}$ upper bounds the unknown heterogeneity $H(\theta)$ with probability at least $1 - \beta$, for any value of θ . I refer to $\hat{\eta}_{1-\beta}$ as a *heterogeneity UCB* at confidence level $1 - \beta$.

Derivation. To derive $\hat{\eta}_{1-\beta}$, I use the representation of $H(\theta) = \sqrt{\theta' Q \theta}$ as a quadratic form in θ . In particular, since $\hat{\theta} \sim N(\theta, \Sigma)$, the definition of Q from (9) implies that the *heterogeneity in estimates* $H(\hat{\theta}) = \sqrt{\hat{\theta}' Q \hat{\theta}}$ satisfies

$$\hat{\theta}' Q \hat{\theta} = (\Sigma^{-1/2} \hat{\theta})' A (\Sigma^{-1/2} \hat{\theta}) \sim \chi_{K-1}^2(H(\theta)), \quad \forall \theta,$$

where $\chi_{K-1}^2(\eta)$ denotes the noncentral chi-squared distribution with $K - 1$ degrees of freedom and noncentrality parameter η^2 . Let $F_{\chi^2}(x; \eta)$ denote its CDF and consider the corresponding pivot function $\eta \mapsto F_{\chi^2}(\hat{\theta}' Q \hat{\theta}; \eta)$. The probability integral transform yields

$$F_{\chi^2}(\hat{\theta}' Q \hat{\theta}; H(\theta)) \sim U(0, 1), \quad \forall \theta.$$

In particular, the probability of observing $F_{\chi^2}(\hat{\theta}' Q \hat{\theta}; H(\theta)) \geq \beta$ is equal to $1 - \beta$. For each $x > 0$, the function $\eta \mapsto F_{\chi^2}(x; \eta)$ is strictly decreasing (Sun, Baricz and Zhou, 2010). Therefore, when $F_{\chi^2}(\hat{\theta}' Q \hat{\theta}; 0) > \beta$, I define $\hat{\eta}_{1-\beta}$ as the unique solution to $F_{\chi^2}(\hat{\theta}' Q \hat{\theta}; \hat{\eta}_{1-\beta}) = \beta$. Otherwise, when

¹³Allcott (2015) is one example where confidentiality restrictions preclude the reporting of $(\hat{\theta}, \Sigma)$, but not of $(\hat{\tau}_w, \sigma_w)$. Even absent such restrictions, researchers still often focus on reporting statistics of the form $(\hat{\tau}_w, \sigma_w)$, as discussed in Athey and Imbens (2023). This reporting convention may be due to communication costs that induce researchers to report low-dimensional statistics for their readers.

$F_{\chi^2}(\hat{\theta}'Q\hat{\theta};0) \leq \beta$, I define $\hat{\eta}_{1-\beta} = 0$.¹⁴ In summary,

$$\hat{\eta}_{1-\beta} = \begin{cases} 0, & F_{\chi^2}(\hat{\theta}'Q\hat{\theta};0) \leq \beta, \\ F_{\chi^2}^{-1}(\hat{\theta}'Q\hat{\theta};\beta), & F_{\chi^2}(\hat{\theta}'Q\hat{\theta};0) > \beta. \end{cases}$$

This construction depends only on $\hat{\theta}'Q\hat{\theta}$, which is conveniently obtained in (11) as the residual sum of squares from the GLS regression of the estimates $\hat{\theta}$ on a constant $\mathbf{1}$ under weighting matrix Σ^{-1} . The inversion for $\hat{\eta}_{1-\beta}$ is computationally efficient, since it amounts to finding the root of a strictly monotone function.

The next result shows that the above $\hat{\eta}_{1-\beta}$ is a valid UCB for $H(\theta)$. Moreover, it is *quantile-unbiased* under heterogeneity in the sense that its (weak) overestimation probability for $H(\theta)$ is equal to $1-\beta$ when $H(\theta) > 0$. Note that when $H(\theta) = 0$, the overestimation probability is equal to one, as would be the case for any nonnegative estimator of $H(\theta)$.

Proposition 2. *The above $\hat{\eta}_{1-\beta}$ satisfies (13). Moreover, under heterogeneous θ , the confidence level $1-\beta$ is attained:*

$$\mathbb{P}_\theta\{H(\theta) \leq \hat{\eta}_{1-\beta}\} = 1-\beta, \quad \forall \theta: H(\theta) > 0. \quad (14)$$

Under homogeneous θ , the coverage probability is equal to one.

Proof. See Appendix C.2. □

The quantile-unbiasedness in Proposition 2 facilitates a variety of inference procedures for $H(\theta)$. First, $[0, \hat{\eta}_{1-\beta}]$ is a one-sided upper confidence interval for $H(\theta)$, with exact coverage rate $1-\beta$ under heterogeneous θ . Next, the one-sided lower interval $[\hat{\eta}_\beta, \infty)$ satisfies

$$\mathbb{P}_\theta\{H(\theta) \geq \hat{\eta}_\beta\} = 1-\beta, \quad \forall \theta: H(\theta) > 0. \quad (15)$$

Likewise, one can construct a two-sided interval $[\hat{\eta}_{\beta/2}, \hat{\eta}_{1-\beta/2}]$ that satisfies

$$\mathbb{P}_\theta\{\hat{\eta}_{\beta/2} \leq H(\theta) \leq \hat{\eta}_{1-\beta/2}\} = 1-\beta, \quad \forall \theta: H(\theta) > 0. \quad (16)$$

Finally, one can construct a point estimator $\hat{\eta}_{1/2}$ that is median-unbiased in the sense that

$$\mathbb{P}_\theta\{H(\theta) \leq \hat{\eta}_{1/2}\} = \mathbb{P}_\theta\{H(\theta) \geq \hat{\eta}_{1/2}\} = \frac{1}{2}, \quad \forall \theta: H(\theta) > 0. \quad (17)$$

¹⁴This construction is consistent with the general approach in Pfanzagl (1994, Section 5.3), which constructs confidence bounds for one-parameter distributions satisfying appropriate monotonicity conditions.

That is, the median of $\hat{\eta}_{1/2}$ is equal to $H(\theta)$ under heterogeneity.¹⁵ Given my focus on upper bounding differences in weighted estimands, I focus on properties (13) and (14). Nevertheless, properties (15)-(17) illustrate the versatility of $\hat{\eta}_{1-\beta}$ for inference on parameter heterogeneity. I conclude this section with comparative statics and an optimality result for $\hat{\eta}_{1-\beta}$.

Comparative Statics. For a given η , the CDF $F_{\chi^2}(x; \eta)$ is increasing in x and decreasing in its degrees of freedom (Sun, Baricz and Zhou, 2010). Thus, $\hat{\eta}_{1-\beta}$ is increasing in $\hat{\theta}'Q\hat{\theta}$ given a fixed K , and decreasing in K given a fixed $\hat{\theta}'Q\hat{\theta}$. In this sense, $\hat{\eta}_{1-\beta}$ is increasing in heterogeneity of the estimates relative to the number of groups: $H(\hat{\theta})/\sqrt{K}$. Notice that, under homoskedasticity $\Sigma = \sigma^2 I$, the expression $\hat{\theta}'Q\hat{\theta}/K$ reduces to the empirical variance of the estimates relative to the common sampling variance σ^2 of the estimates:

$$\frac{\hat{\theta}'Q\hat{\theta}}{K} = \frac{\frac{1}{K} \sum_{k=1}^K \hat{\theta}_k^2 - \left(\frac{1}{K} \sum_{k=1}^K \hat{\theta}_k \right)^2}{\sigma^2}, \quad \Sigma = \sigma^2 I.$$

Thus, the proposed heterogeneity measure $\hat{\eta}_{1-\beta}$ has intuitive comparative statics in terms of $\hat{\theta}'Q\hat{\theta}/K$. Finally, holding $\hat{\theta}'Q\hat{\theta}$ and K fixed, a larger confidence level $1 - \beta$ implies a larger $\hat{\eta}_{1-\beta}$. This highlights a tradeoff when constructing $\hat{\eta}_{1-\beta}$: to upper bound $H(\theta)$ with higher confidence, one must tolerate a larger UCB, corresponding to a lower significance level β .

Optimality. Let $\bar{\Theta}^\eta = \{\theta : H(\theta) = \eta\}$ denote the set of parameters for which the heterogeneity is equal to η , and consider the class of potentially randomized estimators $\tilde{\eta}_{1-\beta}$ that are quantile-unbiased in the sense of (14). The following result shows that $\hat{\eta}_{1-\beta}$ is minimax *most accurate* in the sense that it minimizes the worst-case probability of overestimating $H(\theta)$, for any degree of underlying heterogeneity η .

Proposition 3. *For any quantile-unbiased estimator $\tilde{\eta}_{1-\beta}$ and any degree of heterogeneity η , the heterogeneity measure $\hat{\eta}_{1-\beta}$ yields a lower worst-case probability of overestimating $H(\theta)$:*

$$\sup_{\theta \in \bar{\Theta}^\eta} \mathbb{P}_\theta \{ \hat{\eta}_{1-\beta} \geq H(\theta) + \varepsilon \} \leq \sup_{\theta \in \bar{\Theta}^\eta} \mathbb{P}_\theta \{ \tilde{\eta}_{1-\beta} \geq H(\theta) + \varepsilon \}, \quad \forall (\eta, \varepsilon) > 0.$$

Proof. See Appendix C.3. □

The above notion of optimality is a minimax analogue of the notion of a uniformly most accurate confidence bound (Lehmann and Romano, 2024, Section 3.5). In Appendix A, I show that $\hat{\eta}_{1-\beta}$ is uniformly most accurate in the class of quantile-unbiased estimators that depend

¹⁵For point estimation of heterogeneity, one can also consider mean-unbiased estimation of $\theta'Q\theta$, following a similar construction to, e.g., Kline, Saggio and Sølvesten (2020). In particular, one can show that $\hat{\theta}'Q\hat{\theta} - (K-1)$ is a mean-unbiased estimator of $\theta'Q\theta$. However, the quantile-unbiased approach adopted here is more amenable to the construction of confidence bounds for my objects of interest.

on $\hat{\theta}$ through $\hat{\theta}'Q\hat{\theta}$. In fact, $\hat{\eta}_{1-\beta}$ minimizes expected loss in this class under any quasiconvex loss function that attains its minimum at $H(\theta)$, for any heterogeneous θ . The latter optimality statement is based on results from Pfanzagl (1994) on optimal quantile-unbiased estimation in models with monotone likelihood ratios.

4 Robust Inference Procedures

Given the maximum distance between weights $\max_{\lambda \in \Lambda} \|\lambda - w\|_{\Sigma}$ and the heterogeneity UCB $\hat{\eta}_{1-\beta}$, I define the bias UCB

$$\hat{B}_w^{\beta}(\Lambda) = \hat{\eta}_{1-\beta} \max_{\lambda \in \Lambda} \|\lambda - w\|_{\Sigma}.$$

The following result shows that $\hat{B}_w^{\beta}(\Lambda)$ is indeed a valid bias UCB.

Proposition 4. *For any $w \in \mathcal{W}$, the above $\hat{B}_w^{\beta}(\Lambda)$ satisfies*

$$\mathbb{P}_{\theta} \left\{ \max_{\lambda \in \Lambda} |\tau_{\lambda}(\theta) - \tau_w(\theta)| \leq \hat{B}_w^{\beta}(\Lambda) \right\} \geq 1 - \beta, \quad \forall \theta.$$

Moreover, when $\Lambda \neq \{w\}$, there exists θ for which the coverage probability is equal to $1 - \beta$.

Proof. See Appendix C.4. □

In words, $\hat{B}_w^{\beta}(\Lambda)$ upper bounds the maximum bias with probability at least $1 - \beta$, for any value of θ . This provides an avenue for addressing the bias and undercoverage of the baseline estimator $\hat{\tau}_w$ and CI_w for inference on alternative estimands $\tau_{\lambda}(\theta)$ across $\lambda \in \Lambda$. To this end, I develop the robust estimator in Section 4.1 and the robust CI in Section 4.2.

4.1 Robust Estimator

Since $\hat{B}_w^{\beta}(\Lambda)$ is a valid bias UCB under any $w \in \mathcal{W}$, it provides a criterion for optimizing w . In the optimization, I allow for consideration sets $\mathcal{W}^* \subseteq \mathcal{W}$ that are nonempty, closed, and convex. I distinguish the weights in the optimization problem from the researcher's baseline weights, denoting the former by $\bar{w} \in \mathcal{W}^*$ and the latter by $w \in \mathcal{W}$. I allow for classes Λ that depend on the baseline w , but not on the optimizer variable $\bar{w}$.

Given a consideration set $\mathcal{W}^*$, I define the *robust estimator* $\hat{\tau}^*$ as the weighted estimator induced by the *minimax-bias weights* $w^* \in \mathcal{W}^*$, given by

$$\hat{\tau}^* = \hat{\tau}_{w^*}, \quad w^* = \arg \min_{\bar{w} \in \mathcal{W}^*} \max_{\lambda \in \Lambda} \|\lambda - \bar{w}\|_{\Sigma}.$$

Let $\hat{B}_{\min}^{\beta}(\Lambda) = \hat{\eta}_{1-\beta} \max_{\lambda \in \Lambda} \|\lambda - w^*\|_{\Sigma}$ denote the corresponding minimax-bias UCB.

Proposition 5. *The minimax-bias weights w^* exist uniquely and satisfy*

$$\hat{B}_{\min}^{\beta}(\Lambda) \leq \hat{B}_{\bar{w}}^{\beta}(\Lambda), \quad \forall \bar{w} \in \mathcal{W}^*.$$

Moreover, for any parameter space $\Theta^{\eta} = \{\theta : H(\theta) \leq \eta\}$ with a bound $\eta > 0$ on heterogeneity, the minimax-bias weights w^* solve

$$w^* = \arg \min_{\bar{w} \in \mathcal{W}^*} \max_{\lambda \in \Lambda} \max_{\theta \in \Theta^{\eta}} |\mathbb{E}_{\theta}[\hat{\tau}_{\bar{w}}] - \tau_{\lambda}(\theta)|, \quad \forall \eta > 0.$$

In this sense, the robust estimator $\hat{\tau}^* = \hat{\tau}_{w^*}$ is optimal for minimizing worst-case bias under any bound on heterogeneity.

Proof. See Appendix C.5. □

Proposition 5 establishes optimality of the minimax-bias weights w^* from two perspectives: first, at the observed data, w^* minimizes the bias $\hat{B}_{\bar{w}}^{\beta}(\Lambda)$ that is inferred *ex post*; and second, across hypothetical data realizations, w^* minimizes the bias that is possible *ex ante*. Conveniently, the maximum bias from either perspective is proportional to the maximum distance: w^* does not depend on the estimated heterogeneity $\hat{\eta}_{1-\beta}$ or on the heterogeneity bound η . In the above setup, the optimality of w^* among weights $\bar{w} \in \mathcal{W}^*$ is equivalent to optimality of the robust estimator $\hat{\tau}^*$ among weighted estimators $\hat{\tau}_{\bar{w}}$.

By default, I take the consideration set to be the class of alternative weights: $\mathcal{W}^* = \Lambda$. In this case, w^* can be interpreted as the weights that minimize worst-case disagreement across the class of alternative weights $\bar{w} \in \Lambda$. Moreover, because the minimax-bias weights depend only on the class of alternative weights, they provide a natural default for researchers facing ambiguity over their initial choice of baseline weights. I now analyze the structure and interpretation of the minimax-bias weights w^* for the example classes of Λ from Section 2.2.

Example (Bounded Variance, continued). Let $w^*(r)$ denote the minimax-bias weights under the bounded variance class $\Lambda_{\sigma}(r)$. Following equation (10), one can show that

$$\hat{\tau}^*(r) = \hat{\tau}_{\text{GLS}}, \quad w^*(r) = w_{\text{GLS}} = \frac{\Sigma^{-1}\mathbf{1}}{\mathbf{1}'\Sigma^{-1}\mathbf{1}}, \quad \min_{\bar{w} \in \Lambda_{\sigma}} \max_{\lambda \in \Lambda_{\sigma}} \|\lambda - \bar{w}\|_{\Sigma} = \sqrt{r^2\sigma_w^2 - \sigma_{\text{GLS}}^2}. \quad (18)$$

The corresponding standard deviation is $\sigma^*(r) = \sigma_{\text{GLS}}$. In summary, the minimax-bias weights coincide with the GLS weights so that the robust estimator and corresponding variance reduce to the GLS estimator and variance, for any choice of standard deviation ratio bound r .¹⁶ Since

¹⁶GLS weights have appeared in discussions about how to aggregate CATEs—Li, Morgan and Zaslavsky (2018) call them overlap weights while Goldsmith-Pinkham, Hull and Kolesár (2024) call them easiest-to-estimate weights. Established properties include variance-efficiency (Crump et al., 2006; Li, Morgan and Zaslavsky, 2018; Goldsmith-Pinkham, Hull and Kolesár, 2024) and interpretation as policy effect weights from a marginal

the GLS weights minimize variance, $\sigma_{\text{GLS}}^2 = \sigma_{\min}^2$, it follows that the robust estimator is optimal for both bias *and* variance under the bounded variance class.¹⁷ From Proposition 5, it follows that $\hat{\tau}^* = \hat{\tau}_{\text{GLS}}$ is minimax-optimal for estimating $\tau_\lambda(\theta)$ over $(\lambda, \theta) \in \Lambda \times \Theta^n$ under squared error in the class of weighted estimators.¹⁸

When Σ is diagonal, the GLS weights are guaranteed to be strictly positive, weighting each group k by the precision $1/\sigma_k^2$ of its estimate relative to the other groups:

$$w_{\text{GLS},k} = \frac{1/\sigma_k^2}{\sum_{k=1}^K 1/\sigma_k^2} > 0, \quad \Sigma = \begin{pmatrix} \sigma_1^2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_K^2 \end{pmatrix}.$$

This occurs, for example, when $\hat{\theta}$ is a vector of estimates from statistically independent sites k , or when $\hat{\theta}$ is a vector of CATE estimates computed on independent observations across covariate cells k . But when Σ is non-diagonal, w_{GLS} may place negative weight on some groups in order to minimize variance. To prevent this negative weighting, one can restrict the bounded variance class to the simplex as in (6). In this case, generally $w^* \neq w_{\text{GLS}}$. $\triangle$

Example (Truncated Simplex, continued). Let $w^*(1) = \arg\min_{\bar{w} \in \mathcal{W}_+} \max_{\lambda \in \mathcal{W}_+} \|\lambda - \bar{w}\|_\Sigma$ denote the minimax-bias weights under $\Lambda_+(1) = \mathcal{W}_+$ and $w^*(0) = w$ the minimax-bias weights under $\Lambda_+(0) = \{w\}$. Following equation (12), one can show that

$$w^*(\epsilon) = (1 - \epsilon)w^*(0) + \epsilon w^*(1), \quad \min_{\bar{w} \in \Lambda_+} \max_{\lambda \in \Lambda_+} \|\lambda - \bar{w}\|_\Sigma = \epsilon \min_{\bar{w} \in \mathcal{W}_+} \max_{\lambda \in \mathcal{W}_+} \|\lambda - \bar{w}\|_\Sigma. \quad (19)$$

Thus, the minimax-bias weights $w^*(\epsilon)$ are a fraction $1 - \epsilon$ consistent with the baseline weights and a fraction ϵ consistent with the weights that minimize disagreement over the unrestricted simplex. The robust estimator is therefore $\hat{\tau}^*(\epsilon) = (1 - \epsilon)\hat{\tau}^*(0) + \epsilon\hat{\tau}^*(1)$, which is between the baseline estimator $\hat{\tau}^*(0) = \hat{\tau}_w$ and the simplex-robust estimator $\hat{\tau}^*(1)$. $\triangle$

Example (Covariate Balance, continued). There does not appear to be a transparent formula for $w^*(\bar{c})$ under $\Lambda_X(\bar{c})$. However, in the case of $\bar{c} = 0$, one can interpret $\Lambda_X(0)$ in (5) as the set

increase in the log odds of treatment (Kennedy, 2019; Zhou and Opacic, 2022). I establish a complementary property: GLS weights minimize worst-case bias when the only consensus over weighting schemes is that they should yield estimators with bounded variance.

¹⁷This double-optimality of $\hat{\tau}_{\text{GLS}}$ parallels recent results in Adusumilli (2026) and Sarfati and Vilfort (2026), which show that variance-efficient estimators are also bias-optimal under unrestricted—but bounded—forms of model misspecification. In the weighted estimand context, bounding heterogeneity in the GLS metric and the alternative weights in the standard deviation norm can be viewed as unrestricted bounds on the misspecification of a target estimand $\tau_\lambda(\theta)$. If one imposes further restrictions, such as intersecting Λ_σ with the simplex, then the robust estimator generally differs from GLS.

¹⁸Recent work studies how to optimally aggregate treatment effects from a decision-theoretic perspective (Armstrong and Kolesár, 2021a; de Chaisemartin, 2021; Kwon and Sun, 2025; Lau, 2026). In contrast to this work, my framework allows for ambiguity and disagreement over the target estimand.

of solutions to a synthetic control problem (Abadie and Gardeazabal, 2003; Abadie, Diamond and Hainmueller, 2010). In particular, if the baseline w represents synthetic control weights derived under the max-norm and standardized covariates, the corresponding predictor means $w'\mathbf{X}$ produce the same synthetic control fit as each alternative $\lambda \in \Lambda_X(0)$, meaning that $\Lambda_X(0)$ is a set of synthetic control weights.¹⁹ The minimax-bias weights $w^*(0)$ under $\Lambda_X(0)$ can then be viewed as a point of centrality among the set of synthetic control weights. $\triangle$

4.2 Robust CI

I define the robust CI centered at $w \in \mathcal{W}$ as

$$CI_w^* = \begin{cases} [\hat{\tau}_w \pm \text{cv}_{1-\alpha}(\hat{B}_w^\beta(\Lambda)/\sigma_w)\sigma_w], & \text{two-sided,} \\ (-\infty, \hat{\tau}_w + z_{1-\alpha}\sigma_w + \hat{B}_w^\beta(\Lambda)], & \text{one-sided (upper),} \\ [\hat{\tau}_w - z_{1-\alpha}\sigma_w - \hat{B}_w^\beta(\Lambda), \infty), & \text{one-sided (lower),} \end{cases}$$

where the critical value function $\text{cv}_{1-\alpha}(b)$ gives the $(1-\alpha)$ -quantile of the folded normal distribution $|N(b,1)|$. Thus, the robust CI_w^* parallels the baseline CI_w , but uses critical values that adjust for the inferred bias $\hat{B}_w^\beta(\Lambda)$. This adjustment widens the baseline endpoints: $CI_w \subseteq CI_w^*$, with equality if and only if $\hat{B}_w^\beta(\Lambda) = 0$. Provided that $\Lambda \neq \{w\}$, this equality holds if and only if the heterogeneity UCB is insignificant at level β in the sense that $\hat{\eta}_{1-\beta} = 0$.

Proposition 6. CI_w^* provides uniformly valid coverage at confidence level $1 - (\alpha + \beta)$:

$$\mathbb{P}_\theta\{\tau_\lambda(\theta) \in CI_w^*\} \geq 1 - (\alpha + \beta), \quad \forall \lambda \in \Lambda, \quad \forall \theta. \quad (20)$$

Proof. See Appendix C.6. $\square$

Proposition 6 shows that CI_w^* covers $\tau_\lambda(\theta)$ at confidence level $1 - (\alpha + \beta)$, uniformly over $\lambda \in \Lambda$.²⁰ This uniform coverage differs from the conventional coverage in (2). The latter provides inference guarantees for one given w , while the former extends such guarantees to every $\lambda \in \Lambda$. Relative to conventional coverage, the price of uniform coverage is (i) a higher error level from using $\hat{\eta}_{1-\beta}$ to infer heterogeneity and (ii) longer intervals to obtain coverage that is robust to alternative weights $\lambda \in \Lambda$.

¹⁹See Liu (2025) for a recent identification framework predicated on such ambiguity sets.

²⁰ $CI_w \subseteq CI_w^*$ implies $\mathbb{P}_\theta\{\tau_w(\theta) \in CI_w^*\} \geq \mathbb{P}_\theta\{\tau_w(\theta) \in CI_w\} = 1 - \alpha$ for all θ . Thus, the robust CI_w^* still covers the baseline estimand $\tau_w(\theta)$ at conventional confidence levels. The additional error level β when considering alternative estimands $\tau_\lambda(\theta)$ is a Bonferroni correction to account for the first-step inference on heterogeneity; such Bonferroni corrections have been used for two-step construction of test statistics and CIs in other contexts, such as Romano, Shaikh and Wolf (2014) and McCloskey (2017).

Interpretation. Intuitively, if $\lambda \in \Lambda$ indexes the preferred weights of different readers, then the robust CI_w^* provides valid coverage for each reader’s estimand $\tau_\lambda(\theta)$. For example, suppose the baseline CI_w excludes zero, leading the researcher to reject the w -null of no average effect $H_{0,w} : \tau_w(\theta) = 0$ at significance level α . If zero remains excluded from the robust CI_w^* , then a reader with alternative weights $\lambda \in \Lambda$ can reject the λ -null of no average effect $H_{0,\lambda} : \tau_\lambda(\theta) = 0$ at significance level $\alpha + \beta$. In this sense, CI_w^* facilitates robust inference.

Additional Coverage Properties. In Appendix B.1, I show that CI_w^* also provides a form of simultaneous coverage—but at a lower confidence level in the two-sided case. I compare this to the uniform coverage in (20). In comparing the two coverage notions, I draw connections to the literature on inference for partially identified parameters (Molinari, 2020). In Appendix B.2, I derive an upper bound on the coverage rate of CI_w^* . The bound is strictly less than one for any $\lambda \in \Lambda$ under any degree of heterogeneity. This implies that CI_w^* has nontrivial coverage for $\tau_\lambda(\theta)$ across all values of $\lambda \in \Lambda$ and θ .

Choice of Centering Weights. The robust CI_w^* can be centered at any $w \in \mathcal{W}$ while maintaining uniform coverage over $\lambda \in \Lambda$. By default, I center the robust CIs at the minimax-bias weights w^* and denote $CI^* = CI_{w^*}^*$. In particular, given the minimax-bias UCB $\hat{B}_{\min}^\beta(\Lambda)$, robust estimator $\hat{\tau}^*$, and standard deviation $\sigma^* = \sqrt{(w^*)' \Sigma w^*}$, the robust CIs become

$$CI^* = \begin{cases} \left[\hat{\tau}^* \pm cv_{1-\alpha} \left(\hat{B}_{\min}^\beta(\Lambda) / \sigma^* \right) \sigma^* \right], & \text{two-sided,} \\ \left(-\infty, \hat{\tau}^* + z_{1-\alpha} \sigma^* + \hat{B}_{\min}^\beta(\Lambda) \right], & \text{one-sided (upper),} \\ \left[\hat{\tau}^* - z_{1-\alpha} \sigma^* - \hat{B}_{\min}^\beta(\Lambda), \infty \right), & \text{one-sided (lower).} \end{cases}$$

This choice of center yields the smallest bias UCB $\hat{B}_{\min}^\beta(\Lambda)$ across the consideration set weights. In this sense, w^* prioritizes the component of robust CI length that comes from the ambiguity or disagreement over Λ . However, the resulting standard deviation σ^* may be larger than the baseline σ_w , so the overall effect on length is ambiguous ex ante.²¹ Note that when $\hat{\eta}_{1-\beta}$ is small enough, it is even possible for CI^* to be shorter than CI_w , and hence shorter than CI_w^* . For example, since the bounded variance class yields $w^* = w_{\text{GLS}}$, and GLS minimizes variance, the corresponding CI^* must be weakly shorter than CI_w in realizations where $\hat{\eta}_{1-\beta} = 0$.

Relationship to Bias-Aware CIs. My robust CIs have a similar structure to the bias-aware CIs advanced in Armstrong and Kolesár (2018, 2020, 2021a,b), but there are important differences in model setup and inferential objective. The bias-aware CIs are designed to provide

²¹ Nevertheless, when the maximum bias is fixed and positive, the bias UCB is asymptotically non-negligible while the variance shrinks to zero with the sample size, which supports choosing w^* as the default center.

valid inference for a fixed target estimand $\lambda'\theta$ subject to bounds on the parameter space for θ , such as bounds on heterogeneity—see [Kwon and Sun \(2025\)](#) for results of this flavor. By contrast, my robust CIs are designed to provide valid inference for a class of target estimands $\{\lambda'\theta: \lambda \in \Lambda\}$ and impose no bound on heterogeneity, opting instead to infer the heterogeneity using $\hat{\eta}_{1-\beta}$. A downside of bounding heterogeneity is the potential for undercoverage when the bound is incorrect, while a downside of inferring heterogeneity is the additional error level β in the uniform coverage statement [\(20\)](#).

5 Uniform Asymptotic Validity

The foregoing results are developed under normally distributed estimates $\hat{\theta} \sim N(\theta, \Sigma)$, a known covariance matrix Σ , known baseline weights w , and a known class of alternative weights Λ . In this section, I establish the uniform asymptotic validity of my procedures under asymptotically normal estimates, consistent covariance matrix estimators, consistent estimators for the baseline weights, and consistent estimators for the class of alternative weights. Sections [5.1](#) and [5.2](#) formalize the asymptotic setup and assumptions. Sections [5.3](#), [5.4](#), and [5.5](#) establish asymptotic results for the class of alternative weights, maximum distance, and robust estimator. Sections [5.6](#) and [5.7](#) establish asymptotic results for the heterogeneity UCB and robust CI. For self-contained practical implementations of my inference procedures, see [Section 6](#).

5.1 Environment

There is a sample of size n drawn from some unknown distribution $P_n \in \mathcal{P}_n$, where $\mathcal{P}_n$ is a class of distributions for samples of size n . Based on this data, the researcher constructs a vector of estimates $\hat{\theta}_n \in \mathbb{R}^K$ and a positive definite covariance matrix estimator $\tilde{\Sigma}_n \in \mathbb{R}^{K \times K}$. Let $\hat{\Sigma}_n = n\tilde{\Sigma}_n$ denote the normalized covariance matrix estimator. Under distribution P_n , the sample objects $(\hat{\theta}_n, \hat{\Sigma}_n)$ have population analogues $(\theta(P_n), \Sigma(P_n))$. The number of groups $K \geq 2$ is fixed while the sample size $n \rightarrow \infty$ grows. I leave the case of growing K to future work.

I assume that the normalized vector of estimates $\sqrt{n}(\hat{\theta}_n - \theta(P_n))$ is uniformly asymptotically normal in the sense that it converges in bounded Lipschitz (BL) metric to the normal distribution $N(0, \Sigma(P_n))$, uniformly over $P_n \in \mathcal{P}_n$. Moreover, I assume that the vector of parameters $\theta(P_n)$ is uniformly bounded (over $P_n \in \mathcal{P}_n$ and $n \geq 1$).

Assumption U1. *For the set $BL_1(\mathbb{R}^K)$ of functions $f: \mathbb{R}^K \rightarrow \mathbb{R}$ that are bounded by one, i.e., $|f(x)| \leq 1$ for all $x \in \mathbb{R}^K$, and have Lipschitz constant bounded by one, i.e., $|f(x) - f(z)| \leq \|x - z\|$ for all $x, z \in \mathbb{R}^K$,*

$$\lim_{n \rightarrow \infty} \sup_{P_n \in \mathcal{P}_n} \sup_{f \in BL_1(\mathbb{R}^K)} \left| \mathbb{E}_{P_n} \left[f(\sqrt{n}(\hat{\theta}_n - \theta(P_n))) \right] - \mathbb{E}[f(Z_n)] \right| = 0, \quad Z_n \sim N(0, \Sigma(P_n)).$$

Moreover, there exists constant $\bar{C}_\theta > 0$ such that $\sup_{P_n \in \mathcal{P}_n} \|\theta(P_n)\| \leq \bar{C}_\theta$ for all n .

Uniform convergence in BL metric is a standard way to formalize uniform convergence in distribution. For example, when the components of $\hat{\theta}_n$ are regression coefficients or sample averages computed over unit-level observations, Assumption U1 follows from bounds on the moments of the observations and bounds on the dependence across observations.

Next, I assume that the normalized sample covariance matrix $\hat{\Sigma}_n = n\tilde{\Sigma}_n$ is uniformly $\sqrt{n}$ -consistent for the population covariance matrix $\Sigma(P_n)$, and that the eigenvalues of $\Sigma(P_n)$ are uniformly bounded above and away from zero.

Assumption U2. *For each $\varepsilon > 0$, there exists constant $C_\varepsilon > 0$ such that*

$$\limsup_{n \rightarrow \infty} \sup_{P_n \in \mathcal{P}_n} \mathbb{P}_{P_n} \left\{ \sqrt{n} \left\| \hat{\Sigma}_n - \Sigma(P_n) \right\| > C_\varepsilon \right\} \leq \varepsilon,$$

where $\|\Sigma\|$ denotes the matrix operator norm. Moreover, there exists constant $\bar{e} > 0$ such that

$$1/\bar{e} \leq \inf_n \inf_{P_n \in \mathcal{P}_n} e_{\min}(\Sigma(P_n)) \leq \sup_n \sup_{P_n \in \mathcal{P}_n} e_{\max}(\Sigma(P_n)) \leq \bar{e},$$

where $e_{\min}(\Sigma)$ and $e_{\max}(\Sigma)$ denote the minimum and maximum eigenvalues of a matrix Σ .

Assumption U2 is a rate condition on the accuracy of covariance matrix estimation. In iid settings, it can be obtained from higher-moment bounds, while in dependent settings it requires moment and dependence conditions strong enough to yield a uniform $\sqrt{n}$ rate. The uniform eigenvalue bounds ensure that $N(0, \Sigma(P_n))$ is uniformly tight and nondegenerate. Note that Assumption U2 implies, for each $\varepsilon > 0$,

$$\lim_{n \rightarrow \infty} \sup_{P_n \in \mathcal{P}_n} \mathbb{P}_{P_n} \left\{ \left\| \hat{\Sigma}_n - \Sigma(P_n) \right\| > \varepsilon \right\} = 0.$$

That is, $\hat{\Sigma}_n$ is uniformly consistent for $\Sigma(P_n)$ under Assumption U2. The stronger $\sqrt{n}$ -consistency is imposed to obtain uniformly valid inference on heterogeneity in Section 5.6.

Remark 3 (Analogy to the Normal Model). The normal model maintains exact normality $\hat{\theta} \sim N(\theta, \Sigma)$ and known Σ . The asymptotic environment instead has normal approximation $\hat{\theta}_n \stackrel{a}{\sim} N(\theta(P_n), \Sigma(P_n)/n)$ and estimated $\tilde{\Sigma}_n = \hat{\Sigma}_n/n \stackrel{a}{\approx} \Sigma(P_n)/n$. Thus, the objects $(\hat{\theta}, \Sigma, \theta)$ in the normal model are analogous to the objects $(\hat{\theta}_n, \tilde{\Sigma}_n, \theta(P_n))$ in the asymptotic environment. $\triangle$

5.2 Baseline and Alternative Weights

The researcher constructs a vector of baseline weights $\hat{w}_n \in \mathcal{W}$ and a nonempty, compact, and convex class of alternative weights $\hat{\Lambda}_n \subseteq \mathcal{W}$. The population analogues are $(w(P_n), \Lambda(P_n))$. The baseline and alternative estimators are $\hat{\tau}_{\hat{w}_n, n} = \hat{w}_n' \hat{\theta}_n$ and $\hat{\tau}_{\lambda, n} = \lambda' \hat{\theta}_n$, where $\lambda \in \hat{\Lambda}_n$. The baseline and alternative estimands are $\tau_{w_n}(P_n) = w(P_n)' \theta(P_n)$ and $\tau_\lambda(P_n) = \lambda' \theta(P_n)$, where $\lambda \in \Lambda(P_n)$. If the baseline and alternative weights are known, then I define $\hat{w}_n = w(P_n)$ and $\hat{\Lambda}_n = \Lambda(P_n)$.

I assume that the sample baseline weights $\hat{w}_n$ are uniformly consistent for the population baseline weights $w(P_n)$. Moreover, I assume that $w(P_n)$ is uniformly bounded—i.e., the baseline weighting scheme cannot place arbitrarily large negative weight on any group.

Assumption U3. For each $\varepsilon > 0$,

$$\lim_{n \rightarrow \infty} \sup_{P_n \in \mathcal{P}_n} \mathbb{P}_{P_n} \{ \|\hat{w}_n - w(P_n)\| > \varepsilon \} = 0.$$

Moreover, there exists constant $\bar{C}_w > 0$ such that $\sup_{P_n \in \mathcal{P}_n} \|w(P_n)\| \leq \bar{C}_w$ for all n .

For example, under Assumption U2, it follows that Assumption U3 holds for GLS weights

$$w_{\text{GLS}}(P_n) = \arg \min_{w \in \mathcal{W}} w' \Sigma(P_n) w = \frac{\Sigma(P_n)^{-1} \mathbf{1}}{\mathbf{1}' \Sigma(P_n)^{-1} \mathbf{1}}. \quad (21)$$

Assumption U3 also holds for weights considered in event studies, such as those in Callaway and Sant'Anna (2021) under uniform versions of their asymptotic linearity assumptions, and those in Sun and Abraham (2021) under uniform versions of their moment conditions.

I suppose that the class of alternative weights $\hat{\Lambda}_n$ depends on the data through statistics $\hat{S}_n \in \mathcal{S}$ taking values in a metric space $(\mathcal{S}, d_{\mathcal{S}})$. In particular, $\hat{\Lambda}_n = \Lambda(\hat{S}_n)$ and $\Lambda(P_n) = \Lambda(S(P_n))$, where $S(P_n)$ is the population analogue of $\hat{S}_n$. I assume the sample statistics $\hat{S}_n$ are uniformly consistent for the population statistics $S(P_n)$. Moreover, $S(P_n)$ is contained in a compact set $\mathbb{S} \subseteq \mathcal{S}$ and $\hat{S}_n$ is contained in $\mathbb{S}$ with probability uniformly approaching one.

Assumption U4. For each $\varepsilon > 0$,

$$\lim_{n \rightarrow \infty} \sup_{P_n \in \mathcal{P}_n} \mathbb{P}_{P_n} \{ d_{\mathcal{S}}(\hat{S}_n, S(P_n)) > \varepsilon \} = 0.$$

Moreover, $S(P_n) \in \mathbb{S}$ for all $P_n \in \mathcal{P}_n$ and n , and

$$\lim_{n \rightarrow \infty} \sup_{P_n \in \mathcal{P}_n} \mathbb{P}_{P_n} \{ \hat{S}_n \notin \mathbb{S} \} = 0.$$

For a bounded variance class with $\hat{S}_n = (\hat{\Sigma}_n, \hat{w}_n)$, Assumption U4 follows from Assumptions U2 and U3, where $\mathbb{S}$ can be defined with constants $2\bar{\varepsilon}$ and $2\bar{C}_w$; see the example below for details. For a truncated simplex class with $\hat{S}_n = \hat{w}_n \in \mathcal{W}_+$, Assumption U4 follows from Assumption U3 and $\mathbb{S} = \mathcal{W}_+$. For a covariate balance class with $\hat{S}_n = (\hat{w}_n, \hat{\mathbf{X}}_n) \in \mathcal{W}_+ \times \mathbb{R}^{K \times M}$, where $\hat{\mathbf{X}}_n$ is an estimated covariate mean matrix, Assumption U4 follows from Assumption U3 and uniform consistency of $\hat{\mathbf{X}}_n$ to a population matrix $\mathbf{X}(P_n)$ with $\|\mathbf{X}(P_n)\| \leq \bar{C}_X$ uniformly for some constant $\bar{C}_X > 0$ and $\min_m \text{sd}(\mathbf{X}_m(P_n)) \geq \varsigma$ uniformly for some constant $\varsigma > 0$, where the corresponding component of $\mathbb{S}$ can be defined with constants $2\bar{C}_X$ and $\varsigma/2$; see the example below for details.

Let $\text{diam}(A) = \sup_{a,b \in A} \|a - b\|$ denote the diameter of a given set $A \subseteq \mathbb{R}^K$. I assume that the class of alternative weights has the following structure.

Assumption U5. *For a compact and convex set $\mathbb{W} \subseteq \mathcal{W}$ with positive diameter $\text{diam}(\mathbb{W}) = \max_{\lambda, w \in \mathbb{W}} \|\lambda - w\| > 0$, there exists a function $g: \mathbb{W} \times \mathbb{S} \rightarrow \mathbb{R}$ with constants $L_g, \delta_g > 0$ such that*

(i) *for each $S \in \mathbb{S}$, the class of alternative weights can be represented as*

$$\Lambda(S) = \{\lambda \in \mathbb{W} : g(\lambda, S) \leq 0\},$$

where the functions $\{\lambda \mapsto g(\lambda, S) : S \in \mathbb{S}\}$ are continuous and convex on $\mathbb{W}$;

(ii) *the functions $\{S \mapsto g(\lambda, S) : \lambda \in \mathbb{W}\}$ are uniformly L_g -Lipschitz on $\mathbb{S}$ in the sense that*

$$\max_{\lambda \in \mathbb{W}} |g(\lambda, S_1) - g(\lambda, S_2)| \leq L_g d_S(S_1, S_2), \quad \forall S_1, S_2 \in \mathbb{S};$$

(iii) *for each $S \in \mathbb{S}$, there exists $\lambda^\circ(S) \in \mathbb{W}$ satisfying the Slater condition*

$$g(\lambda^\circ(S), S) \leq -\delta_g.$$

Condition U5(i) requires that, on the set $\mathbb{S}$ containing the population statistics $S(P_n)$, the classes of interest can be represented as inequality constraints of continuous and convex functions $\lambda \mapsto g(\lambda, S)$ on a compact $\mathbb{W}$, which implies that the population classes $\Lambda(P_n) \subseteq \mathbb{W}$ are compact and convex. Condition U5(ii) requires that the functions $S \mapsto g(\lambda, S)$ in the representation are uniformly Lipschitz on $\mathbb{S}$, which helps in establishing continuity of $S \mapsto \Lambda(S)$ in the Hausdorff metric defined below in Section 5.3. Condition U5(iii) requires that the inequality constraints are uniformly strictly feasible, which helps in establishing continuity of $S \mapsto \Lambda(S)$ and implies that the population classes $\Lambda(P_n)$ are nonempty. Moreover, it implies classes are nondegenerate in the sense that the maximum distance between weights is uniformly bounded away from zero. In particular, the maintained assumptions combined with Lemma 4 yield

$$\inf_n \inf_{P_n \in \mathcal{P}_n} \inf_{w \in \mathcal{W}} \max_{\lambda \in \Lambda(P_n)} \|\lambda - w\|_{\Sigma(P_n)} \gtrsim \text{diam}(\mathbb{W}) > 0.$$

Thus, condition U5(iii) rules out asymptotic regimes in which the class of alternatives collapses to a singleton, which helps in formulating general arguments for the asymptotic validity of the robust CIs. I now map the conditions of Assumption U5 to the example classes of interest.

Example (Bounded Variance, continued). Let $S = (\Sigma, w)$, where Σ is a positive definite matrix and $w \in \mathcal{W}$. Given $r \geq 1$, define the bounded variance class

$$\Lambda_\sigma(\Sigma, w) = \{\lambda \in \mathcal{W} : g(\lambda, \Sigma, w) \leq 0\}, \quad g(\lambda, \Sigma, w) = \sqrt{\lambda' \Sigma \lambda} - r \sqrt{w' \Sigma w}.$$

Take the eigenvalue and norm bounds in Assumptions U2 and U3 as given and define

$$\mathbb{S} = \left\{ (\Sigma, w) : 1/(2\bar{e}) \leq e_{\min}(\Sigma) \leq e_{\max}(\Sigma) \leq 2\bar{e}, w \in \mathcal{W}, \|w\| \leq 2\bar{C}_w \right\}.$$

Choose a finite constant $\bar{C}_\Lambda > 4r\bar{e}\bar{C}_w$ and define

$$\mathbb{W} = \left\{ \lambda \in \mathcal{W} : \|\lambda\| \leq \bar{C}_\Lambda \right\}.$$

If $\lambda \in \Lambda_\sigma(\Sigma, w)$ and $(\Sigma, w) \in \mathbb{S}$, then

$$\|\lambda\|^2 / (2\bar{e}) \leq \lambda' \Sigma \lambda \leq r^2 w' \Sigma w \leq 2r^2 \bar{e} \bar{C}_w^2,$$

which implies $\|\lambda\| \leq 2r\bar{e}\bar{C}_w \leq \bar{C}_\Lambda$. This means $\Lambda_\sigma(\Sigma, w) \subseteq \mathbb{W}$ for each $(\Sigma, w) \in \mathbb{S}$. One can thus consider $g(\lambda, \Sigma, w)$ over $(\lambda, \Sigma, w) \in \mathbb{W} \times \mathbb{S}$, which is continuous and convex in λ , and Lipschitz in (Σ, w) . For the Slater condition, consider

$$w_{\text{GLS}}(\Sigma) = \frac{\Sigma^{-1} \mathbf{1}}{\mathbf{1}' \Sigma^{-1} \mathbf{1}}, \quad \sigma_{\min}(\Sigma) = \min_{w \in \mathcal{W}} \sqrt{w' \Sigma w} = \sqrt{w_{\text{GLS}}(\Sigma)' \Sigma w_{\text{GLS}}(\Sigma)} = \frac{1}{\sqrt{\mathbf{1}' \Sigma^{-1} \mathbf{1}}}.$$

If $r > 1$, then $\lambda^\circ(\Sigma, w) = w_{\text{GLS}}(\Sigma)$ is a Slater point:

$$g(\lambda^\circ(\Sigma, w), \Sigma, w) \leq (1-r)\sigma_{\min}(\Sigma) \leq -\delta_g, \quad \delta_g = (r-1)/\sqrt{2\bar{e}K} > 0.$$

For $r=1$, it suffices to assume that w is bounded away from $\lambda^\circ(\Sigma, w) = w_{\text{GLS}}(\Sigma)$. But since this may be difficult to verify in practice, a simple alternative is to use $r=1+\delta$ for some small fixed $\delta > 0$.²² The same logic applies to the boundary cases in the examples below. $\triangle$

Example (Truncated Simplex, continued). Let $S = w \in \mathcal{W}_+$ and $\mathbb{S} = \mathbb{W} = \mathcal{W}_+$. Given $\epsilon \in [0, 1]$, define the truncated simplex class

$$\Lambda_+(w) = \{ \lambda \in \mathbb{W} : g(\lambda, w) \leq 0 \}, \quad g(\lambda, w) = \max_{k \in \{1, \dots, K\}} ((1-\epsilon)w_k - \lambda_k), \quad w \in \mathbb{S}.$$

The function $g(\lambda, w)$ is continuous and convex in λ and Lipschitz in w . For the Slater condition, consider the mixture $\lambda^\circ(w) = (1-\epsilon)w + \epsilon \mathbf{1}/K$. If $\epsilon > 0$, then

$$g(\lambda^\circ(w), w) = -\delta_g, \quad \delta_g = \epsilon/K > 0.$$

$\triangle$

Example (Covariate Balance, continued). Let $S = (w, \mathbf{X})$, where $\mathbf{X} = (\mathbf{X}_1, \dots, \mathbf{X}_M) \in \mathbb{R}^{K \times M}$

²²If one wishes to allow for data-dependent δ and r , it suffices to embed these parameters into $\mathbb{S}$ and then investigate the conditions of Assumption U5 (and Assumption U4) relative to that new class.

and $w \in \mathcal{W}_+$. Given $\bar{c} \geq 0$, define the balance gap measure

$$c_\lambda(w, \mathbf{X}) = \max_{m \in \{1, \dots, M\}} \frac{|(\lambda - w)' \mathbf{X}_m|}{\text{sd}(\mathbf{X}_m)}.$$

Take $\mathbb{W} = \mathcal{W}_+$ and define the covariate balance class

$$\Lambda_X(w, \mathbf{X}) = \{\lambda \in \mathbb{W} : g(\lambda, w, \mathbf{X}) \leq 0\}, \quad g(\lambda, w, \mathbf{X}) = c_\lambda(w, \mathbf{X}) - \bar{c}.$$

Define the set

$$\mathbb{S} = \left\{ (w, \mathbf{X}) : w \in \mathcal{W}_+, \|\mathbf{X}\| \leq 2\bar{C}_X, \min_{m \in \{1, \dots, M\}} \text{sd}(\mathbf{X}_m) \geq \varsigma/2 \right\}.$$

Over $(w, \mathbf{X}) \in \mathbb{S}$, the standard deviation functions $\mathbf{X}_m \mapsto \text{sd}(\mathbf{X}_m)$ are Lipschitz and bounded away from zero and the functions $\{(w, \mathbf{X}_m) \mapsto |(\lambda - w)' \mathbf{X}_m| : \lambda \in \mathbb{W}\}$ are (uniformly) Lipschitz. Thus, the function $g(\lambda, w, \mathbf{X})$ is Lipschitz in $(w, \mathbf{X})$. It is also continuous and convex in λ . For the Slater condition, consider $\lambda^\circ(w, \mathbf{X}) = w$. If $\bar{c} > 0$, then

$$g(\lambda^\circ(w, \mathbf{X}), w, \mathbf{X}) = -\delta_g, \quad \delta_g = \bar{c} > 0.$$

△

Remark 4 (Mapping to Intersections). Finite intersections of the above classes can be handled by considering the maximum of their corresponding g -functions, yielding

$$\Lambda(S) = \bigcap_{j=1}^J \{\lambda \in \mathbb{W} : g_j(\lambda, S) \leq 0\} = \{\lambda \in \mathbb{W} : g(\lambda, S) \leq 0\}, \quad g(\lambda, S) = \max_{j \in \{1, \dots, J\}} g_j(\lambda, S).$$

If each function $g_j(\lambda, S)$ is continuous and convex in λ and Lipschitz in S , then so is $g(\lambda, S)$. The Slater condition holds when there is a common point $\lambda^\circ(S) \in \mathbb{W}$ satisfying

$$g_j(\lambda^\circ(S), S) \leq -\delta_g, \quad \forall j \in \{1, \dots, J\}.$$

The above observations can be used to map the conditions of Assumption U5 to the bounded variance simplex class and the truncated covariate balance class. For example, the bounded variance simplex $\Lambda_\sigma^+(\Sigma, w) = \{\lambda \in \mathcal{W}_+ : \sqrt{\lambda' \Sigma \lambda} - r \sqrt{w' \Sigma w} \leq 0\}$ can be mapped to $\mathbb{W} = \mathcal{W}_+$, $\mathbb{S} = \{(\Sigma, w) : 1/(2\bar{e}) \leq e_{\min}(\Sigma) \leq e_{\max}(\Sigma) \leq 2\bar{e}, w \in \mathcal{W}_+\}$, $g(\lambda, \Sigma, w) = \sqrt{\lambda' \Sigma \lambda} - r \sqrt{w' \Sigma w}$, and Slater point $\lambda^\circ(\Sigma, w) = \arg\min_{\lambda \in \mathcal{W}_+} \sqrt{\lambda' \Sigma \lambda}$ for $r > 1$ given by the simplex-constrained variance minimizing weights. For $r = 1$, the aforementioned caveats apply. △

Notation for Population Objects. When convenient, I will use the shorthands

$$(\theta_n, \Sigma_n, w_n, \Lambda_n, S_n) = (\theta(P_n), \Sigma(P_n), w(P_n), \Lambda(P_n), S(P_n)),$$

and likewise for other objects that depend on P_n .

5.3 Consistency of the Class of Alternatives

To formulate uniform consistency of the estimated class $\hat{\Lambda}_n$ for the population class Λ_n , I use a standard notion of distance between sets. Given two nonempty sets $A \subseteq \mathbb{R}^K$ and $B \subseteq \mathbb{R}^K$, let $\text{dist}(a, B) = \inf_{b \in B} \|a - b\|$ and $\text{dist}(b, A) = \inf_{a \in A} \|b - a\|$ denote the distance from $a \in A$ to B and from $b \in B$ to A , respectively. The Hausdorff distance between A and B is defined as

$$d_H(A, B) = \max \left\{ \sup_{a \in A} \text{dist}(a, B), \sup_{b \in B} \text{dist}(b, A) \right\}.$$

Under Assumption U5, Lemma 1 shows that $S \mapsto \Lambda(S)$ is Lipschitz continuous in the Hausdorff metric d_H :

$$d_H(\Lambda(S_1), \Lambda(S_2)) \leq \frac{\text{diam}(\mathbb{W})}{\delta_g} L_g d_S(S_1, S_2), \quad \forall S_1, S_2 \in \mathbb{S}.$$

Based on this continuity, the following result shows that $\hat{\Lambda}_n = \Lambda(\hat{S}_n)$ is a uniformly consistent estimator of $\Lambda_n = \Lambda(S_n)$ in the Hausdorff metric.

Proposition U1. *Under Assumptions U4 and U5, for each $\varepsilon > 0$,*

$$\lim_{n \rightarrow \infty} \sup_{P_n \in \mathcal{P}_n} \mathbb{P}_{P_n} \left\{ d_H(\hat{\Lambda}_n, \Lambda_n) > \varepsilon \right\} = 0.$$

Proof. See Appendix D.1. □

5.4 Consistency of the Maximum Distance

The estimated and population maximum distances between weights are

$$\max_{\lambda \in \hat{\Lambda}_n} \|\lambda - \hat{w}_n\|_{\hat{\Sigma}_n} = \max_{\lambda \in \hat{\Lambda}_n} \sqrt{(\lambda - \hat{w}_n)' \hat{\Sigma}_n (\lambda - \hat{w}_n)}, \quad \max_{\lambda \in \Lambda_n} \|\lambda - w_n\|_{\Sigma_n} = \max_{\lambda \in \Lambda_n} \sqrt{(\lambda - w_n)' \Sigma_n (\lambda - w_n)}.$$

Intuitively, $\|\lambda - w_n\|_{\Sigma_n} / \sqrt{n}$ is the standard deviation for $(\lambda - w_n)' \hat{\theta}_n$ under the normal approximation $\hat{\theta}_n \stackrel{a}{\sim} N(\theta_n, \Sigma_n/n)$, while $\|\lambda - \hat{w}_n\|_{\hat{\Sigma}_n} / \sqrt{n} = \|\lambda - \hat{w}_n\|_{\hat{\Sigma}_n}$ is the plug-in standard error. The following result shows that the estimated maximum distance is uniformly consistent for the population maximum distance.

Proposition U2. Under Assumptions U2-U5, for each $\varepsilon > 0$,

$$\lim_{n \rightarrow \infty} \sup_{P_n \in \mathcal{P}_n} \mathbb{P}_{P_n} \left\{ \left| \max_{\lambda \in \hat{\Lambda}_n} \|\lambda - \hat{w}_n\|_{\hat{\Sigma}_n} - \max_{\lambda \in \Lambda_n} \|\lambda - w_n\|_{\Sigma_n} \right| > \varepsilon \right\} = 0.$$

Proof. See Appendix D.2. □

5.5 Optimality of the Robust Estimator

The estimated and population minimax-bias weights are

$$\hat{w}_n^* = \arg \min_{\bar{w} \in \hat{\Lambda}_n} \max_{\lambda \in \hat{\Lambda}_n} \|\lambda - \bar{w}\|_{\hat{\Sigma}_n}, \quad w_n^* = \arg \min_{\bar{w} \in \Lambda_n} \max_{\lambda \in \Lambda_n} \|\lambda - \bar{w}\|_{\Sigma_n}.$$

The following result shows that the minimax-bias weights satisfy Assumption U3.

Proposition U3. Under Assumptions U2, U4, and U5, for each $\varepsilon > 0$,

$$\lim_{n \rightarrow \infty} \sup_{P_n \in \mathcal{P}_n} \mathbb{P}_{P_n} \{ \|\hat{w}_n^* - w_n^*\| > \varepsilon \} = 0.$$

Moreover, there exists constant $\bar{C}_w > 0$ such that $\sup_{P_n \in \mathcal{P}_n} \|w_n^*\| \leq \bar{C}_w$ for all n .

Proof. See Appendix D.3. □

To establish asymptotic optimality of the robust estimator

$$\hat{\tau}_n^* = \hat{\tau}_{\hat{w}_n^*, n} = \hat{w}_n^{*'} \hat{\theta}_n,$$

I first characterize the asymptotic maximum bias of weighted estimators $\hat{\tau}_{\hat{w}_n, n} = \hat{w}_n' \hat{\theta}_n$ under the following asymptotic uniform integrability (UI) condition:

$$\lim_{C \rightarrow \infty} \limsup_{n \rightarrow \infty} \sup_{P_n \in \mathcal{P}_n} \left(\mathbb{E}_{P_n} \left[\|\hat{\theta}_n - \theta_n\|^2 \mathbb{1} \{ \|\hat{\theta}_n - \theta_n\|^2 > C \} \right] \right) = 0. \quad (22)$$

UI condition (22) controls the tails of the squared estimation errors $\|\hat{\theta}_n - \theta_n\|^2$ and $\|\hat{w}_n - w_n\|^2$, ensuring that the corresponding mean squared errors uniformly converge to zero under Assumptions U1 and U3. Similar to the discussions for Assumptions U1 and U3, UI condition (22) follows from moment and dependence bounds on the underlying observations used to construct $(\hat{\theta}_n, \hat{w}_n)$.²³

For the population heterogeneity in parameters $H_n(\theta_n) = \min_{\gamma \in \mathbb{R}} \|\theta_n - \mathbf{1}\gamma\|_{\Sigma_n^{-1}}$, denote the largest heterogeneity over $\mathcal{P}_n$ as $\bar{\eta}(\mathcal{P}_n) = \sup_{P_n \in \mathcal{P}_n} H_n(\theta_n)$, which is uniformly bounded over n

²³By Markov's inequality, UI condition (22) follows from bounds strong enough to yield uniformly bounded $L^{2+\delta}(P_n)$ norms on the estimation errors for some $\delta > 0$; e.g., see Van der Vaart (2000, Example 2.21).

under the eigenvalue bounds on Σ_n and the norm bound on θ_n . For each $w \in \mathcal{W}$, $P_n \in \mathcal{P}_n$, and n , let $P_n^\dagger(w)$ denote a distribution where $\Sigma(P_n^\dagger(w)) = \Sigma_n$, $S(P_n^\dagger(w)) = S_n$, $w(P_n^\dagger(w)) = w$, and

$$\theta(P_n^\dagger(w)) = \mathbf{1}\{\Lambda_n \neq \{w\}\} \bar{\eta}(\mathcal{P}_n) \Sigma_n(\lambda_n^*(w) - w) / \|\lambda_n^*(w) - w\|_{\Sigma_n}, \quad \lambda_n^*(w) \in \arg\max_{\lambda \in \Lambda_n} \|\lambda - w\|_{\Sigma_n},$$

which satisfies $H_n(\theta(P_n^\dagger(w))) = \bar{\eta}(\mathcal{P}_n)$ whenever $\Lambda_n \neq \{w\}$. For example, if $P_n^\dagger(w_n) \in \mathcal{P}_n$ for all $P_n \in \mathcal{P}_n$ and n , then the class of distributions will be rich enough for the bias bound below to be sharp, in analogy to the sharpness of the bias bound from Proposition 5. In this sense, the distributions $P_n^\dagger(w_n)$ are least-favorable. In what follows, I denote $\tau_{\lambda,n} = \lambda' \theta_n$.

Proposition U4. *Under Assumptions U1-U5 and UI condition (22),*

$$\limsup_{n \rightarrow \infty} \left(\sup_{P_n \in \mathcal{P}_n} \max_{\lambda \in \Lambda_n} |\mathbb{E}_{P_n}[\hat{\tau}_{\hat{w}_n,n}] - \tau_{\lambda,n}| \right) \leq \limsup_{n \rightarrow \infty} \left(\bar{\eta}(\mathcal{P}_n) \sup_{P_n \in \mathcal{P}_n} \max_{\lambda \in \Lambda_n} \|\lambda - w_n\|_{\Sigma_n} \right),$$

where the right-hand side is finite. Moreover, equality holds if $P_n^\dagger(w_n) \in \mathcal{P}_n$ for all $P_n \in \mathcal{P}_n$ and n .

Proof. See Appendix D.4. □

The following result establishes asymptotic optimality of the robust estimator, serving as the asymptotic analogue of Proposition 5 from the normal model.

Proposition U5. *Let Assumptions U1, U2, U4, and U5 be satisfied. Moreover, suppose that $\hat{w}_n^*$ satisfies UI condition (22) and $P_n^\dagger(\bar{w}) \in \mathcal{P}_n$ for all $\bar{w} \in \Lambda_n$, $P_n \in \mathcal{P}_n$, and n . Then for any estimator $\hat{\tau}_{\hat{w}_n,n}$ with weights $\hat{w}_n$ satisfying Assumption U3, UI condition (22), and $w_n \in \Lambda_n$ for all $P_n \in \mathcal{P}_n$ and n , the robust estimator $\hat{\tau}_n^* = \hat{w}_n^{*'} \hat{\theta}_n$ has a lower asymptotic maximum bias in the sense that*

$$\limsup_{n \rightarrow \infty} \left(\sup_{P_n \in \mathcal{P}_n} \max_{\lambda \in \Lambda_n} |\mathbb{E}_{P_n}[\hat{\tau}_n^*] - \tau_{\lambda,n}| \right) \leq \limsup_{n \rightarrow \infty} \left(\sup_{P_n \in \mathcal{P}_n} \max_{\lambda \in \Lambda_n} |\mathbb{E}_{P_n}[\hat{\tau}_{\hat{w}_n,n}] - \tau_{\lambda,n}| \right),$$

where the right-hand side is finite.

Proof. See Appendix D.5. □

This optimality result presumes that $\hat{w}_n^*$ satisfies UI condition (22). This holds, for example, when the class of alternatives $\Lambda(S)$ is uniformly bounded over $S \in \mathcal{S}$ —not just $S \in \mathbb{S}$ —as with any subset of the simplex. In other cases, it may be possible to argue the UI condition directly. For example, $\hat{w}_n^*$ reduces to the GLS weights in formula (21) under the bounded variance class, from which it suffices to impose appropriate UI conditions on $\hat{\Sigma}_n$. Note that even without UI conditions, $\hat{w}_n^*$ is still uniformly consistent for the weights w_n^* that minimize the population maximum distance, as established in Proposition U3.

5.6 Validity of the Heterogeneity UCB

The *sample heterogeneity in parameters* is the square root of the residual sum of squares from the GLS regression of the parameters θ_n on a constant $\mathbf{1}$ weighted by $\hat{\Sigma}_n^{-1}$:

$$\hat{H}_n(\theta_n) = \min_{\gamma \in \mathbb{R}} \|\theta_n - \mathbf{1}\gamma\|_{\hat{\Sigma}_n^{-1}} = \sqrt{\theta_n' \hat{Q}_n \theta_n}, \quad \hat{Q}_n = \hat{\Sigma}_n^{-1/2} \hat{A}_n \hat{\Sigma}_n^{-1/2}, \quad \hat{A}_n = I - \frac{\hat{\Sigma}_n^{-1/2} \mathbf{1} \mathbf{1}' \hat{\Sigma}_n^{-1/2}}{\mathbf{1}' \hat{\Sigma}_n^{-1} \mathbf{1}},$$

where $\hat{A}_n$ is the annihilator matrix for $\hat{\Sigma}_n^{-1/2} \mathbf{1}$. The function $\hat{H}_n$ used to measure heterogeneity is random due to $\hat{\Sigma}_n$. A different way to measure heterogeneity is the *population heterogeneity in parameters*, which replaces $\hat{\Sigma}_n$ with Σ_n :

$$H_n(\theta_n) = \min_{\gamma \in \mathbb{R}} \|\theta_n - \mathbf{1}\gamma\|_{\Sigma_n^{-1}} = \sqrt{\theta_n' Q_n \theta_n}, \quad Q_n = \Sigma_n^{-1/2} A_n \Sigma_n^{-1/2}, \quad A_n = I - \frac{\Sigma_n^{-1/2} \mathbf{1} \mathbf{1}' \Sigma_n^{-1/2}}{\mathbf{1}' \Sigma_n^{-1} \mathbf{1}}.$$

The *heterogeneity in estimates* is $\hat{H}_n(\hat{\theta}_n)$. The *heterogeneity UCB* at significance level β is

$$\hat{\eta}_{1-\beta,n} = \frac{\tilde{\eta}_{1-\beta,n}}{\sqrt{n}}, \quad \tilde{\eta}_{1-\beta,n} = \begin{cases} 0, & F_{\chi^2}(n\hat{H}_n^2(\hat{\theta}_n); 0) \leq \beta, \\ F_{\chi^2}^{-1}(n\hat{H}_n^2(\hat{\theta}_n); \beta), & F_{\chi^2}(n\hat{H}_n^2(\hat{\theta}_n); 0) > \beta. \end{cases}$$

The following result establishes uniform asymptotic validity of $\hat{\eta}_{1-\beta,n}$ for inference on the sample heterogeneity in parameters $\hat{H}_n(\theta_n)$, which will turn out to be sufficient for establishing uniform asymptotic validity of the robust CI—see Remark 5 below for why I consider inference on the sample measure $\hat{H}_n(\theta_n)$ instead of the population measure $H_n(\theta_n)$.

Proposition U6. *Under Assumptions U1 and U2,*

$$\lim_{n \rightarrow \infty} \sup_{P_n \in \mathcal{P}_n} \left| \mathbb{P}_{P_n} \left\{ \hat{H}_n(\theta_n) \leq \hat{\eta}_{1-\beta,n} \right\} - (1-\beta) \right| \mathbf{1}\{H_n(\theta_n) > 0\} = 0. \quad (23)$$

Proof. See Appendix D.6. □

Note that when $H_n(\theta_n) = 0$, the UCB $\hat{\eta}_{1-\beta,n} \geq 0$ always covers $\hat{H}_n(\theta_n)$, so multiplication by the indicator $\mathbf{1}\{H_n(\theta_n) > 0\}$ restricts attention to sequences where heterogeneity is nontrivial, in analogy to quantile-unbiasedness criterion (14) from the normal model. Thus, (23) implies

$$\liminf_{n \rightarrow \infty} \inf_{P_n \in \mathcal{P}_n} \mathbb{P}_{P_n} \left\{ \hat{H}_n(\theta_n) \leq \hat{\eta}_{1-\beta,n} \right\} \geq 1 - \beta.$$

In this sense, $\hat{\eta}_{1-\beta,n}$ is a uniformly asymptotically valid UCB, in analogy to criterion (13).

Remark 5 (Sample versus Population Heterogeneity). Consider a sequence where the normalized population heterogeneity diverges: $\sqrt{n}H_n(\theta_n) \rightarrow \infty$. Given the norm bound on θ_n and

the $\sqrt{n}$ -consistency of $\hat{\Sigma}_n$, this implies $\sqrt{n}\hat{H}_n(\theta_n) \xrightarrow{p} \infty$. To see what this means for validity of $\hat{\eta}_{1-\beta,n}$, consider noncentral chi-squared asymptotics for the heterogeneity in estimates $\hat{H}_n(\hat{\theta}_n)$ relative to the sample heterogeneity in parameters $\hat{H}_n(\theta_n)$:

$$\frac{\sqrt{n}(\hat{H}_n^2(\hat{\theta}_n) - \hat{H}_n^2(\theta_n))}{2\hat{H}_n(\theta_n)} = \frac{\hat{Z}_n' \hat{Q}_n \hat{Z}_n}{2\sqrt{n}\hat{H}_n(\theta_n)} + \frac{(\hat{A}_n \hat{\Sigma}_n^{-1/2} \theta_n)'}{\|\hat{A}_n \hat{\Sigma}_n^{-1/2} \theta_n\|} \hat{\Sigma}_n^{-1/2} \hat{Z}_n, \quad \hat{Z}_n = \sqrt{n}(\hat{\theta}_n - \theta_n). \quad (24)$$

In this asymptotic regime, the first term is $o_p(1)$ while the second term converges in distribution to a standard normal. The asymptotic validity of $\hat{\eta}_{1-\beta,n}$ follows from expression (24) combined with bounds from [Seri \(2015\)](#) that imply the uniform convergence of a noncentral chi-squared CDF with diverging noncentrality parameter to a normal CDF; see [Appendix D.6](#) for details. By comparison, the population heterogeneity analogue of (24) yields

$$\frac{\sqrt{n}(\hat{H}_n^2(\hat{\theta}_n) - H_n^2(\theta_n))}{2H_n(\theta_n)} = \frac{\hat{Z}_n' \hat{Q}_n \hat{Z}_n}{2\sqrt{n}H_n(\theta_n)} + \frac{(\hat{A}_n \hat{\Sigma}_n^{-1/2} \theta_n)'}{\|A_n \Sigma_n^{-1/2} \theta_n\|} \hat{\Sigma}_n^{-1/2} \hat{Z}_n + \frac{\sqrt{n}(\hat{H}_n^2(\theta_n) - H_n^2(\theta_n))}{2H_n(\theta_n)},$$

where the first two terms behave like before, but now there is a third term where

$$\left| \frac{\sqrt{n}(\hat{H}_n^2(\theta_n) - H_n^2(\theta_n))}{2H_n(\theta_n)} \right| \lesssim_p \frac{\sqrt{n} \|\hat{\Sigma}_n - \Sigma_n\|}{H_n(\theta_n)} \lesssim_p \frac{1}{H_n(\theta_n)},$$

which need not be $o_p(1)$. [Proposition U6](#) avoids this term altogether. $\triangle$

Remark 6 (Analogy to Heterogeneity in the Normal Model). As noted in [Remark 3](#), objects $(\hat{\theta}, \Sigma, \theta)$ in the normal model are analogous to objects $(\hat{\theta}_n, \tilde{\Sigma}_n, \theta_n)$ in the asymptotic environment, where $\tilde{\Sigma}_n = \hat{\Sigma}_n/n$. The corresponding GLS weighting matrices are Σ^{-1} and $\tilde{\Sigma}_n^{-1} = n\hat{\Sigma}_n^{-1}$. Thus, using $\hat{\eta}_{1-\beta}$ for inference on $H(\theta)$ in the normal model is analogous to using $\tilde{\eta}_{1-\beta,n} = \sqrt{n}\hat{\eta}_{1-\beta,n}$ for inference on $\tilde{H}_n(\theta_n) = \sqrt{n}\hat{H}_n(\theta_n)$ in the asymptotic environment. $\triangle$

5.7 Validity of the Robust CI

I now establish the asymptotic validity of robust CIs centered at weight vectors $\hat{w}_n$ that satisfy [Assumption U3](#), which include the minimax-bias weights $\hat{w}_n^*$ in view of [Proposition U3](#).

For centering weights $\hat{w}_n$, the bias UCB is

$$\hat{B}_{\hat{w}_n,n}^\beta(\hat{\Lambda}_n) = \hat{\eta}_{1-\beta,n} \max_{\lambda \in \hat{\Lambda}_n} \|\lambda - \hat{w}_n\|_{\hat{\Sigma}_n} = \tilde{\eta}_{1-\beta,n} \max_{\lambda \in \hat{\Lambda}_n} \|\lambda - \hat{w}_n\|_{\tilde{\Sigma}_n}, \quad \tilde{\Sigma}_n = \hat{\Sigma}_n/n.$$

For standard error $\tilde{\sigma}_{\hat{w}_n,n} = \sqrt{\hat{w}_n' \tilde{\Sigma}_n \hat{w}_n} = \sqrt{\hat{w}_n' \hat{\Sigma}_n \hat{w}_n} / \sqrt{n} = \hat{\sigma}_{\hat{w}_n,n} / \sqrt{n}$, the robust CI is

$$CI_{\hat{w}_n,n}^* = \begin{cases} \left[\hat{\tau}_{\hat{w}_n,n} \pm \text{cv}_{1-\alpha} \left(\hat{B}_{\hat{w}_n,n}^\beta(\hat{\Lambda}_n) / \tilde{\sigma}_{\hat{w}_n,n} \right) \tilde{\sigma}_{\hat{w}_n,n} \right], & \text{two-sided,} \\ \left(-\infty, \hat{\tau}_{\hat{w}_n,n} + z_{1-\alpha} \tilde{\sigma}_{\hat{w}_n,n} + \hat{B}_{\hat{w}_n,n}^\beta(\hat{\Lambda}_n) \right], & \text{one-sided (upper),} \\ \left[\hat{\tau}_{\hat{w}_n,n} - z_{1-\alpha} \tilde{\sigma}_{\hat{w}_n,n} - \hat{B}_{\hat{w}_n,n}^\beta(\hat{\Lambda}_n), \infty \right), & \text{one-sided (lower).} \end{cases}$$

The validity of $CI_{\hat{w}_n,n}^*$ depends on its coverage for weights in the population class $\Lambda_n \subseteq \mathbb{W}$, where $\mathbb{W} \subseteq \mathcal{W}$ is the compact convex set with positive diameter from Assumption U5. The uniform consistency in Proposition U1 shows that the estimated class $\hat{\Lambda}_n$ is close to the population class Λ_n with high probability as $n \rightarrow \infty$. However, this is not the same as $\hat{\Lambda}_n$ containing the alternatives from Λ_n with high probability. To see this, recall that Assumption U5 yields the inequality constraint representations

$$\hat{\Lambda}_n = \Lambda(\hat{S}_n) = \{ \lambda \in \mathbb{W} : g(\lambda, \hat{S}_n) \leq 0 \}, \quad \Lambda_n = \Lambda(S_n) = \{ \lambda \in \mathbb{W} : g(\lambda, S_n) \leq 0 \},$$

where the former holds with probability uniformly approaching one under Assumption U4. The issue is at the boundary of the inequality constraint, where small estimation errors can lead to exclusion, $g(\lambda, \hat{S}_n) > 0$, of population alternatives where $g(\lambda, S_n) = 0$. Since membership is discontinuous at the boundary, uniform consistency of $\hat{S}_n$ to S_n generally does not ensure that $\hat{\Lambda}_n$ contains the alternatives in Λ_n . However, as I show in examples further below, there will typically exist a subclass $\Lambda_0(S) \subseteq \Lambda(S)$ such that, for $\Lambda_{0,n} = \Lambda_0(S_n)$,

$$\lim_{n \rightarrow \infty} \inf_{P_n \in \mathcal{P}_n} \inf_{\lambda \in \Lambda_{0,n}} \mathbb{P}_{P_n} \{ \lambda \in \hat{\Lambda}_n \} = 1. \quad (25)$$

For subclasses $\Lambda_{0,n}$ satisfying containment condition (25), the robust CI constructed with $\hat{\Lambda}_n$ provides uniformly asymptotically valid coverage for the target alternatives $\lambda \in \Lambda_{0,n}$.

Proposition U7. *Let Assumptions U1-U5 be satisfied and suppose that $\Lambda_{0,n} \subseteq \Lambda_n$ satisfies containment condition (25). Then*

$$\liminf_{n \rightarrow \infty} \inf_{P_n \in \mathcal{P}_n} \inf_{\lambda \in \Lambda_{0,n}} \mathbb{P}_{P_n} \{ \tau_{\lambda,n} \in CI_{\hat{w}_n,n}^* \} \geq 1 - (\alpha + \beta). \quad (26)$$

Proof. See Appendix D.7. □

The asymptotic coverage in (26) is analogous to the coverage in (20) from the normal model, but stated for the subclass $\Lambda_{0,n}$ rather than the population class Λ_n . This suggests viewing the class $\Lambda_{0,n}$ in Proposition U7 as a target class, $\Lambda_n \supseteq \Lambda_{0,n}$ as an enlargement of the target class, and $\hat{\Lambda}_n$ as an estimator of the enlargement Λ_n . In other words, to ensure asymptotic coverage

for target alternatives $\Lambda_{0,n}$, it suffices to use an enlarged estimator $\hat{\Lambda}_n \supseteq \hat{\Lambda}_{0,n}$ to construct the robust CI rather than using the sample analogue $\hat{\Lambda}_{0,n}$ of the target. For example, if the class $\hat{\Lambda}_{0,n} = \Lambda(\hat{S}_n; \kappa_0)$ increases in a scalar parameter κ , this suggests using $\hat{\Lambda}_n = \Lambda(\hat{S}_n; \kappa)$ with a larger $\kappa > \kappa_0$. Importantly, unlike the enlarged Λ_n , I do not require the target $\Lambda_{0,n}$ to satisfy the conditions of Assumption U5, so long as it satisfies containment condition (25) relative to Λ_n . For example, a target bounded variance class with $r_0 = 1$ may not satisfy the Slater condition, but nevertheless satisfy containment condition (25) relative to an enlargement with $r > 1$ that does satisfy the Slater condition; see the examples below.

Remark 7 (Caveats Regarding Class Estimation Error). If there is no class estimation error (i.e., $\hat{\Lambda}_n = \Lambda_n$), then containment condition (25) holds for $\Lambda_{0,n} = \Lambda_n$ and Proposition U7 applies with $\Lambda_n = \Lambda_{0,n}$. However, Proposition U7 still assumes that Λ_n satisfies Assumption U5. Thus, even when there is no class estimation error, it can still be useful to apply Proposition U7 to $\Lambda_n \supseteq \Lambda_{0,n}$ when the target class $\Lambda_{0,n}$ may violate the conditions of Assumption U5, as with the aforementioned bounded variance class with $r = 1$ and the Slater condition. $\triangle$

To establish containment condition (25), a practical approach is to show that the inequality constraint from Λ_n is uniformly slack when evaluated over $\Lambda_{0,n}$. In particular, it suffices to show the existence of a constant $\nu > 0$ such that for all $P_n \in \mathcal{P}_n$ and n ,

$$\sup_{\lambda \in \Lambda_{0,n}} g(\lambda, S_n) \leq -\nu. \quad (27)$$

By Lemma 5, slack condition (27) implies containment condition (25) under Assumptions U4 and U5. I now show how to derive $\Lambda_{0,n}$ and ν for the benchmark classes of interest.

Example (Bounded Variance, continued). For $S_n = (\Sigma_n, w_n)$ and $r_0 \geq 1$, the target class is

$$\Lambda_{0,n} = \{\lambda \in \mathcal{W} : g(\lambda, \Sigma_n, w_n; r_0) \leq 0\} = \left\{ \lambda \in \mathcal{W} : \sqrt{\lambda' \Sigma_n \lambda} \leq r_0 \sqrt{w_n' \Sigma_n w_n} \right\}.$$

For $r > r_0$, the enlarged class is

$$\Lambda_n = \{\lambda \in \mathcal{W} : g(\lambda, \Sigma_n, w_n; r) \leq 0\} = \left\{ \lambda \in \mathcal{W} : \sqrt{\lambda' \Sigma_n \lambda} \leq r \sqrt{w_n' \Sigma_n w_n} \right\}.$$

Observe that

$$\sup_{\lambda \in \Lambda_{0,n}} g(\lambda, \Sigma_n, w_n; r) = \sup_{\lambda \in \Lambda_{0,n}} \sqrt{\lambda' \Sigma_n \lambda} - r \sqrt{w_n' \Sigma_n w_n} \leq -(r - r_0) \sqrt{w_n' \Sigma_n w_n} \leq 0.$$

Since $e_{\min}(\Sigma_n) \geq 1/\bar{e}$ and $\|w_n\| \geq 1/\sqrt{K}$, slack condition (27) is satisfied:

$$\sup_{\lambda \in \Lambda_{0,n}} g(\lambda, \Sigma_n, w_n; r) \leq -\nu, \quad \nu = (r - r_0)/\sqrt{\bar{e}K} > 0.$$

By Proposition U7, a robust CI constructed with estimated class

$$\hat{\Lambda}_n = \left\{ \lambda \in \mathcal{W} : g(\lambda, \hat{\Sigma}_n, \hat{w}_n; r) \leq 0 \right\} = \left\{ \lambda \in \mathcal{W} : \sqrt{\lambda' \hat{\Sigma}_n \lambda} \leq r \sqrt{\hat{w}_n' \hat{\Sigma}_n \hat{w}_n} \right\}$$

provides uniformly asymptotically valid coverage for $\lambda \in \Lambda_{0,n}$. These derivations also apply to the bounded variance simplex by replacing $\mathcal{W}$ with $\mathcal{W}_+$ and restricting to $\hat{w}_n, w_n \in \mathcal{W}_+$. $\triangle$

Example (Truncated Simplex, continued). For $S_n = w_n \in \mathcal{W}_+$ and $\epsilon_0 \in [0, 1)$, the target class is

$$\Lambda_{0,n} = \{ \lambda \in \mathcal{W}_+ : g(\lambda, w_n; \epsilon_0) \leq 0 \} = \{ \lambda \in \mathcal{W}_+ : \lambda \geq (1 - \epsilon_0) w_n \}.$$

For $\epsilon \in (\epsilon_0, 1]$, the enlarged class is

$$\Lambda_n = \{ \lambda \in \mathcal{W}_+ : g(\lambda, w_n; \epsilon) \leq 0 \} = \{ \lambda \in \mathcal{W}_+ : \lambda \geq (1 - \epsilon) w_n \}.$$

Observe that

$$\sup_{\lambda \in \Lambda_{0,n}} g(\lambda, w_n; \epsilon) = \max_{k \in \{1, \dots, K\}} \left((1 - \epsilon) w_{n,k} - \inf_{\lambda \in \Lambda_{0,n}} \lambda_k \right) \leq -(\epsilon - \epsilon_0) \min_{k \in \{1, \dots, K\}} w_{n,k}.$$

Suppose that $w_{n,k}$ is uniformly bounded away from zero:

$$\inf_n \inf_{P_n \in \mathcal{P}_n} w_{n,k} \geq w_{\min} > 0, \quad \forall k \in \{1, \dots, K\}.$$

This holds, for instance, for the equal weights vector $w_n = \mathbf{1}/K$. Given this uniform lower bound on $w_{n,k}$, slack condition (27) is satisfied:

$$\sup_{\lambda \in \Lambda_{0,n}} g(\lambda, w_n; \epsilon) \leq -\nu, \quad \nu = (\epsilon - \epsilon_0) w_{\min} > 0.$$

By Proposition U7, a robust CI constructed with estimated class

$$\hat{\Lambda}_n = \{ \lambda \in \mathcal{W}_+ : g(\lambda, \hat{w}_n; \epsilon) \leq 0 \} = \{ \lambda \in \mathcal{W}_+ : \lambda \geq (1 - \epsilon) \hat{w}_n \}$$

provides uniformly asymptotically valid coverage for $\lambda \in \Lambda_{0,n}$. $\triangle$

Example (Covariate Balance, continued). For $S_n = (w_n, \mathbf{X}_n) \in \mathcal{W}_+ \times \mathbb{R}^{K \times M}$ and $\bar{c}_0 \geq 0$, the target class is

$$\Lambda_{0,n} = \{ \lambda \in \mathcal{W}_+ : g(\lambda, w_n, \mathbf{X}_n; \bar{c}_0) \leq 0 \} = \{ \lambda \in \mathcal{W}_+ : c_\lambda(w_n, \mathbf{X}_n) \leq \bar{c}_0 \}.$$

For $\bar{c} > \bar{c}_0$, the enlarged class is

$$\Lambda_n = \{\lambda \in \mathcal{W}_+ : g(\lambda, w_n, \mathbf{X}_n; \bar{c}) \leq 0\} = \{\lambda \in \mathcal{W}_+ : c_\lambda(w_n, \mathbf{X}_n) \leq \bar{c}\}.$$

Slack condition (27) is satisfied:

$$\sup_{\lambda \in \Lambda_{0,n}} g(\lambda, w_n, \mathbf{X}_n; \bar{c}) = -\nu, \quad \nu = \bar{c} - \bar{c}_0 > 0.$$

By Proposition U7, a robust CI constructed with estimated class

$$\hat{\Lambda}_n = \{\lambda \in \mathcal{W}_+ : g(\lambda, \hat{w}_n, \hat{\mathbf{X}}_n; \bar{c}) \leq 0\} = \{\lambda \in \mathcal{W}_+ : c_\lambda(\hat{w}_n, \hat{\mathbf{X}}_n) \leq \bar{c}\}$$

provides uniformly asymptotically valid coverage for $\lambda \in \Lambda_{0,n}$. $\triangle$

Remark 8 (Reporting Convention for Boundary Alternatives). The takeaway from the above examples and discussions is that asymptotic coverage for boundary alternatives requires some care. For instance, a target bounded variance simplex class

$$\Lambda_{0,n} = \left\{ \lambda \in \mathcal{W}_+ : \sqrt{\lambda' \Sigma_n \lambda} \leq r_0 \sqrt{w_n' \Sigma_n w_n} \right\}, \quad r_0 \geq 1,$$

may contain alternative weights that lie on the boundary of the variance constraint: $\sqrt{\lambda' \Sigma_n \lambda} = r_0 \sqrt{w_n' \Sigma_n w_n}$. Proposition U7 says that a sufficient avenue for asymptotic coverage of the entire $\Lambda_{0,n}$ is to use a robust CI constructed with an enlarged estimated class

$$\hat{\Lambda}_n = \left\{ \lambda \in \mathcal{W}_+ : \sqrt{\lambda' \hat{\Sigma}_n \lambda} \leq r \sqrt{\hat{w}_n' \hat{\Sigma}_n \hat{w}_n} \right\}, \quad r = r_0 + \delta, \quad \delta > 0.$$

The asymptotic coverage in Proposition U7 applies to the r_0 target class, with coverage obtained through a small enlargement, $r = r_0 + \delta$, of the estimated class. However, in the implementation and empirical applications, I will suppress this caveat and talk about coverage as if $\delta = 0$. The reason is that, in practice, replacing $r = r_0$ by $r = r_0 + \delta$ for $\delta = 0.0001$ leaves the reported CIs unchanged at the displayed precision. $\triangle$

6 Practical Implementation

Motivated by the asymptotic results in Section 5, but suppressing the sample size index n for conciseness, I consider asymptotically normal estimates $\hat{\theta} \overset{a}{\sim} N(\theta, \Sigma/n)$, consistent covariance matrix estimator $\tilde{\Sigma} = \hat{\Sigma}/n \overset{a}{\approx} \Sigma/n$, consistent baseline weights estimator $\hat{w} \overset{a}{\approx} w$, and classes of alternative weights $\hat{\Lambda} = \Lambda(\hat{S})$ characterized by consistently estimated statistics $\hat{S} \overset{a}{\approx} S$: e.g., $\hat{S} = (\hat{w}, \hat{\Sigma})$ for bounded variance class, $\hat{S} = \hat{w}$ for truncated simplex class, and $\hat{S} = (\hat{w}, \hat{\mathbf{X}})$ for covariate balance class with estimated covariate mean matrix $\hat{\mathbf{X}}$.

Section 6.1 summarizes my robust inference procedures in terms of the above objects and discusses specification choices for $\hat{\Lambda}$. Sections 6.2 and 6.3 discuss practical implementations of these procedures in the context of event studies and multisite experiments, respectively.

6.1 Feasible Inference Procedures

The GLS regression of $\hat{\theta}$ on a constant $\mathbf{1}$ weighted by $\tilde{\Sigma}^{-1}$ yields

$$\hat{\tau}_{\text{GLS}} = \frac{\mathbf{1}'\tilde{\Sigma}^{-1}\hat{\theta}}{\mathbf{1}'\tilde{\Sigma}^{-1}\mathbf{1}}, \quad \tilde{\sigma}_{\text{GLS}} = \frac{1}{\sqrt{\mathbf{1}'\tilde{\Sigma}^{-1}\mathbf{1}}}, \quad \tilde{H}^2(\hat{\theta}) = (\hat{\theta} - \mathbf{1}\hat{\tau}_{\text{GLS}})'\tilde{\Sigma}^{-1}(\hat{\theta} - \mathbf{1}\hat{\tau}_{\text{GLS}}), \quad (28)$$

where $\hat{\tau}_{\text{GLS}}$ is the GLS estimator, $\tilde{\sigma}_{\text{GLS}}$ is the GLS standard error, $\tilde{H}^2(\hat{\theta})$ is the GLS residual sum of squares, which yields heterogeneity in estimates $\tilde{H}(\hat{\theta})$ with explicit formula

$$\tilde{H}(\hat{\theta}) = \sqrt{\hat{\theta}'\tilde{Q}\hat{\theta}}, \quad \tilde{Q} = \tilde{\Sigma}^{-1/2}\tilde{A}\tilde{\Sigma}^{-1/2}, \quad \tilde{A} = I - \frac{\tilde{\Sigma}^{-1/2}\mathbf{1}\mathbf{1}'\tilde{\Sigma}^{-1/2}}{\mathbf{1}'\tilde{\Sigma}^{-1}\mathbf{1}}. \quad (29)$$

Following the steps of Section 3.2, the heterogeneity UCB is

$$\tilde{\eta}_{1-\beta} = \begin{cases} 0, & F_{\chi^2}(\tilde{H}^2(\hat{\theta}); 0) \leq \beta, \\ F_{\chi^2}^{-1}(\tilde{H}^2(\hat{\theta}); \beta), & F_{\chi^2}(\tilde{H}^2(\hat{\theta}); 0) > \beta. \end{cases} \quad (30)$$

Following the steps of Section 4, the minimax-bias weights and bias UCB are

$$\hat{w}^* = \underset{\bar{w} \in \hat{\Lambda}}{\operatorname{argminmax}} \|\lambda - \bar{w}\|_{\tilde{\Sigma}}, \quad \hat{B}_{\min}^{\beta}(\hat{\Lambda}) = \tilde{\eta}_{1-\beta} \max_{\lambda \in \hat{\Lambda}} \|\lambda - \hat{w}^*\|_{\tilde{\Sigma}}. \quad (31)$$

For standard error $\tilde{\sigma}^* = \sqrt{(\hat{w}^*)'\tilde{\Sigma}\hat{w}^*}$, the robust estimator and corresponding robust CI are

$$\hat{\tau}^* = (\hat{w}^*)'\hat{\theta}, \quad CI^* = \begin{cases} [\hat{\tau}^* \pm \text{cv}_{1-\alpha}(\hat{B}_{\min}^{\beta}(\hat{\Lambda})/\tilde{\sigma}^*)\tilde{\sigma}^*], & \text{two-sided}, \\ (-\infty, \hat{\tau}^* + z_{1-\alpha}\tilde{\sigma}^* + \hat{B}_{\min}^{\beta}(\hat{\Lambda})], & \text{one-sided (upper)}, \\ [\hat{\tau}^* - z_{1-\alpha}\tilde{\sigma}^* - \hat{B}_{\min}^{\beta}(\hat{\Lambda}), \infty), & \text{one-sided (lower)}, \end{cases} \quad (32)$$

where $\text{cv}_{1-\alpha}(b)$ can be computed as the square root of the $(1-\alpha)$ -quantile of the noncentral chi-squared distribution with one degree of freedom and noncentrality parameter b^2 .

Recipe 1. Implementation of Inference Procedures

1. Construct $\tilde{\eta}_{1-\beta}$ as in (30), based on (29) or (28).
2. Construct $\hat{w}^*$ and $\hat{B}_{\min}^{\beta}(\hat{\Lambda})$ as in (31).

3. Construct $\hat{\tau}^*$ and CI^* as in (32).

4. Report $(\hat{\tau}^*, CI^*)$.

Recipe 1 summarizes the implementation of my inference procedures for inputs given by (i) the estimates and covariance matrix estimator $(\hat{\theta}, \tilde{\Sigma})$, for which I describe typical constructions in Section 6.2 for event studies and Section 6.3 for multisite experiments; (ii) the significance levels α and β , which by default can be set to $\alpha = \beta = 0.05$; and (iii) the class of alternatives $\hat{\Lambda}$.

In some cases there may be ambiguity over which class of alternatives to use. In particular, the class $\hat{\Lambda} = \hat{\Lambda}(\kappa)$ may be indexed by a scalar parameter $\kappa \in \mathbb{R}$ over which there is ambiguity. To address this, one can execute Recipe 1 for a range of plausible κ . Alternatively, assuming that $\hat{\Lambda}(\kappa)$ is increasing in κ , one can compute the *breakdown value* κ^* , which is the smallest value of κ at which the robust $CI^*(\kappa)$ includes a given threshold value τ_0 , such as $\tau_0 = 0$.

Example (Bounded Variance, continued). From the derivations in (18), it follows that Recipe 1 is entirely determined by the GLS output (28). In particular, for a lower one-sided CI^* and baseline weights $\hat{w}$, Recipe 1 under $\hat{\Lambda}_\sigma(r) = \{\lambda \in \mathcal{W} : \tilde{\sigma}_\lambda \leq r \tilde{\sigma}_{\hat{w}}\}$ for standard error ratio bound r and standard error $\tilde{\sigma}_{\hat{w}} = \sqrt{\hat{w}' \tilde{\Sigma} \hat{w}}$ yields

$$\hat{\tau}^* = \hat{\tau}_{\text{GLS}}, \quad CI^*(r) = \left[\hat{\tau}_{\text{GLS}} - z_{1-\alpha} \tilde{\sigma}_{\text{GLS}} - \tilde{\eta}_{1-\beta} \sqrt{r^2 \tilde{\sigma}_{\hat{w}}^2 - \tilde{\sigma}_{\text{GLS}}^2}, \infty \right).$$

The breakdown value for a given threshold τ_0 is

$$r^* = \frac{\tilde{\sigma}_{\text{GLS}}}{\tilde{\sigma}_{\hat{w}}} \sqrt{\left(\frac{\max\{T_{\text{GLS}}(\tau_0) - z_{1-\alpha}, 0\}}{\tilde{\eta}_{1-\beta}} \right)^2 + 1}, \quad T_{\text{GLS}}(\tau_0) = \frac{\hat{\tau}_{\text{GLS}} - \tau_0}{\tilde{\sigma}_{\text{GLS}}}.$$

$CI^*(r)$ includes τ_0 when $r \geq r^*$. If $\tau_0 = 0$, this happens when the baseline t -statistic $T_{\text{GLS}}(0)$ is sufficiently large relative to the degree of inferred heterogeneity $\tilde{\eta}_{1-\beta}$. $\triangle$

Example (Truncated Simplex, continued). Given baseline weights $\hat{w}$, one can execute Recipe 1 with $\hat{\Lambda}_+(\epsilon) = \{\lambda \in \mathcal{W}_+ : \lambda \geq (1-\epsilon)\hat{w}\} = \{(1-\epsilon)\hat{w} + \epsilon\lambda : \lambda \in \mathcal{W}_+\}$ for some range of truncation parameters ϵ , or compute breakdown values ϵ^* . To interpret magnitudes of ϵ , consider the equal weights baseline $\hat{w} = w_{\text{EW}} = \mathbf{1}/K$. In this case, ϵ is the maximum possible discrepancy in weights $|\lambda_k - \lambda_{k'}|$ between two groups k and k' . Moreover, note that the maximum amount of weight that can be reallocated across groups when moving from w_{EW} to $\lambda \in \Lambda_+(\epsilon)$ is

$$\left((1-\epsilon) \frac{1}{K} + \epsilon \right) - \frac{1}{K} = \frac{K-1}{K} \epsilon.$$

On the other hand, the amount of weight that is reallocated when moving *all* the weight from $J \in \{1, \dots, K-1\}$ of the groups to the other $K-J$ groups (i.e., a leave- J -out procedure) is equal to J/K . Equating these two reallocation quantities yields the ϵ calibration

$$\frac{K-1}{K}\epsilon = \frac{J}{K} \iff \epsilon = \frac{J}{K-1}. \quad (33)$$

Thus, in the EW baseline, choosing ϵ according to (33) allows for the same maximal reallocation of weight as simply dropping J of the groups. In this sense, ϵ is a rescaled leave- J -out magnitude: $(K-1)\epsilon$ is the corresponding number of groups dropped, and $(K-1)\epsilon/K$ is the corresponding fraction of groups dropped. This calibration allows one to maintain the mathematical structure of $\Lambda_+(\epsilon)$ while casting the logic of ϵ -departures in terms of the potentially more familiar logic of leave- J -out procedures. $\triangle$

Example (Covariate Balance, continued). Given covariates $\hat{\mathbf{X}}$, one can execute Recipe 1 with $\hat{\Lambda}_X(\bar{c}) = \{\lambda \in \mathcal{W}_+ : c_\lambda(\hat{\mathbf{X}}) \leq \bar{c}\}$ for some range of balance gap bounds $\bar{c}$, or compute breakdown values $\bar{c}^*$. To interpret magnitudes of $\bar{c}$, consider the group-level balance gaps

$$c_k(\hat{\mathbf{X}}) = \max_{m \in \{1, \dots, M\}} \frac{|\hat{\mathbf{X}}_{m,k} - \hat{w}' \hat{\mathbf{X}}_m|}{\text{sd}(\hat{\mathbf{X}}_m)}. \quad (34)$$

Intuitively, $c_k(\hat{\mathbf{X}})$ is the balance gap that arises from a population with covariate means equal to those of group k . If one assumes that readers are interested in balance gaps within the range of $\{c_k(\hat{\mathbf{X}})\}_{k=1}^K$, then the deciles $(\bar{c}_d)_{d=1}^9$ of the group-level balance gaps provide one simple way to determine relevant magnitudes of $\bar{c}$. For example, $\bar{c} = \bar{c}_5$ allows one to assess robustness to populations that have covariate profiles consistent with the median group-level balance gap. $\triangle$

6.2 Implementation in Event Studies

Context. Following the exposition in Section 2, an event study has units i treated at different time periods t , allowing one to estimate the dynamic effects of treatment on outcomes Y_{it} . In this setting, two key sources of heterogeneity are (i) across treatment cohorts G_i , which denote the periods of initial treatment for units i ; and (ii) across *event times* $\ell \geq 0$, which denote the number of periods since initial treatment. Given the $\text{ATT}_{g,t}$ object in (1), the average treatment effect ℓ periods after initial treatment for cohort $G_i = g$ is

$$\text{ATT}_{g,g+\ell} = E[Y_{i,g+\ell}(g) - Y_{i,g+\ell}(\infty) | G_i = g].$$

Under event-study assumptions, [Sun and Abraham \(2021\)](#) show that event-time ℓ coefficients from conventional TWFE regression specifications recover estimands of the form

$$\text{TWFE}_\ell(\tilde{w}_\ell) = \sum_{g,\ell'} \tilde{w}_{\ell,(g,\ell')} \text{ATT}_{g,g+\ell'},$$

which allow treatment effects from different event times $\ell' \neq \ell$ and negative weights in $\tilde{w}_\ell$. As surveyed in [Roth et al. \(2023\)](#), the recent event studies literature proposes heterogeneity-robust (HR) approaches that, for weights $w_\ell > 0$, recover event-time estimands of the form

$$\text{ATT}_\ell(w_\ell) = \sum_g w_{\ell,g} \text{ATT}_{g,g+\ell},$$

preventing negative weighting and excluding treatment effects from different event times.

For each target event time ℓ , a researcher may report conventional estimators and CIs for both a TWFE specification and one of the proposed HR approaches and examine the stability of results. There are two ways to map this setup to the notation of [Recipe 1](#).

- Consider TWFE as the baseline and estimate $(\tilde{w}_{\ell,(g,\ell')}, \text{ATT}_{g,g+\ell'})$ across (g, ℓ') .
- Consider an HR approach as the baseline and estimate $(w_{\ell,g}, \text{ATT}_{g,g+\ell})$ across g .

While it is possible to do the former (see [Sun and Abraham \(2021, equations \(13\) and \(27\)\)](#)), the latter has the advantage of automatically shutting off the $\ell' \neq \ell$ channels and reducing the number of objects to be estimated. In fact, the ingredients required for implementing [Recipe 1](#) are available in many applications of HR approaches. For example, the [Sun and Abraham \(2021\)](#) procedure already requires estimation of $(w_{\ell,g}, \text{ATT}_{g,g+\ell})$, and so [Recipe 1](#) can be readily executed in applications that were already using the [Sun and Abraham \(2021\)](#) procedure, or any other procedure that first estimates ATT objects and their corresponding standard errors, such as the doubly robust procedure of [Callaway and Sant'Anna \(2021\)](#), or stacked DiD procedures ([Wing, Freedman and Hollingsworth, 2024](#)).

Implementation. I now use k to index the groups g . For each target event time ℓ , denote the number of available cohorts by K_ℓ . Take an existing HR approach for which one can estimate baseline weights $\hat{w}_\ell \stackrel{a}{\approx} w_\ell$ and ATT objects

$$\hat{\theta}_\ell = \left(\widehat{\text{ATT}}_{k,k+\ell} \right)_{k=1}^{K_\ell} \stackrel{a}{\sim} N(\theta_\ell, \tilde{\Sigma}_\ell),$$

where $\tilde{\Sigma}_\ell$ is a covariance matrix estimator. Let $\tilde{\sigma}_{\hat{w},\ell} = \sqrt{\hat{w}'_\ell \tilde{\Sigma}_\ell \hat{w}_\ell}$ denote the plug-in standard error for $\hat{\tau}_{\hat{w},\ell} = \hat{w}'_\ell \hat{\theta}_\ell$. Following [equation \(6\)](#), consider the bounded variance simplex class

$$\hat{\Lambda}_{\sigma,\ell}^+(r) = \{\lambda \in \mathcal{W}_{+,\ell} : \tilde{\sigma}_{\lambda,\ell} \leq r \tilde{\sigma}_{\hat{w},\ell}\}, \quad \mathcal{W}_{+,\ell} = \{w \in \mathbb{R}^{K_\ell} : \mathbf{1}'w = 1, w \geq 0\}.$$

Executing Recipe 1 with $\hat{\Lambda}_{\sigma,\ell}^+(r)$ produces $(\hat{\tau}_\ell^*(r), CI_\ell^*(r))$. For the choice of standard error ratio bound r , I propose the following implementation.

1. Choose $r = 1$, allowing one to assess robustness to the class of nonnegative weights that yield standard errors no larger than the HR baseline $\tilde{\sigma}_{\hat{w},\ell}$. This is the minimal r for which the baseline weights are included in $\hat{\Lambda}_{\sigma,\ell}^+(r)$.
2. Report breakdown value r_ℓ^* , which gives the smallest value of r at which the robust $CI_\ell^*(r)$ includes a threshold value τ_0 (e.g., $\tau_0 = 0$). If r_ℓ^* is large, then the robust CI only includes τ_0 when it attempts to cover estimands that cannot be estimated precisely.

I apply this implementation to [Lakdawala, Nakasone and Kho \(2023\)](#) in Section 7.1.

6.3 Implementation in Multisite Experiments

Following the exposition in Section 2, a multisite experiment has sites k containing units i over which ATEs are estimated:

$$\hat{\theta} = \left(\widehat{\text{ATE}}_k \right)_{k=1}^K \stackrel{a}{\sim} N(\theta, \tilde{\Sigma}),$$

where $\tilde{\Sigma}$ is a covariance matrix estimator. Given independent sites, it suffices to obtain site-level objects $(\hat{\theta}_k, \tilde{\sigma}_k, \hat{E}_k[X_i])$, where estimated variances $\tilde{\sigma}_k^2$ can be collected into a diagonal matrix $\tilde{\Sigma}$ and estimated covariate means $\hat{E}_k[X_i]$ can be collected into a matrix $\hat{\mathbf{X}} = (\hat{E}_1[X_i], \dots, \hat{E}_K[X_i])'$. Given microdata on units i from the sites, one can directly estimate $\hat{E}_k[X_i]$ with sample means and $(\hat{\theta}_k, \tilde{\sigma}_k)$ via regression. For example, letting $\mathbf{1}\{i \in k\}$ be an indicator for membership in site k , one can estimate the (no-intercept) saturated model

$$Y_i = \sum_{k=1}^K \gamma_{0k} \mathbf{1}\{i \in k\} + \sum_{k=1}^K \theta_k D_i \mathbf{1}\{i \in k\} + u_i$$

with OLS to obtain ATE estimates $\hat{\theta}_k$ and heteroskedasticity-robust standard errors $\tilde{\sigma}_k$.

The baseline weights $\hat{w} \in \mathcal{W}_+$ represent a particular population/distribution over the set of sites. To model departures from the baseline population, one can use the truncated simplex class $\hat{\Lambda}_+(\epsilon)$ defined in (4) or the covariate balance class $\hat{\Lambda}_X(\bar{c})$ defined in (5).

- Departures in $\hat{\Lambda}_+(\epsilon)$ ensure that groups k represented in the baseline population remain represented in the alternative population: $\hat{w}_k > 0 \implies \lambda_k > 0$ for each $\lambda \in \hat{\Lambda}_+(\epsilon)$.
- Departures in $\hat{\Lambda}_X(\bar{c})$ incorporate information about group-level covariates $\hat{\mathbf{X}}$.

To obtain benefits from both approaches, I propose the following implementation based on the truncated covariate balance class $\hat{\Lambda}_X^\epsilon(\bar{c})$ defined in (7).

1. Choose the equal weights baseline $\hat{w} = w_{EW} = \mathbf{1}/K$, which represents an empirical distribution over the sites, under which all groups k will be represented in the alternative populations within $\hat{\Lambda}_X^\epsilon(\bar{c}) = \Lambda_+(\epsilon) \cap \hat{\Lambda}_X(\bar{c})$.
2. Formulate a choice of truncation parameter ϵ . Following (33), one can interpret ϵ in terms of the leave- J -out calibration.
3. Execute Recipe 1 with $\hat{\Lambda}_X^\epsilon(\bar{c})$ for a range of balance gap bounds $\bar{c}$. Following (34), one can use the site-level balance gap deciles $(\bar{c}_d)_{d=1}^9$.

I apply this implementation to Project STAR in Section 7.2.

7 Empirical Applications

I now apply my procedures to the event study in Lakdawala, Nakasone and Kho (2023) and the multisite experiment in Project STAR, following the implementations described in Sections 6.2 and 6.3. Below I use significance levels $\alpha = \beta = 0.05$ and hence confidence levels $1 - \alpha = 95\%$ for conventional CIs, $1 - \beta = 95\%$ for heterogeneity UCBs, and $1 - (\alpha + \beta) = 90\%$ for robust CIs.

7.1 Application to Lakdawala, Nakasone and Kho (2023)

Lakdawala, Nakasone and Kho (2023)—henceforth LNK—study the rollout of school-based internet access in Peruvian public primary schools. The estimating sample contains schools that either gained internet access between 2007 and 2020 or remained unconnected by 2020, with second grade math and reading test scores observed from 2007–2016. LNK use a staggered-adoption event-study design: after absorbing school fixed effects, calendar-time controls, and time-varying school and student controls, the event-time coefficients compare changes in scores around a school’s installation date to contemporaneous changes among schools that are not yet connected or remain unconnected. LNK’s event-study figures report school-level dynamic effects for successive cohorts of second graders, with scores standardized within calendar year and effects interpreted relative to the year before internet installation.

LNK’s Figure 2 shows a delayed achievement response. The TWFE event-study estimates, replicated in Figure 1, are small immediately after internet installation and grow over time. In LNK’s page 234 description of the medium-run effect, “By year 5, scores are 0.110 standard deviations higher for math” and “0.063 standard deviations higher for reading” relative to the year before installation. These dynamics are central to the interpretation of the intervention because LNK argue that schools require time to adapt to new internet access. The empirical target is therefore a dynamic path of treatment effects.

Weighting Issues. Following Section 6.2, let cohorts k index the first year in which a school gains internet access and event times ℓ denote the time relative to first internet access. LNK’s pages 241–42 discuss weighting issues: “Recent work . . . raises some important issues with using

two-way fixed effects estimation when there is staggered timing of treatment. We show that these issues do not explain our main estimates by illustrating robustness to the estimator proposed in Sun and Abraham (2021) . . . The patterns and magnitudes of these estimates are very similar to our main estimates.” The Sun and Abraham (2021)—henceforth SA—estimator corresponds to weights $\hat{w}_{SA,\ell}$. LNK’s Figure A.12, replicated in Figure 1, is therefore a robustness check that replaces the TWFE event-study aggregation with the SA aggregation in order to assess whether results are robust to alternative weights.²⁴ This suggests a sharper robustness question: are the substantive conclusions robust only to the SA weights, or to a broader class of plausible event-time weights? To answer this question, I compute robust estimators and CIs for classes of alternative weights that nest the SA weights.

Measuring Heterogeneity. Table 1 reports the estimated event-time heterogeneity UCBs $\tilde{\eta}_\ell$ used to construct the robust CIs. For event times with $\tilde{\eta}_\ell = 0$, the heterogeneity adjustment in the respective robust CI drops out. The last row reports K_ℓ , the total number of cohorts k at each event time ℓ . Across all but one event time, the heterogeneity UCB is lower for the math cohort-level estimates than for the reading cohort-level estimates, previewing the greater degree of robustness for the math results.

Table 1: Heterogeneity UCB $\tilde{\eta}_\ell$ and number of event-time cohorts K_ℓ

Quantity	$\ell=0$	$\ell=1$	$\ell=2$	$\ell=3$	$\ell=4$	$\ell=5$	$\ell=6$
$\tilde{\eta}_\ell$ for Math	0.00	2.77	2.41	2.93	3.29	2.70	0.00
$\tilde{\eta}_\ell$ for Reading	2.58	2.69	4.42	3.54	5.59	4.53	2.79
Number of cohorts K_ℓ	9	8	7	6	5	4	3

Bounded Variance Simplex Class. Consider the bounded variance simplex

$$\hat{\Lambda}_{\sigma,\ell}^+(r) = \{\lambda \in \mathcal{W}_{+,\ell} : \tilde{\sigma}_{\lambda,\ell} \leq r \tilde{\sigma}_{SA,\ell}\}, \quad \tilde{\sigma}_{SA,\ell} = \sqrt{\hat{w}'_{SA,\ell} \tilde{\Sigma}_\ell \hat{w}_{SA,\ell}},$$

with $r = 1$, which corresponds to the class of simplex-weighted estimands that can be estimated at least as precisely as the SA estimand. For each event time ℓ , I compute the robust $CI_\ell^*(1)$ centered at the robust estimator $\hat{\tau}_\ell^*(1)$. Figure 1 presents three series: the TWFE CI centered at the TWFE estimate, the SA CI centered at the SA estimate, and the robust CI centered at the robust

²⁴LNK use a bootstrap procedure to construct the SA CIs. Their replication files estimate cohort-specific event-study coefficients using the same controls and fixed effects as the main specification, drop cohorts not identified relative to the omitted period, aggregate cohort-event coefficients using the SA event-time weights, and bootstrap the aggregated coefficients using 500 replications stratified by first access year and clustered by school. The CIs reported here instead use the plug-in standard error computed from the estimated covariance matrix $\tilde{\Sigma}_\ell$ for the cohort-event coefficient vector $\hat{\theta}_\ell$, treating the SA weights as fixed. Thus, while I replicate the same SA estimates, my reported SA CIs are not identical to LNK’s bootstrap SA CIs.

estimate. For math, the robust CIs are positive for event times 1 through 6 and only include zero at event time 0. At event time 5, the robust estimate is 0.099 with CI $[0.055, 0.143]$, close to both the TWFE and SA estimates. Thus, LNK’s medium-run math conclusion is robust to alternative simplex weights that yield estimators at least as precise as the SA estimator. For reading, the robust estimates remain close to the SA path, but the conclusion that results are significantly positive is somewhat more sensitive. The robust CIs are positive for event times 1 through 4 and 6, while event time 5 barely includes zero: the robust estimate is 0.042 with CI $[-0.0002, 0.084]$.

Table 2 presents breakdown values r_ℓ^* , which for each ℓ gives the smallest value of r at which the robust $CI_\ell^*(r)$ includes zero for the class $\hat{\Lambda}_{\sigma,\ell}^+(r)$. A value of 1.60, for example, means that the robust CI begins to include zero once the allowable $\tilde{\sigma}$ bound is expanded to about 1.60 times the SA $\tilde{\sigma}$. A dash means that zero is not reached even when considering the simplex without the variance constraint.²⁵ The breakdown values reinforce the visual evidence. For math, event times 2 through 5 tolerate substantial increases in r before the robust CI includes zero, with $r_5^* = 1.60$ at the headline year-5 event time. Event time 6 is even stronger in this sense: zero is not reached over the full simplex. For reading, by contrast, the breakdown values are close to one for event times 1 through 5, and event time 5 already includes zero at $r = 1$. Overall, these results support the baseline math conclusions strongly, but not as strongly for the reading conclusions.

Table 2: Breakdown r_ℓ^* for $0 \in CI_\ell^*(r)$ under $\hat{\Lambda}_{\sigma,\ell}^+(r)$

Outcome	$\ell=0$	$\ell=1$	$\ell=2$	$\ell=3$	$\ell=4$	$\ell=5$	$\ell=6$
r_ℓ^* for Math	0.98	1.18	1.69	1.63	1.54	1.60	–
r_ℓ^* for Reading	0.98	1.06	1.12	1.06	1.02	1.00	1.35

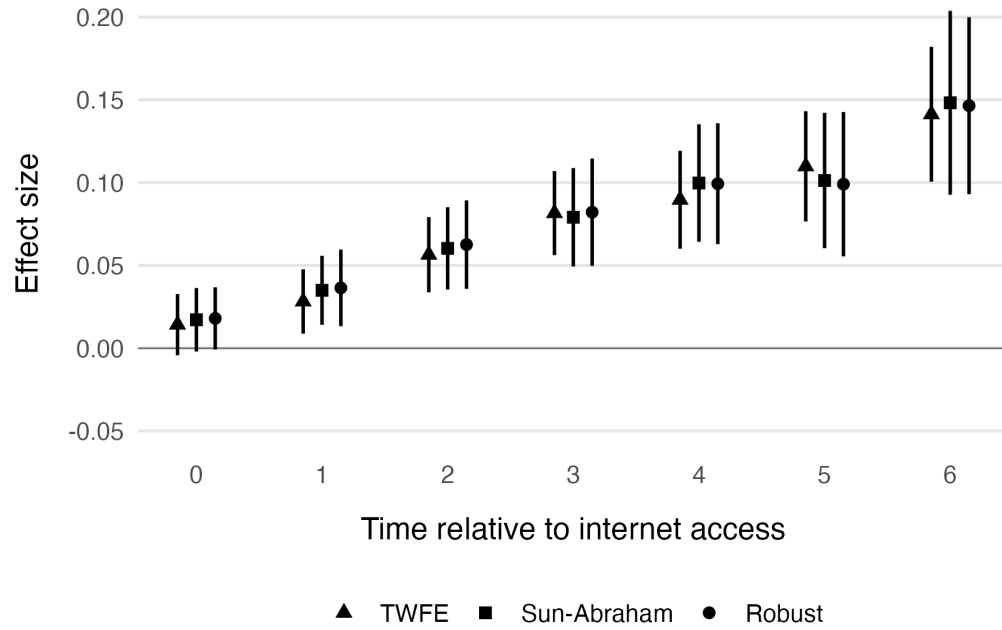
7.2 Application to Project STAR

The Tennessee Project STAR (Student/Teacher Achievement Ratio) experiment randomized students in seventy-nine Tennessee public elementary schools to classes with different numbers of students to estimate the causal effects of class size on test scores (Achilles et al., 2008; Krueger, 1999). The experimental arms were (i) regular-sized classes (20-25 students), small classes (13-17 students), and regular-sized classes with a teacher aide. To define the sample for my analysis, I drop observations in the teacher-aide arm and focus on the effect of being assigned to a small class versus a regular-sized class. Following Goldsmith-Pinkham, Hull and Kolesár (2024), I focus on the effects of treatment in kindergarten, which mitigates attrition and other complications.

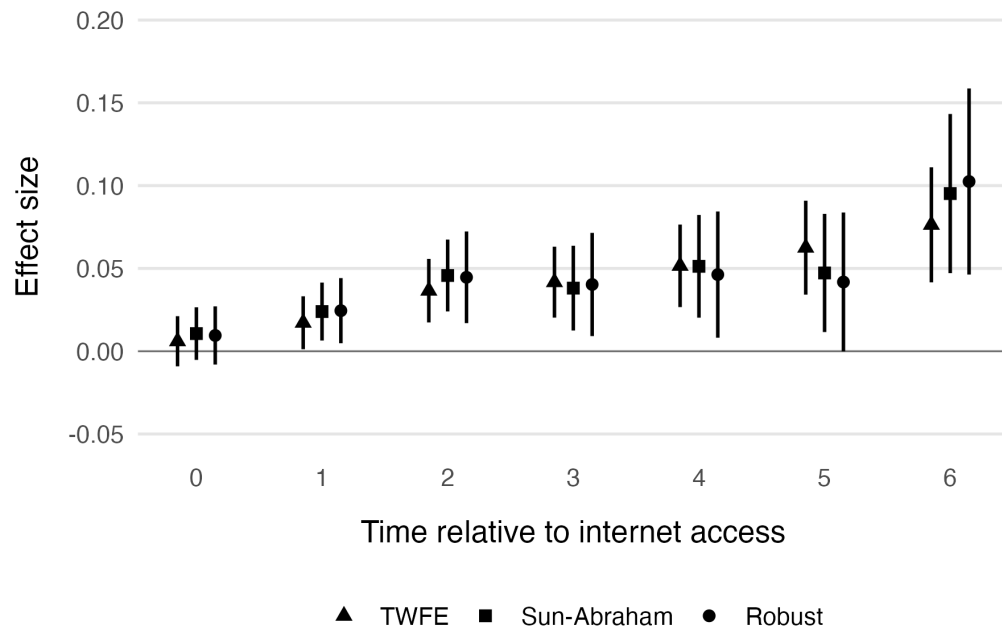
²⁵If $\tilde{\eta}_\ell = 0$, I still compute breakdown values using the same definition: report the smallest feasible value of the robustness parameter at which the robust CI contains zero, and report “–” if zero is never reached over the full class. This convention implies, for example, that math at $\ell=0$ has $r_0^* < 1$ because zero is already included at the smallest feasible bounded-variance simplex class, while math at $\ell=6$ has no finite breakdown value because zero is not reached even when the variance constraint is no longer binding over the simplex.

Figure 1: Impact of Internet Access on Test Scores

Panel A. Standardized Math Scores



Panel B. Standardized Reading Scores



This leaves $K = 78$ schools and a total of $n = 3783$ students. Following [Schanzenbach \(2006\)](#), I define the outcome as the average of a student’s math and reading scores on the SAT, where the scores are demeaned and standardized relative to control-group scores. This puts the outcome into effect size units that have benchmarks in causal studies of preK-12 education interventions on student achievement: <0.05 = small, 0.05 - 0.20 = medium, >0.20 = large ([Kraft, 2020](#)).

In my notation, θ_k is the ATE of being assigned to a small class on the test score outcome at school k , $\hat{\theta}_k$ is the estimated ATE at school k from regressing the outcome on school-specific treatment indicators and fixed effects, $\tilde{\Sigma}$ is the heteroskedasticity-robust covariance matrix from this regression, and the baseline weights are equal weights (EW) $w = w_{EW} = \mathbf{1}/K$ so that $\tau_w(\theta) = \tau_{EW}(\theta)$ is the ATE in the empirical distribution of Project STAR schools. The baseline results establish medium-sized positive effects of small class size on test scores. In particular, the EW estimate is $\hat{\tau}_{EW} = 0.188$ with standard error $\tilde{\sigma}_{EW} = 0.028$, yielding conventional EW CI $[0.133, 0.242]$. My baseline estimate is consistent with the one in [Schanzenbach \(2006, Table 4\)](#). However, my standard error is about 0.01 smaller. In particular, [Schanzenbach \(2006\)](#) clusters her standard errors at the classroom level. By contrast, I follow [Goldsmith-Pinkham, Hull and Kolesár \(2024\)](#) and use heteroskedasticity-robust standard errors since the randomization of students to classrooms was at the individual level.²⁶

Weighting Issues. As reviewed by [Schanzenbach \(2006\)](#), Project STAR has been used in many policy discussions. However, [Schanzenbach \(2006, page 207\)](#) notes external validity concerns:

- “A few aspects of the sample may limit the validity of generalizing the study to other settings. In order to be eligible to participate in the program, schools were required to have a minimum-size cohort of 57 students . . . As a result, the schools that participated were about 25% larger, on average, than other Tennessee schools”;
- “Because of requirements imposed by the legislature for geographic diversity, schools in inner cities were overrepresented, and the students included were more economically disadvantaged and more likely to be African American”;
- “Finally, average school spending in Tennessee was about three-fourths of the nationwide average, and teachers were less likely to have a master’s degree.”

Thus, the Project STAR results may not be representative of Tennessee (or U.S.) schools more broadly. Formally, the EW estimand averages over the empirical distribution of STAR schools, while alternative estimands of interest may place unequal weights across the STAR schools. This suggests a robustness question: are the baseline STAR results robust to alternative choices of weights/distributions? To answer this question, I compute robust estimators and CIs for classes of alternative weights that nest the EW weights, following implementations from [Section 6.3](#).

²⁶This is consistent with the when-to-cluster framework of [Abadie et al. \(2023\)](#).

Measuring Heterogeneity. The estimated heterogeneity UCB is $\tilde{\eta}=16.214$, which is over twice as large as the baseline t -statistic $\hat{\tau}_{EW}/\tilde{\sigma}_{EW}=6.714$. This suggests a meaningful degree of heterogeneity in the school-level ATE estimates, which previews the sensitivity of baseline results to the choice of weights.

Truncated Simplex Class. I first consider the truncated simplex class

$$\Lambda_+(\epsilon) = \{\lambda \in \mathcal{W}_+ : \lambda \geq (1-\epsilon)/K\} = \{(1-\epsilon)w_{EW} + \epsilon\lambda : \lambda \in \mathcal{W}_+\},$$

which collects the set of ϵ -deviations in weight distribution from the STAR empirical distribution. For each $\tau_0 \in \{0, 0.05\}$, Table 3 presents the breakdown value ϵ^* , which is the smallest value of ϵ at which τ_0 is included in the robust $CI^*(\epsilon)$ centered at robust estimate $\hat{\tau}^*(\epsilon)$ with standard error $\tilde{\sigma}(\epsilon^*)$. The breakdown value for $\tau_0 = 0.05$ is $\epsilon^* = 0.0169$, while the breakdown value for $\tau_0 = 0$ is $\epsilon^* = 0.0261$. These are both small departures from EW: even though the corresponding robust estimates are larger than the baseline EW estimate $\hat{\tau}_{EW} = 0.188$, the robust CIs indicate that baseline inferences are robust to only small departures from EW.

Table 3: Breakdown ϵ^* for $\tau_0 \in CI^*(\epsilon)$ under $\Lambda_+(\epsilon)$

Target	ϵ^*	$\hat{\tau}^*(\epsilon^*)$	$\tilde{\sigma}(\epsilon^*)$	$CI^*(\epsilon^*)$
$\tau_0 = 0.05$	0.0169	0.191	0.029	[0.050, 0.332]
$\tau_0 = 0$	0.0261	0.193	0.029	[-0.000, 0.386]

Truncated Covariate Balance Class. I intersect the truncated simplex with the covariate balance class, yielding

$$\hat{\Lambda}_X^\epsilon(\bar{c}) = \{\lambda \in \Lambda_+(\epsilon) : c_\lambda(\hat{\mathbf{X}}) \leq \bar{c}\}, \quad c_\lambda(\hat{\mathbf{X}}) = \max_{m \in \{1, \dots, M\}} \frac{|\lambda' \hat{\mathbf{X}}_m - \mu(\hat{\mathbf{X}}_m)|}{\text{sd}(\hat{\mathbf{X}}_m)}, \quad \mu(\hat{\mathbf{X}}_m) = w'_{EW} \hat{\mathbf{X}}_m.$$

Table 4 reports the Project STAR school-level covariates $\hat{\mathbf{X}}_m$ used in my analysis.

I fix $\epsilon = \epsilon^*(0) = 0.0261$ at the truncated simplex breakdown value for $\tau_0 = 0$ from the previous analysis and consider balance gap bounds $(\bar{c}_d)_{d=1}^9$ equal to the deciles of the school-level balance gaps $c_k(\hat{\mathbf{X}})$. Intuitively, $c_k(\hat{\mathbf{X}})$ is the balance gap from a population that has covariate means $\hat{\mathbf{X}}_{m,k}$ equal to those of school k . The deciles $\bar{c}_d$ presented in Table 5 therefore represent typical values for the balance gaps.

Figure 2 presents robust intervals $CI^*(\bar{c}_d)$ centered at robust estimates $\hat{\tau}^*(\bar{c}_d)$ under $\epsilon^*(0)$. Across every balance gap decile $\bar{c}_d$, the robust CIs include $\tau_0 = 0.05$. Thus, while the covariate balance constraints allow the robust CIs to exclude zero relative to imposing only the truncated simplex constraint, this is not enough to allow one to robustly conclude that effects are

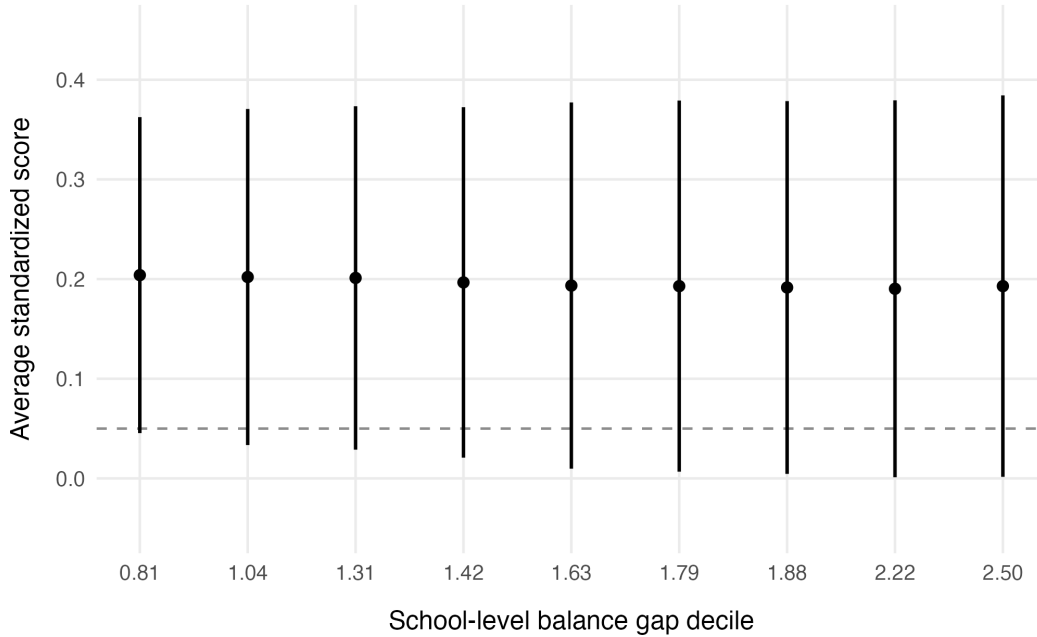
Table 4: Means and standard deviations of Project STAR covariates

Covariate	$\mu(\hat{\mathbf{X}}_m)$	$\text{sd}(\hat{\mathbf{X}}_m)$
Number of students	80.65	25.67
Fraction students female	0.49	0.06
Fraction students White or Asian	0.71	0.39
Average teacher years of experience	9.32	2.67
Fraction students on FRPL	0.47	0.27
Fraction teachers White	0.87	0.25
Fraction teachers with master's degree or above	0.36	0.27

Table 5: Deciles of STAR school-level balance gaps

Quantity	$d=1$	$d=2$	$d=3$	$d=4$	$d=5$	$d=6$	$d=7$	$d=8$	$d=9$
Gap decile $\bar{c}_d$	0.805	1.040	1.310	1.419	1.635	1.792	1.880	2.222	2.497

medium-sized under $\epsilon^*(0)$ -departures from EW. Overall, these results provide further support for the conclusion that baseline STAR inferences are robust to only small departures from EW. There is too much heterogeneity for the conclusions of a baseline CI to be (uniformly) informative for alternative weights that depart meaningfully from the baseline weights.

Figure 2: Robust CIs for $\hat{\Lambda}_X^\epsilon(\bar{c}_d)$ at $\epsilon^*(0)$ 

8 Conclusion

Weighted estimands are ubiquitous in empirical research. The choice of weights is not merely a statistical nuisance: it shapes the empirical and policy relevance of a researcher’s weighted estimand. Conventional inference procedures can therefore be limited when readers entertain alternative weights, leading to ambiguity and disagreement over the choice of weights.

This paper develops inference procedures that account for these issues. The key geometric observation is that the difference between two weighted estimands can be controlled by two objects: the heterogeneity of the underlying parameters and the distance between the corresponding weights. This decomposition leads to two practical tools. First, I derive minimax-bias weights that minimize the maximum distance to a class of alternative weights. The resulting robust estimator provides a natural default when researchers face ambiguity over the choice of weights. Second, I construct robust confidence intervals by combining a confidence bound for the heterogeneity in parameters with the maximum distance between the centering weights and the alternative weights. These intervals provide coverage guarantees for classes of weighted estimands rather than for a single baseline estimand.

My framework accommodates many empirically relevant classes of alternatives. The bounded variance class captures concerns about comparing the baseline estimand to alternatives that can be estimated with reasonable precision. The truncated simplex class captures departures from a baseline target population while ensuring that all represented groups continue to receive positive weight. The covariate balance class captures external validity concerns by restricting attention to alternative populations whose covariate means remain close to the baseline. These classes can also be intersected, allowing researchers to combine precision, positivity, and covariate-balance restrictions in a single robustness analysis.

The empirical applications illustrate that robustness to the choice of weights can be context-specific. For the event study in [Lakdawala, Nakasone and Kho \(2023\)](#), conclusions are broadly robust to classes of weights that yield reasonable estimation precision. For the experiment in Project STAR, however, conclusions are sensitive even to small departures from the baseline weighting scheme. These findings highlight the value of reporting robust estimators and intervals alongside conventional ones: doing so makes explicit whether the conclusions reached under a researcher’s baseline weights are informative for alternative weighted estimands that may be of interest to readers.

References

- Abadie, Alberto, Alexis Diamond, and Jens Hainmueller.** 2010. “Synthetic control methods for comparative case studies: Estimating the effect of California’s tobacco control program.” *Journal of the American statistical Association*, 105(490): 493–505.
- Abadie, Alberto, and Javier Gardeazabal.** 2003. “The economic costs of conflict: A case study of the Basque Country.” *American economic review*, 93(1): 113–132.
- Abadie, Alberto, Jiaying Gu, and Shu Shen.** 2024. “Instrumental variable estimation with first-stage heterogeneity.” *Journal of Econometrics*, 240(2): 105425.
- Abadie, Alberto, Joshua Angrist, Brigham Frandsen, and Jörn-Steffen Pischke.** 2025. “Harvesting Differences-in-Differences and Event-Study Evidence.” National Bureau of Economic Research.
- Abadie, Alberto, Susan Athey, Guido W Imbens, and Jeffrey M Wooldridge.** 2023. “When should you adjust standard errors for clustering?” *The Quarterly Journal of Economics*, 138(1): 1–35.
- Achilles, Charles M, Helen Pate Bain, Fred Bellott, Jayne Boyd-Zaharias, Jeremy Finn, John Folger, John Johnston, and Elizabeth Word.** 2008. “Tennessee’s student teacher achievement ratio (STAR) project.” *Harvard Dataverse*, 1: 2008.
- Adusumilli, Karun.** 2026. “You’ve Got to be Efficient: Ambiguity, Misspecification and Variational Preferences.” *arXiv preprint arXiv:2604.05327*.
- Allcott, Hunt.** 2015. “Site selection bias in program evaluation.” *The Quarterly journal of economics*, 130(3): 1117–1165.
- Andrews, Isaiah, and Jiafeng Chen.** 2025. “Certified Decisions.” *arXiv preprint arXiv:2502.17830*.
- Angrist, Joshua.** 1998. “Estimating the Labor Market Impact of Voluntary Military Service Using Social Security Data on Military Applicants.” *Econometrica*, 66(2): 249–288.
- Angrist, Joshua, and Ivan Fernandez-Val.** 2010. “Extrapolate-ing: External validity and overidentification in the late framework.” National Bureau of Economic Research.
- Armstrong, Timothy B, and Michal Kolesár.** 2018. “Optimal inference in a class of regression models.” *Econometrica*, 86(2): 655–683.
- Armstrong, Timothy B, and Michal Kolesár.** 2020. “Simple and honest confidence intervals in nonparametric regression.” *Quantitative Economics*, 11(1): 1–39.

- Armstrong, Timothy B, and Michal Kolesár.** 2021a. “Finite-Sample Optimal Estimation and Inference on Average Treatment Effects Under Unconfoundedness.” *Econometrica*, 89(3): 1141–1177.
- Armstrong, Timothy B, and Michal Kolesár.** 2021b. “Sensitivity analysis using approximate moment condition models.” *Quantitative Economics*, 12(1): 77–108.
- Armstrong, Timothy B, Patrick Kline, and Liyang Sun.** 2025. “Adapting to misspecification.” *Econometrica*, 93(6): 1981–2005.
- Aronow, Peter M, and Cyrus Samii.** 2016. “Does regression produce representative estimates of causal effects?” *American Journal of Political Science*, 60(1): 250–267.
- Athey, Susan, and Guido W Imbens.** 2022. “Design-based analysis in difference-in-differences settings with staggered adoption.” *Journal of econometrics*, 226(1): 62–79.
- Athey, Susan, and Guido W. Imbens.** 2023. “In Search of the Third Number: What to Report Beyond Point Estimates and Standard Errors.” <https://www.youtube.com/watch?v=i-K9a6UhtsE>, Keynote Lecture at the Berkeley Initiative for Transparency in the Social Sciences (BITSS).
- Ben-Michael, Eli, D James Greiner, Kosuke Imai, and Zhichao Jiang.** 2025. “Safe policy learning through extrapolation: Application to pre-trial risk assessment.” *Journal of the American Statistical Association*, forthcoming.
- Bhuller, Manudeep, and Henrik Sigstad.** 2024. “2SLS with multiple treatments.” *Journal of Econometrics*, 242(1): 105785.
- Blandhol, Christine, John Bonney, Magne Mogstad, and Alexander Torgovitsky.** 2022. “When is TSLS actually late?” National Bureau of Economic Research Cambridge, MA.
- Borusyak, Kirill, Xavier Jaravel, and Jann Spiess.** 2024. “Revisiting event-study designs: robust and efficient estimation.” *Review of Economic Studies*, 91(6): 3253–3285.
- Callaway, Brantly, and Pedro HC Sant’Anna.** 2021. “Difference-in-differences with multiple time periods.” *Journal of econometrics*, 225(2): 200–230.
- Chernozhukov, Victor, Sokbae Lee, Adam M Rosen, and Liyang Sun.** 2025. “Policy Learning with Confidence.” *arXiv preprint arXiv:2502.10653*.

- Chiu, Albert, Xingchen Lan, Ziyi Liu, and Yiqing Xu.** 2026. “Causal panel analysis under parallel trends: lessons from a large reanalysis study.” *American Political Science Review*, 120(1): 245–266.
- Cole, Stephen R, and Elizabeth A Stuart.** 2010. “Generalizing evidence from randomized clinical trials to target populations: the ACTG 320 trial.” *American journal of epidemiology*, 172(1): 107–115.
- Coussens, Stephen, and Jann Spiess.** 2021. “Improving inference from simple instruments through compliance estimation.” *arXiv preprint arXiv:2108.03726*.
- Crump, Richard K, V Joseph Hotz, Guido Imbens, and Oscar Mitnik.** 2006. “Moving the goalposts: Addressing limited overlap in the estimation of average treatment effects by changing the estimand.”
- de Chaisemartin, Clément.** 2021. “Trading-off Bias and Variance in Stratified Experiments and in Matching Studies, Under a Boundedness Condition on the Magnitude of the Treatment Effect.” *arXiv preprint arXiv:2105.08766*.
- De Chaisemartin, Clément, and Xavier d’Haultfoeuille.** 2020. “Two-way fixed effects estimators with heterogeneous treatment effects.” *American economic review*, 110(9): 2964–2996.
- Dehejia, Rajeev, Cristian Pop-Eleches, and Cyrus Samii.** 2021. “From local to global: External validity in a fertility natural experiment.” *Journal of Business & Economic Statistics*, 39(1): 217–243.
- Farahani, Abolfazl, Sahar Voghoei, Khaled Rasheed, and Hamid R Arabnia.** 2021. “A brief review of domain adaptation.” *Advances in data science and information engineering: proceedings from ICDATA 2020 and IKE 2020*, 877–894.
- Gardner, John.** 2022. “Two-stage differences in differences.” *arXiv preprint arXiv:2207.05943*.
- Goldsmith-Pinkham, Paul, Peter Hull, and Michal Kolesár.** 2024. “Contamination bias in linear regressions.” *American Economic Review*, 114(12): 4015–4051.
- Goodman-Bacon, Andrew.** 2021. “Difference-in-differences with variation in treatment timing.” *Journal of econometrics*, 225(2): 254–277.
- Hartman, Erin, Richard Grieve, Roland Ramsahai, and Jasjeet S Sekhon.** 2015. “From sample average treatment effect to population average treatment effect on the treated: combining experimental with observational studies to estimate population treatment effects.” *Journal of the Royal Statistical Society Series A: Statistics in Society*, 178(3): 757–778.

- Heckman, James J, and Edward J Vytlačil.** 2007. “Econometric evaluation of social programs, part II: Using the marginal treatment effect to organize alternative econometric estimators to evaluate social programs, and to forecast their effects in new environments.” *Handbook of econometrics*, 6: 4875–5143.
- Heckman, James J, and Edward Vytlačil.** 2005. “Structural equations, treatment effects, and econometric policy evaluation 1.” *Econometrica*, 73(3): 669–738.
- Henry, Marc, and Alexei Onatski.** 2012. “Set coverage and robust policy.” *Economics Letters*, 115(2): 256–257.
- Hotz, V Joseph, Guido W Imbens, and Julie H Mortimer.** 2005. “Predicting the efficacy of future training programs using past experiences at other locations.” *Journal of econometrics*, 125(1-2): 241–270.
- Huntington-Klein, Nick.** 2020. “Instruments with heterogeneous effects: Bias, monotonicity, and localness.” *Journal of Causal Inference*, 8(1): 182–208.
- Imbens, Guido W, and Charles F Manski.** 2004. “Confidence intervals for partially identified parameters.” *Econometrica*, 72(6): 1845–1857.
- Imbens, Guido W., and Joshua D. Angrist.** 1994. “Identification and Estimation of Local Average Treatment Effects.” *Econometrica*, 62(2): 467–475.
- Kennedy, Edward H.** 2019. “Nonparametric causal effects based on incremental propensity score interventions.” *Journal of the American Statistical Association*, 114(526): 645–656.
- Kline, Patrick, Raffaele Saggio, and Mikkel Sølvsten.** 2020. “Leave-out estimation of variance components.” *Econometrica*, 88(5): 1859–1898.
- Kolesár, Michal.** 2013. “Estimation in an Instrumental Variables Model with Treatment Effect Heterogeneity.” Princeton University, Department of Economics Working Paper 2013-2.
- Kraft, Matthew A.** 2020. “Interpreting effect sizes of education interventions.” *Educational researcher*, 49(4): 241–253.
- Krueger, Alan B.** 1999. “Experimental estimates of education production functions.” *The quarterly journal of economics*, 114(2): 497–532.
- Kwon, Soonwoo, and Liyang Sun.** 2025. “Estimating Treatment Effects Under Bounded Heterogeneity.” *arXiv preprint arXiv:2510.05454*.

- Lakdawala, Leah K, Eduardo Nakasone, and Kevin Kho.** 2023. “Dynamic impacts of school-based internet access on student learning: Evidence from Peruvian public primary schools.” *American Economic Journal: Economic Policy*, 15(4): 222–254.
- Lau, Chun Pong.** 2026. “Aggregating Treatment Effects across Multiple Outcomes.” Job market paper.
- Lehmann, E.L., and Joseph P. Romano.** 2024. *Testing statistical hypotheses*. Springer.
- Li, Fan, Kari Lock Morgan, and Alan M Zaslavsky.** 2018. “Balancing covariates via propensity score weighting.” *Journal of the American Statistical Association*, 113(521): 390–400.
- Liu, Yiqi.** 2025. “Synthetic Parallel Trends.” *arXiv preprint arXiv:2511.05870*.
- Manski, Charles F.** 2021. “Econometrics for decision making: Building foundations sketched by Haavelmo and Wald.” *Econometrica*, 89(6): 2827–2853.
- McCloskey, Adam.** 2017. “Bonferroni-based size-correction for nonstandard testing problems.” *Journal of Econometrics*, 200(1): 17–35.
- Mogstad, Magne, Alexander Torgovitsky, and Christopher R Walters.** 2021. “The causal interpretation of two-stage least squares with multiple instrumental variables.” *American Economic Review*, 111(11): 3663–3698.
- Mogstad, Magne, and Alexander Torgovitsky.** 2024. “Instrumental variables with unobserved heterogeneity in treatment effects.” In *Handbook of Labor Economics*. Vol. 5, 1–114. Elsevier.
- Molinari, Francesca.** 2020. “Microeconometrics with partial identification.” *Handbook of econometrics*, 7: 355–486.
- Pfanzagl, Johann.** 1994. *Parametric statistical theory*. Walter de Gruyter.
- Poirier, Alexandre, and Tymon Słoczyński.** 2024. “Quantifying the internal validity of weighted estimands.” *arXiv preprint arXiv:2404.14603*.
- Romano, Joseph P, Azeem M Shaikh, and Michael Wolf.** 2014. “A practical two-step method for testing moment inequalities.” *Econometrica*, 82(5): 1979–2002.
- Roth, Jonathan, and Pedro HC Sant’Anna.** 2023. “Efficient estimation for staggered rollout designs.” *Journal of Political Economy Microeconomics*, 1(4): 669–709.

- Roth, Jonathan, Pedro HC Sant’Anna, Alyssa Bilinski, and John Poe.** 2023. “What’s trending in difference-in-differences? A synthesis of the recent econometrics literature.” *Journal of Econometrics*, 235(2): 2218–2244.
- Sarfati, Reca, and Vod Vilfort.** 2026. “Integrating Diagnostic Checks into Estimation.” *arXiv preprint arXiv:2604.16690*.
- Schanzenbach, Diane Whitmore.** 2006. “What have researchers learned from Project STAR?” *Brookings papers on education policy*, 205–228.
- Scheffé, Henry.** 1953. “A method for judging all contrasts in the analysis of variance.” *Biometrika*, 40(1-2): 87–110.
- Seri, Raffaello.** 2015. “A tight bound on the distance between a noncentral chi square and a normal distribution.” *IEEE Communications Letters*, 19(11): 1877–1880.
- Sloczyński, Tymon.** 2020. “When should we (not) interpret linear iv estimands as late?” *arXiv preprint arXiv:2011.06695*.
- Sloczyński, Tymon.** 2022. “Interpreting OLS estimands when treatment effects are heterogeneous: Smaller groups get larger weights.” *Review of Economics and Statistics*, 104(3): 501–509.
- Stuart, Elizabeth A, Stephen R Cole, Catherine P Bradshaw, and Philip J Leaf.** 2011. “The use of propensity scores to assess the generalizability of results from randomized trials.” *Journal of the Royal Statistical Society Series A: Statistics in Society*, 174(2): 369–386.
- Sun, Liyang, and Sarah Abraham.** 2021. “Estimating dynamic treatment effects in event studies with heterogeneous treatment effects.” *Journal of econometrics*, 225(2): 175–199.
- Sun, Yin, Árpád Baricz, and Shidong Zhou.** 2010. “On the Monotonicity, Log-Concavity, and Tight Bounds of the Generalized Marcum and Nuttall Q -Functions.” *IEEE Transactions on Information Theory*, 56(3): 1166–1186.
- Van der Vaart, Aad W.** 2000. *Asymptotic statistics*. Vol. 3, Cambridge university press.
- Wing, Coady, Seth M Freedman, and Alex Hollingsworth.** 2024. “Stacked difference-in-differences.” National Bureau of Economic Research.
- Yitzhaki, Shlomo.** 1996. “On using linear regressions in welfare economics.” *Journal of Business & Economic Statistics*, 14(4): 478–486.
- Zhou, Kaiyang, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy.** 2022. “Domain generalization: A survey.” *IEEE transactions on pattern analysis and machine intelligence*, 45(4): 4396–4415.

Zhou, Xiang, and Aleksei Opacic. 2022. “Marginal interventional effects.” *arXiv preprint arXiv:2206.10717*.

Supplemental Appendix to “Robust Inference for Weighted Estimands”

Vod Vilfort

A Optimal Quantile-Unbiased Estimation

If one restricts attention to the class of quantile-unbiased estimators that depend on $\hat{\theta}$ through $\hat{\theta}'Q\hat{\theta}$, then the optimality statement for $\hat{\eta}_{1-\beta}$ in Proposition 3 can be strengthened. To see this, first recall that since $\hat{\theta} \sim N(\theta, \Sigma)$,

$$\hat{\theta}'Q\hat{\theta} = (\Sigma^{-1/2}\hat{\theta})'A(\Sigma^{-1/2}\hat{\theta}) \sim \chi_{K-1}^2(H(\theta)), \quad \forall \theta.$$

The family $\{\chi_{K-1}^2(\eta) : \eta \geq 0\}$ of noncentral chi-squared distributions has monotone likelihood ratios in $\hat{\theta}'Q\hat{\theta}$. These distributions are continuous, so their CDFs satisfy (5.3.10) from Pfanzagl (1994). Consequently, the estimator $\hat{\eta}_{1-\beta}$ satisfies properties 5.4.1(i)-(ii) of Pfanzagl (1994) over the parameter space $\eta \in H = (0, \infty)$. By Pfanzagl (1994, Theorem 5.4.3), $\hat{\eta}_{1-\beta}$ is maximally concentrated around $H(\theta)$ when θ is heterogeneous. That is, for any other potentially randomized estimator $\tilde{\eta}_{1-\beta}$ that satisfies (14) and depends on $\hat{\theta}$ through $\hat{\theta}'Q\hat{\theta}$,

$$\mathbb{P}_\theta\{\underline{\eta} \leq \hat{\eta}_{1-\beta} \leq \bar{\eta}\} \geq \mathbb{P}_\theta\{\underline{\eta} \leq \tilde{\eta}_{1-\beta} \leq \bar{\eta}\}, \quad \forall (\underline{\eta}, \bar{\eta}, \theta) : 0 \leq \underline{\eta} \leq H(\theta) \leq \bar{\eta} \leq \infty, H(\theta) > 0.$$

Considering $\underline{\eta} = 0$ and $\bar{\eta} = H(\theta) + \varepsilon$ yields the statement that $\hat{\eta}_{1-\beta}$ is uniformly most accurate in the class of quantile-unbiased estimators that depend on $\hat{\theta}$ through $\hat{\theta}'Q\hat{\theta}$.

By Pfanzagl (1994, Proposition 2.5.3), the above maximal concentration is equivalent to minimum-risk optimality for every loss function $L(a, \theta)$ that attains its minimum at $a = H(\theta)$ and that is quasiconvex in a for all heterogeneous θ :

$$\mathbb{E}_\theta[L(\hat{\eta}_{1-\beta}, \theta)] \leq \mathbb{E}_\theta[L(\tilde{\eta}_{1-\beta}, \theta)], \quad \forall \theta : H(\theta) > 0.$$

The above loss functions formalize the notion that losses get larger as a moves away from $H(\theta)$. For example, taking $L(a, \theta) = (H(\theta) - a)^2$ to be squared error loss, the above result shows that when θ is heterogeneous, $\hat{\eta}_{1-\beta}$ is optimal under mean squared error for estimating $H(\theta)$ in the class of quantile-unbiased estimators that depend on $\hat{\theta}$ through $\hat{\theta}'Q\hat{\theta}$.

B Additional Coverage Properties

B.1 Simultaneous Coverage

Proposition 6 shows that CI_w^* is robust in the sense of uniform coverage: for each λ -weighted estimand, CI_w^* provides valid coverage. A more general notion of robust coverage is *simultaneous coverage*, which entails coverage of the entire set of λ -weighted estimands: $\Lambda(\theta) = \{\lambda'\theta : \lambda \in \Lambda\}$. The following result shows that CI_w^* also provides such coverage, but at a different confidence level for the two-sided case.

Proposition 7. *CI_w^* provides simultaneous coverage for $\Lambda(\theta) = \{\lambda'\theta : \lambda \in \Lambda\}$ at confidence level (i) $1 - (2\alpha + \beta)$ in the two-sided case and (ii) $1 - (\alpha + \beta)$ in the one-sided case. That is,*

$$\mathbb{P}_\theta\{\Lambda(\theta) \subseteq CI_w^*\} = \mathbb{P}_\theta\{\lambda'\theta \in CI_w^*, \forall \lambda \in \Lambda\} \geq \begin{cases} 1 - (2\alpha + \beta), & \text{two-sided,} \\ 1 - (\alpha + \beta), & \text{one-sided,} \end{cases} \quad \forall \theta. \quad (35)$$

Proof. See Appendix C.7. □

Thus, CI_w^* is robust in the sense of simultaneous coverage, with the caveat that inferences in the two-sided case come at a lower confidence level relative to uniform coverage. In settings where simultaneous inferences are more relevant than uniform ones, this additional cost may be worthwhile. To see how, I draw connections to the literature on partial identification by viewing $\Lambda(\theta)$ as the identified set for an estimand with unknown weights in Λ . From this perspective, uniform coverage (20) corresponds to coverage of points in the identified set, while simultaneous coverage (35) corresponds to coverage of the identified set—for precise definitions in a general partial identification setup, see Molinari (2020, Section 4.3).

Following Imbens and Manski (2004), coverage of points in the identified set is relevant when there is a true estimand $\lambda_0'\theta$. For example, suppose that $\lambda_0'\theta$ represents the ATE for a target population that one has partial information about and Λ represents the class of populations consistent with that information. In such cases, the relevant error is the exclusion of $\lambda_0'\theta$ from CI_w^* , which occurs with probability $\mathbb{P}_\theta\{\lambda_0'\theta \notin CI_w^*\}$. Since CI_w^* provides the uniform coverage in (20), this error rate is uniformly bounded over the potential values of $\lambda_0 \in \Lambda$.

Following Henry and Onatski (2012), coverage of the identified set is relevant when there is concern for robust decision-making. For instance, suppose Λ represents readers λ who will use CI_w^* to make downstream decisions and the researcher is concerned about the worst-case loss from those decisions across readers.²⁷ In such cases, Henry and Onatski (2012) show that the relevant error is the exclusion of *any* reader's estimand $\lambda'\theta$ from CI_w^* , which occurs with probability $\mathbb{P}_\theta\{\Lambda(\theta) \not\subseteq CI_w^*\}$. The simultaneous coverage in (35) bounds this error rate.

²⁷The role of confidence intervals for decision-making has been considered more recently in Manski (2021); Andrews and Chen (2025); Ben-Michael et al. (2025); Chernozhukov et al. (2025).

In summary, the relevant notion of coverage depends on the type of error that one wishes to mitigate. In my setting, uniform coverage appears most appropriate. For example, a setup with a partially known target population λ_0 from the researcher's perspective is consistent with the mapping of my framework to multisite experiments discussed in Section 2.2. Moreover, even in a decision-making setup where Λ is taken to represent readers, it is plausible that each reader is only concerned about the exclusion of their own estimand, in which case simultaneous coverage may be too strong. With these examples in mind, I therefore focus on uniform coverage as the relevant notion of robust coverage.

B.2 Upper Bounding the Coverage Rate

Proposition 6 shows that the coverage rate of CI_w^* for $\tau_\lambda(\theta)$ is bounded below:

$$\mathbb{P}_\theta\{\tau_\lambda(\theta) \in CI_w^*\} \geq 1 - (\alpha + \beta), \quad \forall \lambda \in \Lambda, \quad \forall \theta.$$

A follow-up question is whether the coverage rate of CI_w^* is nontrivial in the sense that there exists $(\lambda, \theta) \in \Lambda \times \mathbb{R}^K$ where the left-hand side is strictly less than one. Below I show that this is indeed the case. To this end, define the function

$$\delta_{w,\Lambda}^\eta(\alpha, \beta) = \Phi \left(- \frac{z_{1-\alpha}\sigma_w + \eta \left(\max_{\lambda \in \Lambda} \|\lambda - w_{\text{GLS}}\|_\Sigma + \max_{\lambda \in \Lambda} \|\lambda - w\|_\Sigma \right) + \sqrt{F_{\chi^2}^{-1}(\beta; \eta)} \|w - w_{\text{GLS}}\|_\Sigma}{\sigma_{\text{GLS}}} \right) \beta.$$

Let $\bar{\Theta}^\eta = \{\theta : H(\theta) = \eta\}$ denote the set of parameters where the heterogeneity is equal to η .

Proposition 8. *For each η , the coverage rate of CI_w^* for $\tau_\lambda(\theta)$ is bounded above as*

$$\mathbb{P}_\theta\{\tau_\lambda(\theta) \in CI_w^*\} \leq \begin{cases} 1 - 2\delta_{w,\Lambda}^\eta(\alpha/2, \beta), & \text{two-sided,} \\ 1 - \delta_{w,\Lambda}^\eta(\alpha, \beta), & \text{one-sided,} \end{cases} \quad \forall \lambda \in \Lambda, \quad \forall \theta \in \bar{\Theta}^\eta.$$

Proof. See Appendix C.8. □

Proposition 8 implies that for any θ , the coverage rate is strictly less than one—of course, this assumes that $\alpha, \beta > 0$, which I maintain throughout. The function $\delta_{w,\Lambda}^\eta(\alpha, \beta)$ depends on η , so the bound is not uniform: $\delta_{w,\Lambda}^\eta(\alpha, \beta) \rightarrow 0$ as $\eta \rightarrow \infty$ unless $\Lambda = \{w\} = \{w_{\text{GLS}}\}$. Under $w = w_{\text{GLS}}$ and $\Lambda = \{\lambda \in \mathcal{W} : \sigma_\lambda \leq r\sigma_w\}$, the function simplifies to $\delta_{w,\Lambda}^\eta(\alpha, \beta) = \Phi(-z_{1-\alpha} - 2\eta\sqrt{r^2 - 1})\beta$.

C Proofs of Results for the Normal Model

C.1 Proof of Proposition 1

Note that $\Sigma^{1/2}(\lambda - w) = A\Sigma^{1/2}(\lambda - w)$ and $A^2 = A$. Thus, the Cauchy-Schwarz inequality yields

$$|\theta'(\lambda - w)| \leq \|A\Sigma^{-1/2}\theta\| \|\Sigma^{1/2}(\lambda - w)\| = \sqrt{\theta'Q\theta} \sqrt{(\lambda - w)' \Sigma (\lambda - w)} = H(\theta) \|\lambda - w\|_{\Sigma},$$

with equality if and only if $A\Sigma^{-1/2}\theta$ and $\Sigma^{1/2}(\lambda - w)$ are linearly dependent. For example, when $\lambda \neq w$, this occurs for $\theta^\dagger = \eta \Sigma(\lambda - w) / \|\lambda - w\|_{\Sigma}$, which satisfies $H(\theta^\dagger) = \eta$.

C.2 Proof of Proposition 2

For θ such that $H(\theta) > 0$, observe that (i) the definition of $\hat{\eta}_{1-\beta}$ yields

$$\{H(\theta) \leq \hat{\eta}_{1-\beta}\} \subseteq \{H(\theta) \leq \hat{\eta}_{1-\beta}, F_{\chi^2}(\hat{\theta}'Q\hat{\theta}; 0) > \beta\} = \{F_{\chi^2}(\hat{\theta}'Q\hat{\theta}; H(\theta)) \geq \beta, F_{\chi^2}(\hat{\theta}'Q\hat{\theta}; 0) > \beta\},$$

and (ii) the monotonicity of $\eta \mapsto F_{\chi^2}(\hat{\theta}'Q\hat{\theta}; \eta)$ yields

$$\{F_{\chi^2}(\hat{\theta}'Q\hat{\theta}; H(\theta)) \geq \beta\} \subseteq \{F_{\chi^2}(\hat{\theta}'Q\hat{\theta}; H(\theta)) \geq \beta, F_{\chi^2}(\hat{\theta}'Q\hat{\theta}; 0) > \beta\}.$$

But since the reverse of the set inclusions in (i) and (ii) hold for any θ , then

$$\mathbb{P}_\theta\{H(\theta) \leq \hat{\eta}_{1-\beta}\} = \mathbb{P}_\theta\{F_{\chi^2}(\hat{\theta}'Q\hat{\theta}; H(\theta)) \geq \beta\} = 1 - \beta, \quad \forall \theta : H(\theta) > 0,$$

where the second equality holds for any θ , by the probability integral transform.

C.3 Proof of Proposition 3

For a given $(\eta, \varepsilon) > 0$, consider the problem of testing null hypothesis $\theta \in \Theta_0 = \{\theta : H(\theta) = \eta + \varepsilon\}$ against alternative hypothesis $\theta \in \Theta_1 = \{\theta : H(\theta) = \eta\}$ in the model $\{N(\theta, \Sigma) : \theta \in \Theta_0 \cup \Theta_1\}$. For the class of potentially randomized tests $\tilde{T}$ with size

$$\sup_{\theta \in \Theta_0} \mathbb{E}_\theta[\tilde{T}] = \beta,$$

a maximin test $\hat{T}$ is one that satisfies, for any other test $\tilde{T}$ with size β , the inequality

$$\inf_{\theta \in \Theta_1} \mathbb{E}_\theta[\hat{T}] \geq \inf_{\theta \in \Theta_1} \mathbb{E}_\theta[\tilde{T}].$$

This testing structure follows the setup in [Lehmann and Romano \(2024, Section 8\)](#).

By quantile-unbiasedness, the tests $\tilde{T} = \mathbb{1}\{\eta + \varepsilon > \tilde{\eta}_{1-\beta}\}$ and $\hat{T} = \mathbb{1}\{\eta + \varepsilon > \hat{\eta}_{1-\beta}\}$ have size β . Note further that $\Theta_1 = \bar{\Theta}^\eta$. Thus, if one can show that $\hat{T} = \mathbb{1}\{\eta + \varepsilon > \hat{\eta}_{1-\beta}\}$ is a maximin test

in the above sense for any given $(\eta, \varepsilon) > 0$, it would then follow that

$$\sup_{\theta \in \Theta^\eta} \mathbb{P}_\theta \{ \hat{\eta}_{1-\beta} \geq H(\theta) + \varepsilon \} \leq \sup_{\theta \in \Theta^\eta} \mathbb{P}_\theta \{ \tilde{\eta}_{1-\beta} \geq H(\theta) + \varepsilon \}, \quad \forall (\eta, \varepsilon) > 0,$$

for any quantile-unbiased $\tilde{\eta}_{1-\beta}$, which would be the desired conclusion.

To solve the maximin problem, I now appeal to invariance arguments. Formally, the above testing problems are invariant to the group of transformations

$$G = \{ \theta \mapsto M\theta + c\mathbf{1} : c \in \mathbb{R}, M\mathbf{1} = \mathbf{1}, M'\Sigma^{-1}M = \Sigma^{-1} \}.$$

Indeed, $M\hat{\theta} + c\mathbf{1} \sim N(M\theta + c\mathbf{1}, \Sigma)$ with $M^{-1} = \Sigma M' \Sigma^{-1}$ and $H(M\theta + c\mathbf{1}) = H(\theta)$ for the above (c, M) . Thus, for maximizing worst-case power in the class of size β tests, it follows from the Hunt-Stein theorem (Lehmann and Romano, 2024, Theorem 8.5.1) and Lehmann and Romano (2024, Lemma 8.4.1) that one can restrict attention to invariant tests. In this setup, an invariant test depends on $\hat{\theta}$ through the maximal invariant $\hat{\theta}'Q\hat{\theta}$. Thus, below I restrict attention to the class of size β tests that depend on $\hat{\theta}$ through $\hat{\theta}'Q\hat{\theta}$.

Now considering tests that depend on $\hat{\theta}$ through $\hat{\theta}'Q\hat{\theta}$, recall that

$$\hat{\theta}'Q\hat{\theta} = (\Sigma^{-1/2}\hat{\theta})'A(\Sigma^{-1/2}\hat{\theta}) \sim \chi_{K-1}^2(H(\theta)), \quad \forall \theta.$$

The family $\{\chi_{K-1}^2(\eta) : \eta \geq 0\}$ of noncentral chi-squared distributions has monotone likelihood ratios in $\hat{\theta}'Q\hat{\theta}$. Combined with the observation that $\hat{T} = \mathbf{1}\{\eta + \varepsilon > \hat{\eta}_{1-\beta}\}$ rejects for small values of $\hat{\theta}'Q\hat{\theta}$, the Neyman-Pearson lemma implies that $\hat{T}$ is most powerful for testing $\chi_{K-1}^2(\eta + \varepsilon)$ against $\chi_{K-1}^2(\eta)$. In particular, for any $\tilde{T} = \mathbf{1}\{\eta + \varepsilon > \tilde{\eta}_{1-\beta}\}$ based on quantile-unbiased $\tilde{\eta}_{1-\beta}$,

$$\mathbb{P}_\theta \{ \eta + \varepsilon > \hat{\eta}_{1-\beta} \} = \mathbb{E}_\theta [\hat{T}] \geq \mathbb{E}_\theta [\tilde{T}] = \mathbb{P}_\theta \{ \eta + \varepsilon > \tilde{\eta}_{1-\beta} \}, \quad \forall \theta : H(\theta) = \eta.$$

This establishes the optimality of $\hat{T} = \mathbf{1}\{\eta + \varepsilon > \hat{\eta}_{1-\beta}\}$ in the maximin problem, from which the optimality result for $\hat{\eta}_{1-\beta}$ follows.

C.4 Proof of Proposition 4

Observe that

$$\{H(\theta) \leq \hat{\eta}_{1-\beta}\} \subseteq \left\{ H(\theta) \max_{\lambda \in \Lambda} \|\lambda - w\|_\Sigma \leq \hat{B}_w^\beta(\Lambda) \right\} \subseteq \left\{ \max_{\lambda \in \Lambda} |(\lambda - w)' \theta| \leq \hat{B}_w^\beta(\Lambda) \right\},$$

where the latter set inclusion follows from Proposition 1 and the former inclusion holds with equality when $H(\theta) > 0$. Proposition 2 shows $\mathbb{P}_\theta \{ H(\theta) \leq \hat{\eta}_{1-\beta} \} \geq 1 - \beta$ for all θ , with equality when $H(\theta) > 0$. Together with the above expressions, this yields the desired coverage bound.

Now consider θ where $H(\theta) > 0$ and suppose that $\Lambda \neq \{w\}$. Then the above implies

$$\mathbb{P}_\theta \left\{ H(\theta) \max_{\lambda \in \Lambda} \|\lambda - w\|_\Sigma \leq \widehat{B}_w^\beta(\Lambda) \right\} = 1 - \beta, \quad \forall \theta: H(\theta) > 0.$$

Moreover, taking $\lambda^* \in \arg\max_{\lambda \in \Lambda} \|\lambda - w\|_\Sigma$ and $\theta^\dagger = \Sigma(\lambda^* - w) / \|\lambda^* - w\|_\Sigma$ yields $H(\theta^\dagger) = 1$ and

$$H(\theta^\dagger) \max_{\lambda \in \Lambda} \|\lambda - w\|_\Sigma = \|\lambda^* - w\|_\Sigma = |(\lambda^* - w)' \theta^\dagger| \leq \max_{\lambda \in \Lambda} |(\lambda - w)' \theta^\dagger| \leq H(\theta^\dagger) \max_{\lambda \in \Lambda} \|\lambda - w\|_\Sigma.$$

Thus, the desired coverage probability is equal to $1 - \beta$ under $\theta^\dagger = \Sigma(\lambda^* - w) / \|\lambda^* - w\|_\Sigma$.

C.5 Proof of Proposition 5

To show existence and uniqueness, it suffices to show

$$w^* = \arg \min_{\bar{w} \in \mathcal{W}^*} f(\bar{w}), \quad f(\bar{w}) = \max_{\lambda \in \Lambda} f_\lambda(\bar{w}), \quad f_\lambda(\bar{w}) = (\lambda - \bar{w})' \Sigma (\lambda - \bar{w}).$$

Each f_λ has Hessian 2Σ , and hence is strongly convex since Σ is positive definite. Because this Hessian does not depend on λ , the pointwise maximum f is also strongly convex, and thus has at most one minimizer over the convex set $\mathcal{W}^*$. Moreover, f is (i) continuous by compactness of Λ and continuity of $(\lambda, \bar{w}) \mapsto f_\lambda(\bar{w})$; and (ii) coercive since $f(\bar{w}) \gtrsim \|\lambda - \bar{w}\|^2$ uniformly over $\lambda \in \Lambda$. Thus, f attains its minimum on nonempty closed set $\mathcal{W}^*$. In totality, w^* exists uniquely.

I now establish the bias-optimality of w^* . For each $\bar{w} \in \mathcal{W}^*$, Proposition 1 implies

$$\max_{\lambda \in \Lambda} \max_{\theta \in \Theta^\eta} |\mathbb{E}_\theta[\hat{\tau}_{\bar{w}}] - \tau_\lambda(\theta)| = \max_{\lambda \in \Lambda} \max_{\theta \in \Theta^\eta} |(\lambda - \bar{w})' \theta| \leq \max_{\lambda \in \Lambda} \max_{\theta \in \Theta^\eta} H(\theta) \|\lambda - \bar{w}\|_\Sigma \leq \eta \max_{\lambda \in \Lambda} \|\lambda - \bar{w}\|_\Sigma.$$

If $\max_{\lambda \in \Lambda} \|\lambda - \bar{w}\|_\Sigma = 0$, the above inequalities become equalities so that bias-optimality follows from the definition of w^* . It therefore suffices to show that the inequalities become equalities in the case of $\max_{\lambda \in \Lambda} \|\lambda - \bar{w}\|_\Sigma > 0$. For $\lambda^* \in \arg\max_{\lambda \in \Lambda} \|\lambda - \bar{w}\|_\Sigma$ and $\theta^\dagger = \eta \Sigma(\lambda^* - \bar{w}) / \|\lambda^* - \bar{w}\|_\Sigma$, observe that

$$H(\theta^\dagger) = \eta, \quad |(\lambda^* - \bar{w})' \theta^\dagger| = \eta \|\lambda^* - \bar{w}\|_\Sigma = \eta \max_{\lambda \in \Lambda} \|\lambda - \bar{w}\|_\Sigma.$$

Thus, since $(\lambda^*, \theta^\dagger) \in \Lambda \times \Theta^\eta$,

$$\max_{\lambda \in \Lambda} \max_{\theta \in \Theta^\eta} |(\lambda - \bar{w})' \theta| \geq |(\lambda^* - \bar{w})' \theta^\dagger| = \eta \max_{\lambda \in \Lambda} \|\lambda - \bar{w}\|_\Sigma.$$

Together with the previous inequalities, this establishes the bias-optimality of w^* .

C.6 Proof of Proposition 6

On $\mathcal{E}_\theta = \{\max_{\lambda \in \Lambda} |(\lambda - w)' \theta| \leq \widehat{B}_w^\beta(\Lambda)\}$, the noncoverage event for $\lambda \in \Lambda$ is $\{\lambda' \theta \notin CI_w^*, \mathcal{E}_\theta\}$. Note $\mathbb{P}_\theta\{\lambda' \theta \notin CI_w^*, \mathcal{E}_\theta^c\} \leq \mathbb{P}_\theta\{\mathcal{E}_\theta^c\} \leq \beta$ for all (λ, θ) , where the last inequality follows from Proposition

4. Thus, it suffices to show that $\mathbb{P}_\theta\{\lambda'\theta \notin CI_w^*, \mathcal{E}_\theta\} \leq \alpha$ for all $\lambda \in \Lambda$ and θ , since then

$$\mathbb{P}_\theta\{\lambda'\theta \notin CI_w^*\} = \mathbb{P}_\theta\{\lambda'\theta \notin CI_w^*, \mathcal{E}_\theta\} + \mathbb{P}_\theta\{\lambda'\theta \notin CI_w^*, \mathcal{E}_\theta^c\} \leq \alpha + \beta, \quad \forall \lambda \in \Lambda, \quad \forall \theta,$$

which would yield the desired conclusion.

For the two-sided CI_w^* , the noncoverage event for $\lambda \in \Lambda$ on $\mathcal{E}_\theta$ is

$$\begin{aligned} \left\{ \left| \lambda'\theta - w'\hat{\theta} \right| > \text{cv}_{1-\alpha} \left(\hat{B}_w^\beta(\Lambda) / \sigma_w \right) \sigma_w, \mathcal{E}_\theta \right\} &\subseteq \left\{ \left| \lambda'\theta - w'\hat{\theta} \right| > \text{cv}_{1-\alpha} \left(\max_{\lambda \in \Lambda} |(\lambda - w)'\theta| / \sigma_w \right) \sigma_w \right\} \\ &\subseteq \left\{ \left| \lambda'\theta - w'\hat{\theta} \right| > \text{cv}_{1-\alpha} (|(\lambda - w)'\theta| / \sigma_w) \sigma_w \right\} \\ &= \left\{ \left| \lambda'\theta - w'\hat{\theta} \right| > \text{cv}_{1-\alpha} ((\lambda - w)'\theta / \sigma_w) \sigma_w \right\}, \end{aligned}$$

where the first line follows from the definition of $\mathcal{E}_\theta$, the second from $\text{cv}_\alpha(|b|)$ increasing in $|b|$, and the third from $\text{cv}_\alpha(|b|) = \text{cv}_\alpha(b)$. Thus, for all $\lambda \in \Lambda$ and θ ,

$$\mathbb{P}_\theta\{\lambda'\theta \notin CI_w^*, \mathcal{E}_\theta\} \leq \mathbb{P}_\theta\left\{ \left| \lambda'\theta - w'\hat{\theta} \right| / \sigma_w > \text{cv}_{1-\alpha} ((\lambda - w)'\theta / \sigma_w) \right\} = \alpha.$$

Now consider the upper one-sided CI_w^* —the proof for the lower case follows analogously. Using similar steps as above, the noncoverage probability is

$$\mathbb{P}_\theta\{\lambda'\theta \notin CI_w^*, \mathcal{E}_\theta\} \leq \mathbb{P}_\theta\left\{ \lambda'\theta - w'\hat{\theta} > z_{1-\alpha} \sigma_w + (\lambda - w)'\theta \right\} = \mathbb{P}_\theta\left\{ w'(\theta - \hat{\theta}) / \sigma_w > z_{1-\alpha} \right\} = \alpha.$$

In conclusion, $\mathbb{P}_\theta\{\lambda'\theta \notin CI_w^*, \mathcal{E}_\theta\} \leq \alpha$ for all $\lambda \in \Lambda$ and θ , as desired.

C.7 Proof of Proposition 7

For each θ , denote

$$\bar{\lambda}_\theta \in \arg\max_{\lambda \in \Lambda} \lambda'\theta, \quad \underline{\lambda}_\theta \in \arg\min_{\lambda \in \Lambda} \lambda'\theta, \quad \mathcal{E}_\theta = \left\{ \max_{\lambda \in \Lambda} |(\lambda - w)'\theta| \leq \hat{B}_w^\beta(\Lambda) \right\},$$

where $\bar{\lambda}_\theta$ and $\underline{\lambda}_\theta$ exist given Assumption 2 and the linearity of $\lambda \mapsto \lambda'\theta$. The noncoverage event on $\mathcal{E}_\theta$ is $\{\Lambda(\theta) \not\subseteq CI_w^*, \mathcal{E}_\theta\}$. Note $\mathbb{P}_\theta\{\Lambda(\theta) \not\subseteq CI_w^*, \mathcal{E}_\theta^c\} \leq \mathbb{P}_\theta\{\mathcal{E}_\theta^c\} \leq \beta$ for all θ . Thus, in the two-sided case, it suffices to show $\mathbb{P}_\theta\{\Lambda(\theta) \not\subseteq CI_w^*, \mathcal{E}_\theta\} \leq 2\alpha$ for all θ , since then

$$\mathbb{P}_\theta\{\Lambda(\theta) \not\subseteq CI_w^*\} = \mathbb{P}_\theta\{\Lambda(\theta) \not\subseteq CI_w^*, \mathcal{E}_\theta\} + \mathbb{P}_\theta\{\Lambda(\theta) \not\subseteq CI_w^*, \mathcal{E}_\theta^c\} \leq 2\alpha + \beta, \quad \forall \theta.$$

Likewise, in the one-sided case, it suffices to show $\mathbb{P}_\theta\{\Lambda(\theta) \not\subseteq CI_w^*, \mathcal{E}_\theta\} \leq \alpha$ for all θ .

For the two-sided CI_w^* , the noncoverage probability on $\mathcal{E}_\theta$ at the upper endpoint is

$$\mathbb{P}_\theta\left\{ \max_{\lambda \in \Lambda} \lambda'\theta > w'\hat{\theta} + \text{cv}_{1-\alpha} \left(\frac{\hat{B}_w^\beta(\Lambda)}{\sigma_w} \right) \sigma_w, \mathcal{E}_\theta \right\} \leq \mathbb{P}_\theta\left\{ \frac{\bar{\lambda}'_\theta \theta - w'\hat{\theta}}{\sigma_w} > \text{cv}_{1-\alpha} \left(\max_{\lambda \in \Lambda} \frac{|(\lambda - w)'\theta|}{\sigma_w} \right) \right\}$$

$$\begin{aligned}
&\leq \mathbb{P}_\theta \left\{ \frac{\bar{\lambda}'_\theta \theta - w' \hat{\theta}}{\sigma_w} > \text{cv}_{1-\alpha} \left(\frac{|(\bar{\lambda}_\theta - w)' \theta|}{\sigma_w} \right) \right\} \\
&= \mathbb{P}_\theta \left\{ \frac{\bar{\lambda}'_\theta \theta - w' \hat{\theta}}{\sigma_w} > \text{cv}_{1-\alpha} \left(\frac{(\bar{\lambda}_\theta - w)' \theta}{\sigma_w} \right) \right\} \\
&\leq \mathbb{P}_\theta \left\{ \left| \frac{\bar{\lambda}'_\theta \theta - w' \hat{\theta}}{\sigma_w} \right| > \text{cv}_{1-\alpha} \left(\frac{(\bar{\lambda}_\theta - w)' \theta}{\sigma_w} \right) \right\} \\
&= \alpha,
\end{aligned}$$

where the first line follows from the definition of $\mathcal{E}_\theta$, the second from $\text{cv}_\alpha(|b|)$ increasing in $|b|$, and the third from $\text{cv}_\alpha(|b|) = \text{cv}_\alpha(b)$. Likewise, at the lower endpoint,

$$\mathbb{P}_\theta \left\{ \min_{\lambda \in \Lambda} \lambda' \theta < w' \hat{\theta} - \text{cv}_{1-\alpha} \left(\frac{\hat{B}_w^\beta(\Lambda)}{\sigma_w} \right) \sigma_w, \mathcal{E}_\theta \right\} \leq \mathbb{P}_\theta \left\{ \left| \frac{\lambda'_\theta \theta - w' \hat{\theta}}{\sigma_w} \right| > \text{cv}_{1-\alpha} \left(\frac{(\lambda_\theta - w)' \theta}{\sigma_w} \right) \right\} = \alpha.$$

Since $\{\Lambda(\theta) \not\subseteq CI_w^*, \mathcal{E}_\theta\}$ is the union of the above two LHS events, a union bound yields

$$\mathbb{P}_\theta \{\Lambda(\theta) \not\subseteq CI_w^*, \mathcal{E}_\theta\} \leq \alpha + \alpha = 2\alpha.$$

Thus, in the two-sided case, $\mathbb{P}_\theta \{\Lambda(\theta) \not\subseteq CI_w^*, \mathcal{E}_\theta\} \leq 2\alpha$ for all θ , as desired.

Now consider the upper one-sided CI_w^* —the proof for the lower case is symmetric. Using similar steps as above, the probability of noncoverage is

$$\mathbb{P}_\theta \{\Lambda(\theta) \not\subseteq CI_w^*, \mathcal{E}_\theta\} \leq \mathbb{P}_\theta \left\{ \frac{\bar{\lambda}'_\theta \theta - w' \hat{\theta}}{\sigma_w} > z_{1-\alpha} + \frac{(\bar{\lambda}_\theta - w)' \theta}{\sigma_w} \right\} = \mathbb{P}_\theta \left\{ \frac{w' \theta - w' \hat{\theta}}{\sigma_w} > z_{1-\alpha} \right\} = \alpha.$$

Thus, in the one-sided case, $\mathbb{P}_\theta \{\Lambda(\theta) \not\subseteq CI_w^*, \mathcal{E}_\theta\} \leq \alpha$ for all θ , as desired.

C.8 Proof of Proposition 8

Fix η , $\theta \in \bar{\Theta}^\eta$, and $\lambda \in \Lambda$. Denote

$$B_w^\theta(\Lambda) = H(\theta) \max_{\lambda \in \Lambda} \|\lambda - w\|_\Sigma, \quad B_{\text{GLS}}^\theta(\Lambda) = H(\theta) \max_{\lambda \in \Lambda} \|\lambda - w_{\text{GLS}}\|_\Sigma, \quad H(\theta) = \eta.$$

In this proof, I make use of the following statements.

(i) $\{H(\theta) \geq \hat{\eta}_{1-\beta}\} \supseteq \{H^2(\hat{\theta}) \leq F_{\chi^2}^{-1}(\beta; \eta)\}$, which follows from

$$\{H(\theta) < \hat{\eta}_{1-\beta}\} = \{\eta < \hat{\eta}_{1-\beta}\} \subseteq \{F_{\chi^2}(H^2(\hat{\theta}); \eta) > \beta\} = \{H^2(\hat{\theta}) > F_{\chi^2}^{-1}(\beta; \eta)\},$$

where the set inclusion follows from the definition of $\hat{\eta}_{1-\beta}$.

(ii) $w'_{\text{GLS}}\hat{\theta} \perp H^2(\hat{\theta})$, since $H^2(\hat{\theta}) = \|(I - \mathbf{1}w'_{\text{GLS}})\hat{\theta}\|_{\Sigma}^2$ and $w'_{\text{GLS}}\hat{\theta} \perp (I - \mathbf{1}w'_{\text{GLS}})\hat{\theta}$. Indeed,

$$\begin{bmatrix} w'_{\text{GLS}}\hat{\theta} \\ (I - \mathbf{1}w'_{\text{GLS}})\hat{\theta} \end{bmatrix} \sim N \left(\begin{bmatrix} w'_{\text{GLS}} \\ I - \mathbf{1}w'_{\text{GLS}} \end{bmatrix} \theta, \begin{bmatrix} \sigma_{\text{GLS}}^2 & w'_{\text{GLS}}\Sigma(I - w_{\text{GLS}}\mathbf{1}') \\ (I - \mathbf{1}w'_{\text{GLS}})\Sigma w_{\text{GLS}} & (I - \mathbf{1}w'_{\text{GLS}})\Sigma(I - w_{\text{GLS}}\mathbf{1}') \end{bmatrix} \right),$$

where $w'_{\text{GLS}}\Sigma(I - w_{\text{GLS}}\mathbf{1}') = (\mathbf{1}'\Sigma^{-1}\mathbf{1})^{-1}\mathbf{1}' - \sigma_{\text{GLS}}^2\mathbf{1}' = 0$, which yields independence.

(iii) $\mathbb{P}_{\theta}\{H^2(\hat{\theta}) \leq F_{\chi^2}^{-1}(\beta; \eta)\} = \beta$, which follows from

$$\mathbb{P}_{\theta}\{H^2(\hat{\theta}) \leq F_{\chi^2}^{-1}(\beta; \eta)\} = \mathbb{P}_{\theta}\{F_{\chi^2}(H^2(\hat{\theta}); \eta) \leq \beta\} = \beta, \quad H^2(\hat{\theta}) \sim \chi_{K-1}^2(\eta),$$

where $F_{\chi^2}(H^2(\hat{\theta}); \eta) \sim U(0, 1)$ follows from the probability integral transform.

I begin by considering the upper one-sided CI_w^* . Observe that

$$\begin{aligned} & \mathbb{P}_{\theta}\{\lambda'\theta \notin CI_w^*\} \\ & \geq \mathbb{P}_{\theta}\{\lambda'\theta > w'\hat{\theta} + z_{1-\alpha}\sigma_w + \hat{B}_w^{\beta}(\Lambda), H(\theta) \geq \hat{\eta}_{1-\beta}\} \\ & \geq \mathbb{P}_{\theta}\{\lambda'\theta > w'\hat{\theta} + z_{1-\alpha}\sigma_w + B_w^{\theta}(\Lambda), H(\theta) \geq \hat{\eta}_{1-\beta}\} \\ & \geq \mathbb{P}_{\theta}\{\lambda'\theta > w'\hat{\theta} + z_{1-\alpha}\sigma_w + B_w^{\theta}(\Lambda), H^2(\hat{\theta}) \leq F_{\chi^2}^{-1}(\beta; \eta)\} \\ & = \mathbb{P}_{\theta}\{\lambda'\theta > w'_{\text{GLS}}\hat{\theta} + z_{1-\alpha}\sigma_w + (w - w_{\text{GLS}})'\hat{\theta} + B_w^{\theta}(\Lambda), H^2(\hat{\theta}) \leq F_{\chi^2}^{-1}(\beta; \eta)\} \\ & \geq \mathbb{P}_{\theta}\{\lambda'\theta > w'_{\text{GLS}}\hat{\theta} + z_{1-\alpha}\sigma_w + H(\hat{\theta})\|w - w_{\text{GLS}}\|_{\Sigma} + B_w^{\theta}(\Lambda), H^2(\hat{\theta}) \leq F_{\chi^2}^{-1}(\beta; \eta)\} \\ & \geq \mathbb{P}_{\theta}\{\lambda'\theta > w'_{\text{GLS}}\hat{\theta} + z_{1-\alpha}\sigma_w + \sqrt{F_{\chi^2}^{-1}(\beta; \eta)}\|w - w_{\text{GLS}}\|_{\Sigma} + B_w^{\theta}(\Lambda), H^2(\hat{\theta}) \leq F_{\chi^2}^{-1}(\beta; \eta)\} \\ & = \mathbb{P}_{\theta}\{\lambda'\theta > w'_{\text{GLS}}\hat{\theta} + z_{1-\alpha}\sigma_w + \sqrt{F_{\chi^2}^{-1}(\beta; \eta)}\|w - w_{\text{GLS}}\|_{\Sigma} + B_w^{\theta}(\Lambda)\} \mathbb{P}_{\theta}\{H^2(\hat{\theta}) \leq F_{\chi^2}^{-1}(\beta; \eta)\} \\ & \geq \mathbb{P}_{\theta}\{\lambda'\theta > w'_{\text{GLS}}\hat{\theta} + z_{1-\alpha}\sigma_w + \sqrt{F_{\chi^2}^{-1}(\beta; \eta)}\|w - w_{\text{GLS}}\|_{\Sigma} + B_w^{\theta}(\Lambda)\} \beta \\ & = \mathbb{P}_{\theta}\{(\lambda - w_{\text{GLS}})'\theta > w'_{\text{GLS}}(\hat{\theta} - \theta) + z_{1-\alpha}\sigma_w + \sqrt{F_{\chi^2}^{-1}(\beta; \eta)}\|w - w_{\text{GLS}}\|_{\Sigma} + B_w^{\theta}(\Lambda)\} \beta \\ & \geq \mathbb{P}_{\theta}\{-B_{\text{GLS}}^{\theta}(\Lambda) > w'_{\text{GLS}}(\hat{\theta} - \theta) + z_{1-\alpha}\sigma_w + \sqrt{F_{\chi^2}^{-1}(\beta; \eta)}\|w - w_{\text{GLS}}\|_{\Sigma} + B_w^{\theta}(\Lambda)\} \beta \\ & = \Phi \left(-\frac{B_{\text{GLS}}^{\theta}(\Lambda) + z_{1-\alpha}\sigma_w + \sqrt{F_{\chi^2}^{-1}(\beta; \eta)}\|w - w_{\text{GLS}}\|_{\Sigma} + B_w^{\theta}(\Lambda)}{\sigma_{\text{GLS}}} \right) \beta \end{aligned}$$

where the third line follows from statement (i), the fifth from Proposition 1, the seventh from statement (ii), the eighth from statement (iii), and the tenth from Proposition 1.

A similar argument for the lower one-sided CI_w^* yields

$$\begin{aligned} & \mathbb{P}_{\theta}\{\lambda'\theta \notin CI_w^*\} \\ & \geq \mathbb{P}_{\theta}\{\lambda'\theta < w'_{\text{GLS}}\hat{\theta} - z_{1-\alpha}\sigma_w - \sqrt{F_{\chi^2}^{-1}(\beta; \eta)}\|w - w_{\text{GLS}}\|_{\Sigma} - B_w^{\theta}(\Lambda)\} \beta \end{aligned}$$

$$\begin{aligned}
&= \mathbb{P}_\theta \left\{ (\lambda - w_{\text{GLS}})' \theta < w'_{\text{GLS}} (\hat{\theta} - \theta) - z_{1-\alpha} \sigma_w - \sqrt{F_{\chi^2}^{-1}(\beta; \eta)} \|w - w_{\text{GLS}}\|_\Sigma - B_w^\theta(\Lambda) \right\} \beta \\
&\geq \mathbb{P}_\theta \left\{ B_{\text{GLS}}^\theta(\Lambda) < w'_{\text{GLS}} (\hat{\theta} - \theta) - z_{1-\alpha} \sigma_w - \sqrt{F_{\chi^2}^{-1}(\beta; \eta)} \|w - w_{\text{GLS}}\|_\Sigma - B_w^\theta(\Lambda) \right\} \beta \\
&= \Phi \left(- \frac{B_{\text{GLS}}^\theta(\Lambda) + z_{1-\alpha} \sigma_w + \sqrt{F_{\chi^2}^{-1}(\beta; \eta)} \|w - w_{\text{GLS}}\|_\Sigma + B_w^\theta(\Lambda)}{\sigma_{\text{GLS}}} \right) \beta,
\end{aligned}$$

which is the same final bound as the upper one-sided case.

Finally, for the case of a two-sided CI_w^* , observe that

$$\begin{aligned}
\mathbb{P}_\theta \{ \lambda' \theta \notin CI_w^* \} &\geq \mathbb{P}_\theta \{ |\lambda' \theta - w' \hat{\theta}| > z_{1-\alpha/2} \sigma_w + \hat{B}_w^\beta(\Lambda) \} \\
&= \mathbb{P}_\theta \{ \lambda' \theta > w' \hat{\theta} + z_{1-\alpha/2} \sigma_w + \hat{B}_w^\beta(\Lambda) \} + \mathbb{P}_\theta \{ \lambda' \theta < w' \hat{\theta} - z_{1-\alpha/2} \sigma_w - \hat{B}_w^\beta(\Lambda) \} \\
&\geq 2\Phi \left(- \frac{B_{\text{GLS}}^\theta(\Lambda) + z_{1-\alpha/2} \sigma_w + \sqrt{F_{\chi^2}^{-1}(\beta; \eta)} \|w - w_{\text{GLS}}\|_\Sigma + B_w^\theta(\Lambda)}{\sigma_{\text{GLS}}} \right) \beta,
\end{aligned}$$

where the first line follows from the bound $\text{cv}_{1-\alpha}(|b|) \leq z_{1-\alpha/2} + |b|$, and the second from the above derivations in the one-sided case. Thus, the desired coverage bounds hold.

D Proofs of Results for the Uniform Asymptotics

D.1 Proof of Proposition U1

Under Assumptions U4 and U5, there is an event whose probability uniformly approaches one as $n \rightarrow \infty$ on which $\hat{S}_n \in \mathbb{S}$ and Lemma 1 implies

$$d_H(\hat{\Lambda}_n, \Lambda_n) \leq \frac{\text{diam}(\mathbb{W})}{\delta_g} L_g d_S(\hat{S}_n, S_n).$$

Thus, since $d_S(\hat{S}_n, S_n)$ uniformly converges in probability to zero, so does $d_H(\hat{\Lambda}_n, \Lambda_n)$.

D.2 Proof of Proposition U2

Under Assumptions U2-U5, there is an event whose probability uniformly approaches one as $n \rightarrow \infty$ on which $1/(2\bar{e}) \leq e_{\min}(\hat{\Sigma}_n) \leq e_{\max}(\hat{\Sigma}_n) \leq 2\bar{e}$, $\hat{S}_n \in \mathbb{S}$, and Lemma 2 implies

$$\begin{aligned}
\left| \max_{\lambda \in \hat{\Lambda}_n} \|\lambda - \hat{w}_n\|_{\hat{\Sigma}_n} - \max_{\lambda \in \Lambda_n} \|\lambda - w_n\|_{\Sigma_n} \right| &\leq \left| \max_{\lambda \in \hat{\Lambda}_n} \|\lambda - \hat{w}_n\|_{\hat{\Sigma}_n} - \max_{\lambda \in \hat{\Lambda}_n} \|\lambda - w_n\|_{\hat{\Sigma}_n} \right| \\
&\quad + \left| \max_{\lambda \in \hat{\Lambda}_n} \|\lambda - w_n\|_{\hat{\Sigma}_n} - \max_{\lambda \in \Lambda_n} \|\lambda - w_n\|_{\hat{\Sigma}_n} \right| \\
&\quad + \left| \max_{\lambda \in \Lambda_n} \|\lambda - w_n\|_{\hat{\Sigma}_n} - \max_{\lambda \in \Lambda_n} \|\lambda - w_n\|_{\Sigma_n} \right|
\end{aligned}$$

$$\begin{aligned}
&\leq \sqrt{2\bar{e}}\|\hat{w}_n - w_n\| \\
&+ \sqrt{2\bar{e}}d_H(\hat{\Lambda}_n, \Lambda_n) \\
&+ \frac{\sqrt{2\bar{e}}}{2}\|\hat{\Sigma}_n - \Sigma_n\|\left(\max_{\lambda \in \mathbb{W}}\|\lambda\| + \bar{C}_w\right).
\end{aligned}$$

Thus, since the right-hand side (RHS) uniformly converges in probability to zero, so does the left-hand side (LHS).

D.3 Proof of Proposition U3

The boundedness claim follows because $w_n^* \in \Lambda_n \subseteq \mathbb{W}$ is contained in the compact set $\mathbb{W}$ for all $P_n \in \mathcal{P}_n$ and n under the maintained assumptions. I now prove the consistency claim. Under Assumptions U2 and U4, there is an event whose probability uniformly approaches one as $n \rightarrow \infty$ on which $1/(2\bar{e}) \leq e_{\min}(\hat{\Sigma}_n) \leq e_{\max}(\hat{\Sigma}_n) \leq 2\bar{e}$ and $\hat{S}_n \in \mathbb{S}$. The following arguments are made on this event under Assumptions U2, U4, and U5. Define the Euclidean projections

$$\pi_n(\hat{w}_n^*) = \arg \min_{\lambda \in \Lambda_n} \|\hat{w}_n^* - \lambda\|, \quad \hat{\pi}_n(w_n^*) = \arg \min_{\lambda \in \hat{\Lambda}_n} \|w_n^* - \lambda\|.$$

Since $\hat{w}_n^* \in \hat{\Lambda}_n$ and $w_n^* \in \Lambda_n$, the definition of Hausdorff distance yields the bounds

$$\|\hat{w}_n^* - \pi_n(\hat{w}_n^*)\| = \text{dist}(\hat{w}_n^*, \Lambda_n) \leq d_H(\hat{\Lambda}_n, \Lambda_n), \quad \|w_n^* - \hat{\pi}_n(w_n^*)\| = \text{dist}(w_n^*, \hat{\Lambda}_n) \leq d_H(\hat{\Lambda}_n, \Lambda_n).$$

By triangle inequality and Lemma 3,

$$\begin{aligned}
\|\hat{w}_n^* - w_n^*\| &\leq \|\hat{w}_n^* - \pi_n(\hat{w}_n^*)\| + \|\pi_n(\hat{w}_n^*) - w_n^*\| \\
&\leq d_H(\hat{\Lambda}_n, \Lambda_n) + \sqrt{2(2\bar{e})^{3/2} \text{diam}(\mathbb{W}) \left(\max_{\lambda \in \Lambda_n} \|\lambda - \pi_n(\hat{w}_n^*)\|_{\Sigma_n} - \max_{\lambda \in \Lambda_n} \|\lambda - w_n^*\|_{\Sigma_n} \right)}.
\end{aligned}$$

I now look to bound the second term. To this end, first note that Lemma 2 implies

$$\max_{\lambda \in \Lambda_n} \|\lambda - \pi_n(\hat{w}_n^*)\|_{\Sigma_n} \leq \max_{\lambda \in \Lambda_n} \|\lambda - \hat{w}_n^*\|_{\Sigma_n} + \sqrt{2\bar{e}}d_H(\hat{\Lambda}_n, \Lambda_n).$$

By Lemma 2 again,

$$\max_{\lambda \in \Lambda_n} \|\lambda - \hat{w}_n^*\|_{\Sigma_n} \leq \max_{\lambda \in \hat{\Lambda}_n} \|\lambda - \hat{w}_n^*\|_{\hat{\Sigma}_n} + \sqrt{2\bar{e}}d_H(\hat{\Lambda}_n, \Lambda_n) + \frac{\sqrt{2\bar{e}}}{2}\|\hat{\Sigma}_n - \Sigma_n\| \text{diam}(\mathbb{W}).$$

By optimality of $\hat{w}_n^*$ and Lemma 2 again,

$$\max_{\lambda \in \hat{\Lambda}_n} \|\lambda - \hat{w}_n^*\|_{\hat{\Sigma}_n} \leq \max_{\lambda \in \hat{\Lambda}_n} \|\lambda - \hat{\pi}_n(w_n^*)\|_{\hat{\Sigma}_n}$$

$$\leq \max_{\lambda \in \Lambda_n} \|\lambda - \hat{\pi}_n(w_n^*)\|_{\Sigma_n} + \sqrt{2\bar{e}}d_H(\hat{\Lambda}_n, \Lambda_n) + \frac{\sqrt{2\bar{e}}}{2} \|\hat{\Sigma}_n - \Sigma_n\| \text{diam}(\mathbb{W}).$$

Note that Lemma 2 implies

$$\max_{\lambda \in \Lambda_n} \|\lambda - \hat{\pi}_n(w_n^*)\|_{\Sigma_n} \leq \max_{\lambda \in \Lambda_n} \|\lambda - w_n^*\|_{\Sigma_n} + \sqrt{2\bar{e}}d_H(\hat{\Lambda}_n, \Lambda_n).$$

Combining the above steps yields the inequality

$$\max_{\lambda \in \Lambda_n} \|\lambda - \pi_n(\hat{w}_n^*)\|_{\Sigma_n} - \max_{\lambda \in \Lambda_n} \|\lambda - w_n^*\|_{\Sigma_n} \leq 4\sqrt{2\bar{e}}d_H(\hat{\Lambda}_n, \Lambda_n) + \sqrt{2\bar{e}} \|\hat{\Sigma}_n - \Sigma_n\| \text{diam}(\mathbb{W}).$$

In summary, under Assumptions U2, U4, and U5, there is an event whose probability uniformly approaches one as $n \rightarrow \infty$ on which

$$\|\hat{w}_n^* - w_n^*\| \leq d_H(\hat{\Lambda}_n, \Lambda_n) + \sqrt{2(2\bar{e})^{3/2} \text{diam}(\mathbb{W}) \left(4\sqrt{2\bar{e}}d_H(\hat{\Lambda}_n, \Lambda_n) + \sqrt{2\bar{e}} \|\hat{\Sigma}_n - \Sigma_n\| \text{diam}(\mathbb{W}) \right)}.$$

Thus, since the RHS uniformly converges in probability to zero, so does the LHS.

D.4 Proof of Proposition U4

Let $\check{\theta}_n = \hat{\theta}_n - \theta_n$ and $\check{w}_n = \hat{w}_n - w_n$. By the triangle and Cauchy-Schwarz inequalities,

$$\begin{aligned} |\mathbb{E}_{P_n}[\hat{\tau}_{\hat{w}_n, n}] - \tau_{\lambda, n}| &\leq |\mathbb{E}_{P_n}[\check{w}_n' \check{\theta}_n]| + |\mathbb{E}_{P_n}[w_n' \check{\theta}_n]| + |\mathbb{E}_{P_n}[\check{w}_n' \theta_n]| + |(\lambda - w_n)' \theta_n| \\ &\lesssim \sqrt{\mathbb{E}_{P_n}[\|\check{w}_n\|^2] \mathbb{E}_{P_n}[\|\check{\theta}_n\|^2]} + \sqrt{\mathbb{E}_{P_n}[\|\check{\theta}_n\|^2]} + \sqrt{\mathbb{E}_{P_n}[\|\check{w}_n\|^2]} + |(\lambda - w_n)' \theta_n|. \end{aligned}$$

I show at the end of this proof that the UI condition (22) implies

$$\lim_{n \rightarrow \infty} \sup_{P_n \in \mathcal{P}_n} \left(\mathbb{E}_{P_n}[\|\check{\theta}_n\|^2], \mathbb{E}_{P_n}[\|\check{w}_n\|^2] \right) = 0. \quad (36)$$

Combined with Proposition 1, this yields

$$\limsup_{n \rightarrow \infty} \left(\sup_{P_n \in \mathcal{P}_n} \max_{\lambda \in \Lambda_n} |\mathbb{E}_{P_n}[\hat{\tau}_{\hat{w}_n, n}] - \tau_{\lambda, n}| \right) \leq \limsup_{n \rightarrow \infty} \left(\bar{\eta}(\mathcal{P}_n) \sup_{P_n \in \mathcal{P}_n} \max_{\lambda \in \Lambda_n} \|\lambda - w_n\|_{\Sigma_n} \right),$$

where the RHS is finite under Assumptions U1-U5.

For the equality statement, observe that the reverse triangle inequality implies

$$\begin{aligned} \max_{\lambda \in \Lambda_n} |\mathbb{E}_{P_n}[\hat{\tau}_{\hat{w}_n, n}] - \tau_{\lambda, n}| &\geq \max_{\lambda \in \Lambda_n} |(\lambda - w_n)' \theta_n| - |\mathbb{E}_{P_n}[\check{w}_n' \check{\theta}_n] + \mathbb{E}_{P_n}[w_n' \check{\theta}_n] + \mathbb{E}_{P_n}[\check{w}_n' \theta_n]| \\ &\geq \max_{\lambda \in \Lambda_n} |(\lambda - w_n)' \theta_n| - \sup_{P_n \in \mathcal{P}_n} \left(|\mathbb{E}_{P_n}[\check{w}_n' \check{\theta}_n]| + |\mathbb{E}_{P_n}[w_n' \check{\theta}_n]| + |\mathbb{E}_{P_n}[\check{w}_n' \theta_n]| \right), \end{aligned}$$

where (36) and the previous Cauchy-Schwarz bound implies

$$\lim_{n \rightarrow \infty} \sup_{P_n \in \mathcal{P}_n} \left(\left| \mathbb{E}_{P_n} [\check{w}'_n \check{\theta}_n] \right| + \left| \mathbb{E}_{P_n} [w'_n \check{\theta}_n] \right| + \left| \mathbb{E}_{P_n} [\check{w}'_n \theta_n] \right| \right) = 0.$$

If $P_n^\dagger(w_n) \in \mathcal{P}_n$ for all $P_n \in \mathcal{P}_n$ and n , then for all $P_n \in \mathcal{P}_n$ and n ,

$$\begin{aligned} \sup_{P_n \in \mathcal{P}_n} \max_{\lambda \in \Lambda_n} |(\lambda - w_n)' \theta_n| &\geq \max_{\lambda \in \Lambda_n} |(\lambda - w_n)' \theta(P_n^\dagger(w_n))| = |(\lambda_n^*(w_n) - w_n)' \theta(P_n^\dagger(w_n))| \\ &= \bar{\eta}(\mathcal{P}_n) \|\lambda_n^*(w_n) - w_n\|_{\Sigma_n} \\ &= \bar{\eta}(\mathcal{P}_n) \max_{\lambda \in \Lambda_n} \|\lambda - w_n\|_{\Sigma_n}. \end{aligned}$$

Taking the supremum over $P_n \in \mathcal{P}_n$ and combining with the previous displays yields

$$\limsup_{n \rightarrow \infty} \left(\sup_{P_n \in \mathcal{P}_n} \max_{\lambda \in \Lambda_n} |\mathbb{E}_{P_n} [\hat{\tau}_{\hat{w}_n, n}] - \tau_{\lambda, n}| \right) \geq \limsup_{n \rightarrow \infty} \left(\bar{\eta}(\mathcal{P}_n) \sup_{P_n \in \mathcal{P}_n} \max_{\lambda \in \Lambda_n} \|\lambda - w_n\|_{\Sigma_n} \right),$$

from which the equality statement follows.

I now verify (36). Note that $\check{\theta}_n$ and $\check{w}_n$ are uniformly consistent for zero, given Assumptions U1 and U3. Focusing on $\check{w}_n$ for the moment, observe that for any $C > 0$,

$$\mathbb{E}_{P_n} [\|\check{w}_n\|^2] \leq \left| \mathbb{E}_{P_n} [\|\check{w}_n\|^2] - \mathbb{E}_{P_n} [\min(\|\check{w}_n\|^2, C)] \right| + \left| \mathbb{E}_{P_n} [\min(\|\check{w}_n\|^2, C)] \right|.$$

Since $x \mapsto \min(\|x\|^2, C)$ is bounded and continuous, uniform convergence in probability of $\check{w}_n$ to zero implies

$$\lim_{n \rightarrow \infty} \sup_{P_n \in \mathcal{P}_n} \left| \mathbb{E}_{P_n} [\min(\|\check{w}_n\|^2, C)] \right| = 0.$$

Since $\left| \mathbb{E}_{P_n} [\|\check{w}_n\|^2] - \mathbb{E}_{P_n} [\min(\|\check{w}_n\|^2, C)] \right| \leq \mathbb{E}_{P_n} [\|\check{w}_n\|^2 \mathbf{1}\{\|\check{w}_n\|^2 > C\}]$, then

$$\limsup_{n \rightarrow \infty} \sup_{P_n \in \mathcal{P}_n} \mathbb{E}_{P_n} [\|\check{w}_n\|^2] \leq \limsup_{n \rightarrow \infty} \sup_{P_n \in \mathcal{P}_n} \mathbb{E}_{P_n} [\|\check{w}_n\|^2 \mathbf{1}\{\|\check{w}_n\|^2 > C\}].$$

The above arguments also apply to $\check{\theta}_n$. Taking $C \rightarrow \infty$, the UI condition (22) implies

$$\limsup_{n \rightarrow \infty} \sup_{P_n \in \mathcal{P}_n} \left(\mathbb{E}_{P_n} [\|\check{w}_n\|^2], \mathbb{E}_{P_n} [\|\check{\theta}_n\|^2] \right) \leq 0.$$

In particular, (36) holds.

D.5 Proof of Proposition U5

Under these conditions, Proposition U4 can be applied to both $\hat{w}_n^*$ and $\hat{w}_n$. In particular,

$$\begin{aligned} \limsup_{n \rightarrow \infty} \left(\sup_{P_n \in \mathcal{P}_n} \max_{\lambda \in \Lambda_n} |\mathbb{E}_{P_n}[\hat{\tau}_n^*] - \tau_{\lambda,n}| \right) &= \limsup_{n \rightarrow \infty} \left(\bar{\eta}(\mathcal{P}_n) \sup_{P_n \in \mathcal{P}_n} \max_{\lambda \in \Lambda_n} \|\lambda - w_n^*\|_{\Sigma_n} \right) \\ &\leq \limsup_{n \rightarrow \infty} \left(\bar{\eta}(\mathcal{P}_n) \sup_{P_n \in \mathcal{P}_n} \max_{\lambda \in \Lambda_n} \|\lambda - w_n\|_{\Sigma_n} \right) \\ &= \limsup_{n \rightarrow \infty} \left(\sup_{P_n \in \mathcal{P}_n} \max_{\lambda \in \Lambda_n} |\mathbb{E}_{P_n}[\hat{\tau}_{\hat{w}_n,n}] - \tau_{\lambda,n}| \right), \end{aligned}$$

where the second inequality follows from $\max_{\lambda \in \Lambda_n} \|\lambda - w_n^*\|_{\Sigma_n} \leq \max_{\lambda \in \Lambda_n} \|\lambda - w_n\|_{\Sigma_n}$.

D.6 Proof of Proposition U6

If (23) does not hold, then there exists a $p > 0$, a subsequence $\{n_s\} \subseteq \{n\}$, and a corresponding sequence of distributions $P_{n_s} \in \mathcal{P}_{n_s}$ such that

$$\left| \mathbb{P}_{P_{n_s}} \left\{ \hat{H}_{n_s}(\theta_{n_s}) \leq \hat{\eta}_{1-\beta, n_s} \right\} - (1-\beta) \right| \mathbf{1} \{ H_{n_s}(\theta_{n_s}) > 0 \} > p, \quad \forall s.$$

Thus, to prove Proposition U6, it suffices to show that the above yields a contradiction under Assumptions U1 and U2. To do so, I will iteratively pass to (sub)subsequences of $\{n_s\}$ along which variables of interest converge under Assumptions U1 and U2.

For conciseness, I will keep using the index s when passing to subsequences, and will also use s to abbreviate the subscript n_s . In this notation, the above becomes

$$\left| \mathbb{P}_{P_s} \left\{ \sqrt{n_s} \hat{H}_s(\theta_s) \leq \tilde{\eta}_{1-\beta, s} \right\} - (1-\beta) \right| \mathbf{1} \{ \sqrt{n_s} H_s(\theta_s) > 0 \} > p, \quad \forall s.$$

This implies $H_s(\theta_s) > 0$ and hence $\hat{H}_s(\theta_s) > 0$ for all s . Thus, using similar arguments to those in Appendix C.2,

$$\left| \mathbb{P}_{P_s} \left\{ F_{\chi^2}(n_s \hat{H}_s^2(\hat{\theta}_s); \sqrt{n_s} \hat{H}_s(\theta_s)) < \beta \right\} - \beta \right| > p, \quad \forall s. \quad (37)$$

It suffices to derive a contradiction from (37). Let $\hat{Z}_s = \sqrt{n_s}(\hat{\theta}_s - \theta_s)$. Under Assumptions U1 and U2, I pass to a subsequence where

$$\Sigma_s \rightarrow \Sigma_\infty, \quad \|\Sigma_\infty\| \in (0, \infty), \quad \hat{\Sigma}_s \xrightarrow{p} \Sigma_\infty, \quad \hat{Z}_s \xrightarrow{d} Z_\infty \sim N(0, \Sigma_\infty).$$

By continuous mapping theorem (CMT), analogous convergences hold for $(\hat{A}_s, A_s, \hat{Q}_s, Q_s)$ relative to (A_∞, Q_∞) , etc. Let $v_s = A_s \Sigma_s^{-1/2} \sqrt{n_s} \theta_s$ so that $\sqrt{n_s} H_s(\theta_s) = \|v_s\|$. By compactness,

I pass to a further subsequence where

$$v_s \rightarrow v_\infty \in [-\infty, \infty]^K, \quad \|v_s\| = \sqrt{n_s} H_s(\theta_s) \rightarrow \eta_\infty = \|v_\infty\| \in [0, \infty], \quad \frac{v_s}{\|v_s\|} \rightarrow u_\infty, \quad \|u_\infty\| = 1.$$

Finally, let $\hat{v}_s = \hat{A}_s \hat{\Sigma}_s^{-1/2} \sqrt{n_s} \theta_s$ so that $\sqrt{n_s} \hat{H}_s(\theta_s) = \|\hat{v}_s\|$. To establish a contradiction of (37), it suffices to establish contradictions in both the case of $\eta_\infty < \infty$ and the case of $\eta_\infty = \infty$.

D.6.1 Case 1: Bounded Normalized Heterogeneity

For the case of $\eta_\infty < \infty$, I appeal to CMT. In particular, if I can show that $\hat{v}_s \xrightarrow{p} v_\infty$, then since $\hat{Q}_s = \hat{\Sigma}_s^{-1/2} \hat{A}_s \hat{\Sigma}_s^{-1/2}$ with $\hat{A}_s \hat{A}_s = \hat{A}_s$ and $\hat{A}_s \hat{v}_s = \hat{v}_s$, this would yield

$$\begin{aligned} n_s \hat{H}_s^2(\hat{\theta}_s) &= (\hat{Z}_s + \sqrt{n_s} \theta_s)' \hat{Q}_s (\hat{Z}_s + \sqrt{n_s} \theta_s) = \left\| \hat{A}_s (\hat{\Sigma}_s^{-1/2} \hat{Z}_s + \hat{v}_s) \right\|^2 \\ &\xrightarrow{d} \left\| A_\infty (\Sigma_\infty^{-1/2} Z_\infty + v_\infty) \right\|^2 \sim \chi_{K-1}^2(\|v_\infty\|^2). \end{aligned}$$

Thus, since $(x, \eta) \mapsto F_{\chi^2}(x; \eta)$ is continuous, another application of CMT implies

$$F_{\chi^2}(n_s \hat{H}_s^2(\hat{\theta}_s); \sqrt{n_s} \hat{H}_s(\theta_s)) \xrightarrow{d} F_{\chi^2}(\chi_{K-1}^2(\|v_\infty\|^2); \|v_\infty\|) \sim U(0, 1),$$

which contradicts (37).

I now show that $\hat{v}_s \xrightarrow{p} v_\infty$. For $\hat{\Sigma}_s^{1/2} \hat{v}_s = \sqrt{n_s} \theta_s - \mathbf{1} \hat{\gamma}_s$, where $\hat{\gamma}_s = \mathbf{1}' \hat{\Sigma}_s^{-1} \sqrt{n_s} \theta_s / (\mathbf{1}' \hat{\Sigma}_s^{-1} \mathbf{1})$ is the coefficient from regressing $\hat{\Sigma}_s^{-1/2} \sqrt{n_s} \theta_s$ on $\hat{\Sigma}_s^{-1/2} \mathbf{1}$, and likewise for $\Sigma_s^{1/2} v_s = \sqrt{n_s} \theta_s - \mathbf{1} \gamma_s$ and $\gamma_s = \mathbf{1}' \Sigma_s^{-1} \sqrt{n_s} \theta_s / (\mathbf{1}' \Sigma_s^{-1} \mathbf{1})$, observe that

$$\begin{aligned} \|\hat{v}_s - v_s\| &= \left\| \hat{\Sigma}_s^{-1/2} (\sqrt{n_s} \theta_s - \mathbf{1} \hat{\gamma}_s) - \Sigma_s^{-1/2} (\sqrt{n_s} \theta_s - \mathbf{1} \gamma_s) \right\| \\ &= \left\| (\hat{\Sigma}_s^{-1/2} - \Sigma_s^{-1/2}) (\sqrt{n_s} \theta_s - \mathbf{1} \gamma_s) - \hat{\Sigma}_s^{-1/2} \mathbf{1} (\hat{\gamma}_s - \gamma_s) \right\| \\ &\lesssim_p \left\| \hat{\Sigma}_s - \Sigma_s \right\| \left\| \sqrt{n_s} \theta_s - \mathbf{1} \gamma_s \right\| + |\hat{\gamma}_s - \gamma_s|. \end{aligned}$$

Note $\left\| \sqrt{n_s} \theta_s - \mathbf{1} \gamma_s \right\| = \left\| \Sigma_s^{1/2} v_s \right\| \lesssim \|v_s\| = O(1)$ and

$$|\hat{\gamma}_s - \gamma_s| = \left| \frac{\mathbf{1}' \hat{\Sigma}_s^{-1} (\sqrt{n_s} \theta_s - \mathbf{1} \gamma_s)}{\mathbf{1}' \hat{\Sigma}_s^{-1} \mathbf{1}} \right| = \left| \frac{\mathbf{1}' (\hat{\Sigma}_s^{-1} - \Sigma_s^{-1}) (\sqrt{n_s} \theta_s - \mathbf{1} \gamma_s)}{\mathbf{1}' \hat{\Sigma}_s^{-1} \mathbf{1}} \right| \lesssim_p \left\| \hat{\Sigma}_s - \Sigma_s \right\| \|v_s\| = o_p(1).$$

Altogether, $\|\hat{v}_s - v_s\| = o_p(1)$. But since $v_s \rightarrow v_\infty$, it then follows that $\hat{v}_s \xrightarrow{p} v_\infty$.

D.6.2 Case 2: Unbounded Normalized Heterogeneity

For the case of $\eta_\infty = \infty$, I appeal to noncentral chi-squared asymptotics. For setup, define the standardized noncentral chi-squared quantile

$$R(\beta; \eta) = \frac{F_{\chi^2}^{-1}(\beta; \eta) - (\eta^2 + (K-1))}{\sqrt{4\eta^2 + 2(K-1)}},$$

which satisfies

$$R(\beta; \eta) = G^{-1}(\beta; \eta), \quad G(x; \eta) = F_{\chi^2} \left(\eta^2 + (K-1) + \sqrt{4\eta^2 + 2(K-1)}x; \eta \right),$$

where [Seri \(2015, III.2\)](#) implies

$$\kappa(\eta) = \sup_x |G(x; \eta) - \Phi(x)| = \sup_x \left| F_{\chi^2}(x; \eta) - \Phi \left(\frac{x - (\eta^2 + (K-1))}{\sqrt{4\eta^2 + 2(K-1)}} \right) \right| = O \left(\frac{1}{\sqrt{\eta^2 + (K-1)}} \right).$$

Moreover, $R(\beta; \eta) \rightarrow z_\beta$ as $\eta \rightarrow \infty$. Indeed, $|G(x; \eta) - \Phi(x)| \leq \kappa(\eta)$ at $x = \Phi^{-1}(\beta \pm \kappa(\eta))$ implies $G(\Phi^{-1}(\beta - \kappa(\eta)); \eta) \leq \beta \leq G(\Phi^{-1}(\beta + \kappa(\eta)); \eta)$ for $\kappa(\eta) < \min\{\beta, 1 - \beta\}$, and therefore $\Phi^{-1}(\beta - \kappa(\eta)) \leq R(\beta; \eta) \leq \Phi^{-1}(\beta + \kappa(\eta))$ since $x \mapsto G(x; \eta)$ is continuous and strictly increasing. Thus, since $\kappa(\eta) \rightarrow 0$ as $\eta \rightarrow \infty$ and $\beta \mapsto \Phi^{-1}(\beta) = z_\beta$ is continuous at $\beta \in (0, 1)$, this implies $R(\beta; \eta) \rightarrow z_\beta$ as $\eta \rightarrow \infty$.

Letting $\hat{c}_{1s} = \|\hat{v}_s\|^2 + (K-1)$ and $\hat{c}_{2s} = \sqrt{4\|\hat{v}_s\|^2 + 2(K-1)}$, observe that

$$\begin{aligned} \mathbb{P}_{P_s} \left\{ F_{\chi^2}(n_s \hat{H}_s^2(\hat{\theta}_s); \sqrt{n_s} \hat{H}_s(\theta_s)) < \beta \right\} &= \mathbb{P}_{P_s} \left\{ n_s \hat{H}_s^2(\hat{\theta}_s) < F_{\chi^2}^{-1}(\beta; \sqrt{n_s} \hat{H}_s(\theta_s)) \right\} \\ &= \mathbb{P}_{P_s} \left\{ \frac{\|\hat{A}_s \hat{\Sigma}_s^{-1/2} \hat{Z}_s + \hat{v}_s\|^2 - \hat{c}_{1s}}{\hat{c}_{2s}} < R(\beta; \|\hat{v}_s\|) \right\}. \end{aligned}$$

If I can show that $\|\hat{v}_s\| \xrightarrow{P} \infty$ and $\hat{v}_s / \|\hat{v}_s\| \xrightarrow{P} u_\infty$, then

$$\frac{\|\hat{A}_s \hat{\Sigma}_s^{-1/2} \hat{Z}_s + \hat{v}_s\|^2 - \|\hat{v}_s\|^2}{2\|\hat{v}_s\|} = \frac{\|\hat{A}_s \hat{\Sigma}_s^{-1/2} \hat{Z}_s\|^2}{2\|\hat{v}_s\|} + \left\langle \frac{\hat{v}_s}{\|\hat{v}_s\|}, \hat{\Sigma}_s^{-1/2} \hat{Z}_s \right\rangle \xrightarrow{d} \left\langle u_\infty, \Sigma_\infty^{-1/2} Z_\infty \right\rangle \sim N(0, 1),$$

and moreover

$$\left| \frac{\|\hat{A}_s \hat{\Sigma}_s^{-1/2} \hat{Z}_s + \hat{v}_s\|^2 - \|\hat{v}_s\|^2}{2\|\hat{v}_s\|} - \frac{\|\hat{A}_s \hat{\Sigma}_s^{-1/2} \hat{Z}_s + \hat{v}_s\|^2 - \hat{c}_{1s}}{\hat{c}_{2s}} \right| \xrightarrow{P} 0,$$

which then implies

$$\frac{\left\|\hat{A}_s\hat{\Sigma}_s^{-1/2}\hat{Z}_s+\hat{v}_s\right\|^2-\hat{c}_{1s}}{\hat{c}_{2s}}\xrightarrow{d}N(0,1).$$

Furthermore, $R(\beta;\|\hat{v}_s\|)\xrightarrow{p}z_\beta$. Indeed, since $R(\beta;\eta)\rightarrow z_\beta$ as $\eta\rightarrow\infty$, then for each $\varepsilon>0$, I can choose $C_\varepsilon>0$ large so that $|R(\beta;\eta)-z_\beta|\leq\varepsilon$ for any $\eta\geq C_\varepsilon$. Thus, $\|\hat{v}_s\|\xrightarrow{p}\infty$ implies $\mathbb{P}_{P_s}\{|R(\beta;\|\hat{v}_s\|)-z_\beta|>\varepsilon\}\leq\mathbb{P}_{P_s}\{\|\hat{v}_s\|<C_\varepsilon\}\rightarrow 0$, and hence $R(\beta;\|\hat{v}_s\|)\xrightarrow{p}z_\beta$. Altogether,

$$\mathbb{P}_{P_s}\left\{\frac{\left\|\hat{A}_s\hat{\Sigma}_s^{-1/2}\hat{Z}_s+\hat{v}_s\right\|^2-\hat{c}_{1s}}{\hat{c}_{2s}}<R(\beta;\|\hat{v}_s\|)\right\}\rightarrow\mathbb{P}_\theta\{N(0,1)<z_\beta\}=\beta,$$

which is a contradiction of (37).

I now show that $\|\hat{v}_s\|\xrightarrow{p}\infty$ and $\hat{v}_s/\|\hat{v}_s\|\xrightarrow{p}u_\infty$. To show $\|\hat{v}_s\|\xrightarrow{p}\infty$, note that

$$\left|\|\hat{v}_s\|-\|v_s\|\right|\leq\left\|\sqrt{n_s}(\hat{A}_s\hat{\Sigma}_s^{-1/2}-A_s\Sigma_s^{-1/2})\theta_s\right\|\lesssim_p\sqrt{n_s}\left\|\hat{\Sigma}_s-\Sigma_s\right\|\left\|\theta_s\right\|=O_p(1)O(1)=O_p(1).$$

Thus, for each $\varepsilon>0$, there exists $C_\varepsilon>0$ such that $\limsup_s\mathbb{P}_{P_s}\{\|\hat{v}_s\|-\|v_s\|<-C_\varepsilon\}\leq\varepsilon$. As a result, for any $C>0$, a union bound yields

$$\limsup_{s\rightarrow\infty}\mathbb{P}_{P_s}\{\|\hat{v}_s\|\leq C\}\leq\varepsilon+\limsup_{s\rightarrow\infty}\mathbb{P}_{P_s}\{\|v_s\|\leq C_\varepsilon+C\}.$$

Since $\|v_s\|\rightarrow\eta_\infty=\infty$, the second term on the RHS is zero. Taking $\varepsilon\downarrow 0$ implies the LHS is zero for any $C>0$. Thus, $\|\hat{v}_s\|\xrightarrow{p}\infty$. Now it remains to show $\hat{v}_s/\|\hat{v}_s\|\xrightarrow{p}u_\infty$. And indeed,

$$\left\|\frac{\hat{v}_s}{\|\hat{v}_s\|}-\frac{v_s}{\|v_s\|}\right\|=\left\|\frac{\hat{v}_s(\|v_s\|-\|\hat{v}_s\|)+(\hat{v}_s-v_s)\|\hat{v}_s\|}{\|\hat{v}_s\|\|v_s\|}\right\|\lesssim_p\frac{\sqrt{n_s}\left\|\hat{\Sigma}_s-\Sigma_s\right\|\left\|\theta_s\right\|}{\|v_s\|}=\frac{O_p(1)O(1)}{\|v_s\|}\xrightarrow{p}0.$$

Thus, $\hat{v}_s/\|\hat{v}_s\|\xrightarrow{p}u_\infty$ follows from $v_s/\|v_s\|\rightarrow u_\infty$. This concludes the proof.

D.7 Proof of Proposition U7

If (26) does not hold, then there exists $p>0$, subsequence $\{n_s\}\subseteq\{n\}$, sequence of distributions $P_{n_s}\in\mathcal{P}_{n_s}$, and sequence of alternative weights $\lambda_{n_s}\in\Lambda_{0,n_s}$ such that

$$\mathbb{P}_{P_{n_s}}\left\{\lambda'_{n_s}\theta_{n_s}\in CI_{\hat{w}_{n_s},n_s}^*\right\}-(1-\alpha-\beta)<-p,\quad\forall s.$$

Thus, to prove Proposition U7, it suffices to show that the above yields a contradiction under Assumptions U1-U5. To do so, I will iteratively pass to (sub)subsequences of $\{n_s\}$ along which variables of interest converge under Assumptions U1-U5.

For conciseness, I will keep using the index s when passing to subsequences, and will also use s to abbreviate the subscript n_s . In this notation, the above becomes

$$\mathbb{P}_{P_s} \left\{ \lambda'_s \theta_s \notin CI_{\hat{w}_s, s}^* \right\} > p + (\alpha + \beta), \quad \forall s. \quad (38)$$

On event $\mathcal{E}_s = \left\{ \hat{H}_s(\theta_s) \max_{\lambda \in \hat{\Lambda}_s} \|\lambda - \hat{w}_s\|_{\hat{\Sigma}_s} \leq \hat{B}_{\hat{w}_s, s}^\beta(\hat{\Lambda}_s) \right\}$, observe that

$$\begin{aligned} \mathbb{P}_{P_s} \left\{ \lambda'_s \theta_s \notin CI_{\hat{w}_s, s}^* \right\} &\leq \mathbb{P}_{P_s} \left\{ \lambda'_s \theta_s \notin CI_{\hat{w}_s, s}^*, \mathcal{E}_s \right\} + \mathbb{P}_{P_s} \left\{ \mathcal{E}_s^c \right\} \\ &\leq \mathbb{P}_{P_s} \left\{ \lambda'_s \theta_s \notin CI_{\hat{w}_s, s}^*, \lambda_s \in \hat{\Lambda}_s, \mathcal{E}_s \right\} + \mathbb{P}_{P_s} \left\{ \lambda_s \notin \hat{\Lambda}_s \right\} + \mathbb{P}_{P_s} \left\{ \mathcal{E}_s^c \right\}, \end{aligned}$$

where since $\hat{B}_{\hat{w}_s, s}^\beta(\hat{\Lambda}_s) = \hat{\eta}_{1-\beta, s} \max_{\lambda \in \hat{\Lambda}_s} \|\lambda - \hat{w}_s\|_{\hat{\Sigma}_s}$,

$$\mathbb{P}_{P_s} \left\{ \mathcal{E}_s^c \right\} = \mathbb{P}_{P_s} \left\{ \hat{H}_s(\theta_s) \max_{\lambda \in \hat{\Lambda}_s} \|\lambda - \hat{w}_s\|_{\hat{\Sigma}_s} > \hat{\eta}_{1-\beta, s} \max_{\lambda \in \hat{\Lambda}_s} \|\lambda - \hat{w}_s\|_{\hat{\Sigma}_s} \right\} \leq \mathbb{P}_{P_s} \left\{ \hat{H}_s(\theta_s) > \hat{\eta}_{1-\beta, s} \right\}.$$

Proposition U6 and containment condition (25) yield, along any subsequence,

$$\limsup_{s \rightarrow \infty} \mathbb{P}_{P_s} \left\{ \hat{H}_s(\theta_s) > \hat{\eta}_{1-\beta, s} \right\} \leq \beta, \quad \limsup_{s \rightarrow \infty} \mathbb{P}_{P_s} \left\{ \lambda_s \notin \hat{\Lambda}_s \right\} = 0.$$

Thus, to establish a contradiction of (38), it suffices to find a subsequence where

$$\limsup_{s \rightarrow \infty} \mathbb{P}_{P_s} \left\{ \lambda'_s \theta_s \notin CI_{\hat{w}_s, s}^*, \lambda_s \in \hat{\Lambda}_s, \mathcal{E}_s \right\} \leq \alpha. \quad (39)$$

I first consider the upper one-sided $CI_{\hat{w}_s, s}^*$ —the lower one-sided case follows analogously.

$$\begin{aligned} \mathbb{P}_{P_s} \left\{ \lambda'_s \theta_s \notin CI_{\hat{w}_s, s}^*, \lambda_s \in \hat{\Lambda}_s, \mathcal{E}_s \right\} &\leq \mathbb{P}_{P_s} \left\{ \frac{\lambda'_s \theta_s - \hat{w}_s' \hat{\theta}_s}{\tilde{\sigma}_{\hat{w}_s, s}} > z_{1-\alpha} + \max_{\lambda \in \hat{\Lambda}_s} \frac{\hat{H}_s(\theta_s) \|\lambda - \hat{w}_s\|_{\hat{\Sigma}_s}}{\tilde{\sigma}_{\hat{w}_s, s}}, \lambda_s \in \hat{\Lambda}_s \right\} \\ &\leq \mathbb{P}_{P_s} \left\{ \frac{\lambda'_s \theta_s - \hat{w}_s' \hat{\theta}_s}{\tilde{\sigma}_{\hat{w}_s, s}} > z_{1-\alpha} + \max_{\lambda \in \hat{\Lambda}_s} \frac{|(\lambda - \hat{w}_s)' \theta_s|}{\tilde{\sigma}_{\hat{w}_s, s}}, \lambda_s \in \hat{\Lambda}_s \right\} \\ &\leq \mathbb{P}_{P_s} \left\{ \frac{\lambda'_s \theta_s - \hat{w}_s' \hat{\theta}_s}{\tilde{\sigma}_{\hat{w}_s, s}} > z_{1-\alpha} + \frac{|(\lambda_s - \hat{w}_s)' \theta_s|}{\tilde{\sigma}_{\hat{w}_s, s}} \right\} \\ &\leq \mathbb{P}_{P_s} \left\{ \frac{\hat{w}_s' \hat{Z}_s}{\tilde{\sigma}_{\hat{w}_s, s}} < z_\alpha \right\}, \quad \hat{Z}_s = \sqrt{n_s}(\hat{\theta}_s - \theta_s), \end{aligned}$$

where the second line follows from Proposition 1. Under Assumptions U1-U4, I pass to a

subsequence where

$$\begin{pmatrix} \Sigma_s \\ w_s \\ S_s \end{pmatrix} \rightarrow \begin{pmatrix} \Sigma_\infty \\ w_\infty \\ S_\infty \end{pmatrix}, \quad \|\Sigma_\infty\|, \|w_\infty\| \in (0, \infty), \quad S_\infty \in \mathbb{S}, \quad \begin{pmatrix} \hat{\Sigma}_s \\ \hat{w}_s \\ \hat{S}_s \end{pmatrix} \xrightarrow{p} \begin{pmatrix} \Sigma_\infty \\ w_\infty \\ S_\infty \end{pmatrix}, \quad \hat{Z}_s \xrightarrow{d} Z_\infty \sim N(0, \Sigma_\infty).$$

CMT implies $\hat{\sigma}_{\hat{w}_s, s} \xrightarrow{p} \sigma_\infty = \sqrt{w'_\infty \Sigma_\infty w_\infty} \in (0, \infty)$ and $\hat{w}'_s \hat{Z}_s / \hat{\sigma}_{\hat{w}_s, s} \xrightarrow{d} N(0, 1)$, from which

$$\limsup_{s \rightarrow \infty} \mathbb{P}_{P_s} \left\{ \lambda'_s \theta_s \notin CI_{\hat{w}_s, s}^*, \lambda_s \in \hat{\Lambda}_s, \mathcal{E}_s \right\} \leq \limsup_{s \rightarrow \infty} \mathbb{P}_{P_s} \left\{ \frac{\hat{w}'_s \hat{Z}_s}{\hat{\sigma}_{\hat{w}_s, s}} < z_\alpha \right\} = \mathbb{P}\{N(0, 1) < z_\alpha\} = \alpha,$$

which establishes (39), and hence a contradiction of (38).

I now consider the two-sided $CI_{\hat{w}_s, s}^*$, staying within the previous subsequence to maintain the above convergences relative to $(\Sigma_\infty, w_\infty, S_\infty, Z_\infty)$. As in Appendix D.6, let $v_s = A_s \Sigma_s^{-1/2} \sqrt{n_s} \theta_s$ and $\hat{v}_s = \hat{A}_s \hat{\Sigma}_s^{-1/2} \sqrt{n_s} \theta_s$ so that $\sqrt{n_s} H_s(\theta_s) = \|v_s\|$ and $\sqrt{n_s} \hat{H}_s(\theta_s) = \|\hat{v}_s\|$. By compactness, I pass to a further subsequence where

$$v_s \rightarrow v_\infty \in [-\infty, \infty]^K, \quad \|v_s\| = \sqrt{n_s} H_s(\theta_s) \rightarrow \eta_\infty = \|v_\infty\| \in [0, \infty].$$

From here it suffices to, separately for the case of $\eta_\infty < \infty$ and the case of $\eta_\infty = \infty$, find a further subsequence where (39) holds. In what follows, denote

$$\begin{aligned} \hat{\Delta}_s &= (\lambda_s - \hat{w}_s)' \sqrt{n_s} \theta_s = (\lambda_s - \hat{w}_s)' \Sigma_s^{1/2} v_s, & \hat{D}_s &= \max_{\lambda \in \hat{\Lambda}_s} \|\lambda - \hat{w}_s\|_{\hat{\Sigma}_s}, \\ \Delta_s &= (\lambda_s - w_s)' \sqrt{n_s} \theta_s = (\lambda_s - w_s)' \Sigma_s^{1/2} v_s, & D_s &= \max_{\lambda \in \Lambda_s} \|\lambda - w_s\|_{\Sigma_s}. \end{aligned}$$

In this notation,

$$\begin{aligned} \mathbb{P}_{P_s} \left\{ \lambda'_s \theta_s \notin CI_{\hat{w}_s, s}^*, \lambda_s \in \hat{\Lambda}_s, \mathcal{E}_s \right\} &\leq \mathbb{P}_{P_s} \left\{ \left| \frac{\lambda'_s \theta_s - \hat{w}'_s \hat{\theta}_s}{\tilde{\sigma}_{\hat{w}_s, s}} \right| > \text{cv}_{1-\alpha} \left(\frac{\hat{H}_s(\theta_s) \hat{D}_s}{\tilde{\sigma}_{\hat{w}_s, s}} \right), \lambda_s \in \hat{\Lambda}_s \right\} \\ &= \mathbb{P}_{P_s} \left\{ \left| \frac{\hat{\Delta}_s}{\hat{\sigma}_{\hat{w}_s, s}} - \frac{\hat{w}'_s \hat{Z}_s}{\hat{\sigma}_{\hat{w}_s, s}} \right| > \text{cv}_{1-\alpha} \left(\frac{\|\hat{v}_s\| \hat{D}_s}{\hat{\sigma}_{\hat{w}_s, s}} \right), \lambda_s \in \hat{\Lambda}_s \right\}. \end{aligned} \quad (40)$$

D.7.1 Case 1: Bounded Normalized Heterogeneity

For the case of $\eta_\infty < \infty$, observe that

$$|\hat{\Delta}_s - \Delta_s| = |(\hat{w}_s - w_s)' \Sigma_s^{1/2} v_s| \lesssim \|\hat{w}_s - w_s\| \|v_s\| \xrightarrow{p} 0, \quad |\Delta_s| \lesssim \|v_s\| = O(1).$$

By compactness, I pass to a subsequence where $\Delta_s \rightarrow \Delta_\infty \in \mathbb{R}$, from which the above implies $\hat{\Delta}_s \xrightarrow{p} \Delta_\infty$. Using Proposition 1 (under $\lambda_s \in \hat{\Lambda}_s$) to bound (40), it suffices to show that

$$\limsup_{s \rightarrow \infty} \mathbb{P}_{P_s} \left\{ \left| \frac{\hat{\Delta}_s}{\hat{\sigma}_{\hat{w}_s, s}} - \frac{\hat{w}'_s \hat{Z}_s}{\hat{\sigma}_{\hat{w}_s, s}} \right| > \text{cv}_{1-\alpha} \left(\frac{|\hat{\Delta}_s|}{\hat{\sigma}_{\hat{w}_s, s}} \right) \right\} \leq \alpha. \quad (41)$$

And indeed, for $\sigma_\infty = \sqrt{w'_\infty \Sigma_\infty w_\infty}$, CMT implies

$$\left| \frac{\hat{\Delta}_s}{\hat{\sigma}_{\hat{w}_s, s}} - \frac{\hat{w}'_s \hat{Z}_s}{\hat{\sigma}_{\hat{w}_s, s}} \right| - \text{cv}_{1-\alpha} \left(\frac{|\hat{\Delta}_s|}{\hat{\sigma}_{\hat{w}_s, s}} \right) \xrightarrow{d} \left| \frac{\Delta_\infty}{\sigma_\infty} - N(0,1) \right| - \text{cv}_{1-\alpha} \left(\frac{|\Delta_\infty|}{\sigma_\infty} \right),$$

By definition of $\text{cv}_{1-\alpha}(|b|)$, (41) follows with equality (and with $\lim_{s \rightarrow \infty}$). This establishes (39), and hence a contradiction of (38).

D.7.2 Case 2: Unbounded Normalized Heterogeneity

For the case of $\eta_\infty = \infty$, I appeal to noncentral chi-squared asymptotics. For setup, note that $\|v_s\| \rightarrow \eta_\infty = \infty$ implies $\|\hat{v}_s\| \xrightarrow{p} \infty$, as shown in Appendix D.6. Moreover, for $\Lambda_\infty = \Lambda(S_\infty)$, compactness and Proposition U2 imply

$$D_s \rightarrow D_\infty, \quad \hat{D}_s \xrightarrow{p} D_\infty, \quad D_\infty = \max_{\lambda \in \Lambda_\infty} \|\lambda - w_\infty\|_{\Sigma_\infty} \in (0, \infty),$$

where $D_\infty > 0$ follows from Lemma 4 under the maintained assumptions. Thus, $\|\hat{v}_s\| \hat{D}_s \xrightarrow{p} \eta_\infty D_\infty = \infty$. Now consider that (40) can be bounded as

$$\mathbb{P}_{P_s} \left\{ \frac{\left(\frac{\hat{\Delta}_s - \hat{w}'_s \hat{Z}_s}{\hat{\sigma}_{\hat{w}_s, s}} \right)^2 - \left(\frac{\|\hat{v}_s\| \hat{D}_s}{\hat{\sigma}_{\hat{w}_s, s}} \right)^2}{\frac{\|\hat{v}_s\| \hat{D}_s}{\hat{\sigma}_{\hat{w}_s, s}}} > \frac{\text{cv}_{1-\alpha} \left(\frac{\|\hat{v}_s\| \hat{D}_s}{\hat{\sigma}_{\hat{w}_s, s}} \right)^2 - \left(\frac{\|\hat{v}_s\| \hat{D}_s}{\hat{\sigma}_{\hat{w}_s, s}} \right)^2}{\frac{\|\hat{v}_s\| \hat{D}_s}{\hat{\sigma}_{\hat{w}_s, s}}}, \lambda_s \in \hat{\Lambda}_s \right\}. \quad (42)$$

The expression on the RHS of the inequality is equal to

$$\left(\text{cv}_{1-\alpha} \left(\frac{\|\hat{v}_s\| \hat{D}_s}{\hat{\sigma}_{\hat{w}_s, s}} \right) - \frac{\|\hat{v}_s\| \hat{D}_s}{\hat{\sigma}_{\hat{w}_s, s}} \right) \left(\frac{\text{cv}_{1-\alpha} \left(\frac{\|\hat{v}_s\| \hat{D}_s}{\hat{\sigma}_{\hat{w}_s, s}} \right) - \frac{\|\hat{v}_s\| \hat{D}_s}{\hat{\sigma}_{\hat{w}_s, s}}}{\frac{\|\hat{v}_s\| \hat{D}_s}{\hat{\sigma}_{\hat{w}_s, s}}} + 2 \right) \xrightarrow{p} z_{1-\alpha} \left(\frac{z_{1-\alpha}}{\infty} + 2 \right) = 2z_{1-\alpha}.$$

The expression on the LHS of the inequality is equal to

$$\frac{(\hat{w}'_s \hat{Z}_s / \hat{\sigma}_{\hat{w}_s, s})^2}{\|\hat{v}_s\| \hat{D}_s / \hat{\sigma}_{\hat{w}_s, s}} - 2 \frac{\hat{\Delta}_s}{\|\hat{v}_s\| \hat{D}_s} \frac{\hat{w}'_s \hat{Z}_s}{\hat{\sigma}_{\hat{w}_s, s}} + \frac{\hat{\Delta}_s^2 - \hat{D}_s^2 \|\hat{v}_s\|^2}{\hat{D}_s \|\hat{v}_s\| \hat{\sigma}_{\hat{w}_s, s}}.$$

The first term is $o_p(1)$ and the third term satisfies, by Proposition 1 under $\lambda_s \in \hat{\Lambda}_s$,

$$\frac{\hat{\Delta}_s^2 - \hat{D}_s^2 \|\hat{v}_s\|^2}{\hat{D}_s \|\hat{v}_s\| \hat{\sigma}_{\hat{w}_s, s}} \leq \frac{\hat{D}_s^2 \|\hat{v}_s\|^2 - \hat{D}_s^2 \|\hat{v}_s\|^2}{\hat{D}_s \|\hat{v}_s\| \hat{\sigma}_{\hat{w}_s, s}} = 0.$$

Thus, (42) is bounded above by

$$\mathbb{P}_{P_s} \left\{ o_p(1) - 2 \frac{\hat{\Delta}_s}{\|\hat{v}_s\| \hat{D}_s} \frac{\hat{w}'_s \hat{Z}_s}{\hat{\sigma}_{\hat{w}_s, s}} > 2z_{1-\alpha} + o_p(1) \right\}.$$

Note $|\Delta_s / \|v_s\| D_s| \leq 1$ since $\lambda_s \in \Lambda_s$ (due to $\lambda_s \in \Lambda_{0,s} \subseteq \Lambda_s$), so I pass to a further subsequence where $\Delta_s / \|v_s\| D_s \rightarrow c_\infty$, where $c_\infty \in [-1, 1]$. It suffices to show $\hat{\Delta}_s / \|\hat{v}_s\| \hat{D}_s \xrightarrow{p} c_\infty$, since then CMT would yield

$$o_p(1) - 2 \frac{\hat{\Delta}_s}{\|\hat{v}_s\| \hat{D}_s} \frac{\hat{w}'_s \hat{Z}_s}{\hat{\sigma}_{\hat{w}_s, s}} - (2z_{1-\alpha} + o_p(1)) \xrightarrow{d} 2N(0, c_\infty^2) - 2z_{1-\alpha}.$$

Combining this with (40) and (42) yields the desired contradiction:

$$\limsup_{s \rightarrow \infty} \mathbb{P}_{P_s} \left\{ \lambda'_s \theta_s \notin CI_{\hat{w}_s, s}^*, \lambda_s \in \hat{\Lambda}_s, \mathcal{E}_s \right\} \leq \mathbb{P} \left\{ N(0, c_\infty^2) > z_{1-\alpha} \right\} \leq \mathbb{P} \left\{ N(0, 1) > z_{1-\alpha} \right\} = \alpha.$$

It now remains to show $\hat{\Delta}_s / \|\hat{v}_s\| \hat{D}_s \xrightarrow{p} c_\infty$. And indeed,

$$\begin{aligned} \left| \frac{\hat{\Delta}_s}{\|\hat{v}_s\| \hat{D}_s} - \frac{\Delta_s}{\|v_s\| D_s} \right| &= \left| \frac{\hat{\Delta}_s (\|v_s\| D_s - \|\hat{v}_s\| \hat{D}_s) + (\hat{\Delta}_s - \Delta_s) \|\hat{v}_s\| \hat{D}_s}{\|\hat{v}_s\| \hat{D}_s \|v_s\| D_s} \right| \\ &= \left| \frac{\hat{\Delta}_s (\|v_s\| D_s - \|\hat{v}_s\| D_s) + \hat{\Delta}_s (\|\hat{v}_s\| D_s - \|\hat{v}_s\| \hat{D}_s) + (\hat{\Delta}_s - \Delta_s) \|\hat{v}_s\| \hat{D}_s}{\|\hat{v}_s\| \hat{D}_s \|v_s\| D_s} \right| \\ &\lesssim_p \frac{\|\hat{v}_s - v_s\|}{\|v_s\|} + |\hat{D}_s - D_s| + \|\hat{w}_s - w_s\|. \end{aligned}$$

As shown in Appendix D.6, $\|\hat{v}_s - v_s\| = O_p(1)$ in this $\eta_\infty = \infty$ regime. The above RHS terms are therefore $o_p(1)$. Thus, $\hat{\Delta}_s / \|\hat{v}_s\| \hat{D}_s \xrightarrow{p} c_\infty$ follows from $\Delta_s / \|v_s\| D_s \rightarrow c_\infty$.

E Supplementary Lemmas

Lemma 1. *Under Assumption U5, $S \mapsto \Lambda(S)$ is Lipschitz in the Hausdorff metric:*

$$d_H(\Lambda(S_1), \Lambda(S_2)) \leq \frac{\text{diam}(\mathbb{W})}{\delta_g} L_g d_S(S_1, S_2), \quad \forall S_1, S_2 \in \mathbb{S}.$$

Proof. Under Assumption U5, the distance from λ to $\Lambda(S)$ is bounded as

$$\text{dist}(\lambda, \Lambda(S)) \leq \frac{\text{diam}(\mathbb{W})}{\delta_g} \max\{g(\lambda, S), 0\}, \quad \forall \lambda \in \mathbb{W}, \quad \forall S \in \mathbb{S}.$$

Indeed, take any $(\lambda, S) \in \mathbb{W} \times \mathbb{S}$ and consider

$$t = \frac{\max\{g(\lambda, S), 0\}}{\max\{g(\lambda, S), 0\} + \delta_g}, \quad \lambda_t = (1-t)\lambda + t\lambda^\circ(S) \in \mathbb{W}.$$

By convexity,

$$g(\lambda_t, S) \leq (1-t)g(\lambda, S) + tg(\lambda^\circ(S), S) \leq (1-t)\max\{g(\lambda, S), 0\} - t\delta_g = 0.$$

This means $\lambda_t \in \Lambda(S)$, and hence

$$\text{dist}(\lambda, \Lambda(S)) \leq \|\lambda - \lambda_t\| \leq t \text{diam}(\mathbb{W}) \leq \frac{\text{diam}(\mathbb{W})}{\delta_g} \max\{g(\lambda, S), 0\}.$$

Now take any $S_1, S_2 \in \mathbb{S}$. For $\lambda \in \Lambda(S_2)$, observe that

$$g(\lambda, S_1) = g(\lambda, S_1) - g(\lambda, S_2) + g(\lambda, S_2) \leq g(\lambda, S_1) - g(\lambda, S_2) \leq \max_{\lambda \in \mathbb{W}} |g(\lambda, S_1) - g(\lambda, S_2)|.$$

Thus, the distance bound and the Lipschitz condition yield

$$\max_{\lambda \in \Lambda(S_2)} \text{dist}(\lambda, \Lambda(S_1)) \leq \frac{\text{diam}(\mathbb{W})}{\delta_g} \max_{\lambda \in \Lambda(S_2)} \max\{g(\lambda, S_1), 0\} \leq \frac{\text{diam}(\mathbb{W})}{\delta_g} L_g d_S(S_1, S_2).$$

Reversing the roles of S_1 and S_2 gives the same bound on $\max_{\lambda \in \Lambda(S_1)} \text{dist}(\lambda, \Lambda(S_2))$, and hence the desired bound on $d_H(\Lambda(S_1), \Lambda(S_2))$. $\square$

Lemma 2. *For positive definite Σ_1, Σ_2 with $e_{\min}(\Sigma_1), e_{\min}(\Sigma_2) \geq 1/\bar{e} > 0$,*

$$\left| \|x\|_{\Sigma_1} - \|x\|_{\Sigma_2} \right| \leq \frac{\sqrt{\bar{e}}}{2} \|\Sigma_1 - \Sigma_2\| \|x\|, \quad \forall x \in \mathbb{R}^K.$$

For nonempty compact $\Lambda_1, \Lambda_2 \subseteq \mathcal{W}$, $w_1, w_2 \in \mathcal{W}$, and positive definite Σ with $e_{\max}(\Sigma) \leq \bar{e}$,

$$\begin{aligned} \left| \max_{\lambda \in \Lambda_1} \|\lambda - w_1\|_{\Sigma} - \max_{\lambda \in \Lambda_2} \|\lambda - w_1\|_{\Sigma} \right| &\leq \sqrt{\bar{e}} d_H(\Lambda_1, \Lambda_2), \\ \left| \max_{\lambda \in \Lambda_1} \|\lambda - w_1\|_{\Sigma} - \max_{\lambda \in \Lambda_1} \|\lambda - w_2\|_{\Sigma} \right| &\leq \sqrt{\bar{e}} \|w_1 - w_2\|. \end{aligned}$$

Proof. For the first claim, $x=0$ is immediate. For $x \neq 0$, observe that

$$\left| \|x\|_{\Sigma_1} - \|x\|_{\Sigma_2} \right| = \frac{|x'(\Sigma_1 - \Sigma_2)x|}{\|x\|_{\Sigma_1} + \|x\|_{\Sigma_2}} \leq \frac{\|\Sigma_1 - \Sigma_2\| \|x\|^2}{\sqrt{e_{\min}(\Sigma_1)} \|x\| + \sqrt{e_{\min}(\Sigma_2)} \|x\|} \leq \frac{\sqrt{\bar{e}}}{2} \|\Sigma_1 - \Sigma_2\| \|x\|.$$

For the first inequality of the second claim, note that for each $\lambda_1 \in \Lambda_1$, compactness of Λ_2 gives a nearest point $\lambda_2 \in \Lambda_2$ such that

$$\|\lambda_1 - \lambda_2\| = \text{dist}(\lambda_1, \Lambda_2) \leq d_H(\Lambda_1, \Lambda_2).$$

The reverse triangle inequality and max-eigenvalue bound imply

$$|\|\lambda_1 - w_1\|_{\Sigma} - \|\lambda_2 - w_1\|_{\Sigma}| \leq \|\lambda_1 - \lambda_2\|_{\Sigma} \leq \sqrt{\bar{e}} \|\lambda_1 - \lambda_2\|.$$

Combining with the previous inequality yields

$$\|\lambda_1 - w_1\|_{\Sigma} \leq \|\lambda_2 - w_1\|_{\Sigma} + \sqrt{\bar{e}} d_H(\Lambda_1, \Lambda_2) \leq \max_{\lambda \in \Lambda_2} \|\lambda - w_1\|_{\Sigma} + \sqrt{\bar{e}} d_H(\Lambda_1, \Lambda_2)$$

Taking the maximum over $\lambda \in \Lambda_1$ then yields

$$\max_{\lambda \in \Lambda_1} \|\lambda - w_1\|_{\Sigma} \leq \max_{\lambda \in \Lambda_2} \|\lambda - w_1\|_{\Sigma} + \sqrt{\bar{e}} d_H(\Lambda_1, \Lambda_2)$$

Reversing the roles of Λ_1 and Λ_2 then yields the first inequality of the second claim. For the second inequality of the second claim, note that for each $\lambda \in \Lambda_1$,

$$|\|\lambda - w_1\|_{\Sigma} - \|\lambda - w_2\|_{\Sigma}| \leq \|w_1 - w_2\|_{\Sigma} \leq \sqrt{\bar{e}} \|w_1 - w_2\|,$$

and hence

$$\begin{aligned} \|\lambda - w_1\|_{\Sigma} &\leq \|\lambda - w_2\|_{\Sigma} + \sqrt{\bar{e}} \|w_1 - w_2\| \leq \max_{\lambda \in \Lambda_1} \|\lambda - w_2\|_{\Sigma} + \sqrt{\bar{e}} \|w_1 - w_2\|, \\ \|\lambda - w_2\|_{\Sigma} &\leq \|\lambda - w_1\|_{\Sigma} + \sqrt{\bar{e}} \|w_1 - w_2\| \leq \max_{\lambda \in \Lambda_1} \|\lambda - w_1\|_{\Sigma} + \sqrt{\bar{e}} \|w_1 - w_2\|. \end{aligned}$$

Taking the maximum over $\lambda \in \Lambda_1$ on the LHS of both lines then yields the second inequality of the second claim. $\square$

Lemma 3. For nonempty, compact, and convex $\Lambda \subseteq \mathcal{W}$ and positive definite Σ with $e_{\min}(\Sigma) \geq 1/\bar{e} > 0$, the corresponding minimax-bias weights $w^* = \operatorname{argmin}_{\bar{w} \in \Lambda} \max_{\lambda \in \Lambda} \|\lambda - \bar{w}\|_{\Sigma}$ satisfy

$$\|\bar{w} - w^*\|^2 \leq 2\bar{e}^{3/2} \operatorname{diam}(\Lambda) \left(\max_{\lambda \in \Lambda} \|\lambda - \bar{w}\|_{\Sigma} - \max_{\lambda \in \Lambda} \|\lambda - w^*\|_{\Sigma} \right), \quad \forall \bar{w} \in \Lambda.$$

Proof. By Proposition 5, w^* exists uniquely. Below I will show the inequality

$$\frac{\|\bar{w} - w^*\|^2}{\bar{e}} \leq \max_{\lambda \in \Lambda} \|\lambda - \bar{w}\|_{\Sigma}^2 - \max_{\lambda \in \Lambda} \|\lambda - w^*\|_{\Sigma}^2,$$

from which the lemma follows by bounding the RHS as

$$\begin{aligned} \max_{\lambda \in \Lambda} \|\lambda - \bar{w}\|_{\Sigma}^2 - \max_{\lambda \in \Lambda} \|\lambda - w^*\|_{\Sigma}^2 &= \left(\max_{\lambda \in \Lambda} \|\lambda - \bar{w}\|_{\Sigma} - \max_{\lambda \in \Lambda} \|\lambda - w^*\|_{\Sigma} \right) \\ &\quad \times \left(\max_{\lambda \in \Lambda} \|\lambda - \bar{w}\|_{\Sigma} + \max_{\lambda \in \Lambda} \|\lambda - w^*\|_{\Sigma} \right) \\ &\leq \left(\max_{\lambda \in \Lambda} \|\lambda - \bar{w}\|_{\Sigma} - \max_{\lambda \in \Lambda} \|\lambda - w^*\|_{\Sigma} \right) \\ &\quad \times \left(\bar{e}^{1/2} \operatorname{diam}(\Lambda) + \bar{e}^{1/2} \operatorname{diam}(\Lambda) \right). \end{aligned}$$

It now remains to show the inequality. First note that

$$\|(1-t)a + tb\|_{\Sigma}^2 = (1-t)\|a\|_{\Sigma}^2 + t\|b\|_{\Sigma}^2 - t(1-t)\|a - b\|_{\Sigma}^2, \quad \forall a, b \in \mathbb{R}^K, \quad \forall t \in (0, 1).$$

For each $\bar{w} \in \Lambda$, define $\bar{w}_t = (1-t)w^* + t\bar{w} \in \Lambda$. For each $\lambda \in \Lambda$, the above implies

$$\max_{\lambda \in \Lambda} \|\lambda - \bar{w}_t\|_{\Sigma}^2 \leq (1-t) \max_{\lambda \in \Lambda} \|\lambda - w^*\|_{\Sigma}^2 + t \max_{\lambda \in \Lambda} \|\lambda - \bar{w}\|_{\Sigma}^2 - t(1-t) \|\bar{w} - w^*\|_{\Sigma}^2.$$

The LHS is bounded below by $\max_{\lambda \in \Lambda} \|\lambda - w^*\|_{\Sigma}^2$. Applying this bound and manipulating terms yields

$$\max_{\lambda \in \Lambda} \|\lambda - \bar{w}\|_{\Sigma}^2 - \max_{\lambda \in \Lambda} \|\lambda - w^*\|_{\Sigma}^2 \geq (1-t) \|\bar{w} - w^*\|_{\Sigma}^2 \geq (1-t) \frac{\|\bar{w} - w^*\|^2}{\bar{e}}.$$

Taking $t \rightarrow 0$ yields the desired inequality. □

Lemma 4. Under Assumption U5, the diameter of $\Lambda(S)$ is bounded away from zero:

$$\operatorname{diam}(\Lambda(S)) = \max_{\lambda, w \in \Lambda(S)} \|\lambda - w\| \geq \frac{\delta_g}{\delta_g + \bar{C}_g} \operatorname{diam}(\mathbb{W}) > 0, \quad \forall S \in \mathbb{S},$$

where $\bar{C}_g = \max_{S \in \mathbb{S}} \max_{\lambda \in \mathbb{W}} \max\{g(\lambda, S), 0\} < \infty$. Consequently, the maximum distance under

a positive definite Σ is bounded below as

$$\max_{\lambda \in \Lambda(S)} \|\lambda - w\|_{\Sigma} \geq \frac{1}{2} \sqrt{e_{\min}(\Sigma)} \frac{\delta_g}{\delta_g + \bar{C}_g} \text{diam}(\mathbb{W}) > 0, \quad \forall S \in \mathbb{S}, \quad \forall w \in \mathbb{W}.$$

Proof. The continuity, Lipschitz, and compactness properties in Assumption U5 yield

$$\bar{C}_g = \max_{S \in \mathbb{S}} \max_{\lambda \in \mathbb{W}} \{g(\lambda, S), 0\} < \infty.$$

By compactness, there exist $w_a, w_b \in \mathbb{W}$ such that

$$\text{diam}(\mathbb{W}) = \max_{\lambda, w \in \mathbb{W}} \|\lambda - w\| = \|w_a - w_b\|.$$

Given the Slater point $\lambda^\circ(S) \in \mathbb{W}$, define

$$\lambda_a(S) = (1-t)\lambda^\circ(S) + tw_a, \quad \lambda_b(S) = (1-t)\lambda^\circ(S) + tw_b, \quad t = \frac{\delta_g}{\delta_g + \bar{C}_g} \in (0, 1].$$

Since $\mathbb{W}$ is convex, then $\lambda_a(S), \lambda_b(S) \in \mathbb{W}$. Moreover, convexity of $\lambda \mapsto g(\lambda, S)$ yields

$$g(\lambda_a(S), S) \leq (1-t)g(\lambda^\circ(S), S) + tg(w_a, S) \leq -(1-t)\delta_g + t\bar{C}_g = 0,$$

and hence $\lambda_a(S) \in \Lambda(S)$. The analogous argument for w_b yields $\lambda_b(S) \in \Lambda(S)$. Thus,

$$\text{diam}(\Lambda(S)) = \max_{\lambda, w \in \Lambda(S)} \|\lambda - w\| \geq \|\lambda_a(S) - \lambda_b(S)\| = t\|w_a - w_b\| = \frac{\delta_g}{\delta_g + \bar{C}_g} \text{diam}(\mathbb{W}) > 0.$$

Consequently, using $\|x\|_{\Sigma} \geq \|x\| \sqrt{e_{\min}(\Sigma)}$; $\lambda_a(S), \lambda_b(S) \in \Lambda(S)$; and triangle inequality,

$$\frac{2 \max_{\lambda \in \Lambda(S)} \|\lambda - w\|_{\Sigma}}{\sqrt{e_{\min}(\Sigma)}} \geq \|\lambda_a(S) - w\| + \|\lambda_b(S) - w\| \geq \|\lambda_a(S) - \lambda_b(S)\| = \frac{\delta_g}{\delta_g + \bar{C}_g} \text{diam}(\mathbb{W}).$$

□

Lemma 5. *Let Assumptions U4 and U5 be satisfied and suppose that $\Lambda_{0,n} \subseteq \Lambda_n$ satisfies slack condition (27). Then containment condition (25) is satisfied:*

$$\lim_{n \rightarrow \infty} \sup_{P_n \in \mathcal{P}_n} \sup_{\lambda \in \Lambda_{0,n}} \mathbb{P}_{P_n} \{ \lambda \notin \hat{\Lambda}_n \} = 0.$$

Proof. Observe that

$$\sup_{\lambda \in \Lambda_{0,n}} \mathbb{P}_{P_n} \{ \lambda \notin \hat{\Lambda}_n \} \leq \mathbb{P}_{P_n} \left\{ \sup_{\lambda \in \Lambda_{0,n}} g(\lambda, \hat{S}_n) > 0, \hat{S}_n \in \mathbb{S} \right\} + \mathbb{P}_{P_n} \{ \hat{S}_n \notin \mathbb{S} \}.$$

Under Assumption U4, $\lim_{n \rightarrow \infty} \sup_{P_n \in \mathcal{P}_n} \mathbb{P}_{P_n} \{ \hat{S}_n \notin \mathbb{S} \} = 0$. Thus, it suffices to show

$$\lim_{n \rightarrow \infty} \sup_{P_n \in \mathcal{P}_n} \mathbb{P}_{P_n} \left\{ \sup_{\lambda \in \Lambda_{0,n}} g(\lambda, \hat{S}_n) > 0, \hat{S}_n \in \mathbb{S} \right\} = 0.$$

For $\lambda \in \Lambda_{0,n} \subseteq \Lambda_n$ and on an event where $\hat{S}_n \in \mathbb{S}$, Assumption U5 implies

$$g(\lambda, \hat{S}_n) \leq g(\lambda, S_n) + L_g d_{\mathcal{S}}(\hat{S}_n, S_n),$$

and hence slack condition (27) implies

$$\sup_{\lambda \in \Lambda_{0,n}} g(\lambda, \hat{S}_n) \leq -\nu + L_g d_{\mathcal{S}}(\hat{S}_n, S_n).$$

When $L_g d_{\mathcal{S}}(\hat{S}_n, S_n) \leq \nu$, the above implies $\sup_{\lambda \in \Lambda_{0,n}} g(\lambda, \hat{S}_n) \leq 0$. By contraposition,

$$\mathbb{P}_{P_n} \left\{ \sup_{\lambda \in \Lambda_{0,n}} g(\lambda, \hat{S}_n) > 0, \hat{S}_n \in \mathbb{S} \right\} \leq \mathbb{P}_{P_n} \{ L_g d_{\mathcal{S}}(\hat{S}_n, S_n) > \nu, \hat{S}_n \in \mathbb{S} \} \leq \mathbb{P}_{P_n} \{ L_g d_{\mathcal{S}}(\hat{S}_n, S_n) > \nu \}.$$

But since $\hat{S}_n$ is uniformly consistent for S_n under Assumption U4, the above implies

$$\lim_{n \rightarrow \infty} \sup_{P_n \in \mathcal{P}_n} \mathbb{P}_{P_n} \left\{ \sup_{\lambda \in \Lambda_{0,n}} g(\lambda, \hat{S}_n) > 0, \hat{S}_n \in \mathbb{S} \right\} = 0.$$

Altogether, this yields the desired conclusion. □