Robust inference for the treatment effect variance in experiments using machine learning *

Alejandro Sánchez-Becerra
Emory QTM

First Version: February 12, 2022
This version: June 5, 2023

Abstract

Experimenters often collect baseline data to study heterogeneity. I propose the first valid confidence intervals for the VCATE, the treatment effect variance explained by observables. Conventional approaches yield incorrect coverage when the VCATE is zero. As a result, practitioners could be prone to detect heterogeneity even when none exists. The reason why coverage worsens at the boundary is that all efficient estimators have a locally-degenerate influence function and may not be asymptotically normal. I solve the problem for a broad class of multistep estimators with a predictive first stage. My confidence intervals account for higher-order terms in the limiting distribution and are fast to compute. I also find new connections between the VCATE and the problem of deciding whom to treat. The gains of targeting treatment are (sharply) bounded by half the square root of the VCATE. Finally, I document excellent performance in simulation and reanalyze an experiment from Malawi.

Keywords: Debiased machine learning, treatment effect heterogeneity, experiments, non-standard inference, variance decomposition
JEL Codes C14, C21, C55, C90

1 Introduction

In recent years, there has been a rapid expansion of experiments to evaluate public policy programs and corporate initiatives. There is also more evidence that the effectiveness of a program can vary across individuals. For instance, Dizon-Ross (2019) studies a population of low-income parents in Malawi with large misperceptions about their children's school performance. She finds that a simple intervention can bridge these gaps and that the Conditional Average Treatment Effect (CATE) varies by children's initial scores. In practice, even though researchers collect baseline surveys with many characteristics, the CATE is typically estimated via regressions with one or two interactions, thus underutilizing the full set of variables. The promise of leveraging the vast and readily available baseline data has sparked more applications of supervised machine learning (Crépon et al., 2021; Davis and Heller, 2020; Deryugina et al., 2019). Methods such as LASSO, neural networks, random forests, or boosting are data-driven and allow for more variables and flexibility.

In this paper, I focus on the unconditional variance of the CATE, the VCATE, which measures the dispersion of treatment effects predicted by a set of baseline characteristics. The VCATE has a clear interpretation, even if the CATE is nonlinear or depends on many characteristics. Chernozhukov et al. (2022a) and Ding et al. (2019) separately propose estimators with a misspecification robust interpretation, whereas Levy et al. (2021) propose an efficient estimator. Despite these recent advances, there are currently no valid confidence intervals for the VCATE. In fact, Levy et al. (2021) study the performance of confidence intervals based on the efficient influence function, i.e., the conventional way. They find that coverage degrades near the boundary, reaching a low of 32% in simulations. They speculate that poor coverage is due to the degeneracy of the efficient influence function when the VCATE is zero. As a result, conventional guarantees for $\sqrt{n}$−asymptotic normality do not apply to this part of the parameter space. Moreover, a VCATE close to zero is economically meaningful because it could reflect null effects, low effect heterogeneity, or irrelevant covariates. For such situations, which are common in practice, conventional approaches to inference could be misleading.

This paper provides fresh insights regarding the VCATE and proposes a solution for inference in experiments. I propose (a) novel ways to interpret the VCATE for decision-making by deriving sharp bounds on the population gains of person-

alized treatment assignment, (b) novel estimators of the VCATE that are both misspecification-robust and efficient, combining the best features of previous approaches, and (c) novel confidence intervals that are shape-adaptive and fast to compute.

I break down why conventional confidence intervals have incorrect size. I show that the boundary inference problem can manifest even when the CATE is linear and univariate. I solve the problem for the linear case by proposing adaptive confidence intervals that meet the high-level conditions outlined in Andrews et al. (2020). I then show that these can be readily extended to the class of nonlinear models in Chernozhukov et al. (2022a) that combine regression adjustments and a machine learning first stage. In addition to providing new confidence intervals, my paper has novel implications for estimation by showing that only a subset of these multi-step estimators is efficient. Adaptive inference and regression adjustments work well for various predictive models under weak assumptions.[1] In the fully nonparametric case, I use a conservative procedure with valid coverage over multiple sample splits.[2] I derive the local power curve for the associated tests of homogeneity and their relationship to the tests in Crump et al. (2008) and Ding et al. (2019). I also propose confidence intervals for settings with cluster dependence.

I document excellent root mean square error (RMSE) performance and coverage in simulations using LASSO, even in high dimensions. I benchmark my multi-step approach against a two-step debiased machine learning estimator. As predicted by theory, all approaches are asymptotically normal, efficient, and have good coverage in highly heterogeneous designs. However, when the VCATE is zero or close to zero, coverage of two-step alternatives can be as low as 45%. By contrast, my adaptive intervals produce coverage at the intended 95% level and better RMSE at all regions of the parameter space. I study the robustness of the multi-step approach in both the theory and simulations. I consider situations where the predictive component is

---

[1]While I specialize my results to the VCATE, my inference approach relies on more general principles: I use knowledge of the limiting distribution function, conditional on the cross-fitted estimates. Correct coverage follows from verifying high-level assumptions that can be satisfied by a wide array of machine learning method used in the first step. In principle, my approach could extend to other non-standard inference problems.

[2]This adjustment is in the spirit of (Chernozhukov et al., 2022a), who propose robust t-tests assuming a conditionally normal distribution. Their results are not directly applicable here due to the boundary inference problem. However, I apply the principles behind their "median-parameter" confidence intervals to my adaptive intervals.

misspecified or slow to converge. I also discuss issues related to uniform vs. pointwise coverage.

I apply my approach to data from Dizon-Ross (2019), an information experiment with low-income parents in Malawi who had at least two school-age children. The intervention redesigned the way in which parents received information about their children's school performance. The endline survey measured parental beliefs about student grades and asked parents to allocate tickets to a scholarship between their children. Dizon-Ross (2019) presented graphs with a non-parametric CATE by baseline test scores, which had an approximately linear shape, and separately tested for significance using a regression with an interaction. I use LASSO to compute the VCATE for the two outcomes (parental beliefs and lottery allocations) by different characteristics of students, parents, and households. To make the results interpretable, I focus on the standard deviation of the CATE, i.e., $\sqrt{VCATE}$, and normalize it by the standard deviation of the outcome in the control group.

My approach allows us to quantify the magnitude of effect heterogeneity. I find that the treatment effect heterogeneity explained by test scores is equivalent to 40% of the standard deviation (SD) of the beliefs of the control group, and 16% of the SD of the control group lottery allocation. I also find that the effect heterogeneity collectively explained by other student variables (grade, age, gender, attendance, and educational expenditures) is comparable to 11% of the SD of beliefs in the control group. The VCATE of beliefs by student variables is significant at the 5% level, but the VCATE of lottery outcomes by student variables is not. The combined VCATE associated with student scores and 12 other key characteristics has a similar value to the VCATE with only scores. Despite being conservative, the intervals for the VCATE are short in length in this empirical example. Using my new welfare bounds $(-|ATE| + \sqrt{VCATE + ATE^2})/2$, I predict that targeted interventions using the baseline covariates have a maximum added benefit of 7.9% SD and 7.4% SD (standard deviations of the outcome for the control group) on beliefs and lottery allocations, respectively.

Researchers should focus on the VCATE because it is a model-free quantity with good properties: it is well-defined even if the CATE is continuous or discrete, and it weakly increases when researchers add more covariates to their analysis. Researchers can test for homogeneity by evaluating whether confidence intervals for the VCATE include zero. In addition to testing, by quantifying the VCATE, researchers can

4

compare the magnitude of heterogeneity relative to a benchmark, such as the variance at baseline, the VCATE for different covariates, or experiments in other sites.

1.1 Contribution

My first main contribution is to show that the VCATE provides a bound for the welfare gains of policy targeting. A policymaker might decide to use the information from the CATE to design personalized treatment recommendations (Athey and Wager, 2021; Kitagawa and Tetenov, 2018; Manski, 2004; Mbakop and Tabord-Meehan, 2021). One can measure utilitarian welfare by computing the expected outcome under different policies. Such policies can be further constrained to a class that respects budget limits, incentive compatibility, or fairness considerations (Sun, 2021; Viviano and Bradic, 2023). I show that the difference in mean outcomes between a targeted policy and a non-targeted policy using only the average treatment effect (ATE), is bounded by $\sqrt{VCATE}/2$. For instance, under homogeneity ($VCATE = 0$), there are no gains from targeting. I show that this bound holds in the population regardless of the choice of policy class and the underlying distribution. Furthermore, the bound is *sharp* in the sense that it holds exactly for at least one policy and distribution.

The proposed bound on utilitarian welfare communicates information to practitioners about whether a targeting exercise is even worth pursuing, without needing to solve the targeting problem itself. The VCATE can be a supplemental quantity reported in regression analyses, or a benchmark for analysts choosing the optimal policy. If the VCATE is very low, practitioners may consider expanding the set of covariates in the analysis. To derive the bound, I use a constructive approach to solve the most adversarial distribution. I also prove a more general bound $(-|ATE| + \sqrt{VCATE + ATE^2})/2$, and show that the distribution that leads to a maximum welfare gain is one where the CATE has binary support and mean zero. The gains from targeting easily diminish if the value of the ATE is relatively higher than the VCATE.

My second contribution is related to efficient estimation and robust inference. New theory is required here because of a unique feature of the VCATE: the efficient influence function is degenerate when the CATE is homogeneous (Levy et al., 2021). Classical results by Newey (1990) show that any regular, efficient estimator can be decomposed as $\frac{1}{n}\sum_{i=1}^{n}\varphi_i + R_n$, where n is the sample size, $\{\varphi_i\}_{i=1}^{n}$ are a set of i.i.d.

mean-zero influence functions, and R_n is a residual with higher-order terms that are $o_p(n^{-1/2})$.[3] Conventional approaches assume that $\mathbb{V}(\varphi) > 0$ and in this case, the estimation error converges at $\sqrt{n}$ to $\mathcal{N}(0, \mathbb{V}(\varphi_i))$ by the CLT. However, when the VCATE is zero, $\mathbb{V}(\varphi_i) = 0$ as well. Hence, the limiting distribution is dominated by the higher terms in R_n, which may not be asymptotically normal. Therefore, while $\sqrt{n}$−estimation is still possible, t-tests that plug in an estimate of $\mathbb{V}(\varphi)$ may have incorrect coverage. By contrast, other common quantities such as the average treatment effect (ATE) or the local average treatment effect (LATE) do not have this problem because they satisfy $\mathbb{V}(\varphi_i) > 0$ uniformly (Chernozhukov et al., 2018).

I start by analyzing a simple two step estimator, assuming that the CATE is linear in covariates and can be estimated from a regression. I show that the limiting distribution of the VCATE estimator can be written as a linear combination of a Chi-square that converges at n-rate and a normal distribution that converges at $\sqrt{n}$−rate. The weights are determined by the value of the VCATE, which means that the shape of the distribution changes depending on the region of the parameter space. At the boundary, it behaves like a rescaled chi-square, is $O_p(n^{-1})$, and confidence intervals with normal critical values will have incorrect coverage. For values of the VCATE bounded away from zero, the distribution is asymptotically normal as in the classical results.

In the linear case, I construct adaptive confidence intervals that account for the higher terms of the distribution. I apply the framework of Andrews et al. (2020) to show that this produces uniform, exact coverage when the linear model is correctly specified.[4] The intervals are fast to compute because the expressions are all analytic. When there is a single covariate, I also show that a homogeneity test that evaluates whether zero is contained in the confidence intervals is algebraically identical to (i) a test of whether the interaction in the regression model is equal to zero, and (ii) the single-covariate homogeneity test of Crump et al. (2008). The test is also asymptotically equivalent to Ding et al. (2019). However, these other tests only apply to series estimators and are not nested with mine in the multivariate and non-parametric

---

[3]Many standard estimators can achieve this property, e.g., the "debiased machine learning" estimator (Chernozhukov et al., 2018) or the targeted maximum likelihood estimator in Levy et al. (2021).

[4]This type of strategy has proven effective to deal with other non-standard problems where the shape of the limiting distribution depends on an unknown parameter, such as the AR coefficient in a time series, the effect parameter under weak instruments, or the quasi-likelihood ratio test for nonlinear regression (Andrews et al., 2020).

cases.

I extend my results to the class of nonlinear models proposed by Chernozhukov et al. (2022a). Chernozhukov et al. (2022a) showed that in experiments with known assignment probabilities, their models produce a meaningful pseudo-VCATE even if the functional form is misspecified. The pseudo-VCATE is non-negative, weakly lower than the VCATE, and converges to the true value under mild conditions on the estimated CATE.[5] Chernozhukov et al. (2022a) argue that the pseudo-VCATE might be of independent interest as a measure of model fit.[6] They describe a three-step estimator with a machine learning/prediction first-stage, a regression second-stage, and a sample-variance third stage. Related multi-step estimators have also been considered in other work (Guo et al., 2021).

To the best of my knowledge, there are no existing asymptotic results for the multi-step VCATE estimator proposed in Chernozhukov et al. (2022a). I fill in that gap by proving two sets of results. First, I show that all the estimators in their class converge to the true VCATE at least at $\sqrt{n}$-rate, are $o_p(n^{-1/2})$ at the boundary (as in the simple linear model), and have the convenient property that they are always non-negative. This builds on the asymptotic expansion for the linear case I introduced above. Second, I prove that only a subset of the Chernozhukov et al. (2022a) estimators are efficient, i.e. converge at $\sqrt{n}-$ to an average of i.i.d. efficient influence functions. The key ingredient is to prove a novel finite-sample equivalence result. I find that the first order conditions of the regression step and the bias-correction component of the VCATE influence function are in fact identical, given a particular decomposition of the nuisance functions. The asymptotic results follow from fairly standard assumptions on convergence rates (Belloni et al., 2017; Chernozhukov et al., 2018). To get the limiting distribution, the only meaningful extra assumption is that the estimated CATE has bounded kurtosis (thin tails).

I show that extending the adaptive confidence intervals (CIs) to multi-step estimators is straightforward. The procedure randomly splits the data into subsets or *folds* and estimates the nuisance functions and the VCATE on different folds. To compute the confidence intervals for a particular fold, the researcher can treat the second-stage regression as if the variables were given, and then construct the CIs as in the simple

---

[5]This monotonicity property means that in experiments the pseudo-VCATE will not falsely detect heterogeneity, even if the machine learning stage is misspecified.

[6] Ding et al. (2019) also define a similar pseudo-VCATE based on randomization inference.

case. I construct median confidence intervals (CIs) to aggregate information across multiple folds. I show that the single fold procedure produces uniform, exact coverage for the pseudo-VCATE and point-wise, exact coverage for the VCATE for all points in the parameter space, at a nominal level $1 - \alpha$. The multifold CIs have pointwise conservative coverage.

Furthermore, the probability that the true VCATE is below the confidence interval bounds is uniformly bounded by α in large samples. This result applies to the single and multifold CIs and does not require that the first-stage estimates to converge. Instead it relies on the fact that in experiments the pseudo-VCATE is weakly lower than the true VCATE. Tests for homogeneity (whether zero is contained in the CI) belong to this broader class of tests. Having uniform size control for this class of one-sided tests means that my tests of homogeneity are robust.

This paper is also related to a growing literature on debiased-machine learning (Belloni et al., 2017, 2014; Chernozhukov et al., 2018, 2022b,c), semiparametric efficiency (Newey, 1990), uniform inference for non-standard problems (Andrews et al., 2020), and tests of treatment effect homogeneity (Bitler et al., 2017; Crump et al., 2008; Ding et al., 2019; Heckman et al., 1997). My approach combines results from these literatures by addressing a boundary inference problem with a machine learning stage, and applying techniques of uniform inference. A related literature also focuses on confidence intervals around point-predictions of the CATE (Athey et al., 2019; Semenova and Chernozhukov, 2021), rather than overall measures of dispersion.

Section 2 provides key definitions, introduces the welfare bound, and presents a version of the adaptive confidence intervals for the univariate regression case. Section 3 frames the inference problem in a more general setting, and extends the adaptive confidence intervals for VCATE estimation with a machine learning first stage. Section 4 presents the large sample theory. Section 6 introduces the simulations. Section 7 applies my approach to an empirical example from Malawi. Section 8 concludes.

2 Overview of framework

Consider a program evaluation setting in which an individual is assigned to either a treatment $(D = 1)$ or a control group $(D = 0)$. The outcome of interest Y depends on the treatment status. I denote the potential outcome under treatment and control status as Y_1 and Y_0, respectively, and the treatment effect as $Y_1 - Y_0$. The conditional

average treatment effect (CATE) given covariates X is defined as

$$\tau(x) := \mathbb{E}[Y_1 - Y_0 \mid X = x],$$

and the average treatment effect (ATE) is defined as $\tau_{av} := \mathbb{E}[Y_1 - Y_0]$. This paper proposes an estimator of the *variance of the CATE* (VCATE) defined as

$$V_\tau := \mathbb{V}(\tau(X)).$$

The variance V_τ measures the dispersion of treatment effects that can be attributed to observable characteristics X. The value of V_τ depends on the choice of covariates. To understand how different covariates might impact the VCATE, let $V_\tau' = \mathbb{V}(\mathbb{E}[Y_1 - Y_0 \mid X'])$ be the VCATE for a different set of covariates X'.

Lemma 1. *If X is X'-measurable, then $V_\tau \leqslant V_\tau' \leqslant \mathbb{V}(Y_1 - Y_0)$.*

Lemma 1 shows that the VCATE has the following monotonicity property: if the researcher adds more covariates to the analysis, or breaks down an existing covariate into more categories, then the VCATE will be weakly larger.

The propensity score, $p(x)$, is defined as follows

$$p(x) := \mathbb{P}(D = 1 \mid X = x). \tag{1}$$

I restrict attention to experimental settings where $p(x)$ is known. The CATE can be identified under further assumptions.

Assumption 1. *(i) Stable unit treatment value assumption (SUTVA), $Y = Y_1 D + (1 - D)Y_0$ (ii) Strong overlap, there is a constant $\delta \in (0, 1/2)$ such that $\mathbb{P}(\delta < p(X) < 1 - \delta) = 1$, (iii) Selection on observables, $Y_1, Y_0 \perp\!\!\!\perp D \mid X$.*

Assumption 1.(i) formalizes the idea that the researcher can only observe either Y_1 or Y_0, but not both, for any particular individual. Assumption (ii) holds in randomized controlled trials with treatment probabilities bounded away from $\{0, 1\}$. Assumption (iii) states that an individual's treatment probability depends on X but not their potential outcomes. Let $\mu_d(x)$ be the conditional mean of Y given X and a fixed value of $d \in \{0, 1\}$,

$$\mu_d(x) := \mathbb{E}[Y \mid D = d, X = x]. \tag{2}$$

9

Under Assumption 1, $\mathbb{E}[Y_d \mid X = x] = \mu_d(x)$, and hence $\tau(x) = \mu_1(x) - \mu_0(x)$. This means that the VCATE is identified, with $V_\tau = \mathbb{V}(\mu_1(X) - \mu_0(X))$.

2.1 The VCATE and policy targeting

Practitioners can use estimates of $\tau(X)$ to decide whom to treat in future interventions (Athey and Wager, 2021; Kitagawa and Tetenov, 2018; Manski, 2004). Program managers can target the treatment recipients based on their initial covariates. However, whether targeting can substantially improve average outcomes depends on the dispersion of $\tau(x)$. I show that a simple function of the VCATE bounds the marginal gains of targeting.

Let γ denote the joint distribution of (X, Y_1, Y_0), $\mathcal{X}$ a set containing the support of X, $\tau_\gamma(x) := \mathbb{E}_\gamma[Y_1 - Y_0 \mid X = x]$ the CATE given γ. A function which maps x to a probability of treatment $\pi(x)$ is known as a statistical allocation rule (Manski, 2004). Furthermore, I denote the set of all possible allocation rules by Π, which contains all functions $\{\pi : \mathcal{X} \to [0,1]\}$. The set Π includes many well-known assignment rules. For instance, it includes the "non-targeted" policy which assigns everyone to treatment if $\mathbb{E}_\gamma[Y_1] > \mathbb{E}_\gamma[Y_0]$ and to the control group otherwise. Moreover, the average outcome under rule π is $\mathbb{E}_\gamma[\pi(X)Y_1 + (1 - \pi(X))Y_0]$, and the marginal benefit compared to the non-targeted policy is defined as $\mathcal{U}_\gamma(\pi) := \mathbb{E}_\gamma[\pi(X)Y_1 + (1 - \pi(X))Y_0] - \max\{\mathbb{E}_\gamma[Y_1], \mathbb{E}_\gamma[Y_0]\}$.

Theorem 1. *Let* Γ *denote the set of distributions such that* $\mathbb{V}_\gamma[\tau_\gamma(X)] = V_\tau$. *For all* $\gamma \in \Gamma$ *and* $\pi \in \Pi$,

$$
\mathcal{U}_\gamma(\pi) \leqslant \underbrace{\begin{pmatrix} Welfare \\ Optimal \\ Targeting \end{pmatrix}}_{\sup_{\pi \in \Pi} \mathbb{E}_\gamma[\pi(X)Y_1 + (1 - \pi(X))Y_0]} - \underbrace{\begin{pmatrix} Welfare \\ No \\ Targeting \end{pmatrix}}_{\max\{\mathbb{E}_\gamma[Y_1], \mathbb{E}_\gamma[Y_0]\}} \leqslant \frac{1}{2}\sqrt{V_\tau}.
$$

The bound is sharp in the sense that $\mathcal{U}_\gamma(\pi) = \frac{1}{2}\sqrt{V_\tau}$ *for at least one* $\gamma \in \Gamma$ *and* $\pi \in \Pi$.

Theorem 2. *Consider distributions where* $\mathbb{E}_\gamma[\tau_\gamma(X)] = \tau_{av}$ *and* $\mathbb{V}_\gamma[\tau_\gamma(X)] = V_\tau$, *then* $U_\gamma(\pi) \leqslant \frac{1}{2}\left(-|\tau_{av}| + \sqrt{V_\tau + \tau_{av}^2}\right)$. *This bound is sharp over this subset of distributions.*

Theorem 1 shows the VCATE provides a welfare bound over the superset of policy classes. Furthermore, any type of restrictions on Π such as budget constraints or incentive compatibility will achieve utilitarian welfare gains that are weakly lower than $\frac{1}{2}\sqrt{V_\tau}$. The bound in Theorem 1 and the generalization in Theorem 2 provide simple bounds on the prospective gains of targeting, without needing to solve for $\pi(x)$. The bounds are most informative when V_τ is low. For instance, when $V_\tau = 0$ there is no heterogeneity explained by the observables X and therefore there are no gains from targeting. However, the fact that the bound is sharp does not imply that it is always achievable for every $\gamma \in \Gamma$, and when V_τ is high it is still be necessary to optimize $\pi(x)$ to determine whether personalized offers are worthwhile.

Finding the bound in Theorem 1 relies on two important insights. On one hand, the optimal policy in Π treats an individual if and only if $\tau(x) \geqslant 0$ (Kitagawa and Tetenov, 2018). Substituting the optimal policy, $\sup_{\pi \in \Pi} \mathcal{U}_\gamma(\pi)$ is equal to $\mathbb{E}_\gamma[\max\{\tau_\gamma(X), 0\}] - \max\{\mathbb{E}_\gamma[\tau_\gamma(X)], 0\}$. On the other hand, to avoid optimizing over all $\gamma \in \Gamma$, I break the problem down into equivalence classes based on the moments of the negative, zero, and positive components of the CATE. I use a constructive approach to derive the most "adversarial" distribution. The upper bound is achieved when the CATE has a binary support, which is partly why the bounds in Theorems 1 and 2 have simple closed forms.

Corollary 1. *Let $\kappa_1, \kappa_2 \in \mathbb{R}$ and define a new outcome $\tilde{Y} = \kappa_1 + \kappa_2 Y$. The maximum welfare gain for the transformed outcome is $\frac{|\kappa_2|}{2}(-|\tau_{av}| + \sqrt{V_\tau + (\tau_{av})^2})$.*

Corollary 1 shows that the welfare bound is invariant to location shifts in the outcome, and grows linearly with scale shifts. This result implies that transformations that change the sign, e.g. $\kappa_2 = -1$, do not change the value of the welfare bound. Consequently, the bound applies regardless of whether the welfare objective is to increase a desirable outcome or to decrease an undesirable outcome.

2.2 Inference using regressions

Consider a simple situation where X is real valued, the treatment D is experimentally assigned with constant probability, and U is a mean zero error term. The researcher runs the following linear regression,

$$Y = c_1 + c_2 X + \beta_1 D + \beta_2 DX + U, \quad \mathbb{E}[(1, X, D, DX)'U] = 0. \tag{3}$$

11

Define the auxiliary quantities $\tau^*(x) := \beta_1 + \beta_2 x$ and $V_x := \mathbb{V}(X)$. The pseudo-VCATE is defined as

$$V_\tau^* := \mathbb{V}(\tau^*(X)) = \beta_2^2 V_x. \tag{4}$$

The pseudo-VCATE has a close connection to the VCATE. If the linear model describes the conditional mean $\mu_d(x)$, then $\tau^*(x) = \tau(x)$ and $V_\tau = V_\tau^*$. For instance, in models with binary X, the functional form is correctly specified and $V_\tau = V_\tau^*$. For now, assume that the pseudo-VCATE and the VCATE coincide. In later sections, I analyze models that allow for misspecification.

Consider a sequence of distributions $\{\gamma_n\}_{n=1}^\infty \in \Gamma^\infty$. I index the regression coefficients and model variances by the sample size n as β_{2n} and $(V_{xn}, V_{\tau n})$, respectively. Define an estimator of the VCATE as $\widehat{V}_{\tau n} = \widehat{\beta}_{2n}^2 \widehat{V}_{xn}$, where $\widehat{\beta}_{2n}$ is the least squares estimator of (3) and $\widehat{V}_{xn} := \frac{1}{n} \sum_i X_i^2 - \left[\frac{1}{n} \sum_i X_i\right]^2$. With some algebraic manipulations the estimation error can be decomposed as

$$\widehat{V}_{\tau n} - V_{\tau n}^* = V_{xn}(\widehat{\beta}_{2n} - \beta_{2n})^2 \left(\frac{\widehat{V}_{xn}}{V_{xn}}\right) + 2\beta_{2n} V_{xn}(\widehat{\beta}_{2n} - \beta_{2n}) \left(\frac{\widehat{V}_{xn}}{V_{xn}}\right) + \beta_{2n}^2 V_{xn} \left(\frac{\widehat{V}_{xn}}{V_{xn}} - 1\right). \tag{5}$$

To derive the asymptotic distribution we can apply the central limit theorem to individual components. For generality, I state joint convergence to a normal distribution as an assumption. This holds as a special case if the observations are i.i.d. and key moments of the distribution are bounded, but may also hold under other forms of dependence. I defer stating primitive conditions until Section 3.2.

Assumption 2. *There is a sequence of distributions $\{\gamma_n\}_{n=1}^\infty \in \Gamma^\infty$ with associated quantities $\{V_{xn}, V_{\tau n}^*, \beta_{2n}, \Omega_n\}_{n=1}^\infty$, which are related by the identity $V_{\tau n}^* = \beta_{2n}^2 V_{xn}$, and satisfy the following properties: (i) $V_{xn} > 0$, (ii) $V_{\tau n}^*$ is contained in a bounded subset of $[0, \infty)$, and (iii) Ω_n is a positive definite matrix with eigenvalues bounded way from zero and a finite upper bound. There is a sequence of estimators $\{\widehat{V}_{xn}, \widehat{V}_{\tau n}, \widehat{\beta}_{2n}, \widehat{\Omega}_n\}_{n=1}^\infty$ which satisfy $\widehat{V}_{\tau n} = \widehat{\beta}_{2n}^2 \widehat{V}_{xn}$. As $n \to \infty$, $\widehat{\Omega}_n \to^p \Omega_n$, and*

$$\Omega_n^{-1/2} \sqrt{n} \left(\begin{array}{c} \sqrt{V_{xn}}(\widehat{\beta}_{2n} - \beta_{2n}) \\ \frac{\widehat{V}_{xn}}{V_{xn}} - 1 \end{array} \right) \to^d Z_n \sim \mathcal{N}(0, I_{2 \times 2}). \tag{6}$$

The normalization by V_{xn} is intended to align with the decomposition in (5). The 2×2 matrix Ω_n is an estimator of the covariance matrix. I present Assumption 2 as a

triangular array because it makes it easier to formalize discussions of uniform coverage over the parameter space. Assumption 2 allows for cases where $V_{\tau n}^*$ is arbitrarily close to or includes zero. Let $\Omega^{1/2}$ denote the Cholesky decomposition of a matrix Ω. The estimator of the VCATE converges to the empirical process G, defined as

$$G(n, V_\tau^*, \Omega, z, \zeta) := \frac{(e_1'\Omega^{1/2}z)^2}{n} + 2\zeta\sqrt{\frac{V_\tau^*}{n}}(e_1'\Omega^{1/2}z) + \frac{V_\tau^*}{\sqrt{n}}(e_2'\Omega^{1/2}z), \qquad (7)$$

where $z \in \mathbb{R}^2$, $\zeta \in \{-1, 1\}$, $e_1 = [1, 0]'$ and $e_2 = [0, 1]'$.

Lemma 2. *Suppose that Assumption 2 holds, then* $\widehat{V}_{\tau n} - V_{\tau n}^* = O_p\left(\max\left\{\frac{1}{n}, \sqrt{\frac{V_{\tau n}^*}{n}}\right\}\right)$, *and there exists a sequence of* $\zeta_n \in \{-1, 1\}$, *such that*

$$\widehat{V}_{\tau n} - V_{\tau n}^* = G(n, V_{\tau n}^*, \Omega_n, Z_n, \zeta_n) + o_p\left(\frac{1}{n}\right) + o_p\left(\sqrt{\frac{V_{\tau n}^*}{n}}\right) + o_p\left(\frac{V_{\tau n}^*}{\sqrt{n}}\right). \qquad (8)$$

Lemma 2 shows that the limiting distribution of $\widehat{V}_{\tau n}$ is a linear combination of a Chi-square and a normal, whose weights depend on the value of $V_{\tau n}^*$. The relative magnitude of $V_{\tau n}^*$ determines the fit of the normal approximation. In the heterogeneous case, $V_{\tau n}^* \geqslant \delta > 0$, $\sqrt{n}G$ converges to a normal as $n \to \infty$ because the first term in (7) is asymptotically negligible. However, when $V_{\tau n}^* = 0$, only the first term remains and nG converges to a non-central Chi-Square distribution, which is asymmetric. Using normal critical values here (even if everything else was known) would produce distorted coverage. Furthermore, when $V_{\tau n}^* = 0$ the rate of convergence is n, which is faster than $\sqrt{n}$, and hence the estimator is "super consistent" near the boundary. The error is dominated by the first stage sampling uncertainty in estimating the nuisance parameter β_{2n}^2, which converges at n rate.

In practice, all three components in (7) contribute to the limiting distribution, and this information can be used for inference. I propose an analytic approach based on the quantiles of the empirical process that can deliver exact coverage. Let $F_{n,V_\tau^*,\Omega,\zeta}(v)$ be the conditional CDF of the empirical process, defined as

$$F_{n,V_\tau^*,\Omega,\zeta}(v) = \mathbb{P}(G(n, V_\tau^*, \Omega, Z, \zeta) \leqslant v), \qquad Z \sim \mathcal{N}(0, I_{2\times 2}), \quad v \in \mathbb{R}. \qquad (9)$$

13

Based on this CDF we can construct a test statistic,

$$F_{n,V_\tau^*,\hat{\Omega}_n,\zeta}(\hat{V}_{\tau n} - V_\tau^*),$$

indexed by unknown values of (V_τ^*, ζ) and substituting the estimated covariance matrix $\hat{\Omega}_n$. By construction, the test statistic is contained in $[0, 1]$. Similarly, I construct critical values as functions of the parameters for a nominal level α, as follows

$$\begin{aligned} q_{\alpha/2}(n, V_\tau^*, \Omega, \zeta) &:= \min\left\{\alpha/2, F_{n,V_\tau^*,\Omega,\zeta}(0)\right\} \\ q_{1-\alpha/2}(n, V_\tau^*, \Omega, \zeta) &:= 1 - \alpha + \min\left\{\alpha/2, F_{n,V_\tau^*,\Omega,\zeta}(0)\right\} \end{aligned} \tag{10}$$

The difference in the critical values is $(1 - \alpha)$ to achieve the desired coverage. The lower critical value is the minimum of the $\alpha/2$ percentile and 0. This adjustment is meant to increase the power of tests of homogeneity (see Remark 2). I propose an adaptive confidence interval by substituting the $\hat{\Omega}_n$, n, and $\hat{V}_{\tau n}$ into the following formula

$$\widehat{CI}_{\alpha n} = \Bigg\{ V_\tau^* \in \mathbb{R}_+, \zeta \in \{-1, 1\} : \tag{11}$$
$$F_{n,V_\tau^*,\hat{\Omega}_n,\zeta}(\hat{V}_{\tau n} - V_\tau^*) \in \left[q_{\alpha/2}(n, V_\tau^*, \hat{\Omega}_n, \zeta), q_{1-\alpha/2}(n, V_\tau^*, \hat{\Omega}_n, \zeta) \right] \Bigg\}.$$

The set $\widehat{CI}_{\alpha n}$ can be constructed via a grid search between 0 and an arbitrarily high value, to test whether a particular V_τ^* satisfies the inequality constraints. The procedure achieves correct asymptotic size because the test statistic converges to a uniform random variable in $[0, 1]$ for each value of V_τ^*. In general, the distribution in (8) depends on the value of ζ and I obtain a conservative interval in (11) by considering the union of intervals with different values of ζ. Moreover, if the off-diagonal element of Ω_n is zero, then the distribution of the empirical process in (7) does not depend on the value of ζ. This property is plausible and I introduce primitive conditions that satisfy it in Section 4. Under those conditions the confidence interval has exact asymptotic coverage .

The procedure is fast because at each point in the grid the researcher evaluates the condition in (11), using the same estimate of $(\hat{V}_{\tau n}, \hat{\Omega}_n)$. The critical values can be computed numerically from the quantiles of a generalized Chi-square with distribution F, which are available in most statistical software packages.

Remark 1 (Equivalence of homogeneity test, $\beta_2 = 0$). Researchers can test for homogeneity by evaluating whether $0 \in \widehat{CI}_{\alpha n}$. By definition, $e_1' \hat{\Omega}_n^{1/2} Z = \sqrt{\Omega_{n,11}} Z_1$, where $\Omega_{n,11}$ is the upper-left entry. Under the null, the test statistic is $F_{n,0,\hat{\Omega}_n,\zeta}(v) = \mathbb{P}(\hat{\Omega}_{n,11} Z_1^2 / n \leqslant v)$, which is the CDF of a rescaled Chi-square distribution with one degree of freedom. Furthermore, the critical values are $\{0, 1 - \alpha\}$, given that $F_{n,0,\hat{\Omega}_n,\zeta}(0) = 0$. Neither quantity depends on the choice of ζ. Because of the normalization in (6), we can choose $\hat{\Omega}_{n,11} = \hat{V}_{xn} \hat{\mathbb{V}}(\hat{\beta}_{2n})$, where $\hat{\mathbb{V}}(\hat{\beta}_{2n})$ is an estimate of the asymptotic variance of $\hat{\beta}_{2n}$ such as the robust sandwich estimator. Therefore, evaluating $F_{n,0,\hat{\Omega}_n,\zeta}(\hat{V}_{\tau n}) \in [0, 1 - \alpha]$ is algebraically equivalent to a test of whether $n(\hat{\beta}_{2n}^2 \hat{V}_{xn} / (\hat{V}_{xn} \hat{\mathbb{V}}(\hat{\beta}_{2n}))) = n\hat{\beta}_{2n}^2 / \hat{\mathbb{V}}(\hat{\beta}_{2n})$ exceeds the $1 - \alpha$ quantile of a Chi-square with one degree of freedom. This is identical to a test of $\beta_2 = 0$ in the regression in (3).

Remark 2 (Adjusting critical values). The critical value $q_{\alpha/2}(V_\tau^*, \Omega, n, \zeta)$ in (10) is constructed to guarantee that $\hat{V}_{\tau n} \in \widehat{CI}_{\alpha n}$. In this case, $\hat{V}_{\tau n}$ belongs to the CI if and only if $F_{n,\hat{V}_{\tau n},\hat{\Omega}_n,\zeta}(0)$ is contained in the critical region for some $\zeta \in \{-1, 1\}$. The unadjusted CI with critical values $\{\alpha/2, 1 - \alpha/2\}$ is not guaranteed to contain the test statistic.[7] Another rationale for doing the adjustment in (10), is to increase the power of the test of homogeneity, $0 \in \widehat{CI}_{\alpha n}$, relative to a test based on the unadjusted CI. The unadjusted test has correct size but the rejection region is discontinuous: it rejects when the test statistics is very close to zero or when it exceeds a threshold. Instead, the adjusted test shifts the critical region left and has the form of a Chi-squared test. It only rejects the null if the test statistic is larger than $1 - \alpha$, which is a threshold that is smaller than $1 - \alpha/2$ for the unadjusted CI.

Remark 3 (Comparison to other tests of homogeneity). Crump et al. (2008) suggest estimating $\hat{\mu}_d(x)$ by a series estimator with K terms, for subsamples $D = d \in \{0, 1\}$. They propose a bias-corrected Wald statistic, which takes the form $\hat{T}_n^{series} := \{[\hat{\xi}_1 - \hat{\xi}_0]'[\hat{\mathbb{V}}(\hat{\xi}_1 - \hat{\xi}_0)]^{-1}[\hat{\xi}_1 - \hat{\xi}_0] - (K - 1)\} / \sqrt{2(K - 1)}$, where $(\hat{\xi}_1, \hat{\xi}_0)$ are non-intercept coefficients associated with $\hat{\mu}_1(x)$ and $\hat{\mu}_0(x)$, respectively and K is the number of covariates. For regressions with univariate X as in (3), $K = 2$ and $\hat{T}_n^{series} = (n\hat{\beta}_{2n}^2 / \hat{\mathbb{V}}(\hat{\beta}_{2n}) - 1) / \sqrt{2}$. Essentially this is just a transformation of the test statistic proposed above, which will produce the same acceptance/rejection result for significance level α (using the critical values in their equation 3.11). Ding et al. (2019)

---

[7] For example, suppose that $\hat{V}_{\tau n} = 0$. Then the empirical process has a Chi-square distribution for $V_\tau = 0$. Since the unadjusted critical value is bounded away from zero, $\hat{V}_{\tau n}$ would not be contained in the unadjusted CI.

study a framework with a fixed population where the only source of randomness is the experimental assignment of offers. They propose a similar Wald estimator, but replace estimates of $(\hat{\xi}_1, \hat{\xi}_0)$ and the asymptotic variance with randomization inference counterparts. In samples with large n, this leads to very similar test statistics, but may produce slightly different results in small samples.

The approach that I introduce in the following section differs substantially in the way that I handle multivariate cases. For $K > 2$, the approaches are non-nested because I use sample splitting and consider a wider range of methods to estimate $\hat{\mu}_d(x)$ than series estimators.

3 Inference for nonparametric CATE

In this section I provide an overview of the inference problems associated with efficient estimators of the VCATE and how to solve them for the nonlinear/high-dimensional case. Let $\{Y_i, D_i, X_i\}$ be i.i.d.. As shown in Newey (1990), efficient estimators can be decomposed as

$$\sqrt{n}(\hat{V}_{\tau n} - V_{\tau n}) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \varphi_i + \underbrace{\text{Residual}_n}_{o_p(1)}, \tag{12}$$

where φ_i is an i.i.d. realization from the efficient influence function with mean zero, and the residual becomes asymptotically negligible as $n \to \infty$. The semiparametric lower bound is $\mathbb{V}(\varphi_i)$. Let $\eta(\cdot)$ be a set of nuisance functions defined as

$$\eta(x) := (\tau(x), \mu_0(x), p(x), \tau_{av}). \tag{13}$$

Levy et al. (2021) showed that the efficient influence function for the VCATE is equal to $\varphi_i = \varphi(Y_i, D_i, X_i, \eta) - V_{\tau n}$, where φ is defined as

$$\varphi(y, d, x, \eta) := (\tau(x) - \tau_{av})^2 + 2(\tau(x) - \tau_{av}) \left[\frac{d(y - \mu_0(x) - \tau(x))}{p(x)} - \frac{(1 - d)(y - \mu_0(x))}{1 - p(x)} \right]. \tag{14}$$

By (12), all efficient estimators –regardless of their form– are $\sqrt{n}$–asymptotically equivalent to $\frac{1}{n} \sum_{i=1}^{n} \varphi_i$. Let $\varphi_i = \varphi(Y_i, D_i, X_i, \eta(X_i))$ be a realization of the efficient

influence function. If $\mathbb{V}(\varphi_i) > 0$, then

$$\mathbb{V}(\varphi_i)^{-1} \times \sqrt{n}(\hat{V}_{\tau n} - V_{\tau n}) = \mathbb{V}(\varphi_i)^{-1}\left[\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\varphi_i\right] + o_p(1) \rightarrow^p \mathcal{N}(0,1).$$

In this case, any confidence interval based on normal critical values and a consistent estimator of $\mathbb{V}(\varphi_i)$, produces valid coverage. For common functionals such as the ATE, $\mathbb{V}(\varphi_i) > 0$. However, this cannot be guaranteed for the VCATE.

Lemma 3. *Let* $\sigma_d^2(x) := \mathbb{V}(Y \mid D = d, X = x)$.

$$\mathbb{V}(\varphi_i) = \mathbb{V}((\tau(X) - \tau_{av})^2) + 4\mathbb{E}\left[(\tau(X) - \tau_{av})^2\left(\frac{\sigma_1^2(X)}{p(X)} + \frac{\sigma_0^2(X)}{p(X)}\right)\right]. \tag{15}$$

Both of the inner terms in (15) are multiplied by $(\tau(x) - \tau)$. When the VCATE is zero, $\tau(x) = \tau$ almost surely, and the influence function is degenerate. The condition that $\mathbb{V}(\varphi_i) > 0$ does not hold uniformly over all V_τ in the parameter space. In this case, the distribution of $\sqrt{n}(\hat{V}_{\tau n} - V_{\tau n})$ is dominated by the higher order terms of the residual (12), and the CLT cannot be applied to guarantee normality near the boundary. The linear estimator discussed in the previous section is just one example. Moreover, if the tails of $\tau(x)$ are thin, then the value of $\mathbb{V}(\varphi_i)$ is also small near the boundary.

Corollary 2. *If* $\mathbb{E}[(\tau(X) - \tau)^4] \leqslant \kappa^2 V_\tau^2$ *for* $\kappa \in \mathbb{R}_+$, *then*

$$\mathbb{V}(\varphi_i) \leqslant \kappa^2 V_\tau^2 + 4\kappa V_\tau\sqrt{\mathbb{E}\left[\left(\frac{\sigma_1^2(X)}{p(X)} + \frac{\sigma_0^2(X)}{1 - p(X)}\right)\right]}.$$

Corollary 2 shows that the variance of the efficient influence function is bounded by a quantity that scales up or down proportional to the value of V_τ. Consequently, when $\mathbb{V}(\varphi_i)$ is relatively small, the higher order terms in the residual may still dominate.

3.1 Pseudo-VCATE, regressions, and efficiency

A robust way to introduce nonlinearity is to consider a regression with real-valued basis functions $M(x)$ and $S(x)$. For now, I will leave these unspecified but in the next section I will show how they can be estimated non-parametrically in a first stage.

$$Y = \underbrace{c_0 + M(X)c_1 + [D - p(X)]\beta_1 + [D - p(X)]S(X)\beta_2}_{W(X,D)'\theta} + U \tag{16}$$

with weights $\lambda(X) := [p(X)(1 - p(X))]^{-1}$, regressors $W(X, D)$, and parameters $\theta = [c_0, c_1, \beta_1, \beta_2]$. This specification accommodates experiments with heterogeneous assignment probabilities.[8] Consider the following minimizers:

$$\Theta^* := \arg\min_{\theta \in \mathbb{R}^4} \mathbb{E}[\lambda(X)(Y - W(X,D)'\theta)^2]. \tag{17}$$

Chernozhukov et al. (2022a) showed that if $\mathbb{V}(S(X)) > 0$, $\mathbb{E}[S(X)] = 0$, and the vector (β_1, β_2) are part of the solution to (17), then (β_1, β_2) are also the intercept and slope of the best linear projection of $\tau(X)$ on $S(X)$.[9] Hence the pseudo-VCATE has an upper bound, $V_\tau^* = \beta_2^2 \mathbb{V}(S(X)) \leqslant V_\tau$. Because of this bound, if $V_\tau = 0$, then the pseudo-VCATE (V_τ^*) will not falsely detect heterogeneity even if $S(X)$ is misspecified. If anything, poor choices of $S(X)$ will possibly understate the amount of heterogeneity. When $\tau(X)$ is spanned by $S(X)$, $V_\tau^* = V_\tau$ and the two notions coincide.

To obtain a feasible estimator we define $\widehat{S}(x) := S(x) - \frac{1}{n}\sum_{i=1}^n S(X_i)$, and compute $\widehat{W}(x, d)$ by substituting $\widehat{S}(x)$ in (16). Now consider a value of $\widehat{\theta}_n$ that minimizes $\frac{1}{n}\sum_{i=1}^n \lambda(X_i)(Y_i - \widehat{W}(X_i, D_i)'\theta)^2$, by solving the first order condition

$$\mathcal{Q}(\widehat{\theta}_n) := \frac{1}{n}\sum_{i=1}^n \lambda(X_i)(Y - \widehat{W}(X_i, D_i)'\widehat{\theta}_n)\widehat{W}(X_i, D_i)'. \tag{18}$$

The regression parameters can be used to construct the CATE and other nuisance functions. For a given $\theta \in \mathbb{R}^4$,

$$\tilde{\eta}_\theta(x) := \begin{pmatrix} \tilde{\tau}_\theta(x) \\ \tilde{\mu}_{0,\theta}(x) \\ \tilde{p}_\theta(x) \\ \tilde{\tau}_{av,\theta} \end{pmatrix} = \begin{pmatrix} (\widehat{W}(x,1) - \widehat{W}(x,0))'\theta \\ \widehat{W}(x,0)'\theta \\ p(x) \\ \frac{1}{n}\sum_{i=1}^n \tilde{\tau}_\theta(X_i) \end{pmatrix}. \tag{19}$$

---

[8]When $p(x) = 1/2$ this produces exactly the same coefficients as a regression of Y on $(1, M(X), D, D \times S(X))$, but differs when the probabilities are heterogeneous. If $M(X) = S(X) = X$ as well, this reduces to (3).

[9]When $\mathbb{V}(S(X)) = 0$, β_2 does not have a unique solution in (17), but $V_\tau^* = \beta_2^2 \mathbb{V}(S(X)) = 0 \leqslant V_\tau$ is still the best linear projection, regardless of the value of β_2.

Lemma 4 shows that the sample variance of the estimated CATE can be interpreted as an estimator that plugs-in (19) to the efficient influence function in (14).

Lemma 4. *Define* $\hat{V}_{xn} = \frac{1}{n}\sum_{i=1}^{n}\hat{S}(X_i)^2$. *Let* φ, $\tilde{\eta}_\theta$ *and* $\mathcal{Q}(\theta)$ *be defined as in* (14), (18), *and* (19), *respectively. If* $\hat{\theta}_n = (\hat{c}_{1n}, \hat{c}_{2n}, \hat{\beta}_{1n}, \hat{\beta}_{2n})$ *solves* $\mathcal{Q}(\hat{\theta}_n) = 0$, *then* $\hat{V}_{\tau n} := \hat{\beta}_{2n}^2 \hat{V}_{xn} = \frac{1}{n}\sum_{i=1}^{n}\varphi(Y_i, D_i, X_i, \tilde{\eta}_{\hat{\theta}_n})$.

Mechanically, the influence function can be decomposed into primary and bias-correction components. As an intermediate step for Lemma 4, I show that the bias-correction terms and the fourth component of (18) are proportional to each other. The optimal $\hat{\theta}_n$ implicitly sets the average bias correction to zero. Intuitively, the linear model minimizes the covariate imbalances between the treatment and control group in-sample. Lemma 4 suggests that $\hat{V}_{\tau n}$ could be asymptotically efficient if $\tilde{\eta}_{\hat{\theta}_n}$ is sufficiently close to η. In Section 4.1, I show that my proposed semiparametric estimator can indeed achieve this.

As a preliminary step, it is necessary to determine which $S(x)$ and $M(x)$ ensure that $\tilde{\eta}_\theta = \eta$. Not all choices achieve this property.[10] However, if they are chosen in such a way that $W(x,d)'\theta = \mu_d(x)$ for some $\theta \in \mathbb{R}^4$, then that's sufficient to guarantee that $\eta_\theta = \eta$. Lemma 5 shows that any θ with this property is also a solution to the regression problem, and provides guidance on the choice of $S(x)$ and $M(x)$.

Lemma 5. *Let* Θ^* *be the optimizer set defined in* (17). *If (i)* $\mathbb{E}[S(X)] = 0$ *and (ii)* $W(x,d)'\theta = \mu_d(x)$ *for some* $\theta \in \mathbb{R}^4$, *then* $\theta \in \Theta^*$. *Conditions (i) can be satisfied by setting* $S(x) = \tau(x) - \mathbb{E}[\tau]$. *Condition (ii) can be satisfied by setting* $M(x) = \mu_0(x) + p(x)\tau(x)$. *In this special case,* $\theta = (0, 1, \mathbb{E}[\tau(X)], 1)' \in \Theta^*$.

Lemma 5 provides efficient choices of $S(x)$ and $M(x)$ that can be expressed in terms of conditional moments, and that for this choice, the optimal θ has a known, simple form. In practice, $S(x)$ and $M(x)$ can be estimated non-parametrically.

3.2 Multi-step approach

My proposed procedure randomly partitions the observations $\mathcal{I}_n := \{1, \ldots, n\}$ into K folds of equal size $n_k := n/K$. Denote the observations in each fold by $\mathcal{I}_{nk}$, so that

---

[10]This point highlights that while all regressions of the form in (16) proposed by Chernozhukov et al. (2022a) estimate an interpretable V_τ^* –regardless of the choice of $M(x)$–, not every regression in this class is efficient. The functions $S(x)$, and $M(x)$ in particular, both affect efficiency.

$\bigcup_{k=1}^{K} \mathcal{I}_{nk} = \mathcal{I}_n$, and let $\mathcal{I}_{-nk} := \mathcal{I}_n \backslash \mathcal{I}_{nk}$ be the set of observations that are not in fold k. In a slight abuse of notation, I use $\mathcal{I}_{-nk}$ when defining conditional expectations, to denote the full set of random variables associated with observations not included in fold k. For simplicity, I also label the fold of observation $i \in \mathcal{I}_{nk}$, by k_i.

Let $\hat{\eta}_{-k}(x) := (\hat{\tau}_{-k}(x), \hat{\mu}_{0,-k}(x), p(x), \hat{\tau}_{-k,av})$ denote a prediction of the nuisance function $\eta(x)$ over the set $\mathcal{I}_{-nk}$, using the researcher's preferred prediction algorithm. This could include traditional methods such as linear regression, or more modern "machine learning" approaches such as LASSO, neural networks, or random forests. The only function that is known in advance is the propensity score, since I restrict attention to randomized experiments. Guided by Lemma 5, define

$$
\begin{aligned}
S_{-k}(x) &:= \hat{\tau}_{-k}(x) - \mathbb{E}[\hat{\tau}_{-k}(X_i) \mid \mathcal{I}_{-nk}], \\
M_{-k}(x) &:= \hat{\mu}_{0,-k}(x) + p(x)\hat{\tau}_{-k}(x), \\
\lambda(x) &:= [p(x)(1 - p(x))]^{-1}, \\
W_i' &:= \begin{bmatrix} 1 & M_{-k_i}(X_i) & (D_i - p(X_i)) & (D_i - p(X_i))S_{-k_i}(X_i) \end{bmatrix}
\end{aligned} \tag{20}
$$

Consider a regression with weights $\lambda(X_i)$, parameters $\theta := (c_1, c_2, \beta_1, \beta_2)$, and

$$
Y_i = W_i' \theta_{nk} + U_i, \qquad \mathbb{E}[W_i U_i \mid \mathcal{I}_{-nk}] = 0 \tag{21}
$$

In practice, $\mathbb{E}[\hat{\tau}_{-k}(X_i) \mid \mathcal{I}_{-nk}]$ needs to be estimated, and I use a sample analog:

$$
\begin{aligned}
\hat{S}_{-k}(x) &:= \hat{\tau}_{-k}(x) - \hat{\tau}_{nk,av}, \qquad \hat{\tau}_{nk,av} := \frac{1}{n_k} \sum_{i \in \mathcal{I}_{nk}} \hat{\tau}_{-k}(X_i), \\
\widehat{W}_i' &:= \begin{bmatrix} 1 & M_{-k_i}(X_i) & (D_i - p(X_i)) & (D_i - p(X_i))\hat{S}_{-k_i}(X_i) \end{bmatrix}, \\
\hat{\theta}_{nk} &:= \left[\frac{1}{n} \sum_{i \in \mathcal{I}_{nk}} \widehat{W}_i \widehat{W}_i' \right]^{-1} \left[\frac{1}{n} \sum_{i \in \mathcal{I}_{nk}} \widehat{W}_i Y_i \right].
\end{aligned} \tag{22}
$$

Let $\hat{\theta}_{nk} = (\hat{c}_{1nk}, \hat{c}_{2nk}, \hat{\beta}_{1nk}, \hat{\beta}_{2nk})$ be the estimator over the subsample $\mathcal{I}_{nk}$. The fold-specific variance of $\hat{S}_{-k_i}(x)$ is defined as

$$
\hat{V}_{xnk} := \frac{1}{n_k} \sum_{i \in \mathcal{I}_{nk}} \hat{S}_{-k_i}(X_i)^2. \tag{23}
$$

The estimator of the VCATE for fold k is

$$\widehat{V}_{\tau nk} := \widehat{\beta}_{2nk}^2 \widehat{V}_{xnk}. \tag{24}$$

In this case $\widehat{V}_{xnk}$ can be viewed as a preliminary estimate of the VCATE using the data in $\mathcal{I}_{-nk}$, whereas $\widehat{V}_{\tau nk}$ is a regression-adjusted estimator that fits the sample $\mathcal{I}_{nk}$. This adjustment will produce better results, with a pseudo-VCATE interpretation even if the first step $\widehat{S}_{-k_i}(x)$ function is noisy, misspecified, or slow to converge to $\tau(x)$. The estimator in (23) belongs to the class of multi-step estimators defined in Chernozhukov et al. (2022a). I add a restriction on the choice of $M_{-k}(x)$, guided by Lemma 5, to ensure asymptotic efficiency.

To quantify the uncertainty in $(\widehat{\theta}_{nk}, \widehat{V}_{xnk})$ I compute a robust (sandwich) estimator. I start by defining two auxiliary residuals, $\widehat{T}_i := \widehat{V}_{xnk}^{-1} \widehat{S}_{-k_i}(X_i)^2 - 1$ and $\widehat{U}_i := Y_i - \widehat{W}_i' \widehat{\theta}_{nk}$. Let $\widehat{\Pi}_{nk}$ be a 4×4 diagonal matrix with diagonal entries $(1, 1, 1, \widehat{V}_{xnk}^{-1/2})$. Researchers can compute estimators of the individual components of the sandwich form $\widehat{J}_{nk}$, $\widehat{H}_{nk}$, and a selection matrix Υ defined as follows

$$\widehat{J}_{nk} := \begin{bmatrix} \frac{1}{n_k} \sum_{i \in \mathcal{I}_k} \lambda(X_i) \widehat{\Pi}_{nk} \widehat{W}_i \widehat{W}_i' \widehat{\Pi}_{nk}' & 0 \\ 0 & 1 \end{bmatrix}, \qquad \Upsilon := \begin{bmatrix} 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \tag{25}$$

$$\widehat{H}_{nk} := \frac{1}{n_k} \sum_{i \in \mathcal{I}_k} \begin{bmatrix} \lambda(X_i)^2 \left[\widehat{U}_i^2 \widehat{\Pi}_{nk} \widehat{W}_i \widehat{W}_i' \widehat{\Pi}_{nk}' \right] & \lambda(X_i) \left[\widehat{U}_i \widehat{\Pi}_{nk} \widehat{W}_i \widehat{T}_i \right] \\ \lambda(X_i) \left[\widehat{U}_i \widehat{W}_i' \widehat{\Pi}_{nk}' \widehat{T}_i \right] & \widehat{T}_i^2 \end{bmatrix}. \tag{26}$$

The sandwich covariance estimator is

$$\widehat{\Omega}_{nk} = \Upsilon \widehat{J}_{nk}^{-1} \widehat{H}_{nk} \widehat{J}_{nk}^{-1} \Upsilon'. \tag{27}$$

The population covariance matrix is

$$\Omega_{nk} = \mathbb{V}\left(\begin{bmatrix} \lambda(X_i)(D_i - p(x_i)) V_{xnk}^{-1/2} S_{-k}(X_i) U_i \\ V_{xnk}^{-1} S_{-k}(X_i)^2 \end{bmatrix} \mid \mathcal{I}_{-nk} \right). \tag{28}$$

When the VCATE is zero, then $\widehat{V}_{xnk}$ (as a consistent estimator of $V_{\tau n}$) should converge to zero along the asymptotic sequence. To prevent asymptotic degeneracy, we need to rescale the estimands along the lines of Assumption 2. The random vari-

21

able $V_{xnk}^{-1/2} S_{-k}(X_i)$ is normalized to (conditionally) have variance one by design, even if $S_{-k}(X_i)$ converges to zero. This requires two much weaker conditions: (i) that $V_{xnk} > 0$, i.e. there is some noise in estimating the CATE;[11] (ii) Ω_{nk} has eigenvalues bounded away from zero. To ensure this, the tails of $S_{-k}(X_i)$ need to be thin.[12]

Assumption 3 (Moment Bounds)**.** *Suppose that there exists a constant $\delta \in (0,1)$ such that for each fold k, almost surely, (i) $\mathbb{E}[M_{-k}(X_i)^2 \lambda(X_i) \mid \mathcal{I}_{-nk}] - \mathbb{E}[M_{-k}(X_i)\lambda(X_i) \mid \mathcal{I}_{-nk}]\mathbb{E}[\lambda(X_i) \mid \mathcal{I}_{-nk}] \geqslant \delta$, (ii) $\mathbb{E}[M_{-k}(X_i)^4 \mid \mathcal{I}_{-nk}] \leqslant 1/\delta$, (iii) $\mathbb{E}[U_i^4 \mid \mathcal{I}_{-nk}] \leqslant (1/\delta)$, and (iv) $\mathbb{E}[S_{-k}(X_i)^4 \mid \mathcal{I}_{-nk}] \leqslant (1/\delta)\mathbb{E}[S_{-k}^2 \mid \mathcal{I}_{-nk}]^2$, (v) $\mathbb{E}[\hat{\tau}_{-k}(X)^2 \mid \mathcal{I}_{-nk}] < 1/\delta$.*

Assumption 3.(i) is a rank condition that ensures that the auxiliary regressor $M_{-k}(X_i)$ is not degenerate. Assumption 3.(ii) ensures that the second-moment of the candidate regressor $M_{-k}(X_i)$ is bounded. Assumption 3.(iii) is a standard condition indicating that the fourth moment of the residuals are bounded. Assumption 3.(iv) is a bounded kurtosis condition indicating that the out-of-sample, machine learning predictions of $\tau(x)$ have thin tails. Finally, Assumption 3.(v) is a bound on the variance of the first-stage VCATE.

Assumption 4 (Non-degeneracy)**.** *The following properties hold almost surely over sequences of random data realizations $\{\mathcal{I}_{n1}, \ldots, \mathcal{I}_{nK}\}_{n=1}^{\infty}$. Conditional on $\mathcal{I}_{-nk}$: (i) $V_{xnk} > 0$, (ii) V_{xnk} has a finite upper bound, (iii) $V_{\tau nk}^* := \beta_{2nk}^2 V_{xnk}$ is contained in a bounded subset of $[0, \infty)$, and (iv) Ω_{nk} defined in (28) is a positive definite matrix with bounded eigenvalues.*

Assumption 5 (Random Sampling)**.** *The observations $\{Y_{0i}, Y_{1i}, D_i, X_i\}_i^n$ are i.i.d. across i for fixed n, and drawn from a sequence of data generating processes $\{\gamma_n\}_{n=1}^{\infty}$.*

Theorem 3 shows how these primitive conditions imply an analog of Assumption 2 for the cross-fitted case.

Theorem 3. *Consider a sequence of random data realizations $\{\mathcal{I}_{n1}, \ldots, \mathcal{I}_{nK}\}_{n=1}^{\infty}$ with associated quantities $\{V_{xnk}, V_{\tau nk}^*, \beta_{2nk}, \Omega_{nk}\}_{n=1}^{\infty}$ for each k, as well as a sequence of estimators $\{\hat{\beta}_{2nk}, \hat{V}_{xnk}, \hat{V}_{\tau nk}^*, \hat{\Omega}_{nk}\}_{n=1}^{\infty}$ computed from (22), (23), (24), and (27), respec-*

---

[11]I also propose an extension that allows for $V_{xnk} = 0$ in Remark 6.

[12]One sufficient additional restriction is that $\mathbb{E}[U_i \mid X_i, D_i, \mathcal{I}_{-nk}] = 0$ (the model is correctly specified), $\mathbb{V}(U_i \mid X_i, D_i \mid \mathcal{I}_{-nk})$ is bounded away from zero, and $S_{-k}(X_i)$ has bounded kurtosis. In that case the off-diagonal elements of Ω_{nk} are zero and the diagonals are uniformly bounded. Positive-definiteness may also hold in a neighborhood where the nuisance functions are close to the true value and $\mathbb{E}[U_i \mid X_i, D_i, \mathcal{I}_{-nk}] \approx 0$.

tively. Suppose that these quantities satisfy Assumptions 1.(ii), 3, 4, and 5. Then as $n_k \to \infty$, for all $k \in \{1, \ldots, K\}$, Conditional on a sequence of $\mathcal{I}_{-nk}$,

(i) $\quad \Omega_{nk}^{-1/2} \sqrt{n_k} \begin{pmatrix} \sqrt{V_{xnk}}(\hat{\beta}_{2nk} - \beta_{2nk}) \\ \frac{\hat{V}_{xnk}}{V_{xnk}} - 1 \end{pmatrix} \mid \mathcal{I}_{-nk} \to Z_{nk} \sim \mathcal{N}(0, I_{2\times2}).$

(ii) $\quad \hat{\Omega}_{nk} \to^p \Omega_{nk}.$

Theorem 3 presents a central limit theorem for the components of $\hat{V}_{\tau nk}^* = \hat{\beta}_{2nk}^2 \hat{V}_{xnk}$, properly rescaled and conditional on $\mathcal{I}_{-nk}$. This result holds regardless of whether the nuisance parameters are properly specified and primarily relies on the independence of the folds. By Lemma 2, conditional on $\mathcal{I}_{-nk}$,

$$\hat{V}_{\tau nk} - V_{\tau nk}^* = G(n_k, V_{\tau nk}^*, \Omega_{nk}, Z_{nk}) + o_p\left(\frac{1}{n_k}\right) + o_p\left(\sqrt{\frac{V_{\tau nk}^*}{n_k}}\right) + o_p\left(\frac{V_{\tau nk}^*}{\sqrt{n_k}}\right). \quad (29)$$

Then it is possible to construct adaptive confidence intervals, substituting the sample size n_k and estimated statistics $(\hat{V}_{\tau nk}, \hat{\Omega}_{nk})$.

$$\widehat{CI}_{\alpha nk} = \left\{ V_\tau^* \in \mathbb{R}_+, \zeta \in \{-1, 1\} : \right.$$

$$\left. F_{n_k, V_\tau^*, \hat{\Omega}_{nk}, \zeta}(\hat{V}_{\tau nk} - V_\tau^*) \in \left[q_{\alpha/2}(n_k, V_\tau^*, \hat{\Omega}_{nk}, \zeta), q_{1-\alpha/2}(n_k, V_\tau^*, \hat{\Omega}_{nk}, \zeta) \right] \right\}.$$

$$(30)$$

The confidence intervals take the same form as in the regression case in (11), except that now the inputs are obtained from the cross-fitted regression step. The confidence interval is fast to compute because $(\hat{V}_{\tau nk}, \hat{\Omega}_{nk})$ only needs to be computed once. It is worth noting that because the confidence interval only uses information in fold $\mathcal{I}_{nk}$, the effective sample size is n_k. While this does not affect the nominal asymptotic size of the confidence interval, it may affect the power of tests against specific alternatives.

3.2.1 Ensemble estimator

We can construct an "ensemble" to aggregate across folds, defined as follows

$$\hat{V}_{\tau n} := \frac{1}{K} \sum_{k=1}^{K} \hat{\beta}_{2nk}^2 \hat{V}_{xnk}. \quad (31)$$

In Section 4.1, I show that this ensemble estimator is efficient.

3.2.2 Splitting uncertainty and median intervals

So far in this section we have used the data from a single split or fold of the data. However, the choice of fold k or the particular split may lead to different values of $\hat{V}_{\tau nk}$ and hence distinct confidence intervals. Chernozhukov et al. (2022a) propose an aggregation procedure based on "median parameter" confidence intervals, inspired by false-discovery rate adjustments. Their proposed conditional t-tests are not directly applicable here because $\hat{V}_{\tau nk}$ conditionally converges to a generalized Chi-square. However, I show that the basic idea can still be adapted.

Let K be the total number of folds, obtained across one or more splits of the data. For instance, a 2-fold sample with 10 splits would have $K = 20$, Let $\inf \widehat{CI}_{\alpha nk}$ and $\sup \widehat{CI}_{\alpha nk}$ denote the lower and upper bounds of $\widehat{CI}_{\alpha nk}$, respectively, and $\mathrm{Med}_K\{\cdots\}$ denote the median over a set indexed by $k = \{1, \ldots, K\}$. If K is even then two quantities might be tied for the median, and in that case I compute their midpoint. The multifold confidence interval is defined as

$$\widehat{CI}_{\alpha n}^{\mathrm{multifold}} = \left[\mathrm{Med}_K \left\{ \inf \widehat{CI}_{\frac{\alpha}{2} nk} \right\}, \mathrm{Med}_K \left\{ \sup \widehat{CI}_{\frac{\alpha}{2} nk} \right\} \right]. \tag{32}$$

Intuitively, the K fold-specific intervals "vote" to include a particular value, and $V_\tau^* \in \widehat{CI}_{\alpha n}^{\mathrm{multifold}}$ only if there is a majority vote. The "median" interval $\widehat{CI}_{\alpha n}^{\mathrm{multifold}}$ contains values within the median lower bound and the median upper bound across folds. To control the overall false discovery rate, I adjust the nominal size to $\alpha/2$. This adjustment produces a conservative interval because it assumes a worst-case dependence structure between the folds and the splits, regardless of the size of K. In some instances, the asymptotic coverage probability may be strictly higher than $(1 - \alpha)$, particularly when there is a lot of heterogeneity.[13] At the boundary, with low effect heterogeneity or none at all, it is much harder to asses the dependence structure between the fold-specific estimators. One of the benefits of using a worst-case approach is that it provides coverage guarantees under weak assumptions. Moreover, the empirical example illustrates that even though these intervals are conservative, they may have a short length in practice.

---

[13]For instance, given a single split, Theorem 6 implies that the $\{\sqrt{n_k}(\hat{V}_{\tau nk} - V_\tau)\}_{k=1}^K$ are asymptotically uncorrelated. However, near the boundary, the estimators converge at a rate faster than $\sqrt{n}$ and their relative dependence structure at that rate is unclear.

4 Large Sample Theory

4.1 $\sqrt{n}$-Consistency, Efficiency, and Boundary Rates

Let $\gamma \in \Gamma$ denote a probability distribution over i.i.d observations $(Y_{1i}, Y_{0i}, D_i, X_i)$. I use the notation $\mathbb{E}_\gamma[\cdot]$ and $\mathbb{P}_\gamma(\cdot)$ to denote the expectation and probability under γ, respectively. Let $S_{-k}(x)$ be the function defined in (20). The true value of the CATE and VCATE is given by τ_γ and $V_\tau(\gamma)$, respectively. The pseudo-VCATE is given by

$$V_\tau^*(\gamma, \mathcal{I}_{-nk}) := V_\tau(\gamma) - \inf_{(\beta_1, \beta_2) \in \mathbb{R}^2} \mathbb{E}_\gamma[(\tau_\gamma(X) - \beta_1 - \beta_2 S_{-k}(X))^2 \mid \mathcal{I}_{-nk}]. \qquad (33)$$

Define the estimation error of the CATE in the L_2 norm as

$$\omega(\gamma) := \sqrt{\mathbb{E}_\gamma[\|\hat{\tau}_{-k}(X) - \tau_\gamma(X)\|^2]}. \qquad (34)$$

We can bound the difference between the pseudo-VCATE and its true value:

Theorem 4 (Bias of the pseudo-VCATE). *Under the distribution $\gamma \in \Gamma$,*

$$\mathbb{E}_\gamma[|V_\tau(\gamma) - V_\tau^*(\gamma, \mathcal{I}_{-nk})|] \leqslant \min\left\{16 \times \omega(\gamma)^2, V_\tau(\gamma)\right\}. \qquad (35)$$

Theorem 4 derives a non-asymptotic bound for the VCATE as the minimum of two key quantities: (i) the conditional L_2 error between the candidate function and the true CATE, and (ii) the true value of the VCATE. This proof only relies on the definition in (33). For instance, when $V_\tau(\gamma) = 0$, then $V_\tau(\gamma) - V_\tau^*(\gamma, \mathcal{I}_{-nk}) = 0$, regardless of whether $\hat{\tau}_{-k}(\cdot)$ is properly specified. The difference between the two quantities is also small if $\omega(\gamma)$ is sufficiently close to zero. In the multi-step approach, $\omega(\gamma)$ captures the first-stage uncertainty from estimating the CATE, which decreases with sample size. I consider the following convergence condition.

Assumption 6 (Convergence CATE). $\sqrt{n_k}\omega(\gamma_n)^2 = o(1)$ *as $n \to \infty$.*

Assumption 6 imposes an L_2 consistency condition on the CATE. A large class of machine learning models can meet this requirement. For example, Bickel et al. (2009) and Belloni et al. (2014) evaluate rates of convergence under sparse models, Chen and White (1999) for neural networks, and Wager and Walther (2015) for regression trees and random forest.

Theorem 5 (Faster than $\sqrt{n}$ convergence near boundary). *Consider a sequence of data generating processes $\{\gamma_n\}_{n=1}^{\infty}$ where $V_\tau(\gamma_n) \to 0$ as $n \to \infty$ and Assumptions 1, 3, 4, 5 and 6 hold. Define $\Delta_{nk} := \left(\widehat{V}_{\tau nk} - V_\tau(\gamma_n)\right)$ for the estimator defined in (24). Then (i) $\sqrt{n_k}\Delta_{nk} = o_p(1)$, and (ii) if in addition $n_k^{1/2+\rho}V_\tau(\gamma_n) = o(1)$ for $\rho \in [0, 1/2)$, then $n_k^{1/2+\rho}\Delta_{nk} = o_p(1)$.*

Theorem 5 shows that multi-step estimators of the VCATE converge to zero faster than $\sqrt{n_k}$ near the boundary. I formalize "near" by considering sequences of distributions where the VCATE approaches zero. Theorem 5 relies on the non-asymptotic bound in Theorem 4, the normal approximation in Theorem 3, and the empirical process in Lemma 2. There is no requirement on the rate of convergence of $\widehat{\mu}_{-0k}(\cdot)$ (and consequently on the generated regressor $M_{-k_i}(\cdot)$), only an assumption that $p(x)$ is known and that the CATE is estimated at a sufficiently fast rate. Furthermore, if the true CATE is nearly flat in the sense that for $\rho \in [0, 1/2)$, then $n_k^{1/2+\rho}V_\tau(\gamma_n) = o(1)$ (or even exactly equal to zero), then the estimator has a faster rate guarantee.

To prove efficiency we have the stronger requirement that all the nuisance functions converge to their true value in the L_4 norm and at $n_k^{1/4}$ rate in the L_2 norm.

Assumption 7 (Regularity conditions). *Define the residuals $U_i = Y_i - \mathbb{E}_{\gamma_n}[Y_i \mid D_i, X_i]$. (i) $\mathbb{E}_{\gamma_n}[\|Y_i\|^4]$, $\mathbb{E}_{\gamma_n}[\|U_i\|^4]$, $\mathbb{E}_{\gamma_n}[\|\eta(X_i)\|^4]$, (ii) $\mathbb{E}_{\gamma_n}[\|\widehat{\eta}_{-k}(X_i)\|^4]$ are uniformly bounded, (iii) $\mathbb{E}_{\gamma_n}\left[\|\widehat{\eta}_{-k}(X) - \eta(X_i)\|^4\right] \to 0$, (iv) $\sqrt{n_k}\mathbb{E}_{\gamma_n}[\|\widehat{\eta}_{-k}(X_i) - \eta(X_i)\|^2] = o(1)$ for all $k \in \{1, \ldots, K\}$.*

The next step is to show that the estimation error of the fold-specific VCATE converges at $\sqrt{n_k}$ to an average of efficient influence functions.

Theorem 6 ($\sqrt{n}$ Consistency and Efficiency). *Consider a sequence of data generating processes $\{\gamma_n\}_{n=1}^{\infty}$ where $V_\tau(\gamma_n) \to 0$ as $n \to \infty$ and Assumptions 1, 3, 4, 5, 6, and 7 hold. Then*

$$\sqrt{n_k}(\widehat{V}_{\tau nk} - V_\tau(\gamma_n)) = \frac{1}{\sqrt{n_k}} \sum_{i \in \mathcal{I}_{nk}} \varphi_i + o_p(1).$$

Theorem 6 shows that the fold-specific estimator converges at $\sqrt{n_k}$-rate to an average of i.i.d influence function. This requires standard regularity conditions. The proof of Theorem 6 is non-standard due to the multi-step nature of the procedure. I start by applying Lemma 4, which shows hows to write $\widehat{V}_{\tau nk}$ as an average of estimated influence functions. I break down the proof into sequences where $V_\tau(\gamma_n)$ converges to

26

zero and those where it's bounded away from zero. For the first part, I leverage (a) the boundary convergence result in Theorem 5, (b) the bound for $\mathbb{V}(\varphi_i)$ in Lemma 2. For the second part, I provide a novel decomposition of regression adjusted nuisance functions. The key is to prove that the regression parameters $\hat{\theta}_{nk}$ converge at $n_k^{1/4}$ rate to the values in Lemma 5 for sequences where $V_\tau(\gamma_n) \to V_\tau > 0$. Once in this form, the rest of the proof relies on a traditional Taylor expansion argument.

The ensemble estimator $\hat{V}_{\tau n}$ combines information from the whole sample. By definition $n = n_k \times K$ and K is finite, which means that algebraically $\sqrt{n}(\hat{V}_{\tau n} - V_\tau(\gamma_n)) = \frac{\sqrt{n_k}}{\sqrt{K}} \sum_{k=1}^{K} (\hat{V}_{\tau nk} - V_\tau(\gamma_n))$, and by Theorem 6,

$$\sqrt{n}(\hat{V}_{\tau n} - V_\tau(\gamma_n)) = \left[\frac{1}{\sqrt{n_k K}} \sum_{k=1}^{K} \sum_{i \in \mathcal{I}_{nk}} \varphi_i + \frac{1}{\sqrt{K}} \sum_{k=1}^{K} o_p(1) \right] = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \varphi_i + o_p(1).$$

This means that aggregating the estimators restores full efficiency, satisfying the property described in (12).

4.2 Asymptotic Coverage

I start by showing that the single fold confidence interval has uniform coverage for the pseudo-VCATE, and exact coverage under an additional assumption.

Assumption 8 (Exact coverage condition). *Let $\Omega_{nk,12}$ be the off-diagonal element of Ω_{nk}. For each $t > 0$, $\limsup_{n \to \infty} \sup_{\gamma \in \Gamma} \mathbb{P}_\gamma(\sqrt{V_\tau^*(\gamma, \mathcal{I}_{-nk})} |\Omega_{nk,12}| > t) = 0$.*

Assumption 8 states that the product of the pseudo-VCATE and the off-diagonal element of the limiting covariance matrix in (28) needs to converge to zero uniformly.

Theorem 7 (Uniform Coverage of Pseudo-VCATE). *Let Γ denote a set of distributions, constrained in such a way that Assumptions 1, 3, 4, and 5 hold. Let $\widehat{CI}_{\alpha nk}$ and $V_\tau^*(\gamma, \mathcal{I}_{-nk})$ be defined as in (30) and (33), respectively. Then*

$$1 - \alpha \leqslant \liminf_{n \to \infty} \inf_{\gamma \in \Gamma} \mathbb{P}_\gamma \left(V_\tau^*(\gamma, \mathcal{I}_{-nk}) \in \widehat{CI}_{\alpha nk} \right) \tag{36}$$

If Assumption 8 also holds, then

$$\limsup_{n \to \infty} \sup_{\gamma \in \Gamma} \mathbb{P}_\gamma \left(V_\tau^*(\gamma, \mathcal{I}_{-nk}) \in \widehat{CI}_{\alpha nk} \right) \leqslant 1 - \alpha. \tag{37}$$

Theorem 7 shows that the confidence intervals always have uniform coverage of the pseudo-VCATE of at least $(1 - \alpha)$.[14] The key is to prove that the confidence intervals yield coverage under arbitrary sequences of distributions, which includes cases where $V_{\tau n}^*$ is either equal to zero or approaches zero as $n \to \infty$. The proof builds on the approximation of Lemma 2 and shows that for every sequence, the test statistic for a particular $\zeta_{n_k} \in \{-1, 1\}$ converges to a uniform distribution. This sequential characterization suffices to apply generic results in Andrews et al. (2020), which guarantee uniform coverage even in non standard cases like this one. Coverage over the pseudo-VCATE holds regardless of whether the nuisance functions are slow to converge or even misspecified.

The intervals are in general conservative because we're not plugging in the unknown ζ, and instead define a robust confidence interval as the union of CIs with given $\zeta \in \{-1, 1\}$. However, the key insight is that ζ only affects the coverage when the pseudo-VCATE is bounded away from zero. If Assumption 8 holds, the value of ζ doesn't enter the asymptotic distribution of the estimator. I show that this condition holds automatically if the nuisance functions converge to their true value at a sufficiently fast rate.

Lemma 6 (Verify Exact Coverage). *Let Γ denote a set of distributions that satisfy Assumptions 1, 3, 4, 5, 6, and 7. Then Assumption 8 also holds.*

As a special case, when the model is correctly specified, i.e. $W_i'\theta = \mu_d(x)$ for some $\theta \in \mathbb{R}^4$, then $\Omega_{nk,12} = 0$ by construction. Lemma 6 states that we only need a model that is correctly specified asymptotically, given the rates in Assumptions 6 and 7. Then for non-boundary cases, Ω_{nk} converges to the population analog under correct specification. These conditions also imply point-wise coverage of the true VCATE.

Theorem 8 (Pointwise, Exact Coverage of VCATE). *Let Γ denote a set of distributions that satisfy Assumptions 1, 3, 4, 5, 6, and 7. Then*

$$\inf_{\gamma \in \Gamma} \liminf_{n \to \infty} \mathbb{P}_\gamma \left(V_\tau(\gamma) \in \widehat{CI}_{\alpha nk} \right) = \sup_{\gamma \in \Gamma} \limsup_{n \to \infty} \mathbb{P}_\gamma \left(V_\tau(\gamma) \in \widehat{CI}_{\alpha nk} \right) = 1 - \alpha.$$

Theorem 8 shows that if the nuisance functions converge at a sufficiently fast rate, then the proposed intervals achieve point-wise exact coverage. The confidence

---

[14]The theorem only uses Assumptions 1, 3, 4, and 5 to verify normality in Assumption 2. A broad class of confidence intervals of the form in (11) constructed from regression adjusted estimators will satisfy these uniformity properties.

intervals provide correct size coverage for all regions of the parameter space, including $V_\tau(\gamma) = 0$.

Proving uniform coverage of the VCATE (rather than the pseudo-VCATE) is more challenging in the non-parametric case without much stronger conditions on the convergence rates of the nuisance functions. The lack of uniformity stems from a difficulty in controlling the ratio $\sqrt{n_k}\omega(\gamma_n)/\sqrt{V_\tau(\gamma_n)}$, which measures the relative error in estimating the CATE vs. the overall level of the VCATE. By the bound in (4), this ratio is easy to control when $n_k V_\tau(\gamma_n) = o(1)$ (near homogeneity) or $V_\tau(\gamma_n) \to V_\tau > 0$ (strong heterogeneity). However, it is possible to construct sequences, e.g., $n_k V_\tau(\gamma_n) \to v > 0$, where $(\hat{V}_{\tau nk} - V_\tau^*(\gamma_n, \mathcal{I}_{-nk}))$ converges to zero at a faster or comparable rate to the error of the pseudo-VCATE. There may be distortions in coverage in smaller samples. I illustrate this issue in the simulations.

Remark 4 (Uniform inference for one-sided tests)**.** Uniform inference is only challenging for two-sided tests. If instead, the researcher is only interested in left-sided tests, then uniform inference is still possible. To do so, we can make explicit use of the inequality $V_\tau^*(\gamma, \mathcal{I}_{-nk}) \leqslant V_\tau(\gamma)$. If $V_\tau(\gamma) < \inf_{V_\tau^*} \widehat{CI}_{\alpha nk}$ (the lower bound of the CI), then $V_\tau^*(\gamma, \mathcal{I}_{-nk}) \notin \widehat{CI}_{\alpha nk}$. Therefore, for all $\gamma \in \Gamma$,

$$\mathbb{P}_\gamma\left(V_\tau(\gamma) \geqslant \inf \widehat{CI}_{\alpha nk}\right) \leqslant \mathbb{P}_\gamma\left(V_\tau^*(\gamma, \mathcal{I}_{-nk}) \notin \widehat{CI}_{\alpha nk}\right). \tag{38}$$

I prove a weaker uniformity result for one-sided tests building on Theorem 7.

Corollary 3. *If Assumptions 1, 3, 4, and 5 hold, then*

$$\limsup_{n\to\infty} \sup_{\gamma\in\Gamma} \mathbb{P}_\gamma\left(V_\tau(\gamma) < \inf \widehat{CI}_{\alpha nk}\right) \leqslant \limsup_{n\to\infty} \sup_{\gamma\in\Gamma} \mathbb{P}_\gamma\left(V_\tau^*(\gamma, \mathcal{I}_{-nk}) \notin \widehat{CI}_{\alpha nk}\right) \leqslant \alpha.$$

Corollary 3 is empirically relevant for interpreting confidence intervals that do not include zero. It states that the asymptotic probability of having $V_\tau(\gamma) \in \left[0, \inf \widehat{CI}_{\alpha nk}\right)$ is uniformly less than α. Tests of homogeneity belong to this class and therefore have the correct size when $V_\tau(\gamma) = 0$. Moreover, the result in Corollary 3 is much stronger because it guarantees that a broader class of one-sided tests also has the correct size. It is important to emphasize that I do not impose any assumptions on rates of convergence of $(\hat{\eta}_{-k}(x) - \eta(x))$, but only the inequality on the pseudo-VCATE. Consequently, while estimating $\mu(x)$ and $\tau(x)$ may be important for increasing the power of tests of homogeneity, it is not necessary for controlling their size.

4.3 Multifold Coverage

The multi-fold confidence interval covers the VCATE asymptotically.

Theorem 9. *Let Γ be a set of distributions that satisfy Assumptions 1, 3, 4, 5. Then*

$$\limsup_{n\to\infty}\sup_{\gamma\in\Gamma}\mathbb{P}_\gamma\left(V_\tau(\gamma) < \inf \widehat{CI}_{\alpha nk}^{multifold}\right) \leqslant \alpha.$$

If Assumptions 6, and 7 also hold, then

$$\sup_{\gamma\in\Gamma}\limsup_{n\to\infty}\mathbb{P}_\gamma\left(V_\tau(\gamma) \notin \widehat{CI}_{\alpha n}^{multifold}\right) \leqslant \alpha.$$

The first part of Theorem 9 shows that the multifold CI uniformly controls the size of one-sided tests. The second part shows that if the nuisance functions converge to their true value asymptotically, then the multifold confidence interval provides point-wise size-control for two-sided tests. Coverage of the true parameter will be weakly larger that $(1-\alpha)$ asymptotically.

4.4 Power

The test of homogeneity has power against local alternatives.

Lemma 7. *Consider a sequence of distributions $\{\gamma_n\}_{n=1}^{\infty}$ and $\{\mathcal{I}_{-nk}\}_{n=1}^{\infty}$, where $\Omega_{nk} \to \Omega_\infty$ and $n_k V_\tau^*(\gamma_n, \mathcal{I}_{-nk}) = v + o(1)$, for $v \in [0,\infty)$. Assume that 1, 3, 4, and 5 hold. Let $\Omega_{\infty,11}$ be the upper-left entry of Ω_∞, $\Phi(\cdot)$ be the standard normal CDF, and $z_{1-\alpha}$ be the $(1-\alpha)-$quantile. Then*

$$\lim_{n\to\infty}\mathbb{P}_{\gamma_n}(0 \notin \widehat{CI}_{\alpha nk} \mid \mathcal{I}_{-nk}) = 1 - \Phi\left(z_{1-\alpha} - \frac{\sqrt{v}}{\sqrt{\Omega_{\infty,11}}}\right) + \Phi\left(-\frac{\sqrt{v}}{\sqrt{\Omega_{\infty,11}}}\right).$$

Lemma 7 computes the power curve for a sequence of local alternatives. When $v = 0$ the power is equal to α, whereas when $v \to \infty$ the power tends to one. This shows that tests of homogeneity have local power the null. When the pseudo-VCATE is bounded away from zero, the test rejects with probability approaching one.

5 Extensions

Remark 5 (Clustered Standard Errors)**.** In some cases, assuming that units i are independent may be strong. For example, in Dizon-Ross (2019) units are randomized at the household level, and it is reasonable to expects that units within a household have correlated outcomes and covariates. To deal with this dependence structure, suppose that the sample can be partitioned into C clusters, $c \in \{1, \ldots, C\}$, which are independent and identically distributed. The researcher can compute $\hat{\beta}_{2nk}$, $\hat{V}_{\tau nk}$ and $\widehat{CI}_{\alpha nk}$ via cross-fitting by randomly partitioning entire clusters rather than the individual observations.

Lemma 8. *Let $\{r_{nk}\}_{n=1}^{\infty}$ be a sequence of positive scalars. Suppose that $V_{xnk} > 0$, Ω_{nk} is positive definite with positive eigenvalues, and that conditional on a sequence $\{\mathcal{I}_{-nk}\}_{n=1}^{\infty}$, $r_{nk}^{-1}\hat{\Omega}_{nk} \to^p \Omega_{nk}$, $n_k/r_{nk} \to \infty$, and*

$$\Omega_{nk}^{-1/2} \sqrt{\frac{n_k}{r_{nk}}} \begin{pmatrix} \sqrt{V_{xn}}(\hat{\beta}_{2nk} - \beta_{2nk}) \\ \frac{\hat{V}_{xnk}}{V_{xnk}} - 1 \end{pmatrix} \mid \mathcal{I}_{-nk} \to^d Z_n \sim \mathcal{N}(0, I_{2 \times 2}). \tag{39}$$

Then $\widehat{CI}_{\alpha nk}$, substituting the arguments $(n_k, \hat{V}_{\tau nk}, \hat{\Omega}_{nk})$, satisfies Theorem 7.

Lemma 8 proposes high-level conditions that ensure that confidence intervals have correct coverage. The quantity $\sqrt{n_k/r_{nk}}$ is the effective rate of convergence, which features prominently in problems with cluster dependence (MacKinnon et al., 2022). For example, if the observations are fully correlated within clusters and the clusters have equal size, then r_{nk} is the cluster size, $n_k/r_{nk} = C$, and the estimators in (8) converge at $\sqrt{C}$ rate (the total number of clusters). The analyst does not need to specify the quantity r_{nk} to apply the procedure, but merely specify an estimator of the covariance matrix that meets the rate requirement. Under minor modifications to the existing proofs, we can also prove analogs of Theorems 8 and 9.

We can construct estimators that satisfy Lemma 8. Let $\mathcal{I}_{nkc}$ be the set of units in fold k and cluster c, and let $\mathcal{C}_{nk}$ be the indexes of the clusters selected for fold k.

$$\hat{H}_{nk}^{cluster} := \frac{1}{n_k} \sum_{c \in \mathcal{C}_{nk}} \left(\sum_{i \in \mathcal{I}_{nkc}} \begin{bmatrix} \lambda(X_i)\hat{U}_i\widehat{W}_i \\ \hat{T}_i \end{bmatrix} \right) \left(\sum_{i \in \mathcal{I}_{nkc}} \begin{bmatrix} \lambda(X_i)\hat{U}_i\widehat{W}_i \\ \hat{T}_i \end{bmatrix} \right)'.$$

The clustered standard errors are $\hat{\Omega}_{nk}^{cluster} = \hat{\Upsilon}_{nk}\hat{J}_{nk}^{-1}\hat{H}_{nk}^{cluster}\hat{J}_{nk}^{-1}\hat{\Upsilon}_{nk}'$, where $\hat{J}_{nk}$, $\hat{\Upsilon}_{nk}$ are computed as outlined in (25).

Remark 6 (Confidence intervals when $V_{xnk} = 0$)**.** When the conditional mean is constant, i.e. $\mu_d(x) = \mathbb{E}[Y_d]$, prediction models with corner solutions like LASSO may estimate a constant conditional mean, i.e. $\hat{\mu}_{d,-k}(x) = \hat{\mu}_{d,av}$, $\hat{\tau}_{-k}(x) = \hat{\mu}_{1,-k}(x) - \hat{\mu}_{0,-k}(x)$ is constant, and consequently $V_{xnk} = 0$.[15] This violates Assumption 4.(i), and it is challenging to construct a confidence interval with exact coverage. One alternative is to construct an ensemble of sparse and non-sparse estimators of the CATE in the first-stage. Another alternative is to use degenerate confidence intervals:

$$\widehat{CI}^0_{\alpha nk} = \begin{cases} \widehat{CI}_{\alpha nk} & \text{if } V_{xnk} \neq 0, \\ [0,0] & \text{if } V_{xnk} = 0. \end{cases} \tag{40}$$

The confidence intervals collapse to zero when the $\hat{\tau}_{-k}(x)$ prediction is degenerate. For example, in LASSO researchers can check whether the coefficients are zero, in tree-based methods when there are no splits, or whether $\hat{V}_{xnk} = 0$. We can also define an analogous multifold confidence interval.

$$\widehat{CI}^{0,\text{multifold}}_{\alpha n} = \left[\text{Med}_K \left\{ \inf \widehat{CI}_{\frac{\alpha}{2}nk} \right\}, \text{Med}_K \left\{ \sup \widehat{CI}_{\frac{\alpha}{2}nk} \right\} \right]. \tag{41}$$

I study the asymptotic properties of these confidence intervals.

Lemma 9. *Let Γ denote a set of distributions that satisfy Assumptions 1, 3, and 5. Suppose that Assumption 4 holds, except for the requirement that $V_{xnk} = 0$. Then (i)*

$$\liminf_{n\to\infty} \inf_{\gamma\in\Gamma} \mathbb{P}_\gamma \left(V^*_\tau(\gamma, \mathcal{I}_{-nk}) \in \widehat{CI}^0_{\alpha nk} \right) \geq 1 - \alpha \tag{42}$$

(ii) If Assumptions 6, and 7 also hold, then

$$\inf_{\gamma\in\Gamma} \liminf_{n\to\infty} \mathbb{P}_\gamma \left(V_\tau(\gamma) \in \widehat{CI}^0_{\alpha nk} \right) \geq 1 - \alpha \tag{43}$$

$$\inf_{\gamma\in\Gamma} \liminf_{n\to\infty} \mathbb{P}_\gamma \left(V_\tau(\gamma) \in \widehat{CI}^{0,multifold}_{\alpha nk} \right) \geq 1 - \alpha. \tag{44}$$

To prove this result I focus on the coverage for subsequences where $V_{xnk} = 0$ and $V_{xnk} > 0$, and apply the results for conservative coverage results in Andrews et al. (2020). In subsequences where $V_{xnk} = 0$, then $V^*_\tau(\gamma, \mathcal{I}_{-nk}) = 0$ which means that coverage of the pseudo-VCATE is equal to one. In subsequences where $V_{xnk} > 0$

---

[15]It is still possible to have $V_{xnk} > 0$ almost surely even if $V_\tau = 0$, as long as $\mu_0(x)$ is not constant.

and assuming that Ω_{nk} has eigenvalues bounded away from zero, then we can apply similar arguments as before to prove $1 - \alpha$ coverage. To prove point-wise coverage, I separate the cases where $V_\tau(\gamma) = 0$ and $V_\tau(\gamma) > 0$. In the latter case, I show that V_{xnk} is point-wise bounded away from zero, though not uniformly. The proof of Lemma 9 does not rely on the i.i.d. assumption, and can also accommodate cluster dependence. In the empirical example, I compute confidence intervals with clustered standard errors and degenerate CATE predictions.

When there's more heterogeneity and the nuisance functions are estimated accurately, then $V_{xnk} > 0$ with high probability. However, when $V_{\tau n} \approx 0$ and $\mathbb{V}(\mu_0(X)) \approx 0$, then procedures like LASSO may imply $V_{xnk} = 0$ (Fu and Knight, 2000), which means that marginally heterogeneous CATEs could be estimated as homogeneous. This could be impact the power of tests of homogeneity. The size for two-sided tests is not uniformly bounded. Furthermore, the multifold confidence interval allows for some quantification of uncertainty across folds/splits: the CI is degenerate only if more than half the fold/split-specific CIs are degenerate.

Furthermore, the degenerate CI has correct size control for one-sided tests.

Corollary 4. *Under the assumptions of Lemma 9.(i),*

$$\limsup_{n \to \infty} \sup_{\gamma \in \Gamma} \mathbb{P}_\gamma \left(V_\tau(\gamma) < \inf \widehat{CI}_{\alpha nk}^0 \right) \leqslant \alpha. \tag{45}$$

$$\limsup_{n \to \infty} \sup_{\gamma \in \Gamma} \mathbb{P}_\gamma \left(V_\tau(\gamma) < \inf \widehat{CI}_{\alpha nk}^{0,multifold} \right) \leqslant \alpha. \tag{46}$$

The tests of homogeneity have the correct size when $V_\tau(\gamma) = 0$. Corollary 4 guarantees that the probability of falsely rejecting a class of one-sided test is uniformly bounded in large samples.

Remark 7 (Monotonic Transformations)**.** It may be useful to report the standard deviation of the CATE, which is $\sqrt{VCATE}$. I propose the following confidence interval:

$$\widehat{CI}_{\alpha n}^{0,multifold,sqrt} = \left\{ \sqrt{V_\tau^*} : V_\tau^* \in \widehat{CI}_{\alpha n}^{0,multifold} \right\}. \tag{47}$$

Since the square root is a strictly increasing transformation and the VCATE is non-negative, then $\sqrt{V_\tau(\gamma)} \in \widehat{CI}_{\alpha n}^{0,multifold,sqrt}$ if and only $V_\tau(\gamma) \in \widehat{CI}_{\alpha n}^{0,multifold}$. Since the events are equivalent, the transformed confidence interval preserves the coverage probabilities and will have valid coverage by Lemma 9.

6 Simulations

I use a simulation design to study the properties of the VCATE estimators. The baseline covariates are distributed as $[X_0, X_1] \in \mathcal{N}(0, \Sigma_x)$, where $\rho = 0.5$ and

$$\Sigma_x = \begin{bmatrix} I_{J \times J} & \rho I_{J \times J} \\ \rho I_{J \times J} & I_{J \times J} \end{bmatrix}.$$

The random variables X_0 and X_1 are standard normal vectors of dimension J. The covariance between pairs of components X_{0j} and $X_{1j'}$ is equal to $\rho = 0.5$ for when $j = j'$, but zero otherwise. The outcome is generated from a model where $Y = DY_0 + (1 - D)$, D is generated by a Bernoulli draw with probability 0.5, and

$$\begin{aligned} Y_0 &= c + \beta_0' X_0 + U_0 \sqrt{\tilde{\sigma}_0^2 + \kappa_0' X_0 X_0' \kappa_0} \\ Y_1 &= (c + \tau) + \beta_0' X_0 + \beta_\tau' X_1 + U_1 \sqrt{\tilde{\sigma}_1^2 + \kappa_1' X_1 X_1' \kappa_1}, \end{aligned} \tag{48}$$

where $c, \tau \in \mathbb{R}$, $\beta_0, \beta_\tau, \kappa_0, \kappa_1 \in \mathbb{R}^p$. The errors (U_0, U_1) are independent of the covariates $(U_0, U_1) \perp\!\!\!\perp (X_0, X_1)$, and distributed as standard normals $[U_0, U_1]' \in \mathcal{N}(0_{2 \times 1}, I_2)$. The key model quantities have closed-form expressions. The conditional means at baseline and the CATE are given by $\mu_1(x) = \alpha + \beta_0' x_0$ and $\tau(x) = \tau + \beta_\tau' Z_1$, respectively. The conditional variances are $\sigma_d^2(x) = \kappa_d' x_d x_d' \kappa_d$ for $d \in \{0, 1\}$. This formulation incorporates heteroskedasticity. Covariates that influence the outcomes at baseline may also affect the treatment effects.

The regressors are constructed in such a way that $\mathbb{E}[X_d X_d'] = I_p$ for $d \in \{0, 1\}$. This implies simple expressions for the variances of the model, $\mathbb{V}(U_d) = \tilde{\sigma}_d^2 + \kappa_d' \kappa_d$,

$$V_\tau = \beta_\tau' \beta_\tau, \qquad \mathbb{V}(Y_0) = \beta_0' \beta_0 + \tilde{\sigma}_d^2 + \kappa_0' \kappa_0,$$

$$\mathbb{V}(Y_1) = \beta_0' \beta_0 + \beta_\tau' \beta_\tau + 2(1 - \rho)\beta_0' \beta_\tau + \tilde{\sigma}_1^2 + \kappa_1' \kappa_1,$$

I choose an approximately sparse specification for (48) where the coefficients decay exponentially at a rate of decay of $\lambda = 0.7$. Let $\ell_j = \sqrt{\left(\frac{1-\lambda}{1-\lambda^J}\right) \lambda^{1-j}}$ be a geometric sequence, which satisfies $\sum_{j=1}^{J} \left(\frac{1-\lambda}{1-\lambda^J}\right) \lambda^{1-j} = 1$. Given user-specified parameters $(V_\mu, V_\tau, \sigma_0^2, \sigma_1^2)$, the coefficients for the entries $j \in \{1, \ldots, J\}$ are determined by $\beta_{0,j} = \ell_j \sqrt{V_\mu}$, $\beta_{\tau,j} = \ell_j \sqrt{V_\tau}$, $\kappa_{d,j} = \ell_j \sqrt{\sigma_d^2 - \tilde{\sigma}_d^2}$, for $d \in \{0, 1\}$. Since $\sum_{j=1}^{J} \left(\frac{1-\lambda}{1-\lambda^J}\right) \lambda^{1-j} = 1$, then $\beta_0' \beta_0 = V_\mu$, $\beta_\tau' \beta_\tau = V_\tau$, and $\beta_0' \beta_\tau = \sqrt{V_\mu V_\tau}$. We can
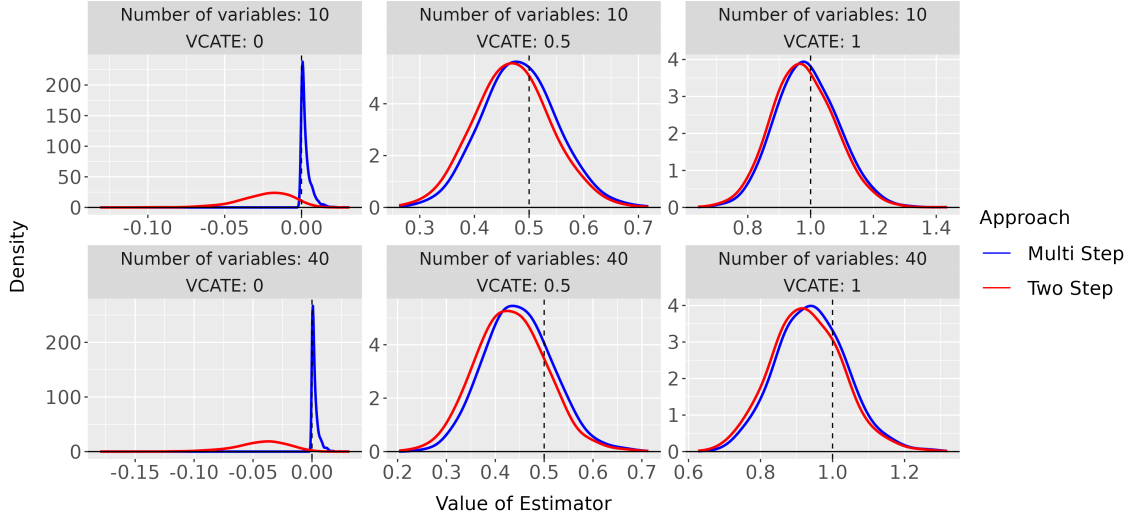
Figure 1: (**Density of estimators**) The figure shows the distributions of $\hat{V}_{\tau n}$ (multi-step) and $\hat{V}_{\tau n}^{\text{two-step}}$, defined in (31) and (49), respectively, for $n = 2500$, $K = 2$ folds, and a single split. The horizontal panels show designs with homogeneity ($V_\tau = 0$, left), moderate heterogeneity ($V_\tau = 0.5$, middle) and high-heterogeneity ($V_\tau = 1$, right). The vertical panels have a low dimension case with $2J = 10$ (top), and a high-dimension case with $2J = 40$ (bottom).

obtain analogous expressions for the variances of the unobserved components, so that $\tilde{\sigma}_d^2 + \kappa_0' \kappa_0 = \sigma_d^2$ for $d \in \{0, 1\}$.

I choose an average effect size of $\tau = 0.15$, that is coherent with the recent meta-analyses of economic experiments in Vivalt (2015). To make sure that the magnitudes are interpretable, I normalize the coefficients so that the variance for the control group is $\mathbb{V}(Y_0) = 1$, by setting set $c = 1$, $\sigma_d = 0.7$, $\tilde{\sigma}_d = 0.21$, and $V_\mu = 0.3$. The design is easy to scale for different values of V_τ and J. My design is similar to that in Belloni et al. (2014) but I choose Σ_x and the sparsity structure in such a way that V_τ has a closed form expression. I use LASSO to estimate $\mu_1(x)$ and $\mu_0(x)$, tuned via cross-validation. The coefficients of this model are consistent given this sparse linear structure, even in high dimensions. I randomly simulate 2000 datasets to compute each of the estimators, and split them into $K = 2$ folds.

Figure 1 considers a simulation with $n = 2500$. The figure displays a density plot for the multi-step estimator, $\hat{V}_{\tau n}$ defined in (31), and a two-step debiased machine learning estimator computed as:

$$\hat{V}_{\tau n}^{\text{two-step}} := \frac{1}{n} \sum_{i=1}^{n} \varphi(Y_i, D_i, X_i, \hat{\eta}_{-k_i}(X_i)), \tag{49}$$

where $\hat{\eta}_{-k}(x) = (\hat{\tau}_{-k}(x), \hat{\mu}_{0,-k}(x), p(x), \hat{\tau}_{n,av})$ and $\hat{\tau}_{n,av} = \frac{1}{n} \sum_{i=1}^{n} \hat{\tau}_{-k_i}(X_i)$. When
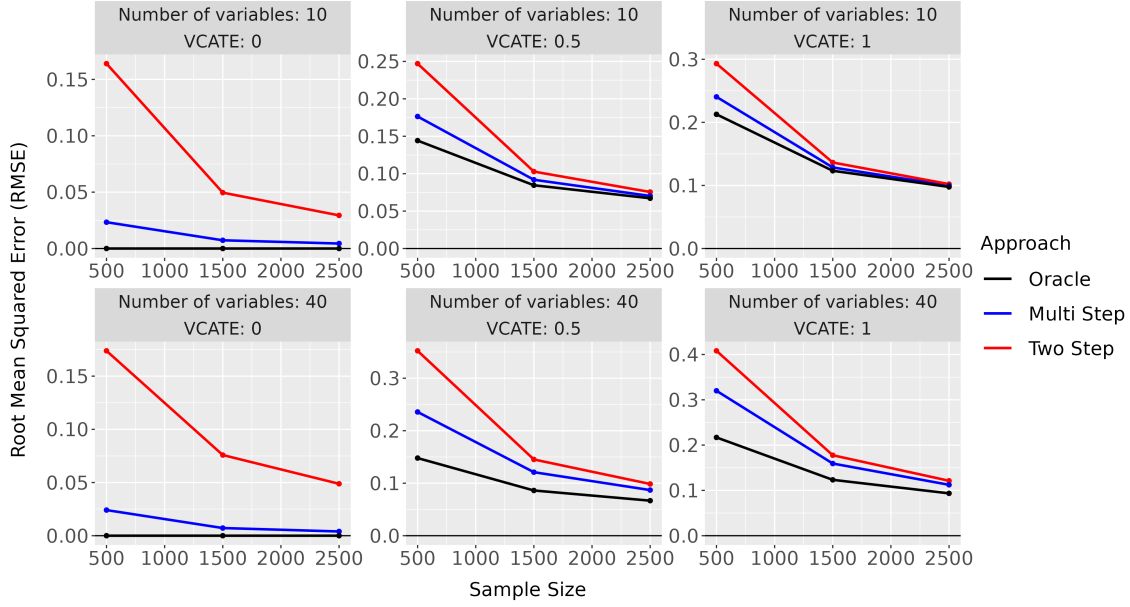
Figure 2: (**Root Mean Square Error**) The figure shows the root mean-square error of $\hat{V}_{\tau n}$ (multi-step) and $\hat{V}_{\tau n}^{\text{two-step}}$, defined in (31) and (49), respectively, for different sample sizes, $K = 2$ folds, and a single split. The 'oracle' estimator is constructed by substituting $\hat{\eta}_{-k}(X_i) = \eta(X_i)$ in (49). The horizontal panels show designs with homogeneity ($V_\tau = 0$, left), moderate heterogeneity ($V_\tau = 0.5$, middle) and high-heterogeneity ($V_\tau = 1$, right). The vertical panels have a low dimension case with $2J = 10$ (top), and a high-dimension case with $2J = 40$ (bottom).

$V_\tau > 0$ the efficient influence function is non-degenerate. In high-heterogeneity regimes both converge to the same limiting distribution.[16] However, when $V_\tau = 0$, the influence function is degenerate and they may converge at different rates. We see that the multi-step approach is much more precise. This can be explained by the fast boundary convergence rates derived in Theorem 5. The two step approach can also produce negative estimates of V_τ, which is an undesirable feature, whereas the multi-step estimator is always non-negative. Both estimators have higher bias when the dimension increases because there is more first-stage noise.

Figure 2 plots the root mean-square error (RMSE) of $\hat{V}_{\tau n}$ and $\hat{V}_{\tau n}^{two-step}$ for different sample sizes. I compute the semiparametric efficiency bound by computing the RMSE of an "oracle" estimator that substitutes $\hat{\eta}_{-k}(X_i) = \eta(X_i)$ in (49). The results show that as the sample size increases, both estimators achieve a higher level of accuracy and their variance approaches the semi-parametric lower bound (the RMSE of the oracle). As expected by Corollary 2, the semiparametric lower bound is zero at the boundary. The differences in RMSE shorten with higher V_τ and in lower dimensional

---

[16]This is shown in Theorem 6 for the multistep approach and can be shown for the two-step using standard arguments, e.g. Chernozhukov et al. (2018).
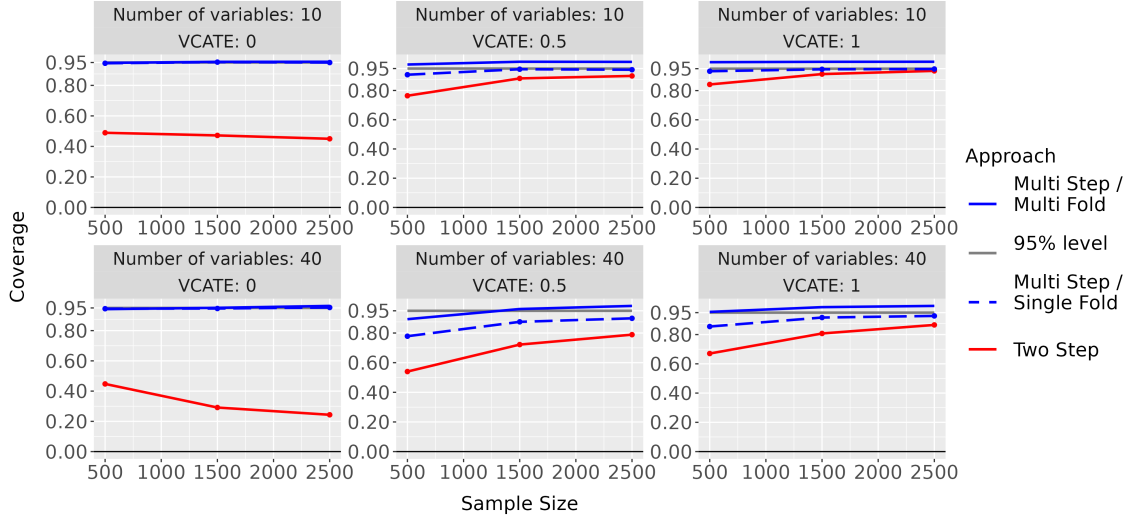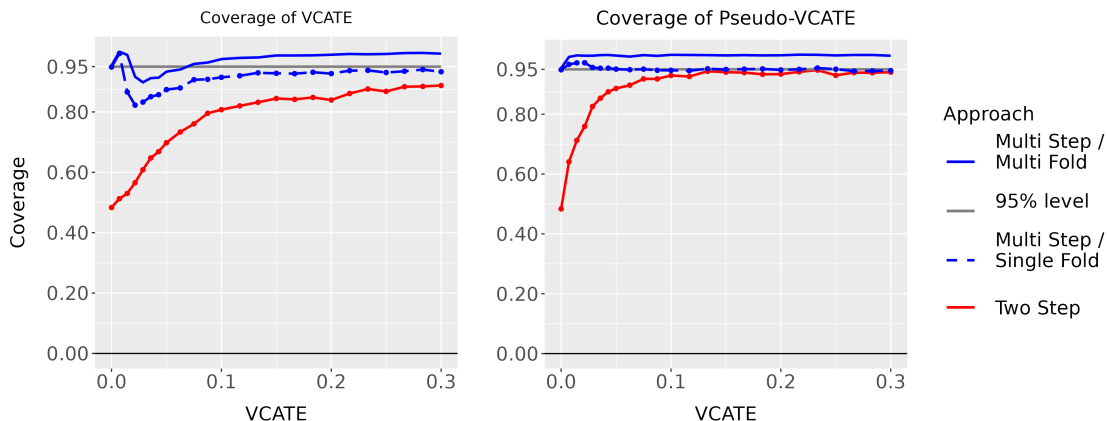
Figure 3: **(Coverage of V_τ)** The figure shows the 95% coverage probabilities at different sample sizes, for the multi-fold CIs and single fold confidence intervals defined in (40) and (41), respectively, for $K = 2$ folds, single split. The two-step CIs are constructed as $(1/n)\sum_{i=1}^{n}\hat{\varphi}_i \pm 1.96\sqrt{\hat{V}_\varphi/n}$, where $\hat{\varphi}_i$ is the summand in (49) and $\hat{V}_\varphi$ is an estimate of its sample variance. The horizontal panels show a regime with homogeneity ($V_\tau = 0$), moderate heterogeneity ($V_\tau = 0.5$) and high-heterogeneity ($V_\tau = 1$). The vertical panels show low dimension case with $2J = 10$ (top), a high-dimension case with $2J = 40$ (bottom). The dotted vertical lines denote the true value of the VCATE.

settings ($2J = 10$) (which have lower first-stage noise).

Figure 3 shows the coverage of V_τ for the different proposed confidence intervals (CIs). For the multi-step approach, I consider the single splits CIs in (40) and the conservative multi-fold CIs from and (41). The two-step CIs are constructed as $\frac{1}{n}\sum_{i=1}^{n}\hat{\varphi}_i \pm 1.96\sqrt{\hat{V}_\varphi/n}$, where $\hat{\varphi}_i$ is the summand in (49) and $\hat{V}_\varphi$ is an estimate of its sample variance. The coverage of the two-step approach is very low under homogeneity, and there is no improvement as sample size increases when $V_\tau = 0$. The coverage of the two-step estimator only improves with higher n, in high heterogeneity designs. By contrast, both multi-step approaches cover the parameter at the intended level, and coverage improves with higher sample size. For fixed n, coverage degrades for both cases when the number of covariates is higher.

Figure 4 explores the differences in covering the VCATE vs the pseudo-VCATE when $n = 2500$ and $2J = 10$ for a fine-grained set of values of V_τ. Panel (a) reflects a dip in coverage close to the boundary. My theory predicts that the multistep CIs have exact coverage when $V_\tau = 0$, but may not cover uniformly close to the boundary (see discussion after Theorem 8). The mulit-fold CIs have conservative coverage. Conversely, Figure 4, Panel (b) shows the multi-step CIs always uniformly cover the pseudo-VCATE, as predicted by theory. This provides a robustness guarantee for

how to interpret the CIs. The two-step approach has much lower coverage and no guarantees when $V_\tau = 0$ in either panel.


Figure 4: (**Coverage of V_τ vs V_τ^***) The figure shows the coverage of the VCATE (left) and the fold-specific pseudo-VCATE (right) for a set of fine-grained values of the true V_τ, given $n = 2500$, $K = 2$, single split, and $2J = 10$. The multi-fold CIs and single fold CIS, defined in (40) and (41), respectively. The two-step CIs are constructed as $(1/n) \sum_{i=1}^n \widehat{\varphi}_i \pm 1.96\sqrt{\widehat{V}_\varphi/n}$, where $\widehat{\varphi}_i$ is the summand in (49) and $\widehat{V}_\varphi$ is an estimate of its sample variance. For the multi-fold and two-step approaches in the right panel, I report coverage of the median $V_{\tau nk}^*$ across folds.

Figure 5.(left) shows the power of tests of homogeneity in a simulation with $n = 2500$. The multi-step, single fold approach has correct size control and has local power, in line with the result of Lemma 7. The power of the test using the multifold approach is similar to using a single fold. The right panel shows the probability that the VCATE is strictly below the CI bounds. As predicted by theory, this probability is uniformly bounded by $\alpha = 0.05$ for the single and multifold approaches (see Corollaries 3 and 4, and Theorem 9). The two-step approach has a non-monotonic power curve with incorrect size. The size of one-sided tests in the right panel is uniformly bounded by $\alpha = 0.05$, though this may be partly the fact that $\widehat{V}_{\tau n}^{two-step}$ can take negative values and has a negative bias (see Figure 1).

7 Empirical Example

In this section, I illustrate my approach using data from a large-scale information experiment conducted by Dizon-Ross (2019). The study, which covered 39 school districts, involved an intervention to provide low-income parents of at least two children with information about their children's school performance. Half the households where assigned to the information intervention and the rest were assigned to
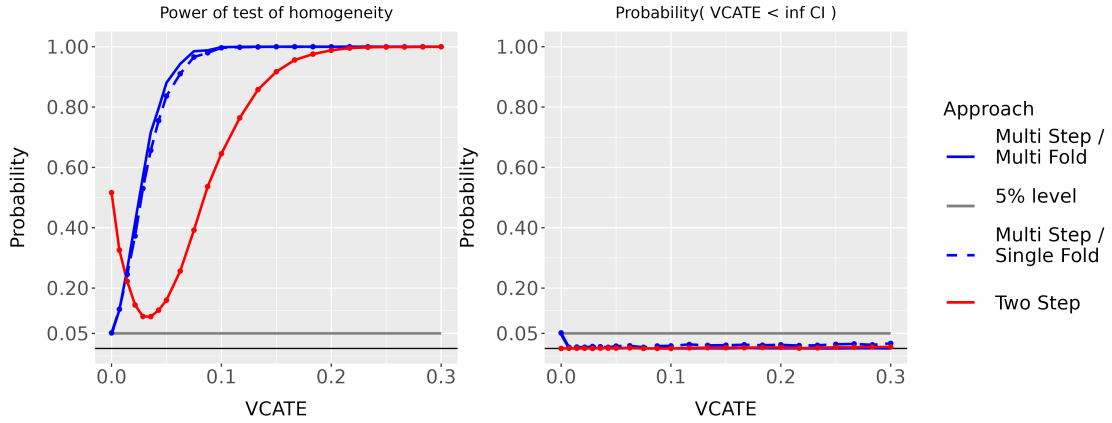
Figure 5: **(Power and size)** Panel(a) shows the power of tests of homogeneity (whether zero is contained in the confidence interval) for a set of fine-grained values of the true V_τ, given $n = 2500$, $K = 2$, single split, and $2J = 10$. Panel (b) shows the size of one-sided tests, i.e., the probability that the VCATE is strictly below the CI bounds. The multi-fold CIs and single fold CIs are defined in (40) and (41), respectively. The two-step CIs are constructed as $(1/n) \sum_{i=1}^{n} \widehat{\varphi}_i \pm 1.96 \sqrt{\widehat{V}_\varphi / n}$, where $\widehat{\varphi}_i$ is the summand in (49) and $\widehat{V}_\varphi$ estimates its sample variance.

the control group. Dizon-Ross (2019) showed that, at baseline, parents faced large information gaps regarding their children's grades and class ranking. Even though schools produced a report card, 60% of parents were unaware of their child's performance. Many parents reported that they did not receive the report card (children either lost them or did not take them home), or had trouble interpreting the report card structure, primarily due to low literacy levels.

The intervention was designed to present details of their children's school performance in an easily accessible way. Dizon-Ross (2019) showed that the information gaps (the difference between believed and true test scores) went down as a result of the intervention, and the amount of updating varied depending on students' initial test scores. Dizon-Ross (2019) also introduced a real-stakes scenario where parents received a series of lottery tickets for a scholarship paying for four years of high school. Parents had to decide how to allocate tickets between two siblings. If there were more than two siblings residing in the household, the survey team selected two at random. The results showed that parents allocated tickets towards their better performing child.

To test for heterogeneity, Dizon-Ross (2019) ran a linear regression of parental beliefs on initial scores, treatment, and an interaction as in (3), and reported estimates $\bar{X} = 46.8$ (on a scale of 100) and $(\widehat{\beta}_1, \widehat{\beta}_2) = (-25.9, 0.40)$ in their Tables 1 and 2, respectively. The coefficient $\widehat{\beta}_2$ captures how much the treatment effects vary (on a

0 to 100 scale) for a each additional point in students initial scores. The VCATE combines information about the coefficient and the initial variability in scores. From the data, I also estimate the variance of the control $\widehat{V}_x = 305.53$, and estimate the VCATE as $\widehat{\beta}_2^2 \widehat{V}_x = 48.89$. Taking the square root and normalizing by the standard deviation of the outcome for the control group ($\widehat{V}_{(Y|D=0)} = 311.72$), produces 0.40. This means that the magnitude of treatment effect heterogeneity explained by scores is comparable to 40% of the standard deviation of beliefs in the control group.

The VCATE can also help us understand the magnitude of treatment effect heterogeneity in the experiment using multiple covariates. I use LASSO for the first-stage predictions, two folds per split with 20 splits, and estimate $\{\widehat{\Omega}_{nk}\}_{k=1}^K$ using clustered standard errors at the household level and the formulas for the confidence intervals defined in (41). The results are not very sensitive to the number of splits. For ease of exposition, I report point-estimates and CIs for $\sqrt{\frac{V_\tau}{\mathbb{V}(Y_0)}}$, which is the standard deviation of the CATE divided by the standard deviation of the outcome for the control group. I compute confidence intervals for the square root via the transformation proposed in (47).

Table 1 computes the ATE and the $\sqrt{VCATE}$ for two outcomes (parental beliefs and lottery allocations) and 8 different sets of covariates. Panel (a) shows that, on average, parents downgrade their beliefs about test scores by 42% of the standard deviation (SD) of the beliefs of the control group. The treatment effect heterogeneity explained by test scores is equivalent to 40% of the standard deviation (SD) of the beliefs of the control group. This is statistically significant at the 5% level and has a comparable magnitude to the ATE. The confidence intervals are relatively short in length. However, applying the bounds from Theorem 2 and Corollary 1 shows that differentiating treatment offers based on scores could further lower beliefs by at most 8.1% SDs of the beliefs in the control group. In this case the ATE is already fairly high compared to $\sqrt{V_\tau}$, so in spite of the large heterogeneity, the marginals gains from targeting would be modest. Panel (b) presents the results for the secondary school lottery. The ATE is estimated precisely at zero, because the lottery tickets had to be divided as a zero sum between the siblings. The VCATE measures how much the dispersion in the allocation depends on the covariates. The standard deviation of the VCATE explained by initial scores is 16% of the SD of the control group lottery allocation, and the maximum welfare gains are around 7.8% SD.

The student variables (grade, age, gender, attendance, and educational expen-

	N	Clusters (HH)	Estimate $\sqrt{V_\tau/\mathbb{V}(Y_0)}$	95 % C.I.	Welfare Bounds Thm. 1	Thm. 2
Panel (a): Endline Beliefs. $\frac{ATE}{\sqrt{\mathbb{V}(Y_0)}}$: -0.42 (0.03)						
Scores	5244	2626	0.40	[0.32, 0.48]	0.201	0.081
Parent Years of Education	5208	2608	0.05	[0.00, 0.13]	0.023	0.001
Scores + Parents' Education	5208	2608	0.40	[0.32, 0.48]	0.199	0.079
Above Median Educ. Expenses	5244	2626	0.06	[0.00, 0.15]	0.028	0.002
Respondent Variables	4722	2365	0.03	[0.00, 0.12]	0.015	0.001
Household Variables	5244	2626	0.02	[0.00, 0.10]	0.009	0.000
Student Variables	4959	2532	0.11	[0.04, 0.20]	0.057	0.007
All variables	4464	2278	0.40	[0.31, 0.48]	0.199	0.079
Panel (b): Secondary School Lottery. $\frac{ATE}{\sqrt{\mathbb{V}(Y_0)}}$: 0.00 (0.00)						
Scores	5258	2629	0.16	[0.07, 0.24]	0.078	0.078
Parent Years of Education	5222	2611	0.00	[0.00, 0.00]	0.000	0.000
Scores + Parents' Education	5222	2611	0.16	[0.07, 0.24]	0.078	0.078
Above Median Educ. Expenses	5258	2629	0.03	[0.00, 0.07]	0.015	0.015
Respondent Variables	4736	2368	0.00	[0.00, 0.00]	0.000	0.000
Household Variables	5258	2629	0.00	[0.00, 0.00]	0.000	0.000
Student Variables	4971	2535	0.06	[0.00, 0.15]	0.029	0.027
All variables	4476	2281	0.15	[0.05, 0.24]	0.076	0.074

Table 1: (**Empirical Estimates**) Each panel computes the ATE and VCATE normalized by the standard deviation of each outcome for the control group. The N varies depending on the missing values for the covariates and the outcome. Each line within panels (a) and (b) considers 8 different sets of covariates measured at baseline which include test scores, years of parental education, an indicator for whether annual educational expenditures the previous year (uniforms, fees, school supplies) are above the median, other respondent variables (gender, age, is literate, is farmer), household variables (number of kids, single-parent), and student variables (grade, age, gender, attendance). I estimated clustered covariance matrices at the household level. The function $\mu_d(x)$ is computed using LASSO with 10-fold cross-validation, and the estimates are computed using 2-fold cross-fitting with 20 splits. The point estimates are the median values of $\widetilde{V}_{\tau n}$ across splits. I compute the bound in Column 7, using the ATE from the corresponding subsample.

ditures) collectively explain 11% of the SD of parental beliefs. This is statistically significant at the 5% level. The magnitude is around a fourth of the variation for test scores, and the maximum welfare gain from targeting is 0.7% of the SD of parental beliefs in the control group. I find that other subsets of covariates do not produce statistically significant estimates of the VCATE at the 5% level. The added welfare of personalizing treatment assignment using these covariates is also very low.

The estimates that use all the covariates are computed over a smaller subsample with non-missing values across all variables. Despite the large number of variables and the smaller sample, the estimates of the VCATE remain relatively stable across

specifications. The VCATE computed from a rich set of respondent, household, and student covariates has a comparable magnitude to the VCATE that only includes student scores. The confidence intervals are also similar. The estimates of the maximum welfare gains from targeting using all covariates are 7.9% SD for beliefs and 7.4% SD for lottery outcomes, respectively, which are similar to the welfare gains computed using only scores.

8 Conclusion

I propose an efficient estimator of the variance of treatment effects that can be attributed to baseline characteristics and propose novel adaptive confidence intervals that produce valid coverage. I analyze issues of non-standard inference that arise in this context, and how to address them. I also explore the economic significance of the VCATE for policymakers and researchers, by showing that the $\sqrt{VCATE}/2$ bounds the marginal gains of targeted policies. Overall, this paper proposes a broadly applicable approach to measure treatment effect heterogeneity in experiments.

References

Andrews, D. W., Cheng, X., Guggenberger, P., 2020. Generic results for establishing the asymptotic size of confidence sets and tests. Journal of Econometrics 218 (2), 496–531.

Athey, S., Tibshirani, J., Wager, S., 2019. Generalized random forests. The Annals of Statistics 47 (2), 1148–1178.

Athey, S., Wager, S., 2021. Policy learning with observational data. Econometrica 89 (1), 133–161.

Belloni, A., Chernozhukov, V., Fernández-Val, I., Hansen, C., 2017. Program evaluation and causal inference with high-dimensional data. Econometrica 85 (1), 233–298.

Belloni, A., Chernozhukov, V., Hansen, C., 2014. Inference on treatment effects after selection among high-dimensional controls. The Review of Economic Studies 81 (2), 608–650.

Bickel, P. J., Ritov, Y., Tsybakov, A. B., 2009. Simultaneous analysis of lasso and dantzig selector. The Annals of statistics 37 (4), 1705–1732.

Bitler, M. P., Gelbach, J. B., Hoynes, H. W., 2017. Can variation in subgroups' average treatment effects explain treatment effect heterogeneity? evidence from a social experiment. Review of Economics and Statistics 99 (4), 683–697.

Chen, X., White, H., 1999. Improved rates and asymptotic normality for nonparametric neural network estimators. IEEE Transactions on Information Theory 45 (2), 682–691.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., Robins, J., 2018. Double/debiased machine learning for treatment and structural parameters.

Chernozhukov, V., Demirer, M., Duflo, E., Fernndez-Val, I., 2022a. Generic machine learning inference on heterogenous treatment effects in randomized experiments.

Chernozhukov, V., Escanciano, J. C., Ichimura, H., Newey, W. K., Robins, J. M., 2022b. Locally robust semiparametric estimation. Econometrica 90 (4), 1501–1535.

Chernozhukov, V., Newey, W. K., Singh, R., 2022c. Automatic debiased machine learning of causal and structural effects. Econometrica 90 (3), 967–1027.

Crépon, B., Duflo, E., Huillery, E., Pariente, W., Seban, J., Veillon, P.-A., 2021. Cream skimming and the comparison between social interventions: Evidence from entrepreneurship programs for at-risk youth in france. Tech. rep.

Crump, R. K., Hotz, V. J., Imbens, G. W., Mitnik, O. A., 2008. Nonparametric tests for treatment effect heterogeneity. The Review of Economics and Statistics 90 (3), 389–405.

Das, A., Geisler, W. S., 2021. A method to integrate and classify normal distributions. Journal of Vision 21 (10), 1–1.

Davis, J. M., Heller, S. B., 2020. Rethinking the benefits of youth employment programs: The heterogeneous effects of summer jobs. Review of Economics and Statistics 102 (4), 664–677.

Deryugina, T., Heutel, G., Miller, N. H., Molitor, D., Reif, J., 2019. The mortality and medical costs of air pollution: Evidence from changes in wind direction. American Economic Review 109 (12), 4178–4219.

Ding, P., Feller, A., Miratrix, L., 2019. Decomposing treatment effect variation. Journal of the American Statistical Association 114 (525), 304–317.

Dizon-Ross, R., 2019. Parents' beliefs about their children's academic ability: Implications for educational investments. American Economic Review 109 (8), 2728–65.

Fu, W., Knight, K., 2000. Asymptotics for lasso-type estimators. The Annals of statistics 28 (5), 1356–1378.

Guo, Y., Coey, D., Konutgan, M., Li, W., Schoener, C., Goldman, M., 2021. Machine learning for variance reduction in online experiments. Advances in Neural Information Processing Systems 34, 8637–8648.

Heckman, J. J., Smith, J., Clements, N., 1997. Making the most out of programme evaluations and social experiments: Accounting for heterogeneity in programme impacts. The Review of Economic Studies 64 (4), 487–535.

Kitagawa, T., Tetenov, A., 2018. Who should be treated? empirical welfare maximization methods for treatment choice. Econometrica 86 (2), 591–616.

Lehmann, E. L., 1999. Elements of large-sample theory. Springer.

Levy, J., van der Laan, M., Hubbard, A., Pirracchio, R., 2021. A fundamental measure of treatment effect heterogeneity. Journal of Causal Inference 9 (1), 83–108.

MacKinnon, J. G., Nielsen, M. Ø., Webb, M. D., 2022. Cluster-robust inference: A guide to empirical practice. Journal of Econometrics.

Manski, C. F., 2004. Statistical treatment rules for heterogeneous populations. Econometrica 72 (4), 1221–1246.

Mbakop, E., Tabord-Meehan, M., 2021. Model selection for treatment choice: Penalized welfare maximization. Econometrica 89 (2), 825–848.

Newey, W. K., 1990. Semiparametric efficiency bounds. Journal of applied econometrics 5 (2), 99–135.

Semenova, V., Chernozhukov, V., 2021. Debiased machine learning of conditional average treatment effects and other causal functions. The Econometrics Journal 24 (2), 264–289.

Sun, L., 2021. Empirical welfare maximization with constraints. arXiv preprint arXiv:2103.15298.

Vivalt, E., 2015. Heterogeneous treatment effects in impact evaluation. American Economic Review 105 (5), 467–70.

Viviano, D., Bradic, J., 2023. Fair policy targeting. Journal of the American Statistical Association, 1–14.

Wager, S., Walther, G., 2015. Adaptive concentration of regression trees, with application to random forests. arXiv preprint arXiv:1503.06388.

Online Appendix

A   Critical Values

Define the correlation $\rho = \Omega_{12}/\sqrt{\Omega_{11}\Omega_{22}}$. By definition of the Cholesky decomposition, $(e_1'\Omega^{1/2}Z) = \sqrt{\Omega_{11}}Z_1$, and $(e_2'\Omega^{1/2}Z) = \sqrt{\Omega_{22}(1-\rho^2)}Z_2 + \rho\sqrt{\Omega_{22}}Z_1$. Substituting these terms into the expression for G

$$
G(n, V_\tau^*, \Omega, Z, \zeta)
$$
$$
= \underbrace{\frac{\Omega_{11}}{n}}_{\nu_1} Z_1^2 + 2\underbrace{\left[\zeta\sqrt{\frac{V_\tau^*\Omega_{11}}{n}} + \frac{V_\tau^*}{2\sqrt{n}}\rho\sqrt{\Omega_{22}} \right]}_{\kappa_1} Z_1 + \underbrace{\frac{V_\tau^*}{\sqrt{n}}\left[\sqrt{\Omega_{22}(1-\rho^2)} \right]}_{\kappa_2} Z_2.
$$

This has a quadratic form, $G(n, V_\tau^*, \Omega, z, \zeta) = \nu_1 \left(Z_1 + \frac{\kappa_1}{2\nu_1} \right)^2 + \kappa_2 Z_2 - \frac{\kappa_1^2}{4\nu_1}$, which fits the form of a generalized chi-square (Das and Geisler, 2021). To compute critical values we compute feasible analogs $(\hat{\nu}_1, \hat{\kappa}_1, \hat{\kappa}_2)$ from an estimate of Ω.

B   Proofs Main Document

Proof of Lemma 1. Define $\tilde{\tau}(X') := \mathbb{E}[Y_1 - Y_0 \mid X']$. Since X is X'-measurable, then by the law of iterated expectations, $\mathbb{E}[\tilde{\tau}(X') \mid X] = \tau(X)$. By the law of total variance $V_\tau' = \mathbb{V}(\tau(X)) + \mathbb{E}[\mathbb{V}(\tilde{\tau}(X') \mid X)] \geq V_\tau$. We can prove the upper bound by setting $X' = Y_1 - Y_0$. $\qquad\square$

Proof of Theorem 1. The result is a special case of Theorem 2. The most adversarial distribution is one where $\tau_{av} = 0$. $\qquad\square$

Proof of Theorem 2. Our goal is to find $\mathcal{R} := \sup_{\gamma\in\Gamma} \sup_{\pi\in\Pi} \mathcal{U}_\gamma(\pi)$, and to prove that $\mathcal{U}_\gamma(\pi) = \mathcal{R}$ for at least one $\gamma \in \Gamma$ and $\pi \in \Pi$.

Define two random variables $T := \mathbb{E}_\gamma[Y_1 - Y_0 \mid X = x]$ and $M := \mathbb{1}\{T > 0\}$. By adding/subtracting $\mathbb{E}_\gamma[Y_0]$ and applying the law of iterated expectations, $\mathcal{U}_\gamma(\pi) = \mathbb{E}_\gamma[\pi(X)T] - \max\{0, \mathbb{E}_\gamma[T]\}$. The optimal policy $\pi^*(X) = \mathbb{1}\{T \geq 0\}$, which belongs to Π, i.e. the "first-best" (Kitagawa and Tetenov, 2018). This means that $\mathcal{R} = \sup_{\gamma\in\Gamma} \mathbb{E}_\gamma[\max\{0, T\}] - \max\{0, \mathbb{E}_\gamma[T]\}$.

Step 1: (Problem Equivalence) The set of distributions Γ is very large. Instead I focus on the problem over a set of equivalence classes. For $m \in \{0, 1\}$, define the moments $p_m = \mathbb{P}_\gamma(M = m)$, $\tau_m = \mathbb{E}_\gamma[T \mid M = m]$, and $\omega_m = \mathbb{V}_\gamma(T \mid M = m)$. By definition, $\mathcal{U}_\gamma(\pi^*) = p_1 \tau_1 - \max\{0, \tau_{av}\}$, where $p_1 \tau_1$ is the proportion of people that benefit from treatment times their conditional mean. Let $\Delta := \tau_1 - \tau_0$ be the mean difference between those that benefit from the program and those that do not. By definition $\Delta > 0$ because $\tau_1 > 0$ and $\tau_0 \leqslant 0$. The conditional treatment effect is $\mathbb{E}_\gamma[T \mid M] = \tau_0 + M\Delta$, and $\tau_{av} = \tau_0 + p_1 \Delta$. Rearranging these expressions, $\tau_1 = \tau_{av} + (1 - p_1)\Delta$ and $\mathcal{U}_\gamma(\pi^*) = p_1(\tau_{av} + (1 - p_1)\Delta) - \max\{0, \tau_{av}\}$. By definition $\mathbb{V}_\gamma(\mathbb{E}_\gamma[T \mid M]) = \Delta^2 p_1(1 - p_1)$. By applying the law of total variance $V_\tau = p_1 \omega_1^2 + (1 - p_1)\omega_0^2 + \Delta^2(p_1)(1 - p_1)$. Then

$$\mathcal{R} = \sup_{\{p_1, \tau_0, \Delta, \omega_0^2, \omega_1^2\}} p_1 \tau_{av} + p_1(1 - p_1)\Delta - \max\{0, \tau_{av}\},$$

$$s.t. \quad p_1 \in [0, 1], \tau_0 \leqslant 0, \tau_0 + \Delta > 0, \Delta > 0, \omega_1^2, \omega_0^2 \geqslant 0, \tau_0 + p_1 \Delta = \tau_{av}, \tag{50}$$

$$p_1 \omega_1^2 + (1 - p_1)\omega_0^2 + p_1(1 - p_1)\Delta^2 = V_\tau.$$

The values of τ_{av} and V_τ impose the following constraints on the feasible set:

	$V_\tau = 0$	$V_\tau > 0$
$\tau_{av} \leqslant 0$	$p_1 = 0$, $\tau_0 = \tau_{av}$, $\omega_0 = 0$	$p_1 \in [0, 1)$
$\tau_{av} > 0$	$p_1 = 1$, $\tau_1 = \tau_{av}$, $\omega_1 = 0$	$p_1 \in (0, 1]$

By the form of the objective, if $p_1 \in \{0, 1\}$ then $\mathcal{R} = 0$. Therefore, if $V_\tau = 0$, then $\mathcal{R} = 0$ and the result of the theorem holds. Any value of the remaining parameters that satisfies the sign constraints will be feasible. Without loss, we focus on $V_\tau > 0$.

Step 2: (Optimum has binary support) Consider a situation where $(p_1, \omega_1^2, \omega_0^2)$ are fixed and $p_1 \in (0, 1)$. From the variance equation, $\Delta^* = \sqrt{\frac{V_\tau - p_1 \omega_1^2 - (1 - p_1)\omega_0^2}{p_1(1 - p_1)}}$, and from the mean $\tau_0^* = \tau_{av} - p_1 \Delta^*$ and $\tau_1^* = \tau_{av} + (1 - p_1)\Delta^*$.

A solution is feasible as long as $\omega_1^2, \omega_0^2 \geqslant 0$, $p_1 \omega_1^2 + (1 - p_1)\omega_0^2 \leqslant V_\tau$, $\tau_0^* \leqslant 0$, $\tau_1^* > 0$, and $\Delta^* = \tau_1^* - \tau_0^* > 0$. The optimal τ_0^* is strictly increasing in (ω_1^2, ω_0^2) and τ_1^* is strictly decreasing. If a given value of (ω_1^2, ω_0^2) is feasible, then another candidate which shrinks it to zero will still satisfy the constraints. Moreover, the objective function is strictly increasing in Δ^*, which in turn is strictly decreasing in (ω_1^2, ω_0^2). Therefore the optimum is $(\omega_1^2, \omega_0^2) = (0, 0)$, i.e. binary support for the CATE, $\Delta^* =$

$\sqrt{V_\tau/p_1(1-p_1)}$, $\tau_0^* = \tau_{av} - \sqrt{p_1/(1-p_1)}\sqrt{V_\tau}$, and $\tau_1^* = \tau_{av} + \sqrt{(1-p_1)/p_1}\sqrt{V_\tau}$.

Only some values of p_1 are feasible, i.e. satisfy the constraints $\Delta^* > 0$, $\tau_0^* \leqslant 0$, and $\tau_1^* > 0$. The function $p_1/(1-p_1)$ is strictly increasing in p with a range in $(0, \infty)$. The feasible values depend on the sign of τ_{av}. If (i) $\tau_{av} \leqslant 0$, then $p_1 \in \left(0, \frac{V_\tau}{\tau_{av}^2 + V_\tau}\right]$. If (ii) $\tau_{av} > 0$, $p_1 \in \left[\frac{\tau_{av}^2}{\tau_{av}^2 + V_\tau}, 1\right)$.

Step 3: (Solve points of support) For case (i) the objective function is $\mathcal{R} = \max_{p_1 \in \left[0, \frac{V_\tau}{\tau^2 + V_\tau}\right]} p_1 \tau_{av} + \sqrt{p_1(1-p_1)}\sqrt{V_\tau}$. The unique solution to the FOC is $p_1^* = (1/2) - (1/2)\sqrt{\tau_{av}^2/(\tau_{av}^2 + V_\tau)}$, and is interior because $\sqrt{\tau_{av}^2/(\tau_{av}^2 + V_\tau)} \geqslant \tau_{av}^2/(\tau_{av}^2 + V_\tau)$. By strict concavity, it is a global optimum, and $\mathcal{R} = \tau_{av}/2 - (\tau_{av}|\tau_{av}|)/(2\sqrt{\tau_{av}^2 + V_\tau}) + V_\tau/(2\sqrt{\tau_{av}^2 + V_\tau})$. Since $\tau_{av} \leqslant 0$, this can be simplified to $\frac{1}{2}(-|\tau_{av}| + \sqrt{\tau_{av}^2 + V_\tau})$. For case (ii), $\tau_{av} > 0$, the objective function is $(p_1 - 1)\tau_{av} + \sqrt{p_1(1-p_1)}\sqrt{V_\tau}$, subject to $1 \geqslant p_1 \geqslant \frac{\tau_{av}^2}{\tau_{av}^2 + V_\tau}$. The unique solution to the FOC is $p_1^* = (1/2) + (1/2)\sqrt{\tau_{av}^2/(\tau_{av}^2 + V_\tau)}$, it is interior because $p_1^* \geqslant \frac{1}{2} + \frac{1}{2}\frac{\tau_{av}^2}{\tau_{av}^2 + V_\tau} \geqslant \frac{\tau_{av}^2}{\tau_{av}^2 + V_\tau}$, and produces the desired $\mathcal{R}$. $\qquad \square$

Proof of Corollary 1. The potential outcomes under a linear transformation are $\kappa_1 + \kappa_2 Y_0$ and $\kappa_1 + \kappa_2 Y_1$, respectively. The treatment effect is $\kappa_2(Y_1 - Y_0)$, which does not depend on κ_1, and the transformed CATE is $\kappa_2 \tau(x)$. The ATE is $\kappa_2 \tau_{av}$ and the VCATE is $\kappa^2 V_\tau$. The result follows from substituting the transformed values into Theorem 2 and factorizing $|\kappa_2|$. $\qquad \square$

Proof of Lemma 2 . We decompose $\widehat{V}_{\tau n} = \widehat{\beta}_{2n}^2 \widehat{V}_{xn}$ into components that map into those of Assumption 2, by centering the key terms.

$$\widehat{V}_{\tau n} - V_{\tau n}^* = \widehat{\beta}_{2n}^2 \widehat{V}_{xn} - \beta_{2n}^2 V_{xn} = (\widehat{\beta}_{2n} - \beta_{2n} + \beta_{2n})^2 \frac{V_{xn}}{V_{xn}}(\widehat{V}_{xn} - V_{xn} + V_{xn}) - \beta_{2n}^2 V_{xn}$$

$$= [\sqrt{V}_{xn}(\widehat{\beta}_{2n} - \beta_{2n})]^2 + 2[\sqrt{V_{xn}}\beta_{2n}][\sqrt{V}_{xn}(\widehat{\beta}_{2n} - \beta_{2n})] + [V_{xn}\beta_2^2]\left(\frac{\widehat{V}_{xn}}{V_{xn}} - 1\right)$$

$$+ [\sqrt{V_{xn}}(\widehat{\beta}_{2n} - \beta_{2n})]^2 \left(\frac{\widehat{V}_{xn}}{V_{xn}} - 1\right) + 2[\sqrt{V_{xn}}\beta_{2n}][\sqrt{V_{xn}}(\widehat{\beta}_{2n} - \beta_{2nk})]\left(\frac{\widehat{V}_{xn}}{V_{xn}} - 1\right).$$
$$\tag{51}$$

Let $\widehat{Z}_n = n^{1/2}\Omega_n^{-1/2}[\sqrt{V}_{xn}(\widehat{\beta}_{2n} - \beta_{2n}), \widehat{V}_{xn}/V_{xn} - 1]'$ be a normalized statistic, and let $e_1 = [1, 0]'$, $e_2 = [0, 1]'$ be vectors that select the first and second coordinates,

respectively. We can always find a $\zeta_n \in \{-1, 1\}$ that solves $\zeta_n\sqrt{V_{\tau n}} = \beta_{2n}\sqrt{V_{xn}}$. If $V_{\tau n} > 0$ this is the sign of β_{2n} but otherwise any $\zeta_n \in \{-1, 1\}$ will solve the equation. By substituting the definition of $\widehat{Z}_n$, $V_{\tau n} = \beta_{2n}^2 V_{xn}$, and

$$
\begin{aligned}
\widehat{V}_{\tau n} - V_{\tau n}^* &= \frac{(e_1'\Omega_n\widehat{Z}_n)^2}{n} + 2\zeta_n\sqrt{V_{\tau n}^*}\left(\frac{e_1'\Omega_n^{1/2}\widehat{Z}_n}{\sqrt{n}}\right) + V_{\tau n}^*\frac{e_2'\Omega_n^{1/2}\widehat{Z}_n}{\sqrt{n}} \\
&\quad + \frac{(e_1'\Omega_n^{1/2}\widehat{Z}_n)^2(e_2'\Omega_n^{1/2}\widehat{Z}_n)}{n^{3/2}} + 2\zeta_n\sqrt{V_{\tau n}^*}\frac{(e_1'\Omega_n^{1/2}\widehat{Z}_n)(e_2'\Omega_n^{1/2}\widehat{Z}_n)}{n}.
\end{aligned}
\tag{52}
$$

By Assumption 2, $\widehat{Z}_n \to Z_n + o_p(1)$. Moreover, since Ω_n has bounded eigenvalues, $\Omega_n^{1/2}\widehat{Z}_n = \Omega_n^{1/2}Z_n + o_p(1)$. This means that

$$
\widehat{V}_{\tau n} - V_{\tau n}^* = \frac{(e_1'\Omega_n^{1/2}Z_n)^2}{n} + 2\zeta_n\sqrt{\frac{V_{\tau n}^*}{n}}e_1'\Omega_n^{1/2}Z_n + \frac{V_{\tau n}^*}{\sqrt{n}}e_2'\Omega_n^{1/2}Z_n + \text{Residual}_n,
\tag{53}
$$

where the residual is $o_p\left(n^{-1}\right) + o_p\left(\sqrt{V_{\tau n}^*/n}\right) + o_p\left(V_{\tau n}^*/\sqrt{n}\right) + O_p(n^{-3/2}) + o_p(n^{-1}\sqrt{V_{\tau n}^*})$. The fourth and fifth terms of the residual are $o_p(n^{-1})$ and $o_p(n^{-1/2}\sqrt{V_{\tau n}^*})$, and hence asymptotically negligible. The leading term in (53) is $O_p\left(\max\left\{1/n, \sqrt{V_{\tau n}^*/n}\right\}\right)$. $\qquad\square$

Proof of Lemma 3. Let $\varphi_i = \varphi(Y_i, D_i, X_i, \eta)$ be a realization of the influence function in (14) and compute $\mathbb{V}[\varphi_i \mid D_i = d, X_i = x] = 4(\tau(x) - \tau_{av})^2\left[\frac{d\sigma_1^2(x)}{p(x)^2} + \frac{(1-d)\sigma_0^2(x)}{(1-p(x))^2}\right]$ and $\mathbb{E}[\varphi_i \mid D_i = d, X_i = x] = (\tau(x) - \tau_{av})^2$.

By the law of iterated expectations $\mathbb{E}[\varphi_i \mid X_i = x] = (\tau(x) - \tau_{av})^2$. By applying the law of total variance recursively, $\mathbb{V}(\varphi_i) = \mathbb{V}(\mathbb{E}[\varphi_i \mid X_i]) + \mathbb{E}[\mathbb{E}[\mathbb{V}(\varphi_i \mid D_i, X_i) \mid X]] + \mathbb{E}[\mathbb{V}(\mathbb{E}[\varphi_i \mid D_i, X_i] \mid X_i)]$. This produces, $\mathbb{V}(\varphi_i) = \mathbb{V}((\tau(X_i) - \tau_{av})^2) + 4\mathbb{E}\left[(\tau(X_i) - \tau_{av})^2\left(\frac{\sigma_1^2(X_i)}{p(X_i)} + \frac{\sigma_0^2(X_i)}{(1-p(X_i))}\right)\right]$. $\qquad\square$

Proof of Corollary 2. $\mathbb{V}((\tau(X_i) - \tau_{av})^2) \leqslant \mathbb{E}[(\tau(X_i) - \tau_{av})^4] \leqslant \kappa^2 V_\tau$. By the Cauchy-Schwarz inequality, the term $\mathbb{E}\left[(\tau(X_i) - \tau_{av})^2\left(\frac{\sigma_1^2(X_i)}{p(X_i)} + \frac{\sigma_0^2(X_i)}{(1-p(X_i))}\right)\right]$ is bounded by $\kappa V_\tau\sqrt{\mathbb{E}\left[\left(\frac{\sigma_1^2(X_i)}{p(X_i)} + \frac{\sigma_0^2(X_i)}{(1-p(X_i))}\right)^2\right]}$. $\qquad\square$

Proof of Lemma 4. Let $\mathcal{Q}(\theta)$ be the Jacobian of the least squares problem, defined in (18). Substituting $\widehat{W}(X_i, D_i)'e_4 = (D_i - p(X_i))\widehat{S}(X_i)$ and $\lambda(X_i) = [p(X_i)(1-p(X_i))]^{-1}$ into the definition, $\beta_2\mathcal{Q}(\theta)e_4 = \frac{1}{n}\sum_{i=1}^n \frac{[D_i - p(X_i)]}{p(X_i)(1-p(X_i))}\beta_2\widehat{S}(X_i)(Y_i - \widehat{W}(X_i, D_i))$.

49

By the definition of the nuisance functions in $\tilde{\eta}_\theta(x)$ in (19), $\widehat{W}(X_i, D_i)'\theta = \tilde{\mu}_{0,\theta}(X_i) + D_i\tilde{\tau}_\theta(X_i)$, $\tilde{\tau}_\theta(x) = \beta_1 + \beta_2\widehat{S}(X_i)$, and $\tilde{\tau}_\theta = \beta_1 + \left[\frac{1}{n}\sum_{i=1}^n \widehat{S}(X_i)\right]\beta_2$. Furthermore, since $\frac{1}{n}\sum_{i=1}^n \widehat{S}(X_i) = 0$, then $\beta_2\widehat{S}(X_i) = \tilde{\tau}_\theta(X_i) - \tilde{\tau}_{av,\theta}$ and

$$\beta_2\mathcal{Q}(\theta)e_4 = \frac{1}{n}\sum_{i=1}^n \frac{[D_i - p(X_i)]}{p(X_i)(1 - p(X_i))}\beta_2\widehat{S}(X_i)(Y_i - \tilde{\mu}_{0,\theta}(X_i) - D_i\tilde{\tau}_\theta(X_i)).$$

Notice that $D_i - p(X_i) = D_i(1 - p(X_i)) + p(X_i)(1 - D_i)$. This implies that $\frac{D_i - p(X_i)}{p(X_i)} = \frac{D_i}{p(X_i)} - \frac{(1-D_i)}{1-p(X_i)}$. The next step is to substitute the estimated parameters. Then $\frac{1}{n}\sum_{i=1}^n \varphi(Y_i, D_i, X_i, \tilde{\eta}_{\hat{\theta}_n}) = \frac{1}{n}\sum_{i=1}^n (\tilde{\tau}_{\hat{\theta}_n}(X_i) - \tilde{\tau}_{av,\hat{\theta}_n})^2 + \hat{\beta}_{2n}\mathcal{Q}(\hat{\theta}_n)e_4$. The first term simplifies to $\hat{\beta}_{2n}^2\left[\frac{1}{n}\sum_{i=1}^n \widehat{S}(X_i)^2\right] = \hat{\beta}_{2n}^2\widehat{V}_{xn}$. Finally, $\mathcal{Q}(\hat{\theta}_n) = 0$ implies that the second term is zero.

$\square$

Proof of Lemma 5. The parameter $\theta = (\alpha_1, \alpha_2, \beta_1, \beta_2) \in \Theta^*$ solves

$$\mathbb{E}\left[\lambda(X)W(X, D)Y\right] - \mathbb{E}\left[\lambda(X)W(X, D)W(X, D)'\right]\theta = 0_{4\times 1}. \tag{54}$$

If there exists a $\theta \in \mathbb{R}^4$ such that $\mathbb{E}[Y \mid D = d, X = x] = \mu_d(x) = W(x, d)'\theta$, then by the law of iterated expectations $\mathbb{E}\left[\lambda(X)W(X, D)Y\right] = \mathbb{E}[\lambda(X)W(X, D)W(X, D)'\theta]$. Such a θ would automatically satisfy (54). Now suppose that $S(x) = \tau(x) - \tau_{av}$, $M(x) = \mu_0(x) + p(x)\tau(x)$, and $\theta = (0, 1, \tau_{av}, 1)$. For this choice, $W(x, d)'\theta = 0 + (\mu_0(x) + p(x)\tau(x)) + (d - p(x))\tau_{av} + (d - p(x))(\tau(x) - \tau_{av})$. This simplifies to $W(x, d)'\theta = \mu_0(x) + d\tau(x) = \mu_d(x)$ and θ solves (54).

$\square$

Proof of Theorem 3.(i). Define a vector $\underline{Y}_{nk} \in \mathbb{R}^{n_k}$ with the outcomes in fold $\mathcal{I}_{nk}$, $\underline{W}_{nk}$ be an $n_k \times 4$ matrix of regressors, $\underline{U}_{nk} \in \mathbb{R}^{n_k}$ be a vector of errors, and Λ_{nk} be a $n_k \times n_k$ diagonal matrix with entries $\{\lambda(X_i)\}$. Let $\widehat{\underline{W}}_{nk}$ be an $n_k \times 4$ matrix of generated regressors defined in (22), and let Π_{nk} be a 4×4 diagonal matrix with entries $(1, 1, 1, V_{xnk}^{-1/2})$. Let $\theta_{nk}^* := (\alpha_{1nk}, \alpha_{2nk}, \beta_{1nk}, \sqrt{V}_{xnk}\beta_{2nk})$, and define an infeasible estimator $\widetilde{\theta}_{nk}^* := \Pi_{nk}^{-1}\widetilde{\theta}_{nk} := (\Pi'_{nk}\underline{W}'_{nk}\Lambda_{nk}\underline{W}_{nk}\Pi_{nk}/n_k)^{-1}(\Pi'_{nk}\underline{W}'_{nk}\Lambda_{nk}Y_{nk}/n_k)$.

50

Define $Q_{ww,nk} := \mathbb{E}[\lambda(X_i)W_i W_i' \mid \mathcal{I}_{-nk}]$. By Lemma 11,

$$
\Pi_{nk}' Q_{ww,nk} \Pi_{nk} = \begin{pmatrix} \mathbb{E}[\lambda(X_i) \mid \mathcal{I}_{-nk}] & \mathbb{E}[\lambda(X_i) M_{-k}(X_i) \mid \mathcal{I}_{-nk}] & 0 & 0 \\ \mathbb{E}[\lambda(X_i) M_{-k}(X_i) \mid \mathcal{I}_{-nk}] & \mathbb{E}[\lambda(X_i) M_{-k}(X_i)^2 \mid \mathcal{I}_{-nk}] & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}.
$$
(55)

By Assumptions 1.(ii), 3.(i) and (ii), $\Pi_{nk}' Q_{ww,nk} \Pi_{nk}$ has bounded eigenvalues. Algebraically, $\frac{\Pi_{nk}' \underline{W}_{nk}' \Lambda_{nk} \underline{W}_{nk} \Pi_{nk}}{n_k} = \frac{1}{n_k} \sum_{i \in \mathcal{I}_{nk}} \Pi_{nk}' \lambda(X_i) W_i W_i' \Pi_{nk}$. Since the terms are conditionally i.i.d, it converges to (55). By Assumptions 3.(iii) and (iv) $\lambda(X_i) \Pi_{nk} W_i U_i$ is bounded in the L_2 norm, and $\sqrt{n_k}(\hat{\theta}_{nk}^* - \theta_{nk}) = (\Pi_{nk}' Q_{ww,nk} \Pi_{nk})^{-1} \frac{\Pi_{nk}' \underline{W}_{nk}' \Lambda_{nk} \underline{U}_{nk}}{\sqrt{n_k}} + o_p(1)$. Also, $V_{xnk}^{-1} \hat{V}_{xnk} - 1 = \frac{1}{n_k} \sum_{i \in \mathcal{I}_{nk}} [V_{xnk}^{-1} S_{-k}(X_i)^2 - 1] - \left[\frac{1}{n_k} \sum_{i \in \mathcal{I}_{nk}} V_{xnk}^{-1/2} S_{-k}(X_i) \right]^2$. The second term is $O_p(n^{-1})$ because the summand has mean zero and unit variance.

$$
A_i := \begin{pmatrix} \lambda(X_i) \Pi_{nk} W_i U_i \\ V_{xnk}^{-1} S_{-k}(X_i) - 1 \end{pmatrix}, \quad J_{nk}^* := \begin{bmatrix} (\Pi_{nk}' Q_{ww,nk} \Pi_{nk})^{-1} & \underline{0}_4 \\ \underline{0}_4' & 1 \end{bmatrix}.
$$

By (55), $(\Pi_{nk}' Q_{ww,nk} \Pi_{nk})^{-1}$ has a block-diagonal structure, with a one in the bottom-right cell, and hence $\Upsilon [J_{nk}^*]^{-1} = \Upsilon$. This means that

$$
\sqrt{n_k} \Upsilon \begin{pmatrix} \tilde{\theta}_n^* - \theta_n^* \\ V_{xnk}^{-1} \hat{V}_{xnk} - 1 \end{pmatrix} = \Upsilon [J_{nk}^*]^{-1} \left[\frac{1}{\sqrt{n_k}} \sum_{i \in \mathcal{I}_{nk}} A_i \right] + o_p(1)
$$
$$
= \frac{1}{\sqrt{n_k}} \sum_{i \in \mathcal{I}_{nk}} \begin{pmatrix} V_{xnk}^{-1/2} \lambda(X_i)(D_i - p(X_i)) S_{-k}(X_i) U_i \\ V_{xnk}^{-1/2} S_{-k}(X_i)^2 - 1 \end{pmatrix} + o_p(1).
$$
(56)

Define $\hat{\tau}_{nk,av} := \frac{1}{n_k} \sum_{i \in \mathcal{I}_{nk}} \hat{\tau}_{-nk}(X_i)$ and $\tau_{nk,av} := \mathbb{E}[\hat{\tau}_{-nk}(X_i) \mid \mathcal{I}_{-nk}]$. Algebraically, $\widehat{\underline{W}}_{nk} = \underline{W}_{nk}(I_{4 \times 4} + \Delta_{nk})$, where $\Delta_{nk} = [\underline{0}_4, \underline{0}_4, \underline{0}_4, e_3(\hat{\tau}_{nk,av} - \tau_{nk,av})]$ and $e_3 = [0, 0, 1, 0]'$. The bias in centering the fourth column of $\widehat{W}_{nk}$ is a scalar times the third regressor. Substituting $\widehat{\underline{W}}_{nk} = \underline{W}_{nk} \Pi_{nk} \Pi_{nk}^{-1}(I + \Delta_{nk})$,

$$
\hat{\theta}_{nk}^* = \left(\Pi_{nk}^{-1}(I + \Delta_{nk}) \Pi_{nk} \right)^{-1} \left(\Pi_{nk}' \underline{W}_{nk}' \Lambda_{nk} \underline{W}_{nk} \Pi_{nk} \right)^{-1} \left(\Pi_{nk}' \underline{W}_{nk}' \Lambda_{nk} \underline{Y}_{nk} \right),
$$
(57)

which simplifies to $\hat{\theta}_{nk}^* = \left(\Pi_{nk}^{-1}(I + \Delta_{nk}) \Pi_{nk} \right)^{-1} \tilde{\theta}_{nk}^*$. Since $\Delta_{nk} \Delta_{nk} = 0_{4 \times 4}$, then

51

$$\left(\Pi_{nk}^{-1}(I + \Delta_{nk})\Pi_{nk} \right)^{-1} = \left(I + V_{xnk}^{-1/2}\Delta_{nk} \right)^{-1} = I - V_{xnk}^{-1/2}\Delta_{nk}, \text{ and}$$

$$\begin{pmatrix} \widehat{\beta}_{2n}^* \\ V_{xnk}^{-1}\widehat{V}_{xnk} \end{pmatrix} := \Upsilon \left(\begin{matrix} \left(\Pi_{nk}^{-1}\tilde{\Delta}_{nk}\Pi_{nk}\right)^{-1} & 0_{4\times 1} \\ 0_{1\times 4} & 1 \end{matrix} \right) \begin{pmatrix} \tilde{\theta}_{nk}^* \\ V_{xnk}^{-1}\widehat{V}_{xnk} \end{pmatrix} = \Upsilon \begin{pmatrix} \tilde{\theta}_{nk}^* \\ V_{xnk}^{-1}\widehat{V}_{xnk} \end{pmatrix}. \quad (58)$$

The second equality follows from the fact that the selection matrix Υ only extracts the fourth and fifth rows of the vector. We can plug-in (58) and the identity $\Upsilon[\theta_{nk}^{*'}, 1]' = [\beta_{2nk}^*, 1]'$ into (56). By Assumptions 4 and 5, the observations are conditionally i.i.d. given $\mathcal{I}_{-nk}$, and Ω_{nk} has bounded eigenvalues. Then by the Lindeberg-Feller CLT,

$$\sqrt{n_k}\Omega_{nk}^{-1/2} \begin{pmatrix} \widehat{\beta}_{2nk}^* - \beta_{2nk}^* \\ V_{xnk}^{-1}\widehat{V}_{xnk} - 1 \end{pmatrix} \mid \mathcal{I}_{-nk} \to^d \mathcal{N}(0_{2\times 1}, I_{2\times 2}). \quad (59)$$

$\square$

Proof of Theorem 3.(ii). Let $\tilde{\Delta}_{nk} = I + \Delta_{nk}$, where the non-zero term in Δ_{nk} is an average with mean zero and variance V_{xnk}. Under the assumptions of part (a), $\widehat{\Pi}_{nk}\Pi_{nk} = I + o_p(1)$ and $\Pi_{nk}\tilde{\Delta}_{nk}\Pi_{nk}^{-1} = I + V_{xnk}^{-1/2}\Delta_{nk} = I + o_p(1)$. Decomposing $\widehat{\Pi}_{nk}\widehat{\underline{W}}'_{nk}\Lambda_{nk}\widehat{\underline{W}}_{nk}\widehat{\Pi}_{nk}$ in a similar way to (57),

$$(\widehat{\Pi}_{nk}\tilde{\Delta}'_{nk}\Pi_{nk}^{-1})(\Pi'_{nk}\underline{W}'_{nk}\Lambda_{nk}\underline{W}_{nk}\Pi_{nk})(\Pi_{nk}^{-1}\tilde{\Delta}_{nk}\widehat{\Pi}_{nk}) = \Pi'_{nk}Q_{ww,nk}\Pi_{nk} + o_p(1)$$

. Hence $\widehat{J}_{nk} = J_{nk}^* + o_p(1)$. Substituting $\widehat{\underline{W}}_{nk}$, into the upper-left block of $\widehat{H}_{nk}$,

$$(\widehat{\Pi}_{nk}\tilde{\Delta}'_{nk}\Pi_{nk}^{-1}) \left[\frac{1}{n_k} \sum_{i\in\mathcal{I}_{-nk}} \widehat{U}_i^2\lambda(X_i)^2\Pi'_{nk}W_iW_i'\Pi_{nk} \right] (\Pi_{nk}^{-1}\tilde{\Delta}_{nk}\widehat{\Pi}_{nk}). \quad (60)$$

As before the outer terms are $I + o_p(1)$. For the inner terms we apply (57), $\widehat{U}_i = Y_i - \widehat{W}_i'\widehat{\theta}_{nk} = Y_i - W_i'\tilde{\Delta}_{nk}\widehat{\theta}_{nk} = Y_i - W_i'\Pi_{nk}\Pi_{nk}^{-1}\tilde{\Delta}_{nk}\Pi_{nk}\Pi_{nk}^{-1}\widehat{\theta}_{nk} = Y_i - W_i'\Pi'_{nk}\widehat{\theta}_{nk}^* = U_i - W_i'\Pi_{nk}(\widehat{\theta}_{nk}^* - \theta_{nk}^*)$. The inner term of (60) is $\mathbb{E}[U_i^2\lambda(X_i)^2\Pi'_{nk}W_iW_i'\Pi_{nk} \mid \mathcal{I}_{-nk}] + o_p(1)$ under Assumption 1.(ii) (overlap) and 3 (bounds on moments of U_i and $\Pi_{nk}W_i$). In particular, 3.(iv) ensures that $\mathbb{E}[\|\Pi_{nk}W_i\|^4 \mid \mathcal{I}_{-nk}]$ is uniformly bounded.

Define $\bar{S}_{nk} := \frac{1}{n_k}\sum_{i\in\mathcal{I}_{-nk}} S_{-k}(X_i)$ and $\widehat{S}_{-k}(X_i) = \widehat{\tau}_{-k}(X_i) - \frac{1}{n_k}\sum_{i\in\mathcal{I}_{-nk}} \widehat{\tau}_{-k}(X_i)$. Adding/subtracting the mean, $\widehat{S}_{-k}(X_i) = S_{-k}(X_i) - \bar{S}_{nk}$. Algebraically, $\frac{1}{n_k}\sum_{i\in\mathcal{I}_{nk}} \widehat{T}_i^2 = \widehat{V}_{xnk}^{-2}\left(\frac{1}{n_k}\sum_{i\in\mathcal{I}_{nk}}[\widehat{S}_{-k}(X_i)^2 - \widehat{V}_{xnk}]^2 \right) = \widehat{V}_{xnk}^{-2}\left(\frac{1}{n_k}\sum_{i\in\mathcal{I}_{nk}} \widehat{S}_{-k}(X_i)^4 - \widehat{V}_{xnk}^2 \right)$. This simpli-

fies to $\widehat{V}_{xnk}^{-2}\left(\frac{1}{n_k}\sum_{i\in\mathcal{I}_{nk}}[S_{-k}(X_i)-\bar{S}_{nk}]^4\right) - 1$. By a binomial expansion,

$$\frac{1}{n_k}\sum_{i\in\mathcal{I}_{nk}}\frac{S_{-k}(X_i)^4}{\widehat{V}_{xnk}^2} \quad + \underbrace{\left(\frac{V_{xnk}}{\widehat{V}_{xnk}}\right)^2}_{\to^p 1}\sum_{\ell=1}^{4}\binom{4}{\ell}(-1)^\ell \underbrace{\frac{\bar{S}_{nk}^\ell}{V_{xnk}^{\ell/2}}}_{o_p(1)}\underbrace{\left[\frac{1}{n_k}\sum_{i\in\mathcal{I}_{nk}}V_{xnk}^{2-\ell/2}S_{-k}(X_i)^{4-\ell}\right]}_{O_p(1)} -1.$$

By Assumption 3.(iv), the bound on V_{xnk}, and the moment bounds for $V_{xnk}^{-1/2}S_{-k}(X_i)$, then for $\ell\in\{1,2,3,4\}$, $\left[\frac{1}{n_k}\sum_{i\in\mathcal{I}_{nk}}V_{xnk}^{2-\ell/2}S_{-k}(X_i)^{4-\ell}\right]$ is $O_p(1)$. By the weak law of large number, $V_{xnk}^{-1/2}\bar{S}_{nk} = o_p(1)$ and by (59), $\widehat{V}_{xnk}/V_{xnk}\to^p 1$. Hence $\frac{1}{n_k}\sum_{i\in\mathcal{I}_{nk}}\widehat{T}_i^2 = \mathbb{E}[V_{xnk}^{-1}S_{-k}(X_i)^4\mid\mathcal{I}_{-nk}] - 1 + o_p(1) = \mathbb{V}(V_{xnk}^{-1/2}S_{-k}(X_i)^2) + o_p(1)$. This shows that the diagonals of $\widehat{H}_{nk}$ converge to their population analogs. The proof of convergence for the off-diagonals is similar. By the form of $Q_{ww,nk}$ in (55), $\Upsilon J_{nk}^{-1} = \Upsilon$. Hence $\Upsilon\widehat{J}_{nk}^{-1}\widehat{H}_{nk}\widehat{J}_{nk}\Upsilon' = \Upsilon J_{nk}^{*-1}H_{nk}J_{nk}^{*}\Upsilon' + o_p(1) = \Upsilon H_{nk}\Upsilon' + o_p(1) = \Omega_{nk} + o_p(1)$.

$\square$

Proof of Lemma 4. We start by proving that $\|V_\tau(\gamma) - V_\tau^*(\gamma,\mathcal{I}_{-nk})\|\leqslant V_\tau$. Since the second term of (33) is always non-negative, $V_\tau^*(\gamma,\mathcal{I}_{-nk})\leqslant V_\tau(\gamma)$. Also, since $\beta_1 = \tau_{\gamma,av}$ and $\beta_2 = 0$ are part of the feasible set, then the second term is at most $V_\tau(\gamma) = \mathbb{E}_\gamma[(\tau_\gamma(X) - \tau_{\gamma,av})^2]$. This shows that $V_\tau^*(\gamma,\mathcal{I}_{-nk})\geqslant 0$.

To prove the second bound, we examine the solution to the best linear projection. Since the folds are independent, conditioning on $\mathcal{I}_{-nk}$ is only important to be able to handle $S_{-k}(x)$ as a deterministic function. For ease of exposition, let F denote the conditional distribution of γ given $\mathcal{I}_{-nk}$, and let $\|g\|_{F,2} = \sqrt{\mathbb{E}_F[\|g(X)\|^2]}$ denote the L_2 norm. Since $\mathbb{E}_F[S_{-k}(X)] = 0$, the optimal solution is $\beta_{1F}^* = \tau_{F,av}$. When $\mathbb{V}_F(S_{-k}(X)) = 0$, the optimal β_{2F}^* is indeterminate, and $V_\tau^*(\gamma,\mathcal{I}_{-nk}) = 0$. Otherwise,

$$(\beta_{2F}^* - 1) = \frac{Cov_F(S_{-k}(X),\tau_F(X))}{\mathbb{V}_F(S_{-k}(X))} - 1 = \frac{Cov_F(S_{-k}(X),\tau_F(X) - S_{-k}(X))}{\mathbb{V}_F(S_{-k}(X))}.$$

By the Cauchy-Schwarz inequality, $(\beta_{2F}^* - 1)^2\mathbb{V}_F(S_{-k}(X))\leqslant \mathbb{E}_F[(\tau_\gamma(X) - \tau_{F,av} - S_{-k}(X))^2]$. Rearranging (33) and substituting the optimum:

$$V_\tau - V_\tau^* = \mathbb{E}_F[(\tau_\gamma(X) - \tau_{av} - S_{-k}(X) - (\beta_2^* - 1)S_{-k}(X))^2].$$

By the triangle inequality for the L_2 norm, and the above Cauchy-Schwartz bound,

$$V_\tau(\gamma) - V_\tau^*(\gamma, \mathcal{I}_{-nk}) \leqslant [2\|\tau_F(X) - \tau_{F,av} - S_{-k}(X)\|_{F,2}]^2. \tag{61}$$

Substituting $S_{-k}(x) = \hat{\tau}_{-k}(x) - \mathbb{E}_F[\hat{\tau}_{-k}(X)]$, (61) can be rewritten as $4\|(\tau_\gamma(X) - \hat{\tau}_{-k}(X)) - (\mathbb{E}_F[\tau_\gamma(X)] - \mathbb{E}_F[\hat{\tau}_{-k}(X)])\|^2 \leqslant 16\|\hat{\tau}_{-k} - \tau_\gamma\|_{F,2}^2$. By the law of iterated expectations $\omega(\gamma) := \mathbb{E}_\gamma[\|\hat{\tau}_{-k}(X) - \tau_\gamma(X)\|^2] = \mathbb{E}_\gamma[\|\hat{\tau}_{-k} - \tau_\gamma\|_{F,2}^2]$. Combining the two bounds and applying Jensen's inequality,

$$\mathbb{E}_\gamma[|V_\tau(\gamma) - V_\tau^*(\gamma, \mathcal{I}_{-nk})|] \leqslant \mathbb{E}_\gamma[\min\{16\|\hat{\tau}_{-k} - \tau\|_{F,2}^2, V_\tau(\gamma)\}] \leqslant \min\{16\omega(\gamma), V_\tau(\gamma)\}.$$

$\square$

Proof of Theorem 5 . Decompose $\Delta_{nk} = \Delta_{nk1} + \Delta_{nk2}$, where $\Delta_{nk1} := V_\tau(\gamma_n) - V_\tau^*(\gamma_n, \mathcal{I}_{-nk})$, and $\Delta_{nk2} := \hat{V}_{\tau nk} - V_\tau^*(\gamma_n, \mathcal{I}_{-nk})$. First, by Lemma 4 and Assumption 6, $\mathbb{E}_{\gamma_n}[\|\Delta_{nk1}\|] \leqslant \min\{16\omega(\gamma_n)^2, V_\tau(\gamma_n)\} = o(n_k^{-1/2})$. By Markov's inequality, $\Delta_{nk1} = o_p(n_k^{-1/2})$. Second, conditional on a sequence $\{\mathcal{I}_{-nk}\}_{n=1}^\infty$, by Theorem 3 and Lemma 2, $\|\Delta_{nk2}\| = O_p\left(\max\left\{\frac{1}{n_k}, \sqrt{V_\tau^*(\gamma_n, \mathcal{I}_{-nk})}/\sqrt{n_k}\right\}\right)$. Since $V_\tau^*(\gamma_n, \mathcal{I}_{-nk}) \leqslant V_\tau(\gamma_n) = o(1)$, then almost surely, for all $\varepsilon > 0$, $\psi_{n,\varepsilon}(\mathcal{I}_{-nk}) := \mathbb{P}_{\gamma_n}(n_k^{1/2}\|\Delta_{nk2}\| > \varepsilon \mid \mathcal{I}_{-nk}) \to 0$. By iterated expectations and the bounded convergence theorem, $\lim_{n_k \to \infty} \mathbb{P}_{\gamma_n}(n_k^{1/2}\|\Delta_{nk2}\| < \varepsilon) = \lim_{n_k \to \infty} \mathbb{E}_{\gamma_n}[\psi_{n,\varepsilon}(\mathcal{I}_{-nk})] = \mathbb{E}_{\gamma_n}[\lim_{n_k \to \infty} \psi_{n,\varepsilon}(\mathcal{I}_{-nk})] = 0$.

If, in addition, $n_k^{1/2+\rho}V_\tau(\gamma_n) \to 0$ for $\rho \in [0, 1/2)$, then $n_k^{1/2+\rho}\Delta_{nk1} = o_p(1)$ (the second bound dominates). Since $n_k^\rho V_\tau(\gamma_n) \to 0$ as well and $\rho < 1/2$, then conditional on $\mathcal{I}_{-nk}$, $n_k^{1/2+\rho}\Delta_{nk2} = o_p(1)$. We can apply the bounded convergence theorem once again to show that $\mathbb{P}_{\gamma_n}(n_k^{1/2+\rho}\|\Delta_{nk2}\| < \varepsilon) = o(1)$.

$\square$

Proof of Theorem 6. Consider a sequence of distributions $\{\gamma_n\}_{n=1}^\infty$, with associated nuisance functions $(\tau_n(x), \mu_{0n}(x), p_n(x), \tau_{n,av})$.

Case 1 (Near Homogeneity): $V_\tau(\gamma_n) \to 0$. By the triangle inequality,

$$\sqrt{n_k}\left\|\hat{V}_{\tau nk} - \frac{1}{n_k}\sum_{i\in\mathcal{I}_{nk}}\varphi_i\right\| \leqslant \underbrace{\sqrt{n_k}\left\|\hat{V}_{\tau nk} - V_\tau(\gamma_n)\right\|}_{\xi_{n1}} + \underbrace{\sqrt{n_k}\left\|\frac{1}{n_k}\sum_{i\in\mathcal{I}_{nk}}\varphi_i - V_\tau(\gamma_n)\right\|}_{\xi_{n2}}.$$

$$\tag{62}$$

By Theorem 5, $\xi_{n1} = o_p(1)$. Since the φ_i are i.i.d with mean $V_\tau(\gamma_n)$, then $\mathbb{E}_{\gamma_n}[\xi_{n2}^2] = \mathbb{V}_{\gamma_n}(\varphi_i)$. Let $\sigma_{dn}^2(x) = \mathbb{V}_{\gamma_n}(Y_i \mid X_i = x, D_i = d)$ and $p_n(x) = \mathbb{P}_{\gamma_n}(D_i \mid X_i = x)$. By Lemma 3 and Corollary 2, $\mathbb{V}_{\gamma_n}(\varphi_i) \leqslant \kappa^2 V_\tau(\gamma_n)^2 + 4\kappa V_\tau(\gamma_n)\sqrt{\mathbb{E}_{\gamma_n}\left[\left(\frac{\sigma_{1n}^2(X_i)}{p_n(X_i)} + \frac{\sigma_{0n}^2(X_i)}{1-p_n(X_i)}\right)\right]}$. Since $V_\tau(\gamma_n) \to 0$ as $n \to \infty$, then $\mathbb{E}[\xi_{n2}^2] \to 0$ and hence $\xi_{n2} = o_p(1)$.

Case 2 (Strong Heterogeneity): $V_\tau(\gamma_n) \to V_\tau \in (0, \infty)$. For fixed (x, d), the cross-fitted regressors are $\widehat{W}(x, d) = [1, \hat{\mu}_{0,-k}(x) + p_n(x)\hat{\tau}_{-k}(x), (d - p_n(x)), (d - p(x))\widehat{S}_{-k}(x)]$. Let e_ℓ be a 4×1 vector with a one in the ℓth coordinate and zero otherwise. Analogous to (19), define the regression adjusted nuisance functions as

$$\hat{\eta}_{-k,\hat{\theta}_{nk}}(x) = [e_1(\widehat{W}(x,1) - \widehat{W}(x,0))' + e_2\widehat{W}(x,0)']\hat{\theta}_{nk} + e_3 p_n(x) + e_4\hat{\tau}_{nk,av}. \quad (63)$$

By applying Lemma 4 to the subsample in $i \in \mathcal{I}_{nk}$,

$$\widehat{V}_{\tau nk} = \frac{1}{n_k}\sum_{i \in \mathcal{I}_k} \varphi(Y_i, D_i, X_i, \hat{\eta}_{-k,\hat{\theta}_{nk}}),$$

where $\eta_{-k,\hat{\theta}_{nk},r}(x) := \eta(x) + r(\eta_{-k,\hat{\theta}_{nk}}(x) - \eta(x))$. The true nuisance function η depends on the distribution indexed by n, but we drop the subscript to simplify the notation. By a second-order term in the Taylor expansion around $r = 0$, for some $\tilde{r} \in (0, 1)$,

$$\begin{aligned}
\sqrt{n_k}(\widehat{V}_{\tau nk} - V_\tau(\gamma_n)) &= \frac{1}{\sqrt{n_k}}\sum_{i \in \mathcal{I}_{nk}} [\varphi(Y_i, D_i, X_i, \eta) - V_\tau(\gamma_n)] \\
&+ \frac{1}{\sqrt{n_k}}\sum_{i \in \mathcal{I}_{nk}} \frac{\partial\varphi(Y_i, D_i, X_i, \eta)}{\partial\eta'}(\eta_{-k,\hat{\theta}_{nk}}(X_i) - \eta(X_i)) \\
&+ \frac{1}{n_k}\sum_{i=1}^n \sqrt{n_k}(\eta_{-k,\hat{\theta}_{nk}}(X_i) - \eta(X_i))'\frac{\partial^2\varphi(Y_i, D_i, X_i, \eta_{-k,\hat{\theta}_{nk},\tilde{r}})}{\partial\eta'}(\eta_{-k,\hat{\theta}_{nk}}(X_i) - \eta(X_i)).
\end{aligned}$$
$$(64)$$

Our ultimate goal is to show that the second and third terms of the expansion are $o_p(1)$. To keep the notation concise, let $\varphi_i(\eta) := \varphi(Y_i, D_i, X_i, \eta)$. One of the main challenges is that the nuisance functions are estimated in multiple steps, combining information from the $\mathcal{I}_{nk}$ and $\mathcal{I}_{-nk}$ subsamples. The key is to decompose these different sources of uncertainty. Define $(\hat{\lambda}_{nk} - \lambda_{nk}) := [(\hat{\theta}_{nk} - \theta_n), \hat{\tau}_{nk,av}(\hat{\theta}_{nk} - \theta_n), (\hat{\tau}_{nk,av} - \tau_{n,av})]'$,

where $\theta_n = (0, 1, \tau_{n,av}, 1)$. By Lemma 14, there exist matrices $\Psi_{-nk}(x)$ and $\Delta_{-nk}(x)$ that are $(x, \mathcal{I}_{-nk})$-measurable, such that for all $x \in \mathcal{X}$,

$$\hat{\eta}_{-k,\hat{\theta}_{nk}}(x) - \eta(x) = \Psi_{-nk}(x)(\hat{\lambda}_{nk} - \lambda_{nk}) + \Delta_{-nk}(x). \tag{65}$$

Let e_ℓ be a 4×1 elementary basis vector. Lemma 14 shows that $e_3'[\hat{\eta}_{-k,\hat{\theta}_{nk}}(x) - \eta(x)] = 0$ and $e_3'\Delta_{-nk}(x) = 0$ because the experimental propensity scores are known. Lemma 14 also shows that $e_4'\Delta_{-nk}(x) = 0$ by construction. Substituting (65),

$$\frac{1}{\sqrt{n_k}} \sum_{i \in \mathcal{I}_{nk}} \frac{\partial \varphi_i(\eta)}{\partial \eta'} (\eta_{-k,\hat{\theta}_{nk}}(X_i) - \eta(X_i)) = \frac{1}{\sqrt{n_k}} \sum_{i \in \mathcal{I}_{nk}} \sum_{\ell=1}^{4} \frac{\partial \varphi_i(\eta)}{\partial \eta_\ell} e_\ell'(\eta_{-k,\hat{\theta}_{nk}}(X_i) - \eta(X_i))$$

$$= \underbrace{\left[\frac{1}{\sqrt{n_k}} \sum_{i \in \mathcal{I}_{nk}} \sum_{\ell \neq 3} \frac{\partial \varphi_i(\eta)}{\partial \eta_\ell} e_\ell' \Psi_{-nk}(X_i) \right] (\hat{\lambda}_{nk} - \lambda_{nk})}_{\Lambda_{1nk}} + \underbrace{\left[\frac{1}{\sqrt{n_k}} \sum_{i \in \mathcal{I}_{nk}} \sum_{\ell=1}^{2} \frac{\partial \varphi_i(\eta)}{\partial \eta_\ell} e_\ell' \Delta_{-nk}(X_i) \right]}_{\Lambda_{2nk}}.$$

Lemma 15 implies that $\hat{\lambda}_{nk} - \lambda_{nk} = o_p(1)$.

(i) Prove that $\Lambda_{1nk} = O_p(1)$: By Lemma 13.(a) the conditional mean of $\frac{\partial \varphi_i(\eta)}{\partial(\eta_1, \eta_2, \eta_4)'}$ given $(x, \mathcal{I}_{-nk})$ is $[0, 0, -2(\tau(x) - \tau_{n,av})]$ and by Lemma 14, $e_4'\Psi_{-nk} = c'$. Also, $\mathbb{E}_{\gamma_n}[\tau_n(X_i) \mid \mathcal{I}_{-nk}] = \mathbb{E}_{\gamma_n}[\tau_n(X_i)] = \tau_{n,av}$ by fold independence. By the law of iterated expectations, $\mathbb{E}_{\gamma_n}\left[\frac{\partial \varphi_i(\eta)}{\partial \eta_\ell} e_\ell' \Psi_{-nk}(X_i) \mid \mathcal{I}_{-nk} \right] = 0$ for $\ell \in \{1, 2, 4\}$. By Assumption 7, $\mathbb{E}_{\gamma_n}[\|\eta(X_i)\|^4]$ and $\mathbb{E}_{\gamma_n}[\|U_i\|^4]$ are uniformly bounded by a constant $C < \infty$. By Assumption 1.(ii), $p_n(x)$ is contained in $[\delta, 1 - \delta]$. Applying Lemma 13.(b), $\mathbb{E}_{\gamma_n}\left[\|\partial \varphi_i(\eta)/\partial \eta_\ell\|^4 \right]^{1/4} \leqslant (16/\delta)C < \infty$. By Lemma 14.(d), and the triangle inequality, $\mathbb{E}_{\gamma_n}\left[\|e_\ell' \Psi_{-nk}(X_i)\|^4 \right]^{1/4} \leqslant C[1 + \mathbb{E}_{\gamma_n}[\|\hat{\eta}_{-k}(X_i)\|^4]^{1/4}]$. By the bound in Assumption 7.(iii) and the Cauchy-Schwartz inequality, $\mathbb{E}_{\gamma_n}\left[\|\frac{\partial \varphi_i(\eta)}{\partial \eta_\ell} e_\ell' \Psi_{-nk}(X_i)\| \right] < \infty$. We can combine these moment bounds to apply Lemma 12.(a), hence $\Lambda_{1nk} = O_p(1)$.

(ii) Prove that $\Lambda_{2nk} = o_p(1)$: By Lemma 13.(a), $\mathbb{E}_{\gamma_n}\left[\frac{\partial \varphi_i(\eta)}{\partial \eta_\ell} e_\ell' \Delta_{-nk}(X_i) \mid \mathcal{I}_{-nk} \right] = 0$ for $\ell \in \{1, 2\}$. Using similar arguments $\mathbb{E}_{\gamma_n}\left[\|\partial \varphi_i(\eta)/\partial \eta_\ell\|^4 \right] < C$. By Lemma 14.(b), $\|\Delta_{nk}(x)\| \leqslant C \times \|\hat{\eta}_{-k}(x) - \eta(x)\|$. By Assumption 7.(ii), $\mathbb{E}_{\gamma_n}[\|\hat{\eta}_{-k}(x) - \eta(x)\|^4] = o(1)$, which means that $\mathbb{E}_{\gamma_n}\left[\|\frac{\partial \varphi_i(\eta)}{\partial \eta_\ell} e_\ell' \Psi_{-nk}(X_i)\|^2 \right]$ is $o(1)$. By Lemma 12.(b), $\Lambda_{2nk} = o_p(1)$.

Let $\Xi = [e_1, e_2, e_4]$ be a 4×3 matrix such that $\|\Xi\| \leqslant 1$. Let Λ_{3nk} be the second-

56

order on the right-hand side of (64). Since the propensity score is known,

$$\Lambda_{3nk} :=$$

$$\frac{1}{n_k} \sum_{i \in \mathcal{I}_{nk}} \sqrt{n_k} (\eta_{-k,\hat{\theta}_{nk}}(X_i) - \eta(X_i))' \Xi \frac{\partial^2 \varphi_i(\eta_{-k,\hat{\theta}_{nk},\tilde{r}})}{\partial(\eta_1,\eta_2,\eta_4)\partial(\eta_1,\eta_2,\eta_4)'} \Xi'(\eta_{-k,\hat{\theta}_{nk}}(X_i) - \eta(X_i)).$$

(iii) Prove that $\Lambda_{3nk} = o_p(1)$: By Lemma 13.(d), $\left\| \frac{\partial^2 \varphi_i(\eta_{-k,\hat{\theta}_{nk},\tilde{r}})}{\partial(\eta_1,\eta_2,\eta_4)\partial(\eta_1,\eta_2,\eta_4)'} \right\| \leqslant 18 \times \sqrt{3}/\delta$. For scalars, $a, b \in \mathbb{R}$, $(a+b)^2 \leqslant 4(a^2 + b^2)$. By (65) and the triangle inequality, $\|\Lambda_{3nk}\| \leqslant 4 \times \frac{18\sqrt{3}}{\delta} \left(\sqrt{n_k} \|\hat{\lambda}_{nk} - \lambda_{nk}\|^2 \left[\frac{1}{n_k} \sum_{i \in \mathcal{I}_{nk}} \|\Psi_{-nk}(X_i)\|^2 \right] + \left[\frac{1}{n_k} \sum_{i \in \mathcal{I}_{nk}} \sqrt{n_k} \|\Delta_{-nk}(X_i)\|^2 \right] \right)$.
By Lemma 15, $\sqrt{n_k} \|\hat{\lambda}_{nk} - \lambda_{nk}\|^2 = o_p(1)$. By Lemma 14.(d), $\|\Psi_{-nk}(x)\| \leqslant C[1 + 2\|\hat{\eta}_{-k}(x)\|]$. By the moment bound in Assumption 7.(i) and Markov's inequality, then $\frac{1}{n_k} \sum_{i \in \mathcal{I}_{nk}} \|\Psi_{-nk}(X_i)\|^2 = O_p(1)$. By Lemma 14.(e), $\|\Delta_{-nk}(x)\| \leqslant C \times \|\hat{\eta}_{-k}(x) - \eta(x)\|$. By Assumption 7.(iii), $\sqrt{n_k} \mathbb{E}_{\gamma_n}[\|\hat{\eta}_{-k}(X_i) - \eta(X_i)\|^2] = o(1)$, then by Lemma 12.(c), $\frac{1}{n_k} \sum_{i \in \mathcal{I}_{nk}} \sqrt{n_k} \|\Delta_{-nk}(X_i)\|^2 = o_p(1)$. Combining these results, $\Lambda_{3nk} = o_p(1)$.

$\square$

Proof of Theorem 7. Let $\rho(\gamma, \mathcal{I}_{-nk}) = \mathbb{P}_\gamma(V_\tau^*(\gamma, \mathcal{I}_{-nk}) \in \widehat{CI}_{\alpha nk} \mid \mathcal{I}_{-nk})$ denote the conditional probability that the pseudo-VCATE is contained in the confidence interval. Let $\mathcal{F}(\gamma)$ be the support of $\mathcal{I}_{-nk}$, and let $\mathcal{F}(\gamma, t) \subseteq \mathcal{F}(\gamma)$. Almost surely,

$$\inf_{\mathcal{I}_{-nk} \in \mathcal{F}(\gamma)} \rho(\gamma, \mathcal{I}_{-nk}) \leqslant \mathbb{E}_\gamma[\rho(\gamma, \mathcal{I}_{-nk})] \leqslant \sup_{\mathcal{I}_{-nk} \in \mathcal{F}(\gamma,t)} \rho(\gamma, \mathcal{I}_{-nk}) + \mathbb{P}_\gamma(\mathcal{I}_{-nk} \notin \mathcal{F}(\gamma, t)).$$

(66)

The left inequality considers the worst-case coverage. The right inequality applies the law of iterated expectations by the event $\mathcal{I}_{-nk} \in \mathcal{F}^*(\gamma, t)$, then bounds $\mathbb{P}_\gamma(\mathcal{I}_{-nk} \in \mathcal{F}(\gamma, t)) \leqslant 1$ and $\rho(\gamma, \mathcal{I}_{-nk}) \leqslant 1$ to simplify the expressions. Applying limits,

$$\liminf_{n \to \infty} \inf_{\gamma \in \Gamma} \inf_{\mathcal{I}_{-nk} \in \mathcal{F}(\gamma)} \rho(\gamma, \mathcal{I}_{-nk}) \leqslant \liminf_{n \to \infty} \inf_{\gamma \in \Gamma} \mathbb{P}_\gamma \left(V_\tau^*(\gamma, \mathcal{I}_{-nk}) \in \widehat{CI}_{\alpha nk} \right),$$

(67)

and an analogous result for the upper bound. The data in $\mathcal{I}_{-nk}$ only affects fold k through the estimated nuisance functions $\hat{\eta}_{-k}(x)$. To prove uniform coverage we will derive the bounds for a class of distributions where the (Y_i, D_i, X_i) in fold k is distributed as γ and the plug-in nuisance functions are deterministic. We will define $\mathcal{F}(\gamma, t)$ as the set where $|V_\tau^*(\gamma, \mathcal{I}_{-nk})\Omega_{nk,12}| \leqslant t$ for some fixed $t > 0$.

Let $\{\gamma_n\}_{n=1}^\infty$ and $\{\mathcal{I}_{-nk}\}_{n=1}^\infty$ denote a sequence of distributions and data realizations, respectively. Theorem 3 verifies that under Assumptions 1, 3, 4, and 5, the conditional CDFs $\{F_{\gamma_n|\mathcal{I}_{-nk}}\}_{n=1}^\infty$ satisfy the high-level Assumption 2, almost surely. Furthermore, $\widehat{CI}_{\alpha nk}$ satisfies the form in (11), substituting $(\hat{V}_{\tau nk}, \hat{\Omega}_{nk})$. Therefore, Lemma 10.(a) shows that the uniform lower bound on the asymptotic coverage probability is $(1-\alpha)$. If Assumption 8 also holds, then for all $t > 0$, $\limsup\limits_{n\to\infty} \sup_{\gamma\in\Gamma} \mathbb{P}_\gamma(\mathcal{I}_{-nk} \notin \mathcal{F}^*(\gamma, t)) = 0$. Hence by Lemma 10.(ii), the upper bound is $(1-\alpha)$ plus a term that can be made arbitrarily small by choosing t close to zero.

$\square$

Proof of Lemma 6. Consider an arbitrary sequence of distributions $\{\gamma_n\}_{n=1}^\infty \in \Gamma^\infty$ and a convergent subsequence where $V_\tau(\gamma_{n_\ell}) \to V_\tau$. Since $V_\tau^*(\gamma_{n_\ell}, \mathcal{I}_{-n_\ell k}) \leqslant V_{\tau n_\ell}$ and $\Omega_{n_\ell k}$ has bounded eigenvalues, if $V_\tau = 0$ then $\lim_{\ell\to\infty} \mathbb{P}_{\gamma_{n_\ell}}(\sqrt{V_\tau^*(\gamma_{n_\ell}, \mathcal{I}_{-n_\ell k})}|\Omega_{n_\ell k,12}| > t) = 0$. Now suppose that $V_\tau > 0$. By Assumption 6, the nuisance functions converge to their true value. Then $\hat{\tau}_{-k}(x)$ converges point-wise to $\tau(x)$, and by the moment bound in 3.(v) and the dominated convergence theorem, $V_{xn_\ell k} = \mathbb{V}_{\gamma_{n_\ell}}(\hat{\tau}_{-k}(X) \mid \mathcal{I}_{-n_\ell k}) = V_\tau(\gamma_{n_\ell}) + o_p(1) \geqslant V_\tau + o_p(1)$. By (28),

$$\Omega_{n_\ell k} = \mathbb{V}_{\gamma_{n_\ell}} \left(\begin{bmatrix} \lambda_{n_\ell}(X_i)(D_i - p_{n_\ell}(x_i))V_{xn_\ell k}^{-1/2}S_{-k}(X_i)U_i \\ V_{xn_\ell k}^{-1}S_{-k}(X_i)^2 \end{bmatrix} \mid \mathcal{I}_{-n_\ell k} \right), \qquad (68)$$

where $U_{in_\ell} = Y_i - W_i'\theta_{n_\ell k}$. By Assumption 7, for fixed $\{D_i = d, X_i = x\}$, W_i' point-wise converges to $W_i^{*'} := [1, \mu_{0n_\ell}(x) + p_n(x)\tau_{n_\ell}(x), (d - p_{n_\ell}(x)), (d - p_{n_\ell}(x))\tau_{n_\ell}(x)]$. By Lemma 15, $\theta_{n_\ell k} \to \theta_{n_\ell} := [0, 1, \tau_{n_\ell,av}, 1]$. By Lemma 5, $W_i^{*'}\theta_{n_\ell} = \mu_{d,n_\ell}(x)$, and hence $U_{in_\ell}^* := Y_i - W_i^{*'}\theta_{n_\ell}$. Since $\Omega_{n_\ell k}$ is almost surely bounded by Assumption 3 over random partitions $\mathcal{I}_{-n_\ell k}$, then applying the dominated convergence theorem,

$$\Omega_{n_\ell k} = \mathbb{V}_{\gamma_{n_\ell}} \left(\begin{bmatrix} \lambda_{n_\ell}(X_i)(D_i - p_{n_\ell}(x_i))V_\tau^{-1/2}(\tau_{n_\ell}(X_i) - \tau_{n_\ell,av})U_{in_\ell}^* \\ V_\tau^{-1}(\tau_{n_\ell}(X_i) - \tau_{n_\ell,av})^2 \end{bmatrix} \mid \mathcal{I}_{-n_\ell k} \right) + o_p(1), \qquad (69)$$

Since $\mathbb{E}_{\gamma_n}[U_{in_\ell}^* \mid D_i = d, X_i = x, \mathcal{I}_{-n_\ell k}] = 0$ and the second component of (69) only depends on X_i. By iterated expectations, $\Omega_{n_\ell k,12} = o_p(1)$ and the limiting probability is $\lim_{\ell\to\infty} \mathbb{P}_{\gamma_{n_\ell}}(\sqrt{V_\tau^*(\gamma_{n_\ell}, \mathcal{I}_{-n_\ell k})}|\Omega_{n_\ell k,12}| > t) = 0$. Hence we verified Assumption B^* in Andrews et al. (2020). Uniform consistency follows from their Corollary 2.1.

$\square$

Proof of Theorem 8. We break-down the proof into cases.

Case (i): When $V_\tau(\gamma) = 0$, then $V_\tau(\gamma, \mathcal{I}_{-nk}) = 0$ almost surely. Therefore, $n_k(\hat{V}_{\tau nk} - V_\tau^*(\gamma_n, \mathcal{I}_{-nk})) = n_k(\hat{V}_{\tau nk}^* - V_\tau(\gamma))$. Define $\rho(\gamma, \mathcal{I}_{-nk})$ as in Theorem 7. Then $\liminf_{n\to\infty} \inf_{\mathcal{I}_{-nk} \in \mathcal{F}(\gamma)} \rho(\gamma, \mathcal{I}_{-nk}) \leqslant \liminf_{n\to\infty} \mathbb{P}_\gamma \left(V_\tau^*(\gamma, \mathcal{I}_{-nk}) \in \widehat{CI}_{\alpha nk} \right)$. We can prove an analogous upper bound. This implies that we only need to derive coverage bounds under sequences of conditional distributions where $V_\tau^*(\gamma_n, \mathcal{I}_{-nk}) = 0$. Exact coverage follows from the proof of the near-homogeneity case in Lemma 10.

Case (ii) When $V_\tau(\gamma) > 0$. By Assumption 6, $\sqrt{n_k}\omega(\gamma)^2 \to 0$ as $n_k \to \infty$. Then by Lemma 4, $\sqrt{n_k}|V_\tau(\gamma) - V_\tau^*(\gamma, \mathcal{I}_{-nk})| = o_p(1)$, and by the continuous mapping theorem, $\sqrt{V_\tau^*(\gamma, \mathcal{I}_{-nk})} \to^p \sqrt{V_\tau(\gamma)} > 0$. By (69) in the proof of Lemma 6 and for fixed $\gamma_n = \gamma$, $\Omega_{nk} = \Omega + o_p(1)$, where Ω is a population covariance matrix that does not depend on $\mathcal{I}_{-nk}$. By applying similar limits to the mild heterogeneity case in Lemma 10 we can show that $(\hat{V}_{\tau nk} - V_\tau(\gamma))$ is $\sqrt{n_k}$ asymptotically equivalent to an empirical process indexed by the oracle $V_\tau(\gamma)$. We obtain exact coverage due to Lemma 6. $\qquad\square$

Proof of Corollary 3. By the first part of Theorem 7,

$$\limsup_{n\to\infty} \sup_{\gamma\in\Gamma} \mathbb{P}_\gamma \left(V_\tau^*(\gamma, \mathcal{I}_{-nk}) \notin \widehat{CI}_{\alpha nk} \right) = 1 - \liminf_{n\to\infty} \inf_{\gamma\in\Gamma} \mathbb{P}_\gamma \left(V_\tau^*(\gamma, \mathcal{I}_{-nk}) \in \widehat{CI}_{\alpha nk} \right) \leqslant \alpha.$$

The result follows taking limits on either side of the inequality in (38). $\qquad\square$

Proof of Theorem 9 . By the definition in (32),

$$V_\tau(\gamma) \geqslant \inf \widehat{CI}_{\alpha n}^{\text{multifold}} \iff \frac{1}{K} \sum_{k=1}^{K} \mathbb{1}\left\{ V_\tau(\gamma) \geqslant \inf \widehat{CI}_{\alpha nk} \right\} \geqslant \frac{1}{2}.$$

By negating the statement, computing expectations, and applying Markov's inquality,

$$\mathbb{P}_\gamma(V_\tau(\gamma) < \inf \widehat{CI}_{\alpha n}^{\text{multifold}}) = \mathbb{E}_\gamma \left[\mathbb{1}\left\{ \left(\frac{1}{K} \sum_k \mathbb{1}\{V_\tau(\gamma) < \inf \widehat{CI}_{\frac{\alpha}{2}nk}\} \right) > 1/2 \right\} \right]$$

$$\leqslant 2\frac{1}{K} \sum_k \mathbb{E}_\gamma[\mathbb{1}\{V_\tau(\gamma) < \inf \widehat{CI}_{\frac{\alpha}{2}nk}\}]$$

$$\leqslant 2\mathbb{P}_\gamma \left(V_\tau(\gamma) < \inf \widehat{CI}_{\frac{\alpha}{2}nk} \right) \leqslant 2\mathbb{P}_\gamma \left(V_\tau(\gamma, \mathcal{I}_{-nk}^*) \notin \widehat{CI}_{\frac{\alpha}{2}nk} \right)$$

The last line follows from the fact that the folds are split at random and the inequality in (38) holds almost surely. By Theorem 7 the asymptotic size is uniformly

bounded by $2(\alpha/2) = \alpha$. By construction, $\mathbb{P}_\gamma(V_\tau(\gamma) \notin \widehat{CI}_{\alpha n}^{\text{multifold}}) = \mathbb{P}_\gamma(V_\tau(\gamma) < \inf \widehat{CI}_{\alpha n}^{\text{multifold}}) + \mathbb{P}_\gamma(V_\tau(\gamma) > \sup \widehat{CI}_{\alpha n}^{\text{multifold}})$. Applying similar arguments as before, $\mathbb{P}_\gamma(V_\tau(\gamma) \notin \inf \widehat{CI}_{\alpha n}^{\text{multifold}}) \leqslant 2\mathbb{P}_\gamma\left(V_\tau(\gamma) \notin \widehat{CI}_{\frac{\alpha}{2}nk}\right)$. If Assumptions 6, and 7 also hold, then Theorem 8 implies that the right-hand side is point-wise bounded by α.

$\square$

Proof of Lemma 7. Under the null hypothesis, $G(n_k, 0, \hat{\Omega}_{nk}, Z, \zeta) = (e_1'\hat{\Omega}_{nk}^{1/2}Z)^2/n_k$, where $\hat{\Omega}_{nk,11}^{1/2}Z$, where $Z \sim \mathcal{N}(0,1)$. The adjusted critical values are $q_{\alpha/2}(n_k, 0, \hat{\Omega}_{nk}, \zeta) = 0$ and $q_{1-\alpha/2}(n_k, 0, \hat{\Omega}_{nk}, \zeta) = \hat{\Omega}_{nk,11}z_{1-\alpha}^2/n_k$. Then $0 \in \widehat{CI}_{\alpha nk}$ if and only if $0 \leqslant \hat{V}_{\tau nk} - 0 \leqslant \hat{\Omega}_{nk,11}z_{1-\alpha}^2/n_k$. Following similar steps to the "near homogeneity" regime in Theorem 7, $n_k(\hat{V}_{\tau nk} - V_\tau^*(\gamma_n, \mathcal{I}_{-nk})) = (\Omega_{\infty,11}^{1/2}Z_{nk} + \sqrt{v})^2 - v + o_p(1)$, where $Z_{nk} \sim \mathcal{N}(0,1)$. Consequently, $n_k\hat{V}_{\tau nk} = (\Omega_{\infty,11}^{1/2}Z_{nk} + \sqrt{v})^2 + o_p(1)$. Then

$$
\begin{aligned}
\mathbb{P}_{\gamma_n}(0 \in \widehat{CI}_{\alpha nk} \mid \mathcal{I}_{-nk}) &= \mathbb{P}_{\gamma_n}\left(0 \leqslant n_k\hat{V}_{\tau nk} \leqslant \hat{\Omega}_{nk,11}z_{1-\alpha}^2\right) \\
&= \mathbb{P}_{\gamma_n}\left(0 \leqslant (\Omega_{\infty,11}^{1/2}Z_{nk} + \sqrt{v})^2 \leqslant \Omega_{\infty,11}z_{1-\alpha}^2\right) + o(1) \\
&= \mathbb{P}_{\gamma_n}\left(-\frac{\sqrt{v}}{\sqrt{\Omega_{\infty,11}}} \leqslant Z_{nk} + o_p(1) \leqslant z_{1-\alpha} - \frac{\sqrt{v}}{\sqrt{\Omega_{\infty,11}}}\right) + o(1).
\end{aligned}
$$

and hence $\lim_{n\to\infty} \mathbb{P}_{\gamma_n}(0 \in \widehat{CI}_{\alpha nk} \mid \mathcal{I}_{-nk}) = \Phi\left(z_{1-\alpha} - \sqrt{v}/\sqrt{\Omega_{\infty,11}}\right) - \Phi\left(-\sqrt{v}/\sqrt{\Omega_{\infty,11}}\right)$. The final result is obtained by $1 - \mathbb{P}_{\gamma_n}(0 \in \widehat{CI}_{\alpha nk} \mid \mathcal{I}_{-nk})$. $\square$

Proof of Lemma 8. The first part of the proof is identical to that of Theorem 7 in terms of setting up the problem, and defining a sequence the conditional distributions $\{F_{\gamma_n \mid \mathcal{I}_{-nk}}\}_{n=1}^\infty$. By equation (39) this sequence satisfies Assumption 2 almost surely with an effective sample size $\tilde{n} = n/r_n$ and a particular choice of covariance estimator. To complete the proof, we develop a modified version of Lemma 10 to prove uniform coverage under cluster dependence. Consider a sequence of distributions $\{\gamma_n\}_{n=1}^\infty \in \Gamma^\infty$ and a subsequence $\{n_\ell\}_{\ell=1}^\infty$, where $\frac{n_\ell}{r_{n_\ell}}V_{\tau n_\ell}^* \to^p v \in [0,\infty)$, $V_{\tau n_\ell}^* \to^p 0$, and $\Omega_{n_\ell} \to \Omega$ as $n_\ell \to \infty$. Applying Lemma 2, and factorizing terms as in Lemma 10,

$$
\frac{n_\ell}{r_{n_\ell}}(\hat{V}_{\tau n_\ell}^* - V_{\tau n_\ell}^*) = (e_1'\Omega^{1/2}Z_{n_\ell} + \sqrt{v})^2 - v + o_p(1). \tag{70}
$$

Now we need to show that the quantiles of the empirical process converge to the same

limit. The quantity $\frac{n_\ell}{r_{n_\ell}} G(n_\ell, V^*_{\tau n_\ell}, \widehat{\Omega}_{n_\ell}, Z)$ is equal to

$$n_\ell \frac{(e_1' r_{n_\ell}^{-1/2} \widehat{\Omega}_{n_\ell}^{1/2} Z)^2}{n_\ell} + 2\sqrt{\frac{n_\ell V^*_{\tau n_\ell}}{r_{n_\ell}}} (e_1' r_{n_\ell}^{-1/2} \widehat{\Omega}_{n_\ell}^{1/2} Z) + \sqrt{\frac{n_\ell V^*_{\tau n_\ell}}{r_{n_\ell}}} \sqrt{V^*_{\tau n_\ell}} (e_2' r_{n_\ell}^{-1/2} \widehat{\Omega}_{n_\ell}^{1/2} Z). \quad (71)$$

In the first term the n_ℓ components cancel out and $r_{n_\ell}^{-1} \widehat{\Omega}_{n_\ell} \to^p \Omega$. Similarly, since $\frac{n_\ell}{r_{n_\ell}} V^*_{\tau n_\ell} \to^p v$ the second term of (71) converges to $2\sqrt{v} e_1' \Omega^{1/2} Z$, and a suitable factorization with the first and second the expression in (70). Since $n_\ell / r_{n_\ell} \to \infty$ by assumption, then $V^*_{\tau n_\ell} \to 0$ and the third term of (71) is $o_p(1)$. Therefore the estimated quantiles are consistent. Proving consistency of the quantiles for the mild hereterogeneity proceeds analogously. Once we prove that the quantiles are asymptotically correct, the rest of the proof is the same as in Lemma 10. $\qquad \square$

Proof of Theorem 9 . The first part of the proof is identical to that of Theorem 7 in terms of setting up the problem. In this case the sequence of conditional distributions $\{F_{\gamma_n | \mathcal{I}_{-nk}}\}_{n=1}^\infty$ only satisfies Assumption 2 for subsequences where $V_{xnk} = 0$. Instead, I will modify the first part of the proof of Lemma 10.(i) for a class of regression-adjusted CIs with possible degeneracy. Consider a sequence of distributions $\{\gamma_n\}_{n=1}^\infty \in \Gamma^\infty$ and let $h_n := (nV^*_{\tau n}, V^*_{\tau n}, vec(\Omega_n), \zeta_n, V_{xn})$ be a sequence of parameters where V_{xn} is the variance of $S_{-k}(X_i)$. Our goal is to show that for $h \in \mathcal{H}$ and all subsequences $\{n_\ell\}_{\ell=1}^\infty$ where $h_{n_\ell} \to h \in \mathcal{H}$,

$$\lim_{n_\ell \to \infty} \mathbb{P}_{\gamma_{n_\ell}} \left(V^*_{\tau n_\ell} \in \widehat{CI}^0_{\alpha nk} \right) \geqslant 1 - \alpha.$$

Partition the subsequences in such a way that h_{n_ℓ} either has $V_{xn_\ell} = 0$ or $V_{xn_\ell} > 0$. When $V_{xn_\ell} = 0$, then $\widehat{V}^*_{\tau n_\ell}$ is exactly degenerate and the confidence interval (40) covers the pseudo-VCATE with probability one. For the sequences where $V_{xn_\ell} > 0$, we can apply the remaining cases from Lemma 10 which have coverage $1 - \alpha$. This satisfies Assumption B in Andrews et al. (2020) and we can prove uniform conservative coverage of the pseudo-VCATE applying their Corollary 2.1

We prove point-wise coverage by cases. When $V_\tau(\gamma) = 0$, then $V^*_\tau(\gamma, \mathcal{I}_{-nk}) = 0$ almost surely. By applying the result above and the near homogeneity case in Theorem 7 we find that coverage is at least $(1 - \alpha)$. Now consider the case where $V_\tau(\gamma)$ is bounded away from zero. By Assumption 6, the nuisance functions converge to their true value. Then $\widehat{\tau}_{-k}(x)$ converges point-wise to $\tau(x)$, and by the moment bound in 3.(v) and the dominated convergence theorem, $\mathbb{V}_{\gamma_n}(\widehat{\tau}_{-k}(X)) = V_\tau(\gamma) + o(1)$,

which is bounded away from zero. Then we can apply the mild heterogeneity results in Theorem 7 to prove (43). The proof of (44) is identical to that of Theorem 9, substituting the confidence intervals $\widehat{CI}^0_{\alpha nk}$ instead of $\widehat{CI}_{\alpha nk}$. Point-wise coverage of the fold-specific confidence intervals holds by (43).

$\square$

Proof of Corollary 4. The proof of (45) and (46) follows a similar structure to Corollary 3 and Theorem 9, respectively. In each case the only thing that changes is that we apply the uniformity result for degenerate CIs in (42) (Lemma 9), rather than the non-degenerate uniformity result in Theorem 7.

$\square$

C Supporting Lemmas and Proofs

Lemma 10. *Suppose that Γ is a set of distributions constrained in such a way that Assumption 2 holds. Let $V^*_\tau(\gamma) = \beta_2(\gamma)^2 V_x(\gamma)$ and $\Omega(\gamma)$ be the pseudo-VCATE and covariance matrix associated with γ, respectively. If $\widehat{CI}_{\alpha n}$ is a confidence interval obtained by substituting $(\hat{V}_{\tau n}, \hat{\Omega}_n)$ into (11), then*

(i) $1 - \alpha \leqslant \liminf\limits_{n \to \infty} \inf_{\gamma \in \Gamma} \mathbb{P}_\gamma \left(V^*_\tau(\gamma) \in \widehat{CI}_{\alpha nk} \right)$

(ii) If in addition, $\Omega_{12}(\gamma) V_\tau(\gamma) \leqslant t$, then $\limsup\limits_{n \to \infty} \sup_{\gamma \in \Gamma} \mathbb{P}_\gamma \left(V^*_\tau(\gamma, \mathcal{I}_{-nk}) \in \widehat{CI}_{\alpha nk} \right) \leqslant (1 - \alpha) + \tilde{\alpha}(t)$, where $\tilde{\alpha}(t) \geqslant 0$ and $\lim_{t \to 0} \tilde{\alpha}(t) = 0$.

Proof. Let $\{\gamma_n \in \Gamma : n \in \mathbb{N}\}$ denote a sequence of distributions. Our goal is to verify that the confidence interval satisfies the assumptions of Corollary 2.1(c) in Andrews et al. (2020). Define a sequence of parameters, $h_n := (nV^*_{\tau n}, V^*_{\tau n}, vec(\Omega_n), \zeta_n)$. By Assumption 2, each element is contained in $\mathcal{H}$, a subset of the extended Euclidean space in which Ω_n is positive-definite with bounded eigenvalues. The quantity $nV^*_{\tau n}$ is positive but unbounded, and can converge to $+\infty$. Assumption B in Andrews et al. (2020) is stated in terms of subsequences and the first step is to write the problem in this way. We show that for $h \in \mathcal{H}$ and all subsequences $\{n_\ell\}_{\ell=1}^\infty$ where $h_{n_\ell} \to h \in \mathcal{H}$,

$$1 - \alpha \leqslant \lim_{n_\ell \to \infty} \mathbb{P}_{\gamma_{n_\ell}} \left(V^*_{\tau n_\ell} \in \widehat{CI}_{\alpha n} \right) \leqslant (1 - \alpha) + \tilde{\alpha}(t), \qquad \tilde{\alpha}(t) \geqslant 0, \alpha \in [0, 1]$$

We break down the proof by cases. *(a) Near homogeneity case:* Suppose that $n_\ell V^*_{\tau n_\ell} \to^p v \in [0, \infty)$, $V^*_{\tau n_\ell} \to^p 0$, $\zeta_{n_\ell} \to \zeta^* \in \{0, 1\}$ and $\Omega_{n_\ell} \to \Omega$ as $n \to \infty$,

62

where Ω is positive-definite. For this case, $\sqrt{n_\ell} V^*_{\tau n_\ell} = o(1)$. By applying Lemma 2,

$$
\begin{aligned}
n_\ell(\hat{V}^*_{\tau n_\ell} - V^*_{\tau n_\ell}) &= (e_1' \Omega^{1/2} Z_{n_\ell})^2 + 2\zeta^* \sqrt{v} e_1' \Omega^{1/2} Z_{n_\ell} + o_p(1) \\
&= (e_1' \Omega^{1/2} Z_{n_\ell} + \zeta^* \sqrt{v})^2 - v + o_p(1).
\end{aligned}
$$

Let $Z \sim \mathcal{N}(0, I_{2 \times 2})$ (independent of n_ℓ). Since $\hat{\Omega}_{n_\ell} = \Omega + o_p(1)$, then the estimated empirical process at $V^*_{\tau n_\ell}$ has the same limiting distribution as the estimator. For a fixed $\zeta \in \{-1, 1\}$ (that may differ from ζ^*),

$$
n_\ell G(n, V^*_{\tau n_\ell}, \hat{\Omega}_{n_\ell}, Z, \zeta) = (e_1' \Omega^{1/2} Z + \zeta \sqrt{v})^2 - v + o_p(1).
$$

Define the limiting CDF as $H_{\Omega, v, \zeta}(\tilde{v}) := \mathbb{P}_{\gamma_{n_\ell}} \left((e_1' \Omega^{1/2} Z + \zeta \sqrt{v})^2 - v \leqslant \tilde{v} \right)$, where $Z \sim \mathcal{N}(0, I_{2 \times 2})$. Since Z has mean zero, $H_{\Omega, v, \zeta=1}(\tilde{v}) = H_{\Omega, v, \zeta=-1}(\tilde{v}) = H_{\Omega, \nu}(\tilde{v})$, which does not depend on ζ. Since Ω is positive-definite, the function $H_{\Omega, v}(\tilde{v})$ is continuous and strictly increasing. Let $\tilde{F}_{n_\ell, V^*_{\tau n_\ell}, \hat{\Omega}_{n\ell}, \zeta_{n_\ell}}(\tilde{v}) = F_{n_\ell, V^*_{\tau n_\ell}, \hat{\Omega}_{n_\ell}, \zeta_{n_\ell}}(\tilde{v}/n_\ell) \to H_{\Omega, v}(\tilde{v})$. Since $H_{\Omega, v}$ is continuous, then (Lehmann, 1999, Theorem 2.6.1) implies that this convergence is uniform in $\tilde{v}$, and since the limiting CDF is strictly increasing,

$$
F_{n_\ell, V^*_\tau, \hat{\Omega}_n, \zeta}(\hat{V}_{\tau n_\ell} - V^*_\tau) = \tilde{F}_{n_\ell, V^*_{\tau n_\ell}, \hat{\Omega}_n, \zeta}(n_\ell(\hat{V}_{\tau n_\ell} - V^*_\tau)) \to^d U_{n_\ell},
$$

for all $\zeta \in \{-1, 1\}$ where $U_{n_\ell} \sim\sim Uniform[0, 1]$. The test statistic converges to a fixed distribution regardless of the choice of ζ. Similarly, $q_{\alpha/2}(n_\ell, V^*_{\tau n_\ell}, \Omega, \zeta) \to^p \min\{\alpha/2, H_{\Omega, v}(0)\}$. Define a random variable, $\hat{R}_{n_\ell, \zeta} := F_{n_\ell, V^*_\tau, \hat{\Omega}_n, \zeta}(\hat{V}_{\tau n_\ell} - V^*_{\tau n_\ell}) - q_{\alpha/2}(n_\ell, V^*_{\tau n_\ell}, \hat{\Omega}_{n_\ell}, \zeta) \to^d U_{n_\ell} - \min\{\alpha/2, H_{\Omega, v}(0)\}$. By definition, $V^*_{\tau n_\ell} \in \widehat{CI}_{\alpha n} \iff \bigcup_{\zeta \in \{-1, 1\}} \left\{ \hat{R}_{n_\ell, \zeta} \in [0, 1 - \alpha] \right\}$. As $n_\ell \to \infty$,

$$
\begin{aligned}
\mathbb{P}_{\gamma_{n_\ell}} \left(V^*_{\tau n_\ell} \in \widehat{CI}_{\alpha n k} \right) &\geqslant \max_{\zeta \in \{-1, 1\}} \mathbb{P}_{\gamma_{n_\ell}} \left(\hat{R}_{n_\ell, \zeta} \in [0, 1 - \alpha] \right) \geqslant \mathbb{P}_{\gamma_{n_\ell}} \left(\hat{R}_{n_\ell, \zeta_{n_\ell}} \in [0, 1 - \alpha] \right) \\
&= \mathbb{P}_{\gamma_{n_\ell}} \left(0 \leqslant U_{n_\ell} - \min\{\alpha/2, H_{\Omega, v}(0)\} \leqslant 1 - \alpha \right) + o(1) \\
&= (1 - \alpha) + o(1).
\end{aligned}
$$

Since the limiting distribution doesn't depend on ζ, we can apply the continuous mapping theorem to show that $\hat{R}^{\max}_{n_\ell} := \min_{\zeta \in \{-1, 1\}} \hat{R}_{n_\ell, \zeta}$ and $\hat{R}^{\min}_{n_\ell} := \max_{\zeta \in \{-1, 1\}} \hat{R}_{n_\ell, \zeta}$

both converge to the same limit, $U_{n_\ell} - \min\{\alpha/2, H_{\Omega,v}(0)\}$. As $n_\ell \to \infty$,

$$\mathbb{P}_{\gamma_{n_\ell} k}\left(V_{\tau n_\ell}^* \in \widehat{CI}_{\alpha n k}\right) \leqslant \mathbb{P}_{\gamma_{n_\ell}}\left(0 \leqslant \hat{R}_{n_\ell}^{\max}, \hat{R}_{n_\ell}^{\min} \leqslant 1 - \alpha\right) = (1 - \alpha) + o(1). \quad (72)$$

For this class of subsequences the confidence interval produces exact coverage.

Mild heterogeneity case: Suppose that $n_\ell V_{\tau n_\ell}^* \to \infty$ and $V_{\tau n_\ell}^* \to V_\tau^*$, where $V_\tau \in [0, \infty)$, $\zeta_{n_\ell} \to \zeta^*$, and $\Omega_{n_\ell} \to \Omega$. Then we can rescale by $\sqrt{n_\ell/V_{\tau n_\ell}^*}$.

$$\sqrt{\frac{n_\ell}{V_{\tau n_\ell}^*}}(\hat{V}_{\tau n_\ell}^* - V_{\tau n_\ell}^*) = \frac{(e_1'\Omega_{n_\ell}^{1/2}Z_{n_\ell})^2}{\sqrt{n_\ell V_{\tau n_\ell}^*}} + 2\zeta_{n_\ell}e_1'(\Omega_{n_\ell}^{1/2}Z_{n_\ell}) + \sqrt{V_{\tau n_\ell}^*}(e_2'\Omega_{n_\ell}^{1/2}Z_{n_\ell}) + o_p(1)$$

$$= 2\zeta^*(e_1'\Omega^{1/2}Z_{n_\ell}) + \sqrt{V_\tau^*}(e_2'\Omega^{1/2}Z_{n_\ell}) + o_p(1).$$

The limiting distribution is normal. For convenience, we write this as $\sqrt{n_\ell/V_{\tau n_\ell}^*}(\hat{V}_{\tau n_\ell}^* - V_{\tau n_\ell}^*) = \sigma(\zeta^*)\tilde{Z}_{n_\ell} + o_p(1)$, where $\tilde{Z}_{n_\ell} \sim \mathcal{N}(0,1)$ and $\sigma(\zeta)^2 := \Omega_{11} + V_\tau^*\Omega_{22} + \zeta\sqrt{V_\tau^*}\Omega_{12}$ for $\zeta \in \{1, -1\}$. Since the norm of $[1, \zeta\sqrt{V_\tau^*}]'$ is larger than one, it follows that $\sigma(\zeta)^2 \geqslant \lambda_{\min}$, where $\lambda_{\min}$ is the smallest eigenvalue of Ω. In this case, the limiting CDF is $H_{\Omega,V_\tau^*,\zeta}(\tilde{v}) := \Phi(\tilde{v}/\sigma(\zeta))$ where $\Phi(\cdot)$ is the CDF of a standard normal. Let $z_\alpha = \Phi^{-1}(\alpha)$ denote the $\alpha-$quantile, and $\phi(\cdot)$ the marginal of a standard normal.

$$\mathbb{P}_{\gamma_{n_\ell}}(\hat{R}_{n_\ell,\zeta} \in [0, 1-\alpha]) = \mathbb{P}\left(-\alpha/2 \leqslant \Phi\left(\frac{\sigma(\zeta^*)}{\sigma(\zeta)}\tilde{Z}_{n_\ell}\right) \leqslant 1 - \alpha/2\right) + o(1)$$

$$= \underbrace{\Phi\left(\frac{\sigma(\zeta^*)}{\sigma(\zeta)}z_{1-\alpha/2}\right) - \Phi\left(\frac{\sigma(\zeta^*)}{\sigma(\zeta)}z_{-\alpha/2}\right)}_{\kappa(\sigma(\zeta),\sigma(\zeta^*))} + o(1). \quad (73)$$

To obtain the lower bound, $\mathbb{P}_{\gamma_{n_\ell}}\left(V_{\tau n_\ell}^* \in \widehat{CI}_{\alpha n k}\right) \geqslant \mathbb{P}_{\gamma_{n_\ell}}\left(\hat{R}_{n_\ell,\zeta^*} \in [0, 1-\alpha]\right) = (1 - \alpha) + o(1)$. To obtain the upper bound, I write down a Taylor expansion. There is a $\tilde{\sigma} \geqslant \sqrt{\lambda_{\min}}$ between $\sigma(\zeta^*)$ and $\sigma(\zeta)$ such that

$$\|\kappa(\sigma_\zeta, \sigma_{\zeta^*}) - (1 - \alpha)\| \leqslant \left\|\left[\phi\left(\frac{\tilde{\sigma}z_{1-\alpha/2}}{\sigma(\zeta)}\right)z_{1-\alpha/2} - \phi\left(\frac{\tilde{\sigma}z_{-\alpha/2}}{\sigma(\zeta)}\right)z_{-\alpha/2}\right]\frac{[\sigma(\zeta^*) - \sigma(\zeta)]}{\sigma(\zeta)}\right\|$$

$$\leqslant \frac{|z_{-\alpha/2}|}{\sqrt{2\pi\lambda_{min}}}\|\sigma(\zeta^*) - \sigma(\zeta)\|.$$

By another Taylor expansion, $\|\sigma(\zeta^*) - \sigma(\zeta)\| \leqslant \frac{1}{2\sqrt{\lambda_{\min}}}\|\sigma(\zeta^*)^2 - \sigma(\zeta)^2\|$. If in addition, $\|\sqrt{V_\tau^*}\Omega_{12}\| \leqslant t$, then $\|\sigma(\zeta^*) - \sigma(\zeta)\| \leqslant t$ and we can define a non-negative function

$\tilde{\alpha}(t) = t\|z_{-\alpha/2}\|/(2\lambda_{\min}\sqrt{2\pi})$, which satisfies $\lim_{t\to 0}\tilde{\alpha}(t) = 0$. Therefore we have bounded the coverage over an exhaustive class of subsequences of distributions. This satisfies Assumption B in Andrews et al. (2020). By their Corollary 2.1,

$$1-\alpha \leqslant \liminf_{n\to\infty} \inf_{\gamma\in\Gamma} \mathbb{P}_\gamma \left(V_\tau^*(\gamma) \in \widehat{CI}_{\alpha n} \right) = \limsup_{n\to\infty} \sup_{\gamma\in\Gamma} \mathbb{P}_\gamma \left(V_\tau^*(\gamma) \in \widehat{CI}_{\alpha n} \right) \leqslant 1-\alpha+\tilde{\alpha}(t).$$

$\square$

Lemma 11. *Let* $(S_{-k}(X_i), M_{-k}(X_i), W_i)$ *and* $\lambda(X_i)$ *be the set of regressors and weights, respectively, that were defined in* (20). *Define* $Q_{ww,nk} := \mathbb{E}[\lambda(X_i)W_iW_i' \mid \mathcal{I}_{-nk}]$ *and let* Π_{nk} *be a* 4×4 *diagonal matrix with entries* $(1,1,1,V_{xnk}^{-1/2})$. *If* (X_i, D_i) *are independent of the data in* $\mathcal{I}_{-nk}$, *then* $\Pi_{nk}'Q_{ww,nk}\Pi_{nk}$ *has the form in* (55).

Proof of Lemma 11. Let $V_i = (D_i - p(X_i))$. By definition, $\lambda(X_i)W_iW_i'$ is

$$\lambda(X_i)\begin{pmatrix} 1 & M_{-k}(X_i) & V_i & V_iS_{-k}(X_i) \\ M_{-k}(X_i) & M_{-k}(X_i)^2 & M_{-k}(X_i)V_i & M_{-k}(X_i)V_iS_{-k}(X_i) \\ V_i & M_{-k}(X_i)V_i & V_i^2 & V_i^2S_{-k}(X_i) \\ V_iS_{-k}(X_i) & M_{-k}(X_i)V_iS_{-k}(X_i) & V_i^2S_{-k}(X_i) & V_i^2S_{-k}(X_i)^2 \end{pmatrix}. \quad (74)$$

Since (X_i, D_i) are independent of the data in $\mathcal{I}_{-nk}$, then $\mathbb{E}[V_i \mid X_i = x, \mathcal{I}_{-nk}]$ does not depend on $\mathcal{I}_{-nk}$ and is equal to $\mathbb{E}[V_i \mid X_i = x] = \mathbb{E}[D_i \mid X = x] - p(x) = 0$. By a similar reasoning, $\mathbb{E}[V_i^2 \mid X_i = x, \mathcal{I}_{-nk}] = p(X_i)(1 - p(X_i)) = \lambda(X_i)^{-1}$. Using both conditional moment results, we can show that $\mathbb{E}[\lambda(X_i)W_iW_i' \mid X_i, \mathcal{I}_{-nk}]$ is equal to

$$Q_{ww,nk} = \mathbb{E}\left[\begin{array}{cccc|c} \lambda(X_i) & \lambda(X_i)M_{-k}(X_i) & 0 & 0 & \\ \lambda(X_i)M_{-k}(X_i) & \lambda(X_i)M_{-k}(X_i)^2 & 0 & 0 & \\ 0 & 0 & 1 & S_{-k}(X_i) & \mathcal{I}_{-nk} \\ 0 & 0 & S_{-k}(X_i) & S_{-k}(X_i)^2 & \end{array}\right].$$

We substitute $\mathbb{E}[S_{-k}(X_i) \mid \mathcal{I}_{-nk}] = 0$ and $\mathbb{E}[S_{-k}(X_i)^2 \mid \mathcal{I}_{-nk}] = V_{xnk}$. Finally, $\Pi_{nk}'Q_{ww,nk}\Pi_{nk}$ only normalizes the lower right corner to 1.

$\square$

Lemma 12 (Convergence of cross-fitted sums). *Consider a sequence of distributions* $\{\gamma_n\}_{n=1}^\infty$ *over a collection of random matrices* $\{Z_{i1}, \ldots, Z_{iL}\}_{i\in\mathcal{I}_{nk}}$ *where L is a finite constant,* $Z_{i\ell} \in \mathbb{R}^M \times \mathbb{R}^B$. *Define* $\hat{\zeta}_{nk} := \sum_{i\in\mathcal{I}_{-nk}} \sum_{\ell=1}^L Z_{i\ell}$. *If for all* $\ell \in \{1, \ldots, L\}$, *(i)*

the observations are i.i.d. conditional on $\mathcal{I}_{-nk}$, (ii) $\mathbb{E}_{\gamma_n}[Z_{i\ell} \mid \mathcal{I}_{-nk}] = \underline{0}_{M \times B}$, and (iii) $\mathbb{E}_{\gamma_n}[\|Z_{i\ell}\|^2]$ has a uniform upper bound and γ_n, then as $n_k \to \infty$, (a) $n_k^{-1/2}\widehat{\zeta}_{nk} = O_p(1)$, (b) If in addition, $\mathbb{E}_{\gamma_n}[\|Z_{i\ell}\|^2] = o(1)$, then $n_k^{-1/2}\widehat{\zeta}_{nk} = o_p(1)$, (c) Now suppose (ii) and (iii) do not necessarily hold, but instead $n_k^r \mathbb{E}_{\gamma_n}[\|Z_{i\ell}\|] = o(1)$ for all ℓ and some $r > 0$, then $n_k^{r-1}\widehat{\zeta}_{nk} = o_p(1)$.

Proof. Since M, L, B are all finite, it suffices to consider $\widehat{\zeta}_{nk\ell mb} := \sum_{i \in \mathcal{I}_{-nk}} Z_{i\ell mb}$, where $Z_{i\ell mb}$ is the coordinate (m, b) of $Z_{i\ell}$. By the law of iterated expectations, $R_{nk\ell mb} := \mathbb{E}_{\gamma_n}[(\widehat{\zeta}_{nk\ell mb} - \mathbb{E}_{\gamma_n}[\widehat{\zeta}_{nk\ell mb} \mid \mathcal{I}_{-nk}])^2]$ is equal to $\mathbb{E}_{\gamma_n}[\mathbb{E}_{\gamma_n}[(\widehat{\zeta}_{nk\ell mb} - \mathbb{E}_{\gamma_n}[\widehat{\zeta}_{nk\ell mb} \mid \mathcal{I}_{-nk}])^2 \mid \mathcal{I}_{-nk}]]$. Substituting the definition of conditional variance,

$$R_{nk\ell mb} = \mathbb{E}_{\gamma_n}\left[\mathbb{V}_{\gamma_n}[\widehat{\zeta}_{nk\ell mb} \mid \mathcal{I}_{-nk}] + \mathbb{E}_{\gamma_n}[\widehat{\zeta}_{nk\ell mb} \mid \mathcal{I}_{-nk}]^2\right]. \tag{75}$$

The first term of (75) is $O(n_k^{-1})$. By Assumption (i), $\widehat{\zeta}_{nk\ell mb}$ is a sum of n_k variables that are i.i.d. conditional on $\mathcal{I}_{-nk}$, and hence $\mathbb{V}_{\gamma_n}[\widehat{\zeta}_{nk\ell mb} \mid \mathcal{I}_{-nk}] = n_k\mathbb{V}_{\gamma_n}(Z_{i\ell mb} \mid \mathcal{I}_{nk}) = n_k\mathbb{E}_{\gamma_n}[Z_{i\ell mb}^2 \mid \mathcal{I}_{nk}]$. By the law of iterated expectations, $\mathbb{E}_{\gamma_n}\left[\mathbb{V}_{\gamma_n}[\widehat{\zeta}_{nk\ell mb} \mid \mathcal{I}_{-nk}]\right] = n_k\mathbb{E}_{\gamma_n}[Z_{i\ell mb}^2]$. The second term of (75) is zero by Assumption (ii).

To prove (a), we apply Chebyshev's inequality $\mathbb{P}(n_k^{-1}\widehat{\zeta}_{nk\ell mb} > t) \leqslant \mathbb{E}_{\gamma_n}[Z_{i\ell mb}^2]/t^2$ for some $t > 0$. This shows that $n_k^{-1/2}\widehat{\zeta}_{nk\ell mb} = O_p(1)$. To prove part (b), we use the condition that $\mathbb{E}_{\gamma_n}[Z_{i\ell mb}^2] = o(1)$ to show that $n_k^{-1/2}\widehat{\zeta}_{nk\ell mb} = o_p(1)$. To prove (c), we apply the triangle inequality, $\|n_k^{r-1}\widehat{\zeta}_{nk\ell mb}\| \leqslant n_k^{r-1}\sum_{i \in \mathcal{I}_{-nk}} \|Z_{i\ell mb}\|$. By Markov's inequality, $\mathbb{P}(\|n_k^{r-1}\widehat{\zeta}_{nk\ell mb}\| > t) \leqslant n_k^r\mathbb{E}[\|Z_{i\ell mb}\|]/t = o(1)$, hence $n_k^{r-1}\widehat{\zeta}_{nk\ell mb} = o_p(1)$. $\square$

Lemma 13 (Derivatives of Influence Function). *Let $\varphi_i(\eta) := \varphi(Y_i, D_i, X_i, \eta)$, for $i \in \mathcal{I}_{nk}$, and $U_i = Y_i - \mathbb{E}[Y_i \mid D_i, X_i]$. Suppose that $\{Y_i, D_i, X_i\}_{i \in \mathcal{I}_{nk}}$ is independent of $\{Y_i, D_i, X_i\}_{i \in \mathcal{I}_{-nk}}$ for all $k \in \{1, \ldots, K\}$, and consider a set of $\eta \in \mathcal{T}$ where the propensity score is bounded in $[\delta, 1-\delta]$ for $\delta \in (0, 1/2]$. Then (a) $\mathbb{E}\left[\frac{\partial \varphi_i(\eta)}{\partial(\eta_1, \eta_2, \eta_4)'} \mid X = x, \mathcal{I}_{-nk}\right] = [0, 0, -2(\tau(x) - \tau_{av})]$ almost surely, (b) $\mathbb{E}\left[\|\partial\varphi_i(\eta)/\partial\eta_\ell\|^4\right]^{1/4} \leqslant (8/\delta)(\mathbb{E}[\|\eta(X_i)\|^4]^{1/4} + \mathbb{E}\left[\|U_i\|^4\right]^{1/4})$, and (c) $\left\|\frac{\partial^2\varphi_i(\tilde{\eta})}{\partial(\eta_1, \eta_2, \eta_4)\partial(\eta_1, \eta_2, \eta_4)'}\right\| \leqslant 18 \times \sqrt{3}/\delta$ almost surely, for $\tilde{\eta} \in \mathcal{T}$.*

Proof. Part (a): The jacobian of $\varphi_i(\eta)$, $\frac{\partial \varphi_i(\eta)}{\partial(\eta_1, \eta_2, \eta_4)}$ is

$$
\begin{pmatrix}
2(\tau(X_i) - \tau_{av}) \left[1 - \frac{D_i}{p(X_i)} \right] + 2 \left[\frac{D_i(Y_i - \mu_0(X_i) - D_i \tau(X_i))}{p(X_i)} - \frac{(1 - D_i)(Y_i - \mu_0(X_i))}{1 - p(X_i)} \right] \\
2(\tau(X_i) - \tau_{av}) \left[-\frac{D_i}{p(X_i)} + \frac{1 - D_i}{1 - p(X_i)} \right] \\
-2(\tau(X_i) - \tau_{av}) - 2 \left[\frac{D_i(Y_i - \mu_0(X_i) - D_i \tau(X_i))}{p(X_i)} - \frac{(1 - D_i)(Y_i - \mu_0(X_i))}{1 - p(X_i)} \right]
\end{pmatrix}
$$

Part (a): Substituting $\mathbb{E}[D_i \mid X_i = x, \mathcal{I}_{-nl}] = p(x)$ and $\mathbb{E}[D_i Y_i \mid X = x, \mathcal{I}_{-nk}] = p(x)(\mu_0(x) + \tau(x))$, then $\mathbb{E}\left[\partial \varphi_i(\eta)/\partial(\eta_1, \eta_2, \eta_4)' \mid X = x, \mathcal{I}_{-nk}\right] = [0, 0, -2(\tau(x) - \tau)]'$.

Part (b): By construction, $U_i = Y_i - \mu_0(X_i) - D_i \tau(X_i)$. Then the jacobian simplifies to $\partial \varphi_i(\eta)/\partial(\eta_1, \eta_2, \eta_4)' = G_{1i} + G_{2i}$, where $G_{1i} := 2(\tau(X_i) - \tau_{av}) \times [(1 - D_i/p(X_i)), (-D_i/p(X_i) + (1 - D)/(1 - p(X_i))), -1]$ and $G_{2i} := 2(D_i/p(X_i) - (1 - D_i)/(1 - p(X_i))) \times [U_i, 0, -U_i]$. Since $p(x) \in [\delta, 1 - \delta]$, $\|G_{1i}\| \leqslant (4/\delta)\|\tau(X_i) - \tau_{av}\| \leqslant (8/\delta)\|\eta(x)\|$ and $\|G_{2i}\| \leqslant (8/\delta)\|U_i\|$. Therefore, by the triangle inequality, $\mathbb{E}\left[\|\partial \varphi_i(\eta)/\partial \eta_\ell\|^4\right]^{1/4} \leqslant (8/\delta)(\mathbb{E}[\|\eta(X_i)\|^4]^{1/4} + \mathbb{E}\left[\|U_i\|^4\right])$.

Part (c): The hessian of $\varphi_i(\tilde{\eta})$, which I denote by $H(\tilde{\eta})$, is symmetric and

$$
\frac{\partial^2 \varphi_i(\tilde{\eta})}{\partial(\eta_1, \eta_2, \eta_4) \partial(\eta_1, \eta_2, \eta_4)'} = \begin{pmatrix}
2 \left[1 - \frac{D_i}{\tilde{p}(X_i)} \right] - 2\frac{D_i}{\tilde{p}(X_i)} & . & . \\
2 \left[-\frac{D_i}{\tilde{p}(X_i)} + \frac{1 - D_i}{1 - \tilde{p}(X_i)} \right] & 0 & . \\
2 \left[1 - \frac{D_i}{\tilde{p}(X_i)} \right] & -2 \left[\frac{D_i}{\tilde{p}(X_i)} - \frac{(1 - D_i)}{1 - \tilde{p}(X_i)} \right] & -2
\end{pmatrix}
$$

Since $\tilde{p}(x) \in [\delta, 1 - \delta]$, then $\|D_i/\tilde{p}(X_i)\| \leqslant 1/\delta$, $\|1 - D_i/\tilde{p}(X_i)\| \leqslant (1 + 1/\delta) \leqslant 2/\delta$, and $\|D_i/\tilde{p}(X_i) - (1 - D_i)/(1 - \tilde{p}(X_i))\| \leqslant 2/\delta$. This means that each of the entries of $H(\tilde{\eta})$ is bounded by $6/\delta$. By Lemma 16, $\|H(\tilde{\eta})\| \leqslant 3 \times (6/\delta) \times \sqrt{3} = 18 \times \sqrt{3}/\delta$.

$\square$

Lemma 14 (Decomposition of Nuisance Functions). *Define* $\hat{\eta}_{-k,\hat{\theta}_{nk}}(x)$ *as in* (63), $\theta_n := (0, 1, \tau_{n,av}, 1)$, $(\hat{\lambda}_{nk} - \lambda_{nk}) := [(\hat{\theta}_{nk} - \theta_n), \hat{\tau}_{nk,av}(\hat{\theta}_{nk} - \theta_n), (\hat{\tau}_{nk,av} - \tau_{n,av})]'$, *and let* $\{e_j\}_{j=1}^4$ *be* 4×1 *vectors with 1 in the* j^{th} *coordinate and zero otherwise. Then there exist* $(x, \mathcal{I}_{-nk})$−*measurable matrices* $\Psi_{-nk}(x)$, $\Delta_{-nk}(x)$, *such that*

$$
\hat{\eta}_{-k,\hat{\theta}_{nk}}(x) - \eta(x) = \Psi_{-nk}(x)(\hat{\lambda}_{nk} - \lambda_{nk}) + \Delta_{-nk}(x), \tag{76}
$$

and for some constant $C < \infty$, *(a)* $e_3'[\hat{\eta}_{-k,\hat{\theta}_{nk}}(x) - \eta(x)] = 0$, *(b)* $e_3' \Delta_{-nk}(x) = e_4' \Delta_{-nk}(x) = 0$, *(c)* $e_4' \Psi_{-nk} = c$, *for* $c \in \mathbb{R}^9$, *(d)* $\|\Psi_{-nk}(x)\| \leqslant C \times [1 + 2\|\hat{\eta}_{-k}(x)\|]$

a.s., (e) $\|\Delta_{-nk}(x)\| \leqslant C \times \|\widehat{\eta}_{-k}(x) - \eta(x)\|$ a.s.

Proof of Lemma 14. Define $\widehat{B}(x) := [e_1(\widehat{W}(x,1) - \widehat{W}(x,0))' + e_2\widehat{W}(x,0)']$.

$$\widehat{\eta}_{-k,\widehat{\theta}_{nk}}(x)' = \widehat{B}(x) + e_3p_n(x) + e_4\widehat{\tau}_{nk,av},$$
$$\widehat{\eta}_{-k,\theta_n}(x) = \widehat{B}(x)\theta_n + e_3p_n(x) + e_4\widehat{\tau}_{nk,av},$$
$$\eta(x) = e_1\tau_n(x) + e_2\mu_{0n}(x) + e_3p_n(x) + e_4\tau_{n,av}.$$

The estimation error can be decomposed as

$$\widehat{\eta}_{-k,\widehat{\theta}_{nk}}(x) - \eta(x) = \widehat{B}(x)(\widehat{\theta}_{nk} - \theta_n) + e_4[\widehat{\tau}_{nk,av} - \tau_{n,av}] + [\widehat{B}(x)'\theta_n - e_1\tau_n(x) + e_2\mu_{0n}(x)]. \tag{77}$$

By definition $\widehat{B}(x) = \Psi_{1,-nk}(x) + \widehat{\tau}_{nk,av}\Psi_{2,-nk}(x)$, with auxiliary matrices $\Psi_{1,-nk} := e_1[0, 0, 1, \widehat{\tau}_{-k}(x)] + e_2[1, M_{-k}(X), -p_n(x), -p_n(x)\widehat{\tau}_{-k}(x)]$ and $\Psi_{2,-nk} := e_1[0, 0, 0, -1] + e_2[0, 0, 0, p_n(x)]$. Substituting the parameter $\theta_n := (0, 1, \tau_{n,av}, 1)$ and grouping common terms, $\widehat{B}(x)'\theta_n - e_1\tau_n(x) - e_2\mu_{0n}(x) = e_1[(\widehat{\tau}_{-k}(x) - \tau_n(x)) + (\tau_{n,av} - \widehat{\tau}_{nk,av})] + e_2[M_{-k}(x) - \mu_{0n}(x) - p_n(x)\widehat{\tau}_{-k}(x) + p_n(x)(\widehat{\tau}_{nk,av} - \tau_{n,av})]$. We can simplify the second term of this expression by substituting $M_{-k}(x) = \widehat{\mu}_{0,-k}(x) + p_n(x)\widehat{\tau}_{-k}(x)$, which produces $e_2[(\widehat{\mu}_{0n}(x) - \mu_{0n}(x)) + p_n(x)(\widehat{\tau}_{nk,av} - \tau_{n,av})]$. Consequently, the second and third terms of (77) can be written as $\Psi_{3,-nk}(x)(\widehat{\tau}_{nk,av} - \tau_{n,av}) + \Delta_{-nk}(x)$, where $\Psi_{3,-nk}(x) := -e_1 + e_2p_n(x) + e_4$ and $\Delta_{-nk}(x) := e_1(\tau_{-k}(x) - \tau_n(x)) + e_2(\widehat{\mu}_{-k}(x) - \mu_{0n}(x))$.

Define $\Psi_{-nk}(x) := [\Psi_{1,-nl}(x), \Psi_{2,-nk}(x), \Psi_{3,-nk}(x)]$ and the parameter error as $(\widehat{\lambda}_{nk} - \lambda_{nk}) := [(\widehat{\theta}_{nk} - \theta_n)', \widehat{\tau}_{nk,av}(\widehat{\theta}_{nk} - \theta_n)', (\widehat{\tau}_{nk,av} - \tau_{n,av})]'$. Combining the results,

$$\widehat{\eta}_{-k,\widehat{\theta}_{nk}}(x) - \eta(x) = \Psi_{-nk}(x)(\widehat{\lambda}_{nk} - \lambda_{nk}) + \Delta_{-nk}(x).$$

Measurability with respect to $(x, \mathcal{I}_{-nk})$ can be verified by inspection. Property (a) follows from the fact that the propensity score is known, and (b) because $\Psi_{-nk}(x)$ and $\Delta_{-nk}(x)$ depend on vectors e_1, e_2, which are orthogonal to e_3, e_4. Property (c) follows by the fact that $e_4'\Psi_{-nk} = [\underline{0}_{1\times 8}, 1]$. To prove part (d), we apply Lemma 16 to show that $\Psi_{-nk}(x)$ is bounded by $9\sqrt{4}$ times the largest absolute value of the matrix. Since $p_n(x) \leqslant 1$, then the largest value is bounded by $1 + \|\widehat{\mu}_{0,-k}(x)\| + \|\widehat{\tau}_{-k}(x)\|$ which is less than $1 + 2\|\widehat{\eta}_{-k}(x)\|$. This means that $\|\Psi_{-nk}(x)\| \leqslant 12 \times \sqrt{4} \times [1 + 2\|\widehat{\eta}_{-k}(x)\|]$. To prove, part (e) we once again apply Lemma 16. The quantity $\Delta_{-nk}(x)$ is a 4×1 vector, whose individual entries are bounded by $\|\widehat{\mu}_{0,-k} - \mu_{0n}(x)\| + \|\widehat{\tau}_{-k}(x) - \tau_n(x)\|$,

which is weakly less than $2\|\widehat{\eta}_{-k}(x) - \eta(x)\|$.

$\square$

Lemma 15 (Convergence of regression parameters). *Consider a sequence of distributions $\{\gamma_n\}_{n=1}^{\infty}$ that satisfy Assumptions 1, 3, 4, 5, 6, and 7, and that $V_{\tau n} \to V_\tau > 0$. Define oracle regressors, $W_i^* := [1, \mu_{0n}(X_i) + p_n(X_i)\tau_n(X_i), (D_i - p_n(x_i)), (D_i - p_n(X_i))(\tau_n(X_i) - \tau_{av,n})]'$ and $\theta_n := \mathbb{E}_{\gamma_n}[\lambda_n(X_i)W_i^*W_i^{*'}]^{-1}\mathbb{E}_{\gamma_n}[\lambda_n(X_i)W_i^*Y_i]$. Then (a) $\widehat{\tau}_{nk,av} - \tau_{n,av} = o_p(n_k^{-1/4})$, (b) $\theta_n = (0, 1, \tau_{n,av}, 1)'$, and (c) $\widehat{\theta}_{nk} - \theta_n = o_p(n_k^{-1/4})$.*

Proof. Part (a): Decompose, $n_k^{1/4}(\widehat{\tau}_{nk,av} - \tau_{av,n}) = n_k^{-1/4}\left[\frac{1}{\sqrt{n_k}}\sum_{i\in\mathcal{I}_{nk}}\tau_n(X_i) - \tau_{n,av}\right] + n_k^{1/4-1}\sum_{i\in\mathcal{I}_{nk}}[\widehat{\tau}_{-k}(X_i) - \tau_n(X_i)]$. The first term is a centered random variable that is $o_p(1)$. By Assumption 7.(iv), and the Cauchy-Schwartz inequality, $\mathbb{E}_{\gamma_n}[n_k^{1/4}\|\widehat{\tau}_{-k}(X_i) - \tau_n(X_i)\|] \leqslant \mathbb{E}_{\gamma_n}[n_k^{1/2}\|\widehat{\eta}_{-k}(X_i) - \eta(X_i)\|^2]^{1/2} \to o(1)$. By applying Lemma 12.(c), $n_k^{1/4}(\widehat{\tau}_{nk,av} - \tau_{av,n}) = o_p(1)$.

Part (b): For given $\{X_i = x, D_i = d\}$, $W_i^{*'}(0, 1, \tau_{n_a v}, 1)' = \mu_d(x)$. Therefore, by applying Lemma 5, $\theta_n = (0, 1, \tau_{n,av}, 1)'$.

Parts (c): Define $\widehat{Q}_{ww} := \frac{1}{n_k}\sum_{i\in\mathcal{I}_{nk}}\lambda_n(X_i)\widehat{W}_i\widehat{W}_i'$, $Q_{ww} := \mathbb{E}_{\gamma_n}[\lambda_n(X_i)W_i^*W_i^{*'}]$, $M_n(x) = \mu_{0n}(x) + p_n(x)\tau_n(x)$. Following similar derivations to Lemma 11,

$$Q_{ww} = \begin{pmatrix} \mathbb{E}_{\gamma_n}[\lambda(X_i)] & \mathbb{E}_{\gamma_n}[\lambda(X_i)(\mu_n(X_i))] & 0 & 0 \\ \mathbb{E}_{\gamma_n}[\lambda(X_i)(\mu_n(X_i))] & \mathbb{E}_{\gamma_n}[\lambda(X_i)(\mu_n(X_i))^2] & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & V_{\tau n} \end{pmatrix}. \tag{78}$$

The upper left block has bounded eigenvalues by Assumption 3.(i) and $V_{\tau n}$ is asymptotically bounded. Therefore Q_{ww} is positive definite with bounded eigenvalues. Furthermore, $\widehat{Q}_{ww} - Q_{ww}$ can be decomposed as:

$$\left[\frac{1}{n_k}\sum_{i\in\mathcal{I}_{nk}}\lambda_n(X_i)W_i^*W_i^{*'} - \mathbb{E}_{\gamma_n}[\lambda_n(X_i)W_i^*W_i^{*'}]\right] + \left[\frac{1}{n_k}\sum_{i\in\mathcal{I}_{nk}}\lambda_n(X_i)(\widehat{W}_i\widehat{W}_i' - \widehat{W}_i^*\widehat{W}_i^{*'})\right]. \tag{79}$$

The first term of (79) is an average of mean-zero random variables and bounded variance, then it is $O_p(n_k^{-1/2}) = o_p(n_k^{-1/4})$. To prove that it has bounded variance, apply Lemma 16, then $\|W_i^*\| \leqslant \sqrt{4}(1 + \|\mu_{0n}(X_i)\| + \|\tau_n(X_i)\|) \leqslant \sqrt{4}(1 + 2\|\eta(X_i)\|)$. Since $p_n(x) \in [\delta, 1 - \delta]$ and $\lambda_n(X_i) = [p_n(X_i)(1 - p_n(X_i))]^{-1}$, then $\mathbb{E}_{\gamma_n}[\|\lambda_n(X_i)W_i^*\|^2]^{1/2} \leqslant (1/\delta^2)\mathbb{E}_{\gamma_n}[\|W_i^*\|^4]^{1/4} \leqslant (\sqrt{4}/\delta^2)(1 + 2\mathbb{E}_{\gamma_n}[\|\eta(X_i)\|^4]^{1/4})$, which is bounded by Assump-

tion 7.(ii).

To bound the second term of (79), we apply the triangle inequality, $\|\lambda_n(X_i)(\widehat{W}_i\widehat{W}'_i - W_i^*W_i^{*'})\| \leqslant (1/\delta^2)\widehat{\phi}_i$, where $\widehat{\phi}_i := 2\|W_i^{*'}\|\,\|\widehat{\zeta}_i\| + \|\widehat{\zeta}_i\|^2$ and $\widehat{\zeta}_i := \widehat{W}_i - W_i^*$. Our goal is to show that $\frac{1}{n_k}\sum_{i\in\mathcal{I}_{-nk}}\widehat{\phi}_i = o_p(n_k^{-1/4})$. Let e_ℓ be a 4×1 vector with one in the ℓth entry and zero otherwise. We can further decompose $\widehat{\zeta}_i = \widehat{\Delta}_{nk}\widehat{\zeta}_{-nk,1}(X_i) + \widehat{\zeta}_{-nk,2}(X_i)$, into components that map into our assumptions: $\widehat{\Delta}_{nk} := (\widehat{\tau}_{nk,av} - \tau_{n,av})$, $\widehat{\zeta}_{-nk,1}(X_i) := e_4(D_i - p_n(X_i))$ and $\widehat{\zeta}_{-nk,2}(X_i) := e_1(M_{-k}(X_i) - M_n(X_i)) + e_2(D_i - p_n(X_i))(\widehat{\tau}_{-k}(X_i) - \tau_n(X_i))$. By construction, $\|\zeta_{-nk,1}(X_i)\| \leqslant 1$. Applying the triangle inequality and grouping terms, $\widehat{\phi}_i \leqslant \|\widehat{\Delta}_{nk}\|^2 + \|\widehat{\Delta}_{nk}\|\widehat{\phi}_{i1} + \widehat{\phi}_{i2}$, where $\widehat{\phi}_{i1} := (2\|W_i^*\| + 2\|\widehat{\zeta}_{-nk,2}\|)$, and $\widehat{\phi}_{i2} := (2\|W_i^*\|\,\|\widehat{\zeta}_{-nk,2}(X_i)\| + \|\widehat{\zeta}_{-nk,2}(X_i)\|^2)$. By Assumption 7.(iv), $\mathbb{E}_{\gamma_n}[\|\widehat{\zeta}_{-nk,2}(X_i)\|^2]^{1/2} \leqslant 3\mathbb{E}_{\gamma_n}[\|\widehat{\eta}_{-k}(X_i) - \eta(X_i)\|^2]^{1/2} = o(n_k^{-1/4})$. Therefore, by the Cauchy-Schwarz inequality, $\mathbb{E}_{\gamma_n}[\widehat{\phi}_{i\ell}^2]^{1/2} = o(n_k^{-1/4})$ for $\ell \in \{1,2\}$. Then by Lemma 12.(c), $\left[\frac{1}{n_k}\sum_{i\in\mathcal{I}_{-nk}}\widehat{\phi}_{i\ell}\right] = o_p(n_k^{-1/4})$ for $\ell \in \{1,2\}$. Combining terms,

$$\left\|\frac{n_k^{1/4}}{n_k}\sum_{i\in\mathcal{I}_{-nk}}\lambda_n(X_i)(\widehat{W}_i\widehat{W}'_i - \widehat{W}_i^*\widehat{W}_i^{*'})\right\| \leqslant \frac{1}{\delta^2}\sum_{\ell=0}^{2}n_k^{1/4}\|\widehat{\Delta}_{nk}\|^{\ell-2}\left[\frac{1}{n_k}\sum_{i\in\mathcal{I}_{-nk}}\widehat{\phi}_{i\ell}\right] = o_p(1).$$

Define $\widehat{Q}_{wy} := \frac{1}{n_k}\sum_{i\in\mathcal{I}_{nk}}\lambda_n(X_i)\widehat{W}_iY_i$ and $Q_{wy} := \mathbb{E}_{\gamma_n}[\lambda_n(X_i)W_i^*Y_i]$. We can apply similar arguments as above to show that $\widehat{Q}_{wy} - Q_{wy} = o_p(n_k^{-1/4})$.

Substituting the definition, $\theta_n := Q_{ww}^{-1}Q_{wy}$ and rearranging terms, $n_k^{1/4}(\widehat{\theta}_{nk} - \theta_n) = n_k^{1/4}(\widehat{Q}_{ww}^{-1}\widehat{Q}_{wy} - \theta_{nk}) = \widehat{Q}_{ww}^{-1}n_k^{1/4}(\widehat{Q}_{wy} - Q_{wy}) - n_k^{1/4}(Q_{ww}^{-1}Q_{wy} - \widehat{Q}_{ww}^{-1}Q_{wy})$. The first term is $o_p(n_k^{-1/4})$. The second term can be rewritten as $n_k^{1/4}(Q_{ww}^{-1} - \widehat{Q}_{ww}^{-1})Q_{wy} = o_p(n_k^{-1/4})$. To prove this, note that $\|Q_{ww}^{-1} - \widehat{Q}_{ww}^{-1}\| = \|Q_{ww}^{-1}(\widehat{Q}_{ww} - Q_{ww})\widehat{Q}_{ww}^{-1}\| \leqslant \|Q_{ww}^{-1}\|\,\|\widehat{Q}_{ww} - Q_{ww}\|\,\|\widehat{Q}_{ww}^{-1}\|$. The right-hand side is $o_p(n_k^{-1/4})$ since Q_{ww} has eigenvalues bounded away from zero and $\widehat{Q}_{ww}$ converges to its true value at rate $n_k^{1/4}$. Combining the results produces $n_k^{1/4}(\widehat{\theta}_{nk} - \theta_n) = o_p(n_k^{-1/4})$. $\qquad\square$

Lemma 16 (Bound on Operator Norm). *Let H be an $M \times L$ matrix and let $\|H\| = \sup_{\{z\in\mathbb{R}^L:\|z\|=1\}}\|Hz\|$ be corresponding matrix operator norm. The absolute value of the individual entries of H is bounded a constant C. Then $\|H\| \leqslant LC\sqrt{M}$.*

Proof of Lemma 16. Let H_m be the mth row and $H_{m\ell}$ be the (m,ℓ) entry. Then

$$\|H\| = \sup_{\{z\in\mathbb{R}^L:\|z\|=1\}}\sqrt{\sum_{m=1}^{M}[H_mz]^2} = \sup_{\{z\in\mathbb{R}^L:\|z\|=1\}}\sqrt{\sum_{m=1}^{M}\left[\sum_{\ell=1}^{L}H_{m\ell}z_\ell\right]^2}.$$ Since

$|z_\ell| \leqslant 1$ and $\|H_{m\ell}\| \leqslant C$, $\|H\| = \sup_{\{z\in\mathbb{R}^L:\|z\|=1\}}\sqrt{\sum_{m=1}^{M}\left[\sum_{\ell=1}^{L}|H_{m\ell}|\,|z_\ell|\right]^2} \leqslant CL\sqrt{M}.$
$\qquad\square$