Robustness, Heterogeneous Treatment Effects and Covariate Shifts

Pietro Emilio Spini[1]

Latest Version
This draft: April 2022
First draft: May 2021

Abstract

This paper studies the robustness of estimated policy effects to changes in the distribution of covariates. Robustness to covariate shifts is important, for example, when evaluating the external validity of (quasi)-experimental results, which are often used as a benchmark for evidence-based policy-making. I propose a novel scalar robustness metric. This metric measures the magnitude of the smallest covariate shift needed to invalidate a claim on the policy effect (for example, $ATE \geq 0$) supported by the (quasi)-experimental evidence. My metric links the heterogeneity of policy effects and robustness in a flexible, nonparametric way and does not require functional form assumptions. I cast the estimation of the robustness metric as a de-biased GMM problem. This approach guarantees a parametric convergence rate for the robustness metric while allowing for machine learning-based estimators of policy effect heterogeneity (for example, lasso, random forest, boosting, neural nets). I apply my procedure to the Oregon Health Insurance experiment. I study the robustness of policy effects estimates of health-care utilization and financial strain outcomes, relative to a shift in the distribution of context-specific covariates. Such covariates are likely to differ across US states, making quantification of robustness an important exercise for adoption of the insurance policy in states other than Oregon. I find that the effect on outpatient visits is the most robust among the metrics of health-care utilization considered.

[1] Email: pspini@ucsd.edu. University of California, San Diego, 9500 Gilman Drive, LA Jolla, CA. I am indebted towards my advisors Yixiao Sun and Kaspar Wuthrich for their support and guidance. This paper benefited from discussions with James Hamilton, Xinwei Ma, Ying Zhu, Graham Elliott, Jeffrey Clemens, Simone Galperti, Julian Martinez-Iriarte, Alyssa Brown, Davide Viviano, Mitchell Van-Vuren, Tyler Paul, Radhika Goyal, and participants of the Econometrics Research Seminar at UCSD, at EGSC 2021 at Washington University in St. Louis and at the EWMES 2021 hosted by the University of Barcelona School of Economics. Seminar participants from the University of Warwick, UvA, Erasmus University Rotterdam, University of Bristol, University of Surrey, NYUAD, the Philadelphia Federal Reserve and the University of Victoria provided useful comments. All remaining errors are mine.

Keywords: Robustness, Heterogeneous Treatment Effects, KL divergence, Semiparametric estimation, De-biased GMM, Oregon Health Insurance Experiment

JEL codes: C14, C18, C44, C51, C54, D81, I13

1 Introduction

The guiding principle of evidence-based policy-making is to use experimental and (quasi)-experimental studies to guide the adoption of policies in various settings. This approach rests on the premise that the (quasi)-experimental findings are sufficiently robust and generalizable to hold beyond the setting of the (quasi)-experiment. In practice, this premise does not always hold: there are several examples of policies that, when adopted in non-experimental settings, under-performed their own experimental estimates Deaton [2010], Cartwright and Hardie [2012], Williams [2020]. In this paper, I argue that experimental estimates are insufficient to guide policy adoption and should be complemented by a measure of robustness that accounts for how policy recipients differ from the experimental ones. I develop a robustness metric, given by a scalar δ^*, that quantifies how much the characteristics of the recipients would have to change in order to invalidate the (quasi)-experimental findings. My metric summarizes the *out-of-sample* uncertainty[1] that the policy-maker faces regarding the policy recipients' characteristics. As such, my metric complements traditional summaries of *in-sample* uncertainty, like the standard errors, which routinely accompany (quasi)-experimental estimates.

As a motivating example, consider a policy-maker who must decide whether to offer medical insurance coverage to low-income households. The policy-maker has access to the experimental estimates of Finkelstein et al. [2012] which suggest that a similar intervention led to higher health-care utilization and reduced financial strain in Oregon. The target population of insurance recipients could differ from the experimental one in Oregon along important dimensions. Our goal is to quantify how robust the experimental findings would be if relevant characteristics of the recipients are allowed to change. In this paper, I provide a solution to this problem by leveraging the policy effect heterogeneity in the experiment.

When policy effects are heterogeneous across sub-populations with different covariate values, (quasi)-experimental findings are generally not robust to changes in the distribution of the covariates. In such cases, even small changes in the distribution of the covariates could lead to significant aggregate changes in the policy effects. For example, in the Oregon experiment, subsidized health insurance could benefit sicker patients

[1] Quantifying other sources of *out-of-sample* uncertainty has been a central theme in the recent econometric literature including Andrews et al. [2017] for moment conditions, Altonji et al. [2005], Oster [2019], Cinelli and Hazlett [2020] for confounding factors, and the break-down approaches in Horowitz and Manski [1995], Masten and Poirier [2020].

more than healthier patients. Then, the proportion of recipients with a given pre-existing health status, health habits, and/or co-morbidities may strongly influence the overall effect of the policy. Usually, these types of covariates are exclusively collected in the experimental context and not all of them are accessible in the new policy prior to implementation. As a result, the procedures proposed by Hsu et al. [2020] and Hartman [2020] that re-weight sup-population effects by the new environment's entire set of covariates are generally not feasible. Moreover, the heterogeneity of policy effects across sub-populations with different covariates values can be hard to model. This is because while domain knowledge can help select covariates that are predictive of the heterogeneity of policy effects, it usually cannot pin down a specific functional form for this heterogeneity. Because this heterogeneity links covariate shifts to shifts in the magnitudes of the aggregate policy effects, a general approach to robustness must reflect the uncertainty regarding the heterogeneity's functional form.

My robustness metric avoids the need to specify a functional form for the policy effect heterogeneity, letting it instead be flexibly estimated through the (quasi)-experimental data. Many popular existing approaches to robustness, like Altonji et al. [2005], Oster [2019] and Cinelli and Hazlett [2020], take advantage of specific functional forms. When designing a robustness metric for distributional changes, relying on functional form assumptions carries important implications for what type of shifts the metric can detect. If the way we measure a shift does not match the way we model heterogeneity, the resulting measure of robustness may be misleading. Consider, for example, measuring the difference between an arbitrary covariate distribution and the (quasi)-experimental one by reporting the difference in their means. With an unrestricted form for the heterogeneity of policy effects, we can, in general, construct a mean-preserving shift of the covariates' distribution which invalidates the policy-maker's claim. For example, in the Oregon experiment, if higher income recipients have negative effects while lower-income recipients have positive effects, we could construct a mean-preserving spread of the income distribution that induces a negative effect overall. Since their means coincide, such a distribution will have a distance of zero from the experimental covariates. A metric that, in most cases, is equal to zero cannot be very informative for assessing the robustness of (quasi)-experimental findings. This example suggests that a robustness metric should be general enough to accommodate unknown forms of policy effect heterogeneity. My robustness metric allows for arbitrary forms of policy effects heterogeneity, avoiding the limitations of a parametric model. Despite its generality, my metric is still easy to construct and interpret: a one-number summary of heterogeneity which only depends on (quasi)-experimental data.

Measuring robustness to covariate shifts requires choosing a distance between an arbitrary distribution of the covariates and the (quasi)-experimental one. In my approach, I adopt Kullback-Leibler divergence distance (KL distance). The KL distance is a popular choice for sensitivity analysis exercises, appearing recently in Christensen and Connault [2019] who apply it to models defined by moment inequalities and Ho [2020] who uses it in a Bayesian context. It has several advantages in our context. First, it is invariant to smooth invertible transformations of the covariates, hence independent of the covariates' units. Second, it provides a closed-form expression for the proposed global robustness measure, while other popular robustness approaches, like Broderick et al. [2020] rely on local approximations. Leveraging the closed-form solution, I cast estimation of my robustness metric as a GMM problem where the moment equation depends on two components. The first is the observed covariate distribution. The second is a functional parameter capturing the heterogeneity of policy effects, which can be flexibly estimated in the (quasi)-experimental data.

The heterogeneity of policy effects is often sparse: out of the rich set of covariates available in the (quasi)-experiment, just a few are needed to approximate it well. When covariate data is even moderately high-dimensional, it can be hard to select which covariates are important *ex-ante*. Machine-learning estimators, like lasso, random forest and boosting, can exploit the sparsity to automatically select the key covariates, reducing the need for *ad-hoc* procedures. Using machine-learning to estimate policy effect heterogeneity is appealing, but it may result in substantial bias in the estimated robustness metric δ^*, due to regularization and/or model selection. To accommodate machine-learning methods, I construct a de-biased GMM estimator: I derive the nonparametric influence function correction for the GMM parameters and leverage the theory in Chernozhukov et al. [2020] to eliminate the first-order bias from first-step estimators. I show that my metric δ^* can be consistently estimated at $\sqrt{n}$-rate under mild conditions on the first-step estimators of the policy effect heterogeneity. Under these conditions the functional parameter that summarizes heterogeneity can be estimated through modern high-dimensional methods like lasso, random forest, boosting and neural nets.

I apply my robustness procedure to study the Oregon health insurance experiment, whose findings have profound implications for public health Sanger-Katz [2014]. I replicate results in Finkelstein et al. [2012] and compute the robustness measure for several outcomes capturing recipients' heath-care utilization and financial strain. As discussed in Finkelstein et al. [2012] and Finkelstein [2013], the Oregon lottery recipients are older, in worse health, and feature a higher proportion of white individuals compared to the national average. These features invite questions about the robustness of the Oregon experiment's findings and the possibility of using them for policy adoption in other states. The differences in magnitude and sign between the effects of Medicaid expansion in Oregon and Massachusetts have motivated an effort to reconcile the discrepancy by identifying different populations of beneficiaries in the two states Kowalski [2018]. My robustness exercise is complementary to Kowalski [2018]: I compute the smallest change in the distribution of the key covariates relative to the Oregon benchmark, that can eliminate the positive effect of the lottery on recipients' health-care utilization and financial strain outcome measures. I find that the increase in outpatients visits is the most robust outcome among the measures of health-care utilization and financial strain.

This paper is also related to a larger strand of the econometric and statistics literature on robustness and sensitivity analysis originally initiated by Tukey [1960] and Huber [1965]. Recently, there are many other important but distinct robustness approaches: geared towards external validity Meager [2019], Gechter [2015], robustness to dropping a percentage of the sample Broderick et al. [2020], by looking at sub-populations Jeong and Namkoong [2020], or with respect to unobservable distributions like in Christensen and Connault [2019], Armstrong and Kolesár [2021], Bonhomme and Weidner [2018], and Antoine and Dovonon [2020]. My contribution complements this tool-set by giving the policy-maker an explicit measure of robustness to shifts in the covariate distributions. There are two reasons to focus on observable characteristics. First, observable characteristics are readily available to the policy-maker and are likely to be of first-level importance when assessing the robustness of (quasi)-experimental findings. Second, the resulting robustness metric is identified through the (quasi)-experimental data, limiting the need for bounding or partial identification approaches.

The paper is organized as follows: Section 2 introduces the basic setting and the notion of robustness to changes in the covariate distribution. Section 3 presents the main estimator and its asymptotic properties using the de-biased GMM theory recently developed in Chernozhukov et al. [2020]. Section 4 applies the proposed robustness metric to the Oregon health insurance experiment and reports empirical findings. Section 5 briefly concludes. In the Appendix, I provide all the proofs and discuss multiple extensions.

2 A robustness metric for covariate shifts

In this section, I use the potential outcome framework to explicitly link the heterogeneity of policy effects to the notion of robustness outlined in the introduction. The discussion focuses on the average treatment effect (ATE) as the main aggregate policy effect of interest. The policy-maker wants to assess the robustness of a claim on the magnitude (and/or sign) of the ATE, of the form $ATE \geq \tilde{\tau}$. The claim is true in the (quasi)-experiment but may no longer be true if covariates changes too much. The idea is to take advantage of the Conditional Average Treatment Effect (CATE), a functional parameter which links sub-population level treatment effects with the ATE. I use CATE to characterize, among the distributions that invalidate the policy-maker's claim ($ATE \geq \tau$), the one that is closest to the distribution of covariates in the (quasi)-experiment. I label this distribution the *least favorable distribution* because, among the distributions that invalidate the policy-maker's claim it is the hardest to distinguish from the covariates in the (quasi)-experiment. To measure the distance between two covariate distributions I use the Kullback-Leibler divergence distance. The value of the KL distance between the *least favorable distribution* and the (quasi)-experimental covariates will be the proposed robustness metric δ^*. Any covariate distribution that is closer than δ^* from the (quasi)-experimental covariates will be guaranteed to satisfy the policy-maker's claim ($ATE \geq \tilde{\tau}$).

2.1 Notation and Set Up

The policy-maker observes an outcome of interest $Y \in \mathcal{Y}$, a set of covariate measurements $X \in \mathcal{X}$ and a treatment status $D \in \{0, 1\}$. I consider two sets of covariates. The first set includes covariates which are exclusively collected in the (quasi)-experimental data and for which no counterpart exists in census data. For example, in the Oregon health insurance experiment, the recipients' health status and previous health history is available through survey data but such information may not be accessible through census variables in other settings (perhaps other states). The second set includes covariates for which a counterpart exists in the census data in other states, for example participants' race and age. To reflect the division of these two covariate types, X could be partitioned into two sets: $X = X_c \cup X_e$ denoting *census covariates* and *(quasi)-experiment specific covariates* respectively. All variables in X will be used to estimate the treatment effect heterogeneity in the (quasi)-experiment, which is the functional parameter needed to compute the robustness metric. The details are introduced in Section 2.3. If the policy-maker had access to observations on X_c in both the (quasi)-experiment and in the setting where the policy is to be adopted, my robustness metric can be modified to account for this additional information. To lighten the notation, in the main text I consider $X = X_e$ and discuss how to include X_c in the Appendix.

Now I introduce the notation to discuss changes in the distribution of the covariates. I use F_X to denote the distribution of the covariates in the (quasi)-experiment and and P_X to denote its associated probability measure. The propensity score is defined as $\pi(x) = P_X(D = 1|X = x)$. Following the traditional potential outcome framework, I denote Y_d for $d = \{0, 1\}$, the potential outcomes under treated and control status when the distribution of the covariates follows F_X. For example, in the Oregon experiment, Y_1 may represent the financial strain of a recipient if they receive insurance coverage while Y_0 represents the financial strain of the same recipient if they do not receive insurance coverage. In principle the distribution of the potential outcomes depends on the distribution of the covariates. To reflect this, I use Y_d and Y'_d to denote the potential outcomes when the distribution of the covariates follows F_X and F'_X respectively. Finally, for any random variable W, $\mathcal{W}$ denotes its support.

The parameter of interest for the policy-maker is the $ATE := \mathbb{E}[Y_1 - Y_0]$. The Conditional Average Treatment Effect (CATE) defined by $\tau(x) := CATE(x) = \mathbb{E}[Y_1 - Y_0|X = x]$ captures how the average treatment effect changes across sub-populations with covariate value $X = x$. Under unconfounded-ness (Assumption 1 i) below), $\tau(x)$ is nonparametrically identified[2] by $\mathbb{E}[Y|D = 1, X = x] - \mathbb{E}[Y|D = 0, X = x]$ in the (quasi)-experiment Imbens and Rubin [2015].

Assumption 1. *Unconfounded-ness & Overlap*

i) $Y_1, Y_0 \perp\!\!\!\perp D|X$.

ii) For all $x \in \mathcal{X}$ *we have* $0 < \epsilon \leq \pi(x) \leq 1 - \epsilon < 1$

In the case of a randomized control trial, for example when treatment assignment is completely randomized or is randomized conditional on covariates, Assumption 1

[2] If the CATE only partially identified, like in the case on non-compliance based on unobservables, it is possible to follow a bounding approach for my robustness procedure. I leave this interesting case for future research.

holds by design. In the case of (quasi)-experimental studies Assumption 1 i) requires the researcher to carefully evaluate the selection mechanism that governs program participation. Assumption 1 ii) is strict overlap. While strict overlap is not a necessary condition for identification, it will be important in the estimation of the robustness metric in Section 3.

In this paper, the goal is to study the robustness of claims concerning the ATE with respect to changes in the distribution of the covariates. Because the ATE is obtained by averaging $\tau(x)$ with weights proportional to F_X we have the following map between the covariate distributions and the ATE:

$$ATE : F_X \mapsto \int_{\mathcal{X}} \tau_{F_X}(x) dF_X(x) \tag{1}$$

The subscript F_X on $\tau(x)$ indicates that, in general, it's possible that the functional form of CATE depends on F_X. In this case, a change in the distribution of the covariates would effect the magnitude of ATE through two channels: a direct effect thorough the weights of F_X and an indirect effect through changing the functional form of $\tau_{F_X}(x)$. In this paper, I introduce the covariate shift assumption[3] to eliminate the indirect effect.

Assumption 2. ***(Covariate Shift)*** *Let X' denote the covariates in the new environment. Then:*

i *$F_{Y_d'|X'}(y|x) = F_{Y_d|X}(y|x)$ for $d = \{0, 1\}$, for all $x \in \mathcal{X}$ and $y \in \mathcal{Y}_d$ and all distributions of X'.*

ii *$\mathcal{X}' \subseteq \mathcal{X}$*

Assumption 2 i) says that the causal link between the treatment variable D and the potential outcomes of interest Y_1 and Y_0 does not depend on the distribution of the observables. One could think of Assumption 2 as analogous to a policy invariance condition where the invariance in this case is with respect to the distribution of covariates.

Assumption 2 ii) says the support of the covariates in the new environments is contained in the support of the baseline environment. In practice, this limits the extrapolation to environments for which any value of the covariates could have been observed in the (quasi)-experimental setting as well. Because Assumption 2 guarantees that $\tau_{F_X}(x)$,

[3] This assumption appears, for example also in Hsu et al. [2020] and Jeong and Namkoong [2020].

the CATE, does not vary when F_X is replaced by any other distribution $F_{X'}$ it is not necessary to index $\tau(x)$ with F_X.[4] Then, the link between F_X and ATE reduces to integration against a fixed $\tau(x)$:

$$ATE : F_X \mapsto \int_{\mathcal{X}} \tau(x) dF_X(x) \tag{2}$$

To emphasize the dependence of the ATE on an arbitrary distribution of the covariates F_X, I occasionally write $ATE(F_X)$. Before presenting the general framework I give perhaps the simplest nontrivial example of a robustness exercise with respect to the distribution of the covariates.

Example 1. *Consider a binary covariate $X = \{0, 1\}$. D is randomly assigned, trivially satisfying Assumption 1. By unconfounded-ness, $\mathbb{E}[Y_1|x = 0], \mathbb{E}[Y_0|x = 0], \mathbb{E}[Y_1|x = 1], \mathbb{E}[Y_0|x = 1]$ can all be identified. Consequently, the average treatment effect for the sub-populations $x = 0$ and $x = 1$, denoted $\tau(0)$ and $\tau(1)$ are also identified. Because X is Bernoulli, any distribution on $\{0, 1\}$ is fully described by $P_X(x = 1) = p_1$ so automatically $P_X(x = 0) = 1 - p_1$. Suppose that, in the experiment $ATE \geq 0$. Note that:*

$$\begin{aligned}
ATE(F_X) &= \mathbb{E}[Y_1|x = 0] \cdot (1 - p_1) + \mathbb{E}[Y_1|x = 1] \cdot p_1 \\
&- \mathbb{E}[Y_0|x = 0] \cdot (1 - p_1) - \mathbb{E}[Y_0|x = 1] \cdot p_1 \\
&= (\mathbb{E}[Y_1|x = 0] - \mathbb{E}[Y_0|x = 0]) \cdot (1 - p_1) + (\mathbb{E}[Y_1|x = 1] - \mathbb{E}[Y_0|x = 1]) \cdot p_1 \\
&= \tau(0) \cdot (1 - p_1) + \tau(1) \cdot p_1.
\end{aligned}$$

A shift in the covariate distribution is simply a shift in the parameter p_1. Assume the treatment effects are sufficiently heterogeneous, namely $\tau(1) > 0 > \tau(0)$ so one group has positive effects from treatment and the other group has negative effects. What is the closest covariate distribution that invalidates the claim $ATE \geq 0$?

It suffices to find the weights on $x = 0, x = 1$ such that the ATE is 0. Expressing it in terms of p_1:

$$\tau(0) \cdot (1 - p_1^*) + \tau(1) \cdot p_1^* = 0$$

[4] This could be cast as an identification result which follows immediately from the Assumption 2. See Hsu et al. [2020], Lemma 2.1.

A solution is given by:

$$p_1^* = \frac{-\tau(0)}{\tau(1) - \tau(0)} \in [0, 1]$$

so the distance $|p_1^* - p_1| = |\frac{-\tau(0)}{\tau(1)-\tau(0)} - p_1|$ *is largest shift in the covariates that still guarantees that the claim* $ATE \geq 0$ *holds.*

Under what conditions we are always guaranteed to find a solution like p_1^* above? Is it unique? Can we always characterize the distance between p_1^* and p_1? If the space $\mathcal{X}$ is not discrete, a probability distribution on $\mathcal{X}$ cannot be described by a finite dimensional parameter without restricting the class of probability distributions on $\mathcal{X}$. How should one measure the distance between two distributions in general?

I start from this last question by introducing a notion of distance that does not require any parametric restriction on probability distributions.[5] Here I introduce the KL-divergence distance:

Definition 2 (KL-divergence)**.** *Consider the KL-divergence between two distributions* F_X *and* F_X' *given by:*

$$D_{KL}(F_X' || F_X) := \int_{\mathcal{X}} \log\left(\frac{dF_X'}{dF_X}(x)\right) \frac{dF_X'}{dF_X}(x) dF_X(x) \tag{3}$$

where $\frac{dF_X'}{dF_X}$ *is the Radon-Nikodym derivative of the distribution* F_X' *with respect to the experimental distribution* F_X*, provided that* $P_X' \ll P_X$ *for the respective probability measures.*

There are several advantages to using the KL divergence to measure the distance between probability distributions: it is nonparametric, it has useful invariance properties and it delivers a closed form solution for the policy-maker's robustness problem introduced below. Both Ho [2020] and Christensen and Connault [2019] use the KL divergence to measure the distance between probability distributions in different contexts. Appendix H discusses in detail how to use convex analysis to obtain a closed form solution for the policy-maker's robustness problem.

[5] I discuss the details of parametric classes in Appendix B, as special cases of the general procedure.

2.2 The policy-maker's problem: quantifying robustness

After isolating the link between the ATE and the distribution of covariates and choosing a distance measure between probability distributions, we can formalize the policy-maker's robustness problem. Consider the claim given by $ATE \geq \tilde{\tau}$: the ATE is larger than a desired threshold $\tilde{\tau}$. The sign of the inequality is without loss of generality, as claims of the type $ATE \leq \tilde{\tau}$ can be accommodated with an equivalent treatment. The threshold $\tilde{\tau}$ captures a minimal desirable aggregate effect that would make the intervention viable for the policy-maker. It could capture the average cost for the roll-out of the intervention or the value of ATE for a competing policy. In Example 1, $\tilde{\tau}$ was fixed at 0. The policy-maker is interested in the smallest shift from the (quasi)-experimental distribution, F_X, such that the claim $ATE \geq \tilde{\tau}$ is invalidated. Recall $\tau(x) = CATE(x)$. Formally the policy-maker wants to solve the following problem:

$$\inf_{dF'_X:\ dF'_X \ll dF_X; dF'_X(\mathcal{X})=1} D_{KL}(F'_X || F_X) \tag{4}$$

$$s.t. \int_{\mathcal{X}} \tau(x) dF'_X(x) \leq \tilde{\tau} \tag{5}$$

The optimization problem in Equation (4) searches across all distributions of the covariates that invalidate the policy-maker's claim $ATE \geq \tilde{\tau}$ (notice that the ATE for all the distributions in Equation (5) is constrained to be less than $\tilde{\tau}$) and selects, if they exist, the one(s) that are closest to the (quasi)-experimental distribution F_X, according to the KL distance in Equation (4). Notice also that $\tau(x)$ in Equation (5) is not indexed by F'_X because of the covariate shift assumption (Assumption 2). Here, the class of probability measures for the covariates is restricted to be absolutely continuous w.r.t the (quasi)-experimental measure dF_X[6] but no other restriction is imposed: the class of distributions is still nonparametric. Absolute continuity does restrict the distributions F'_X to be supported on $\mathcal{X}$. While it may appear as an unnecessary restriction, I view it as a very reasonable requirement: the feasible distributions in Equation (5) cannot put mass on a sub-population $X = x$ that could not theoretically be observed in the (quasi)-experimental setting. Clearly, treatment effect values for sub-populations with $X = x$ that can never be observed can lead to arbitrarily large average effects and the

[6] This is a refinement of Assumption 1. Namely, with a slight abuse of notation, requiring for instance that $dF_X, dF'_X \ll \lambda$ will deliver absolute continuity of dF'_X w.r.t dF_X. Restricting the support guarantees that dF'_X cannot put mass on areas where dF_X does not put mass.

robustness exercise would not be very informative. We are now ready to define the *least favorable distribution* and the robustness metric.

Definition 3. *i) The least favorable distribution set* $\{F^*_X\}$ *is given by the expression below:*

$$\begin{aligned}\{F^*_X\} = \arg &\min_{P'_X:\ P'_X \ll P_X; P'_X(\mathcal{X})=1} D_{KL}(F'_X||F_X) \\ &s.t. \int_{\mathcal{X}} \tau(x) dF'_X(x) \leq \tilde{\tau}\end{aligned} \tag{6}$$

where the set in Equation (6) is allowed to be the empty set.
ii) For a given $\tilde{\tau} \in \mathbb{R}$ *the robustness metric* $\delta^*(\tilde{\tau})$ *is given by:*

$$\delta^*(\tilde{\tau}) = D_{KL}(F^*_X||F_X). \tag{7}$$

The minimizer of Equation (4) is the *least favorable distribution*, the closest distribution of the covariates that invalidates the target claim. I define the KL-distance between the experimental distribution and the *least favorable distribution* as my metric $\delta^*(\tilde{\tau})$ which quantifies the robustness of the claim $ATE \geq \tilde{\tau}$. Observe that, if the (quasi)-experimental ATE satisfies the constraint in Equation (5), then we can always choose the *least favorable distribution* to be the (quasi)-experimental one, namely $F^*_X = F_X$ since it's feasible and $D_{KL}(F^*_X||F_X) = 0$. In words this means that the policy-maker's claim is already invalidated in the (quasi)-experiment. The problem is non-trivial when the $ATE(F_X) > \tilde{\tau}$ condition is satisfied for the (quasi)-experimental distribution F_X. In such a case, the (quasi)-experimental distribution F_X is excluded from the feasible set of Equation (5). As a result, the value of $D_{KL}(F^*_X||F_X)$ in Equation (4) must be strictly positive. Notice that, in Example 1, we imposed the requirement that the $ATE(p_1)$ in the experiment was larger than 0, to guarantee that the problem was indeed non-trivial.

If $\mathcal{X}$ is a set containing finitely many elements, the covariate distribution is discrete. In practice, there are many empirical applications in which covariates of interest are either discrete or have been discretized for privacy reasons. Any grouping of a continuous variables in finitely many classes, gives rise to discrete distribution. For example, in the Oregon experiment, the recipients income may have been discretized into income groups. When the covariates space is discrete, we can get an important geometric insight in the structure of the robustness problem as formulated by Equations (4) and

(5). The example below illustrates the case where $\mathcal{X}$ contains only 3 points. In this case, a probability distribution on $\mathcal{X}$ can be parametrized by 2 parameters and there is convenient visual representation of the robustness problem contained in Equations (4) and (5).

Example 4. *Consider the case $\mathcal{X} = \{x_1, x_2, x_3\}$ each value representing an income bin: high, medium and low respectively. Here the experimental distribution is represented by a triplet (p_1, p_2, p_3). Because $p_1 + p_2 + p_3 = 1$ the whole space of probability distributions on $\mathcal{X}$ is 2-dimensional: it suffices to choose p_1 and p_2 to fully characterize a distribution. Suppose that conditional treatment effects are highest for lower income participants and are lowest for high income participants: $\tau(x_1) = 1, \tau(x_2) = 2, \tau(x_3) = 3$. The average cost of roll-out is equal to $\tilde{\tau} = 1.8$. The claim is $ATE \geq \tilde{\tau}$ meaning that the ATE should be higher than average cost. In the experiment ATE is equal to $2.4 > 1.8$ which satisfies the claim.*

The policy-maker's robustness problem in Example 4 is depicted in Figure 1. Since the functions in Equations (4) and (5) are differentiable in p_1 and p_2 the finite dimensional problem could be easily solved through the standard Karush-Kuhn-Tucker conditions. The level sets of the KL distance, the feasible set and the *least favorable distribution* are all indicated in Figure 1. The KL level set associated to $\delta^*(\tilde{\tau})$ is highlighted by a green contour. It includes the set of covariate distributions that are guaranteed to satisfy the policy-maker's claim. This region is conservative, in the sense that there exist covariate distributions that satisfy the policy-maker's claim but fall outside of the green contour. This feature reflects the definition of robustness as a minimization problem in Equations (4) and (5).

When $\mathcal{X}$ is not discrete, a representation like Figure 1 may not be possible. Nonetheless one can still show that, given some conditions, a solution for F_X^* like the one in Figure 1 always exists, is unique, and can be characterized by a closed form expression, with virtually little difference from the finite dimensional case. This result also guarantees that the robustness metric $\delta^*(\tilde{\tau})$ is well defined for a wide range of $\tilde{\tau}$ values.

2.3 A closed form solution for quantifying robustness

In this section I characterizes the solution for the policy-maker's robustness problem in Equations (4) and (5) in the general case. Some additional conditions are introduced below.

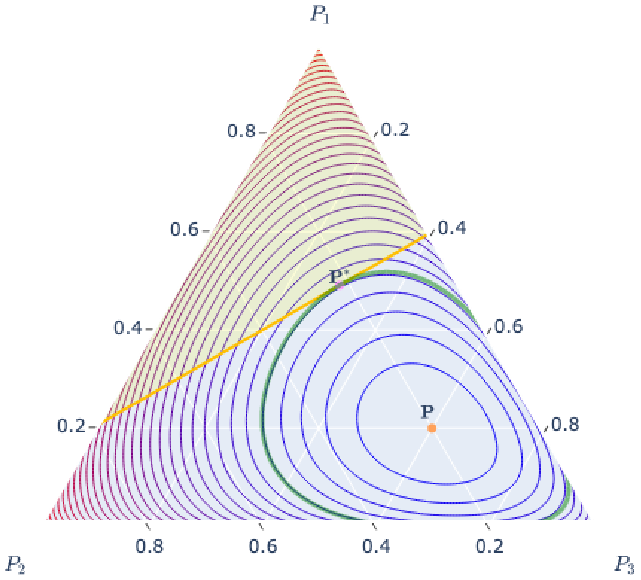

Figure 1: The triangle represents the collection of all arbitrary probability distribution triplets (p_1, p_2, p_3) on the discrete set (x_1, x_2, x_3) represented in barycentric coordinates. P denotes the experimental distribution, given by $(0.2, 0.2, 0.6)$. The $CATE(x_1, x_2, x_3) = (1, 2, 3)$ so the conditional treatment effect is greater in the highest income group. The yellow shaded region is the feasible set: the collection of covariate distributions with an $ATE \leq 1.8$, which invalidate the policy-maker's claim. The solid yellow line is the boundary of the feasible set. The contour lines from blue to red represent the level sets of the KL distance of any distribution in the triangle with respect to the experimental distribution P (bluer indicates a lower value for the KL divergence). The distribution $P^* = (0.491, 0.218, 0.291)$ is the *least favorable distribution*. It is the minimizer of the KL divergence, subject to the feasibility constraint (it lies on the orange line). The green boundary is the level set of KL that corresponds to $\delta^* \approx 0.296$. Any distribution closer than δ^*, within the green boundary is guaranteed to satisfy the policy-maker's claim.

Assumption 3 (Bounded-ness)**.** *The conditional average treatment effect $\tau(X)$ is bounded P_X-almost surely over $\mathcal{X}$. In particular for some $M \in \mathbb{R}_+$ we have:*

$$\mathbb{P}_X\left(|\tau(X)| \leq M\right) = 1$$

Incidentally, for any covariate probability measure that is absolutely continuous w.r.t P_X, Assumptions 3 continues to hold. This is because $P_{X'}$ cannot put mass on the subsets of X that P_X considers negligible, which includes the subset of X where $\tau(x)$ is unbounded. Assumption 3 is automatically satisfied if $\tau(X)$ is bounded on $\mathcal{X}$. Boundedness is not very restrictive in a micro-econometrics framework where virtually all variables are bounded in the cross-section.

Consider the feasible set in Equation (5). While the set is guaranteed to be convex, it may be empty. If that is the case, the value of the minimization problem in Equation (4) is $+\infty$. I avoid such cases by guaranteeing that, for a given claim, an $ATE = \tilde{\tau}$ is attainable, for some distribution F'_X. This amounts to assuming that there is enough variation in $\tau(x)$ to induce an ATE of $\tilde{\tau}$ through changes in the distribution of the covariates. An extreme case where such requirement fails is described below.

Example 5 (Homogeneous treatment effects)**.** *Consider a situation of constant treatment effects. In this case $ATE(F_X) = \int_{\mathcal{X}} c \; dF_X = c$ so that the ATE is equal to c regardless of the distribution of the covariates.*

Not surprisingly, no heterogeneity in treatment effects translates in no threat to robustness. One can freely extrapolate the claim from the (quasi)-experimental environment to any other environment. Constant treatment effects are a rather extreme case. A more realistic example concerns whether the minimal desired magnitude $\tilde{\tau}$ is outside of the range of variation of the heterogenous treatment effects. For example, suppose that $2 \leq \tau(x) \leq 5$ with probability equal to 1. Then, choosing $\tilde{\tau} = 1$ results in an empty feasible set of distributions, since no probability distribution may ever integrate against $\tau(x)$ to an ATE of 1. In this case, since the set of distributions in Equation (5) is empty, the infimum in Equation (4) evaluates to $+\infty$. So we see that enough heterogeneity of treatment effects is a necessary condition for robustness to be non-trivial. For estimation purposes it is convenient to consider a parameter space for the robustness measure that is a subset of $\mathbb{R}$ rather than $\mathbb{R} \cup \{+\infty\}$. The following assumption guarantees that the feasible set is not empty:

Assumption 4. *(Non-emptiness) Denote the interior S° of a set S to be the union of all open sets $O \subseteq S$. Let $L : F_X \to \int_{\mathcal{X}} \tau(x) dF_X(x)$ be the linear map defined on the set of probability distributions on $\mathcal{X}$ that are absolutely continuous w.r.t P_X, denoted as $\mathcal{P}_X \subset \mathcal{M}$. We require $\tilde{\tau} \in L^\circ(\mathcal{P}_X)$, that $\tilde{\tau}$ is in the interior of the range of L.*

Assumption 4 says that $\tilde{\tau}$ is in the interior of the range of the linear map L. In other words, there is enough observable heterogeneity in treatment effects that there exists a distribution of covariates that, when integrated against $\tau(x)$, it induces an $ATE = \tilde{\tau}$. Contrast this to the homogeneous treatment effect case in Example 5, where Assumption 3 fails. There, $L^\circ(\mathcal{P}_X) = \emptyset$. More generally, the length of $L(\mathcal{P}_X)$ measures how rich is the set of ATEs that could be produced by choosing an arbitrary distribution F_X. Assumption 4 is testable. For a given value for $\tilde{\tau}$, one could obtain an estimate of the $\tau(x)$ and test whether $\tilde{\tau}$ is smaller than $\sup_x \tau(x)$ or greater than $\inf_x \tau(x)$, depending on the sign of the claim of interest, using the procedure in Chernozhukov et al. [2013]. Testing Assumption 4 tests for whether treatment effects are sufficiently heterogeneous to invalidate the claim of interest through a covariate shift, which is more general than testing whether any form of treatment effect heterogeneity is present. This is because, along the lines of the discussion above, treatment effects can indeed be heterogeneous but not heterogeneous enough to invalidate the policy-maker's claim. A rejection in the test means implies an infinite value for the robustness metric and signals that the policy-maker's claim can never be invalidated by covariates shifts.

Remark 6. *The interior condition cannot be relaxed. By Assumption 3, the image of $\mathcal{P}_X$ under L is a compact convex subset of $\mathbb{R}$, that is, an interval. If $\tilde{\tau}$ is at a an endpoint of this interval, the feasible set in Equation (5) may consist of only a point mass measure Because such a covariate measure is not absolutely continuous w.r.t. P_X, the feasible set is again empty and will necessarily result in an infinite value for the KL-divergence in Equation (4).*

In Example 1 we imposed the condition $ATE(1) = \tau(0) < 0$ to guarantee that the problem has a solution. In the context of Example 1, $L(\mathcal{P}_X) = [\tau(0), \tau(1)]$, the image of L is the interval between the conditional average treatment effects at $x = 0$ and $x = 1$ since any $ATE(p)$ is a weighted average of $\tau(0)$ and $\tau(1)$. By requiring that $\tau(0) < 0 < \tau(1)$, $\tilde{\tau} = 0 \in L^\circ(\mathcal{P}_X)$ hence satisfies Assumption 4.

With Assumptions 3 and 4 we are now ready to introduce the key result that always delivers a closed form solution for the robustness metric. It says that the *least favorable distribution* set in Definition 3 is nonempty and it contains a unique distribution (P_X-almost everywhere). Moreover the robustness metric $\delta^*(\tilde{\tau})$ is finite and both it and the *least favorable distribution* have a closed form solution:

Lemma 7 (Closed form solution)**.** *Let Assumptions 1, 2, 3 and 4 hold. Then: i) The infimum in Equation* (4) *is achieved. Moreover F_X^*, is characterized, P_X-almost everywhere, by:*

$$\frac{dF_X^*}{dF_X}(x) = \frac{\exp(-\lambda(\tau(x) - \tilde{\tau}))}{\int_{\mathcal{X}} \exp(-\lambda(\tau(x) - \tilde{\tau}))dF_X(x)} \tag{8}$$

where $\frac{dF_X^*}{dF_X}$ is the Radon-Nikodym derivative of dF_X^* with respect to dF_X and λ is the Lagrange multiplier implicitly defined by the equation:

$$\int_{\mathcal{X}} \exp(-\lambda(\tau(x) - \tilde{\tau}))(\tau(x) - \tilde{\tau})dF_X(x) = 0 \tag{9}$$

ii) The value of the robustness metric $\delta^*(\tilde{\tau})$ is given by:

$$\delta^*(\tilde{\tau}) = D_{KL}(F_X^* || F_X) = -\log\left(\int_{\mathcal{X}} \exp(-\lambda(\tau(x) - \tilde{\tau}))dF_X(x)\right) \tag{10}$$

Proof. See Appendix I. □

Lemma 7 greatly simplifies the computation of the robustness metric by essentially showing that the fully general robustness problem that searches over the nonparametric space of probability distribution is no-harder than the parametric cases in Examples 1 and 4. We can compare the closed form solution of Lemma 7 with the KKT solution one could derive for Example 1 and verify that the two solutions are indeed identical.

Example 8. *Return to the example of the discrete variable so $X = \{0, 1\}$. First notice that the dominating measure here is the counting measure on $\{0, 1\}$. We are therefore interested in simply the ratio $\frac{p_1}{p_1^*}$ since it completely characterizes p_1^*. Because the problem is one dimensional, the unique minimizer is the one that satisfies the constraint:*

$$\tau(1) \cdot p_1^* + \tau(0) \cdot (1 - p_1^*) = \tilde{\tau} \implies p_1^* = \frac{\tilde{\tau} - \tau(0)}{\tau(1) - \tau(0)} \tag{11}$$

Recall that in Example 1 $\tilde{\tau}=0$. On the other hand, from the solution provided by 7 we have:

$$\frac{p_1^*}{p_1}=\frac{\exp(-\lambda(\tau(1)-\tilde{\tau}))}{\exp(-\lambda(\tau(1)-\tilde{\tau}))\cdot p_1+\exp(-\lambda(\tau(0)-\tilde{\tau}))\cdot(1-p_1)} \tag{12}$$

where λ is implicitly defined as in Equation (7).

Fact 9. *Equations 12 and 11 are equivalent.*

Proof. See Appendix I. □

Lemma 7 completely characterizes the robustness metric in terms of the (quasi)-experimental distribution $F_X(x)$ and the CATE, $\tau(x)$. This is important because both of them are nonparametrically identified from the (quasi)-experimental data. Hence, to give an answer to the policy-makers robustness problem, it is enough to estimate the treatment effect heterogeneity in $\tau(x)$. This result will deliver a very convenient estimation theory which I discuss in Section 3.

2.4 Locally infeasible problem

We have seen how the restriction in Assumption 4 is key to guarantee that a solution to Equation (4) exists and that the associated $\delta(\tilde{\tau})$ is finite. There is a partial extension to Lemma 7 with respect to a local violation of Assumption 4. Consider a sequence of $\tilde{\tau}_m$ converging to a boundary point $\tilde{\tau}_b$ of the range of $\tau(X)$. An example is depicted in Figure 2. Suppose the policy-maker's claim is given by: $ATE \leq \tilde{\tau}_m$.

For each $\tilde{\tau}_m$ within the range of variation of $\tau(X)$, the policy-maker's problem has a solution, $F^*_{X,m}$ given by Lemma 7. This is because there is a sub-population with covariates x such that $\tau(x) \geq \tilde{\tau}_m$. The *least favorable distribution* will increase the weight on this sub-population. If $\tilde{\tau}$ is on the boundary, for example $\tilde{\tau}=3$ in Figure 2, the only sub-population that has $\tau(x) \geq \tilde{\tau}_b$ is $x=0.6$, concentrated on a singleton. But distributions that put unit mass on singletons are not feasible in the policy-maker's problem. For $\tilde{\tau}=\tilde{\tau}_b$, the feasible set is empty so there is no solution. If one looks at the sequence of *least favorable distributions*, $F^*_{X,m}$, associated to the sequence $\tilde{\tau}_m \to \tilde{\tau}_b$, is there a limiting distribution to which the sequence $F^*_{X,m}$ converges in some sense?

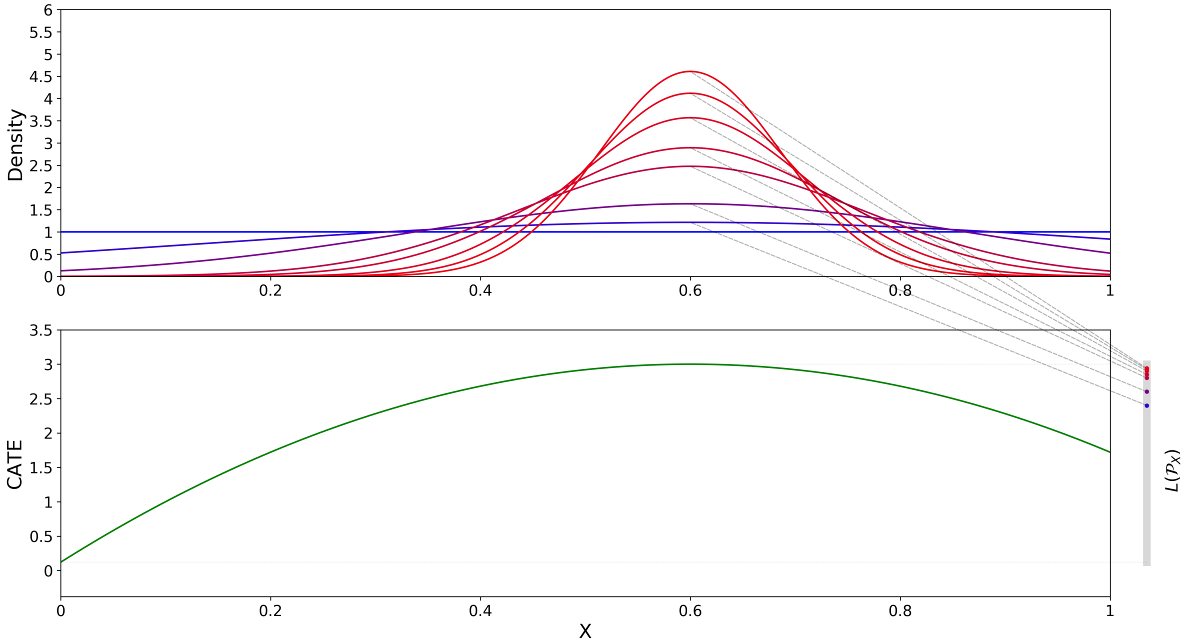

Figure 2: Local to boundary conditions. The lower panel displays the conditional average treatment effect, $\tau(x)$ for a univariate variable X. The experimental distribution is in blue: the uniform distribution. The gray segment on the left labelled $L(\mathcal{P}_X)$ is the image of the collection of probability distributions supported on $\mathcal{X}$ under the map $L : F_X \mapsto \int_{\mathcal{X}} \tau(x) dF_X(x)$. For every point in the interior, Lemma 7 holds and, for each $\tilde{\tau}_m$, there is an associated *least favorable distribution* $F^*_{X,m}$ displayed in the upper panel. As the sequence of $\tilde{\tau}_m$ approaches the boundary of $L(\mathcal{P}_X)$, the distributions concentrate around $x = \arg\max \tau(x) = 0.6$.

Under some additional assumptions, one can show a type of concentration result for the sequence of solutions obtained by applying the closed-from solution formula in Lemma 7. If $\tau(x)$ is a single peaked function, that is, it achieves its maximum (or minimum) at a single point, we obtain convergence in distribution of the sequence $F^*_{X,m}$ to the Dirac distribution at the single peak, δ_{x_b}.

Proposition 10 (Local to boundary $\tilde{\tau}$)**.** *Let Assumptions 1-3 hold and let $\tilde{\tau}_m \to \tilde{\tau}_b \in \partial L(\mathcal{P}_X)$. Assume that the pre-image $\tau^{-1}(\tilde{\tau}_b) = \mathcal{X}_b = \{x_b\} \in \mathcal{X}^{\circ}$ is a singleton. Further, let X be compactly supported, with density $f(x) < M$ on $\mathcal{X}$. Then the sequence of least favorable distributions for the policy-maker's problem with parameter $\tilde{\tau}_m$, denoted $F^*_{X,m}$, converges weakly to δ_{x_b}, the Dirac delta distribution with point mass at x_b, that is:*

$$\lim_{m\to\infty} \int_{\mathcal{X}} g(x) dF^*_{X,m}(x) \to \int_{\mathcal{X}} g(x)\delta_{x_b} := g(x_b)$$

for $g \in C_b(\mathcal{X})$, the space of all continuous, bounded functions on $\mathcal{X}$.

Proof. See Appendix I. □

The point-mass distribution δ_{x_b} is not a solution to the policy-maker's problem with parameter $\tilde{\tau}_b$ because the feasible set never includes point mass distributions unless $\mathcal{X}$ is discrete. In this sense, Proposition 10 delivers the limit of the sequence of solutions in the sense of weak convergence. This is a weaker that the notion of convergence induced by D_{KL}. In particular when $dF_X \ll \lambda_{Leb}$ (the Lebesgue measure on $\mathbb{R}^k$), $D_{KL}(dF^*_{X,m}||\delta_{x_b}) = +\infty$ so the sequence of solutions $F^*_{X,m}$ does not converge to δ_{x_b} in D_{KL}.[7]

2.5 Interpreting robustness

In this section I offer some practical guidance on how to interpret the the robustness metric proposed in Definition 18. The first interpretation links the robustness metric to a bound on the probability of drawing a sample of size n for which the experimental conclusion is false. The second interpretation is a bench-marking exercise using available census covariates.

2.5.1 A probability interpretation using Sanov's theorem

One way to link the magnitude of the robustness metric $\delta^*(\tau)$ to an easily interpretable probability bound is through Sanov's theorem. In this section I provide the intuition through a finite dimensional example and give the interpretation. I discuss more details on the foundations of Sanov's theorem in Appendices F. First, consider the setting of Example 4. Now suppose we collect a sample containing n i.i.d observations. Consider a generic sequence of the data of size n, $x = (x_1, x_2, \cdots, x_n)$. Each sequence is an ordered list of values $(High, Medium, Low)$. Define the *type* P_x of a sequence x as the proportion (relative to n) of realizations of a in x. This is $P_x(a) = \frac{N(a|x)}{n}$ where $N(a|x)$

[7] In fact, Posner [1975] showed that D_{KL} is lower-semicontinuous, that is, if $P_n \to P$ weakly, then $\lim_{n\to\infty} D_{KL}(P_n||Q) \geq D_{KL}(P||Q)$. In this case we have $+\infty > 0$

is the number of times realization a shows up is in sequence x. We denote the collection of types as $\mathcal{P}_n$.[8]

For the present example, a result by Cover [1999] shows that while the number of sequences is of the order of 3^n, the number of types is bounded above by $|\mathcal{P}_n| \leq (n+1)^3$. We can look at the *types* that fall within a specific subset E of probability distributions. For example we can look at all the *types* that invalidate the experimental conclusion on the ATE. In this case $E := \{Q \in \mathcal{P}_X : \int_{\mathcal{X}} \tau(x)dQ \leq \tilde{\tau}\}$, the constraint in Equation 5. Notice that whether a sequence $x \in E$ or not depends only on its *type* P_x. Now, what is the probability that, drawing a sequence x according to P_X, such a sequence invalidates the experimental results, that is $x \in E$?. It turns out that Sanov's theorem provides a link between this probability and the metric of robustness $\delta^*(\tau)$.

Theorem 11. *(Sanov's theorem) Let $X_1, \cdots X_n$ be i.i.d distributed according to F_X. Let E be a convex set of probability distributions. Letting P^n_X be the product measure of n copies of P_X. Then*

$$P^n_X(E \cap \mathcal{P}_n) \leq e^{-nD_{KL}(P^*_X||P_X)}$$

where

$$P^* = \min_{Q \in E} D_{KL}(Q||P)$$

Moreover, if the set E is the closure of its interior then

$$\lim_{n\to\infty} \frac{1}{n} \log(P^n(E)) \to -D_{KL}(P^*|P)$$

Proof. The proof can be found in Cover [1999] Theorem 11.4.1. □

Note that $E := \{Q : \int_{\mathcal{X}} \tau(x)dQ \leq \tilde{\tau}\}$ is obtained through imposing a linear restriction on Q and therefore E is convex. Sanov's theorem remains true for larger classes of probability distributions, not necessarily confined to finitely supported X variables like discussed in Csiszár [1984]. Note that $\delta^*(\tilde{\tau}) = D_{KL}(P^*||P)$ is precisely the metric of

[8] One can think of a *type* P_x as keeping track of the proportion but forgetting the order. So for example the two sequences of size $n = 3$ given by $x = (High, Medium, Medium)$ and $x' = (Medium, High, Medium)$ are distinct: $x \neq x'$. But they have the same type: $P_{x'} = P_x$.

robustness $\delta^*(\tau)$. It captures the smallest distance from the experimental distribution of the covariates that will fail to satisfy the conclusion, hence a bound for the probability that a sequence does not satisfy the policy-maker's conclusion is given by

$$P_X^n(E) \leq e^{-n\delta^*(\tilde{\tau})}$$

The fact that the probability bound depends on $\tilde{\tau}$ should not be surprising since $\tilde{\tau}$ also controls the inequality that defines the constraint set E. Notably the bound is non-asymptotic in that it applies for any n. The bound is monotonically decreasing in the magnitude of $\delta^*(\tau)$ and it becomes trivial when $\delta^*(\tau) = 0$. Of course if $\delta^*(\tau) = \infty$ we know that the set E does not contain any valid distributions, so it is reasonable that $P_X^n(E) = 0$. Below, we may revisit the discrete example to get a sense of the estimate that Sanov's theorem provides.

Example 4 (continuing from p. 14)**.** *Recall X = income, $\mathcal{X} = \{High, Medium, Low\}$ and the experimental distribution is $F_X = (p_1, p_2, p_3) = (0.2, 0.2, 0.6)$. For a given n we can list the types of sequences of size n that can be generated. Here the count of $High$ and $Medium$ income individuals will completely determine the type of a sequence (since for fixed n, $\#Low = n - \#High - \#Medium$. For $n = 3$ for example, there are 10 possible sequence types each corresponding to one of the sequences $(3,0,0)$, $(2,0,1)$, $(2,1,0)$, $(1,2,0)$, $(1,1,1)$, $(1,0,2)$, $(0,3,0)$, $(0,2,1)$, $(0,1,2)$, $(0,0,3)$ divided by 3. Therefore $|\mathcal{P}_3| = 10$. They are displayed below in barycentric coordinates as red points in the 2-simplex. The set E is also displayed in yellow.*

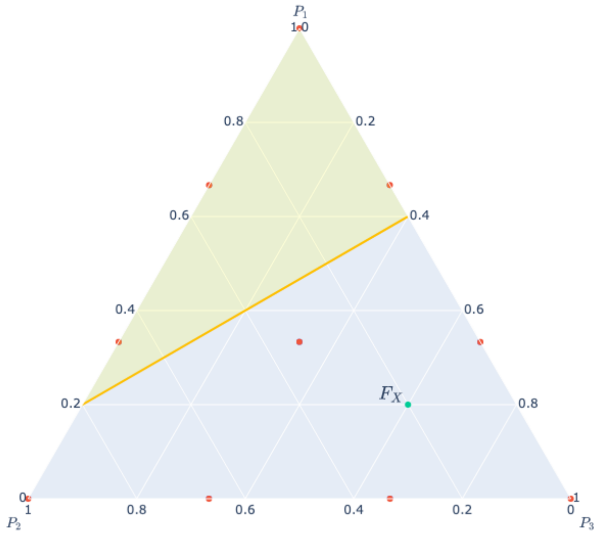

Figure 3: Distribution *types* for the 3 point space, $n = 3$

For $n = 10$ there are 110 distinct sequence types, that is, $|\mathcal{P}_{10}| = 110$. They are displayed in the figure below.

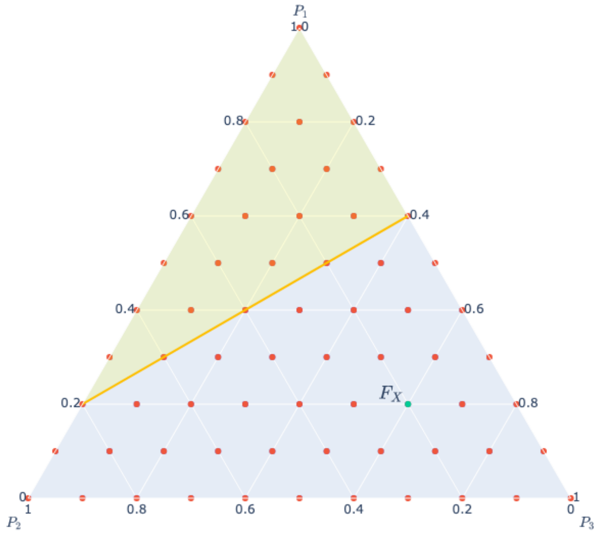

Figure 4: Distribution *types* for the 3 point space, $n = 10$

Note that each of the sequence types may contain many sequences. Because the *draws from the distribution F_X are i.i.d, all sequences of the same type have the same probability under P_X. The result in Sanov's theorem gives a finite sample upper bound on the probability that a sequence $X_l = (X_{1,l}, \cdots, X_{n,l})$, drawn from the joint distribution P_X^n belongs to the set E. For $n = 3$ there are only 4 sequence types that are in E, namely $(3,0,0)$, $(2,0,1)$, $(2,1,0)$, $(1,2,0)$. What is the probability associate to them? $P_X^3(x_l \in E) = 0.128$. On the other hand, we know that $\delta^*(\tilde{\tau}) =$ so Sanov's theorem gives the upper bound $e^{-3 \cdot 0.2492} = 0.474$ so the bound is fairly loose. On the other hand, when $n = 10$, 26 out of 110 sequence types fall in the set E. The total probability associated with those sequences is 0.0174. Sanov's theorem gives an upper bound of 0.0827. Finally for $n = 30$ $P_X^{30}(x \in E) = 0.000083$, while Sanov's bound gives $P_X^{30}(x \in E) \leq 0.00057$. The bound is known to be optimal in the exponent for $\lim n \to \infty$.*

2.5.2 Benchmarking robustness using census covariates

Several papers have proposed benchmarking the sensitivity to unobservable variables, which is often not computable, using observable variables. For example, Cinelli and Hazlett [2020] and Oster [2019] who use the explanatory power of observed covariates to benchmark for the explanatory power of unobserved covariates. This section suggests a similar approach for the robustness problem. In the context of this paper I would like to quantify whether a given value for the robustness parameter, δ^* is high or low. To this end I propose to leverage the subset of covariates in X_c, which are available in both the (quasi)-experimental environment and in the extrapolation environment to benchmark the robustness measure. At the population level it amounts to:

- computing the robustness metric δ^* through Equation (10)
- use the census information to compute $D_{KL}(P'_{X_c}||P_{X_c})$, the KL divergence between the distributions of the X_c covariates in the (quasi)-experimental population and the new population
- compare the two measures

If the variables in X_c collectively differ across the two environments by the same amount as X_e, observing $\delta^* > D_{KL}(P'_{X_c}||P_{X_c})$ suggests that the (quasi)-experimental claim can be extrapolated to the new environment. In words, it says that the distance, measured by the KL divergence between the observable census variables in the two environments would not be large enough to invalidate the claim drawn from the (quasi)-experimental evidence. In principle, one could develop a formal test that uses both δ^* and $D_{KL}(P'_{X_c}||P_{X_c})$ (under the assumption that the true distance in X_e is no larger than the true distance in X_c) to provide a pessimistic policy-maker with a clear rule on when to expand the policy given the census data. For now, transforming the heuristic exercise above in a full fledged two-sample test is beyond the scope of this paper and I leave it to future research.

2.6 A conditional limit theorem interpretation for F_X^*

We have seen that the value of $\delta^*(\tau)$ has a natural interpretation as a probability bound. What about the *least favorable distribution* F_X^*, the minimizer of Equation (4)? It turns out that an extension of the result by Sanov provides a new perspective for it. Adapting a version of Theorem 1 in Csiszár [1984], one obtains a striking result on the joint distribution of the data $(X_1, \cdots X_n)$:

Theorem 12. *(adapted from Csizar, 1984) Let Assumptions 2 - 4 hold. Set* $E = \{Q : \int_{\mathcal{X}} \tau(x) dQ \leq \tilde{\tau}\}$*, let* P_X *be the probability measure of i.i.d data. Denote the empirical distribution of* $X_1, \cdots, X_n$ *as* $\hat{F}_n$*. Then:*

(i) the random variables $X_1, \cdots, X_n$ *are asymptotically quasi-independent*[9] *conditional on the event that the empirical distribution* $\hat{F}_n \in E$

(ii) $P(X_i|\hat{F}_n \in E) \approx P^*(X_i)$ *for* $i = 1, \cdots, n$

Proof. The proof follows straightforwardly from Theorem 1 in Csiszár [1984] noting that, by Assumption 4, condition (2.18) in Csiszár [1984] is satisfied. For finitely supported X, an easier proof is given in Theorem 11.6.2 in Cover [1999]. □

In contrast to Theorem 11 which holds for any n, Theorem 12 is an asymptotic result: the approximation of the conditional law in ii) depends on the sample size n. The interpretation is the following, $P^{*n} := \Pi_{i=1}^{n} P^*$ is the approximate joint law of the covariates $X_1, \cdots X_n$, if we learned that the empirical distribution $\hat{F}_n$ does not satisfy the experimental conclusions. To visualize this, imagine drawing S-many repeated samples

[9] See Definition 2.1 in Csiszár [1984].

of n observations each from the covariate distribution. Then, combining the Sanov theorem in Section 2.5.1 together with the Csiszár [1984] conditional limit theorem tells us that:

(i) $\lim_{S\to\infty} \frac{1}{S}\sum_{l=1}^{S} \mathbb{1}[\hat{F}_{n,l} \in E] \leq e^{-n\delta^*(\tilde{\tau})}$

(ii) $P_X^n(X_i|\hat{F}_{n,l} \in E) \approx P^{*n}(X_i)$ for any $i = 1, \cdots, n$ and $l = 1, \cdots, S$

Part (i) says that the proportion of samples of size n that fail to satisfy the experimental evidence is bounded above by $e^{-n\delta^*(\tilde{\tau})}$. This interpretation is closest to the robustness approach in Broderick et al. [2020] which is based on dropping a percentage of the sample. The difference is that their procedure focuses on a proportion of the fixed sample, whereas this result concerns the proportion all possible samples of size n that could be drawn from the joint distribution of P_X^n. A small value for the robustness metric $\delta^*(\tilde{\tau})$ will not control this probability very well. Part (ii) gives an approximate law for the joint distribution P_X^n of the collection of samples that invalidate the experimental results. This tells us that the F_X^* is not just a by-product of the optimization problem in Equations (4) and (5) but it gives the approximate law of the data if we happen to draw a sample l which does not satisfy the experimental results.

3 Estimation and Asymptotic Results

In this section I introduce a semi-parametric estimator for my robustness metric δ^*, according to Definition 3 ii) and I characterize its asymptotic properties. I show that the robustness metric can be estimated using a GMM criterion function which only depends on the (quasi)-experimental distribution and on the CATE $\tau(x)$, both of which are identified in the quasi experiment. The theory is based on constructing the nonparametric influence function correction for the de-biased GMM procedure in Chernozhukov et al. [2020] to account for flexible nonparametric estimation of $\tau(x)$. The proofs are in the Appendix I.

3.1 An empirical estimate of the robustness metric δ^*

The closed form solution in Lemma 7 suggests a natural estimator based on empirical averages. In particular, one would like to replace Equation (10) with its sample analog using the Generalized Method of Moments (GMM) framework. Consider the quantities:

$$\nu_0 := \int_{\mathcal{X}} \exp(-\lambda_0(\tau(x) - \tilde{\tau}))dF_X(x)$$

where λ_0 is defined implicitly as the unique solution to:

$$\int_{\mathcal{X}} \exp(-\lambda_0(\tau(x) - \tilde{\tau}))(\tau(x) - \tilde{\tau})dF_X(x) = 0$$

The pair of parameters that solves the population moment condition is denoted by $\theta_0 = (\nu_0, \lambda_0)^T$. Ultimately, the robustness measure $\delta^* = -\log(\nu_0)$ is the parameter of interest. The asymptotic theory for δ^* follows directly from establishing the asymptotic theory for $\hat{\theta} = (\hat{\nu}, \hat{\lambda})^T$ hence, I will focus on these parameters in this section. The parameter space $\Theta \subseteq \mathbb{R}^2$ satisfies some constraints. First, observe that if the policy-maker's claim ($ATE \geq \tilde{\tau}$) holds with a strict inequality for the (quasi)-experimental distribution, then the true $\delta^* > 0$. This implies a restriction on $\nu_0 < 1$. Moreover, $\nu_0 > 0$ because by the properties of the exponential, the quantity $\exp(-\lambda(\tau(x) - \tilde{\tau}) > 0$ for all $x \in \mathcal{X}$. Hence, the restriction on ν is $0 \leq \nu_0 \leq 1$.

Let $W = (X, D, Y)$ be the data. Then, as in Newey and McFadden [1994] we can write the moment condition jointly for ν_0 and λ_0 as:

$$\mathbb{E}[g(W, \theta, \tau)] = \mathbb{E}\begin{bmatrix} \exp(-\lambda_0(\tau_0(X) - \tilde{\tau})) - \nu_0 \\ \exp(-\lambda_0(\tau_0(X) - \tilde{\tau}))(\tau_0(X) - \tilde{\tau}) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \tag{13}$$

where $\tau_0(X)$ denotes the true value of CATE. Assumptions 1–4 guarantee that the parameters of interest (λ_0, ν_0) are (globally) identified by Equation (13). Because the true value for $\tau_0(X)$ is an unknown but estimable population quantity, I consider a feasible version of Equation (13) that uses an estimate $\hat{\tau}(X)$ in place of $\tau_0(X)$. One could define the vector $\hat{\theta} = (\hat{\lambda}, \hat{\nu})^T$ is defined as the approximate solution to the empirical moment:

$$\mathbb{E}_n[g(W, \theta, \hat{\tau})] = \begin{bmatrix} \frac{1}{n}\sum_{i=1}^n \exp(-\hat{\lambda}(\hat{\tau}(X_i) - \tilde{\tau})) - \hat{\nu} \\ \frac{1}{n}\sum_{i=1}^n \exp(-\hat{\lambda}(\hat{\tau}(X_i) - \tilde{\tau}))(\hat{\tau}(X_i) - \tilde{\tau}) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \tag{14}$$

where $\hat{\tau}(X)$ is a plug-in estimate of the conditional average treatment effect. While Assumption 1 guarantees nonparametric identification of $\tau_0(X)$, there are many ways that one could estimate it, both parametrically and nonparametrically. For example Athey et al. [2016] uses random forest, Hsu et al. [2020] uses a doubly robust score function.

One caveat of the estimator based on Equation (14) is that the identifying moment conditions provided in Equation (13) are not Neyman orthogonal with respect to the first-step estimator $\hat{\tau}(X)$. As a result, the first-step estimation of $\hat{\tau}(X)$ can, in general, have a first-order effect on the estimator for $\theta_0 = (\nu_0, \lambda_0)^T$, and consequently on the estimator for δ^*, and possibly lead to incorrect inferences on the robustness metric, a general problem discussed in Chernozhukov et al. [2018]. Deriving primitive conditions on this form of the moment condition requires *ad-hoc* conditions on the first-step nonparametric estimator that can be hard or inconvenient to check in practice. As an alternative, I use the debiased-GMM approach in Chernozhukov et al. [2020] that allows to choose flexible estimators for $\tau_0(X)$ while automatically correcting for the first-order bias.

3.2 Nonparametric influence function correction and de-biased GMM estimator

In this section, I derive the nonparametric correction for the GMM estimator of θ based on Equation (14). I map the causal quantities like $\tau(X)$ to the statistical functionals that identify them and then explicitely construct the nonparametric influence function for these functionals. Because these functionals are always implicitly regarded as mapping the distribution function of the data, F, to some space, it is natural to index the functional with a subscript F. For example the $\tau(X) = \tau_F(X)$ because depends of the distribution of the data F. The true distribution of the data will be denoted as F_0 and it is understood that $\tau_0(X) = \tau_{F_0}(X)$. Recall that $\tau_{F_0}(X)$ is a causal parameter which needs to be identified through the distribution of the data. By Assumption 1, $\tau_{F_0}(X)$ can be nonparametrically identified as the difference between the conditional means: $\tau_{F_0}(X) = \gamma_{1,F_0}(X) - \gamma_{0,F_0}(X)$ where $\gamma_{1,F}(X) := \mathbb{E}_F[Y|X, D = 1]$ and $\gamma_{0,F}(X) := \mathbb{E}_F[Y|X, D = 0]$. The left hand side features a causal quantity while the right hand side features two statistical quantities. The first step then has two functions that need to be estimated. For convenience, I gather them into a single vector-valued statistical functional $\gamma_F = (\gamma_{0,F}, \gamma_{1,F})^T$. When considering the de-biasing term to correct for the first-step estimation of $\tau_{F_0}(X)$, we actually need to consider the first-step correction with respect to the full vector γ_F.

Now consider a parametric sub-model for the distribution function, consisting of $F_r := (1-r)\cdot F_0 + rH$ where F_0 is the true baseline distribution function of the data and H is an arbitrary distribution function which satisfies Assumption 1. For any $r \in [0,1], F_r$ is a mixture distribution and hence, it is also a valid distribution function. Moreover, if both F_0 and H satisfy Assumption 1 then F_r does as well. In order to de-bias the moment conditions in $\mathbb{E}[g(W,\theta,\tau_F)]$ with the approach of Chernozhukov et al. [2020] one needs to compute the nonparametric influence function with respect to τ_F. The nonparametric influence function maps infinitesimal perturbations of F in the direction of H in a neighborhood of F_0, to perturbations in $\mathbb{R}^2$ (because there are 2 moment conditions). It does so *linearly* in H. In particular, the nonparametric influence function of $\mathbb{E}[g(W,\theta,\tau_F)]$ with respect to F, labelled $\phi(\cdot)$ is implicitly defined by the equation below:

$$\left.\frac{d\mathbb{E}[g(W,\theta,\gamma_{F_r})]}{dr}\right|_{r=0} = \int \phi(w,\gamma_{F_0},\theta,\alpha)dH(w) \tag{15}$$

Note that, other than the original arguments of $g(\cdot)$, which feature the vector of conditional means γ_{F_0}, $\phi(\cdot)$ is allowed to depend on additional nonparametric components, gathered in $\alpha(\cdot)$. In the next result I derive the nonparametric influence function explicitly.

Proposition 13. *The de-biased GMM nonparametric influence function based on moment function $g(\cdot)$ is:*

$$\begin{aligned}\phi(w,\theta,\gamma_0,\alpha_0) = &\begin{bmatrix} \exp\left(-\lambda\cdot(\gamma_{1,F_0}(x)-\gamma_{0,F_0}(x)-\tilde{\tau})\right)\cdot(-\lambda) \\ \exp\left(-\lambda\cdot(\gamma_{1,F_0}(x)-\gamma_{0,F_0}(x)-\tilde{\tau})\right)\cdot\left(1-\lambda\cdot(\gamma_{1,F_0}(x)-\gamma_{0,F_0}(x)-\tilde{\tau})\right)\end{bmatrix}\\ &\times\left(\frac{d(y-\gamma_{1,F_0}(x))}{\pi_{F_0}(x)} - \frac{(1-d)(y-\gamma_{0,F_0}(x))}{1-\pi_{F_0}(x)}\right)\end{aligned}$$

which could be written in the form:

$$\begin{aligned}\phi(w,\theta,\gamma_0,\alpha_0) = &\begin{bmatrix} \exp\left(-\lambda\cdot(\gamma_{1,F_0}(x)-\gamma_{0,F_0}(x)-\tilde{\tau})\right)\cdot(-\lambda) \\ \exp\left(-\lambda\cdot(\gamma_{1,F_0}(x)-\gamma_{0,F_0}(x)-\tilde{\tau})\right)\cdot\left(1-\lambda\cdot(\gamma_{1,F_0}(x)-\gamma_{0,F_0}(x)-\tilde{\tau})\right)\end{bmatrix}\\ &\times\left(\begin{bmatrix}\alpha_{1,F_0}(x)\\ \alpha_{0,F_0}(x)\end{bmatrix}^T \begin{bmatrix} d(y-\gamma_{1,F_0}(x)) \\ (1-d)(y-\gamma_{0,F_0}(x))\end{bmatrix}\right)\end{aligned}$$

with $\alpha_{F_0}(x) := \begin{bmatrix} \alpha_{1,F_0}(x) \\ \alpha_{0,F_0}(x) \end{bmatrix} = \begin{bmatrix} \frac{1}{\pi_{F_0}(x)} \\ \frac{1}{1-\pi_{F_0}(x)} \end{bmatrix}$.

There are two main multiplicative terms in $\phi(\cdot)$. The first term is the derivative of the moment conditions with respect to the first-step estimator. The second one is the variation of individual treatment effects about their conditional mean, appropriately weighted by the propensity score. One can immediately check that, by the law of iterated expectations, $\mathbb{E}_F[\phi(W, \theta, \gamma_0, \alpha_0)] = 0$ for any θ. Hence we can form the de-biased GMM moment functions by taking:

$$\psi(w, \gamma, \theta, \alpha) = g(w, \theta, \gamma) + \phi(w, \theta, \gamma, \alpha) \tag{16}$$

Notice that $\mathbb{E}_{F_0}[\psi(W, \theta, \gamma_0, \alpha_0)] = 0$ so an estimator for θ that uses the de-biased moment function $\psi(\cdot)$ instead of $g(\cdot)$ will preserve identification. Standard conditions can be given to guarantee $\mathbb{V}[\psi(W, \theta, \gamma_0, \alpha_0)] < \infty$ so that $\psi(W_i, \theta, \gamma_0, \alpha_0)$ is a valid influence function. As emphasized in Chernozhukov et al. [2020] the de-biased GMM form of $\psi(\cdot)$ corrects for the first order bias induced by replacing $\gamma_{1,F_0} - \gamma_{0,F_0}$, the statistical counterpart of the true τ_{F_0}, with a flexibly estimated $\hat{\gamma}_1 - \hat{\gamma}_0$. In particular, for $\sqrt{n}$-consistency of θ, the estimators for $\hat{\gamma}_1$ and $\hat{\gamma}_0$ only need to satisfy mild conditions on the L^2-rate of convergence in Assumption 5 below. This allows to characterize simple inference for the robustness measure $\hat{\delta}^*$ while allowing for flexible nonparametric estimation of γ_{1,F_0} and γ_{0,F_0} using a large collection of machine learning-based estimators which include, among others, random forest, boosting, and neural nets. In practice, machine learning methods can help when the covariate space is high-dimensional but the true $\tau_0(X)$ has a sparse representation.

The key property to guarantee de-biasing is given by the Neyman orthogonality of the new moment conditions with respect to the first-step estimator, established in the result below.

Proposition 14. *Equation* (16) *satisfies Neyman orthogonality.*

Proof. See Appendix I. □

Consider now the empirical version of the de-biased GMM equations:

$$\hat{\psi}(\theta, \hat{\gamma}, \hat{\alpha}) = \frac{1}{K}\sum_{k=1}^{K}\frac{1}{|I_k|}\sum_{i\in I_k}\Big(g(W_i, \theta, \hat{\gamma}_{-k}) + \phi(W_i, \tilde{\theta}, \hat{\gamma}_{-k}, \hat{\alpha}_{-k})\Big)$$

The de-biased GMM estimator takes advantage of a cross-fitting procedure where the sample is split into K many folds. For each fold $k = 1, \cdots, K$, the nonparametric components in $\psi(\cdot)$, that is, the $\gamma(\cdot)$ and $\alpha(\cdot)$ functions, are estimated on the observations in the remaining $(K-1)$ folds which explains the indexing $-k$ in the subscripts of $\gamma(\cdot)$ and $\alpha(\cdot)$. Sample splitting reduces own-observation bias and, together with the Neymann orthogonality property established above, avoids complicated Donsker-type conditions that would potentially not be satisfied for some first-step estimators of $\hat{\gamma}$ and $\hat{\alpha}$, as discussed in Chernozhukov et al. [2020]. Finally note that $\tilde{\theta}$ is a consistent estimator for θ needed to evaluate ϕ. For example one could use the θ from the plug-in GMM which is consistent but may not be $\sqrt{n}$-consistent in general. The de-biased GMM estimator is given by:

$$\hat{\theta} = \arg\min_{\theta\in\Theta}\hat{\psi}(\theta, \hat{\gamma}, \hat{\alpha}) \tag{17}$$

To establish $\sqrt{n}$-convergence of the GMM estimators for θ, some quality conditions on the L_2 rates of convergence of the first-step estimators for γ and α are required.

Assumption 5. *For any k, $\|\hat{\gamma}_{-k} - \gamma_0\|_L^2 = o_P(N^{-\frac{1}{4}})$; $\|\hat{\alpha}_{-k} - \alpha_0\|_L^2 = o_P(1)$.*

In Appendix I, I use Assumptions 1 – 5 to prove the influence function representation for $\hat{\theta}$ to which a standard central limit theorem applies to establish the asymptotic normality of the de-biased GMM estimator for $\theta = (\nu, \lambda)^T$. This, in turn, allows to conduct inference on the parameter of interest, δ^* through a straightforward application of the delta method.

Theorem 15 (Asymptotic normality of θ)**.** *Let Assumptions 1–5. For $\hat{\theta}$ defined in*

Equation (17):

$$\begin{aligned}
\sqrt{n}(\hat{\theta}-\theta_0) &\xrightarrow{d} \mathcal{N}(0,S)\\
S &:= (G)^{-1}\Omega(G')^{-1}\\
G &:= \mathbb{E}[D_\theta\psi(w,\theta,\gamma_0,\alpha_0)]\\
\Omega &:= \mathbb{E}[\psi(w,\theta_0,\gamma_0,\alpha_0)\psi(w,\theta_0,\gamma_0,\alpha_0)^T]
\end{aligned}$$

and $D_\theta\psi(\cdot)$ is the Jacobian of the augmented moment condition with respect to the parameters in θ.

Proof. See Appendix I. □

The parameter of interest follows from a straightforward application of the parametric delta method.

Corollary 16 (Asymptotic normality of δ^*)**.** *Let $\hat{\delta}^* = -\log(\hat{\nu})$. Then*

$$\sqrt{N}(\hat{\delta}^*-\delta^*) \xrightarrow{d} \mathcal{N}\left(0,\frac{S_{11}}{\nu_0^2}\right)$$

where S_{11} is the (1,1) entry of the variance covariance matrix S in Theorem 15.

With the results of Theorem 15 one can obtain a point estimate δ^*, together with a confidence interval for a pre-specified coverage level. Because of the nature of the estimand, the researcher or the policy-maker, are likely to care especially about the lower bound for δ^*. This is because overestimating the δ^* implies that there is a distribution of the covariates within the estimated $\hat{\delta}^*$ that invalidates the policy-maker's claim. This defies the entire purpose of the robustness exercise. On the other hand, underestimating δ^* may result in unduly conservative characterization of the set of distributions for which the claim is valid, but it does not defy the purpose of the robustness exercise. A similar, asymmetric approach is followed by Masten and Poirier [2020] who report a lower confidence region for their breakdown frontier rather than a confidence band.

3.3 Reporting features of the *least favorable distribution*

Lemma 7 gives an explicit formula for the *least favorable distribution* F_X^* and shows that it depends on λ_0 and $\tau(X)$. Because of the interpretation of F_X^* as the conditional law of the data that we have given in Section F, the researcher may be interested in F_X^* directly. If $\mathcal{X} \subseteq \mathbb{R}^d$ is even moderately high dimensional, it may be very inconvenient to look at features of the estimated F_X^*. Moreover, the rate of convergence of as estimator of F_X^* can, in general, be nonparametric. This is because, under some conditions, it inherits the nonparametric rate of $\hat{\tau}(X)$. Rather, the researcher could report particular moments of F_X^* that are of interest. This exercise is analogous to reporting moments of the covariate distribution and compare them across treatment status to gauge at covariate balance, like in Rosenbaum and Rubin [1984]. The researchers may want to report moments of F_X^*, in addition to the robustness metric δ^*. For example they may want to report a vector of covariate means under the *least favorable distribution* F_X^* and compare it with the (quasi)-experimental distribution. In such a case, we would like to construct an estimator for the moments of interest and establish the asymptotic theory of these estimators. I give a convenient extension of Theorem 15, to include an arbitrary, finite dimensional collection $\zeta \in \mathbb{R}^s$ of moments of interest, along with the original θ parameters.

Theorem 17 (De-biased estimator of *least favorable* moments)**.** *Let $u : \mathbb{R}^d \to \mathbb{R}^s$, with $u \in (L^\infty(\mathcal{X}, \mu))^s$ for μ some dominating measure of P_X. Let $\zeta_0 = \mathbb{E}_{F_X^*}[u(X)] \in \mathbb{R}^s$. Define the following estimating equation for the parameters $(\hat{\theta}, \hat{\zeta})$, that is, the original parameters of interest, augmented by ζ, the additional moments of the least favorable distribution:*

$$\hat{\psi}^u(\theta, \zeta, \hat{\gamma}, \hat{\alpha}) := \frac{1}{K} \sum_{k=1}^{K} \frac{1}{|I_k|} \sum_{i \in I_k} \begin{bmatrix} g(W_i, \theta, \hat{\gamma}_{-k}) + \phi(W_i, \theta, \hat{\gamma}_{-k}, \hat{\alpha}_{-k}) \\ g^u(W_i, \theta, \zeta, \gamma_{-k}) + \phi^u(W_i, \theta, \zeta, \hat{\gamma}_{-k}, \hat{\alpha}_{-k}) \end{bmatrix}$$

where $g(\cdot), \phi(\cdot), \gamma(\cdot)$ and $\alpha(\cdot)$ are the same as in Propositions 13 – 27 and $g^u(\cdot)$ and $\phi^u(\cdot)$, whose values are vectors in $\mathbb{R}^s$ are defined below.

$$\begin{aligned} g^u(W_i, \theta, \zeta, \gamma) &= u(X_i) \exp(-\lambda(\tau(X_i) - \tilde{\tau}) - \nu \cdot \zeta \\ \phi^u(W_i, \theta, \zeta, \gamma, \alpha) &= u(X_i) \exp\left(-\lambda(\tau(X_i) - \tilde{\tau})\right) \cdot (-\lambda) \\ &\quad \times \left(\frac{D_i(Y_i - \gamma_1(X_i))}{\pi(X_i)} - \frac{(1 - D_i)(Y_i - \gamma_0(X_i))}{1 - \pi(X_i)} \right) \end{aligned}$$

$$(\hat{\theta}, \hat{\zeta}) := \arg \min_{(\theta,\zeta)\in\mathbb{R}^{s+2}} \hat{\psi}^u(\theta, \zeta, \hat{\gamma}, \hat{\alpha})^T \hat{\psi}^u(\theta, \zeta, \hat{\gamma}, \hat{\alpha}) + o_P(1) \tag{18}$$

Let Assumptions 1–5 hold. Then:

$$\frac{1}{\sqrt{n}} \sum_{k=1}^{K} \sum_{i\in I_k} \psi^u(W_i, \theta, \zeta, \hat{\gamma}_{-k}, \hat{\alpha}_{-k}) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi^u(W_i, \theta, \zeta, \gamma_0, \alpha_0) + o_P(1)$$

Moreover

$$\begin{aligned}
\sqrt{n}\left(\begin{bmatrix}\hat{\theta} - \theta_0 \\ \hat{\zeta} - \zeta_0\end{bmatrix}\right) &\xrightarrow{d} \mathcal{N}(0, S^u) \\
S^u &:= (G^u)^{-1}\Omega^u(G^{u'})^{-1} \\
G^u &:= \mathbb{E}[D_{\theta,\zeta}\ \psi^u(W, \theta, \zeta, \gamma_0, \alpha_0)] \\
\Omega^u &:= \mathbb{E}[\psi^u(w, \theta_0, \gamma_0, \alpha_0)^T \psi^u(w, \theta_0, \gamma_0, \alpha_0)]
\end{aligned}$$

where $D_{\theta,\zeta}$ denotes the Jacobian matrix with respect to the parameters θ and ζ.

Proof. The proof follows the same structure of Theorem 15 and is omitted. □

3.4 Simulation data

I conclude this section with a small Monte-Carlo exercise featuring three different data generating processes (DGPs) with increasing degrees of observable heterogeneity. To capture the idea of possibly high-dimensional experimental data, I consider a setting with $k = 100$ covariates, all independent and each distributed uniformly on $[0, 1]$ so that $\mathcal{X} = [0, 1]^k$. To reflect the fact that only a few out of all available experimental covariates are important to predict the treatment effect, I construct $\tau(x)$ to be sparse: $\tau(x)$ is a function of of only 1,3 and 10 out of 100 covariates in DGP1, DGP2 and DGP3 respectively. In each DGP, the potential outcomes also depend on an additive unobservable noisy error term.[10] To show that it is the heterogeneity that drives the

[10] In particular:

- DGP1: $Y_1 - Y_0 = \exp(X_1) + U_1 - U_0$;
- DGP2: $Y_1 - Y_0 = \exp(X_1) \cdot (X_2 + 0.5) \cdot (X_3 + 0.5) + U_1 - U_0$;
- DGP3: $Y_1 - Y_0 = \exp(X_1) \cdot (X_2 + 0.5) \cdot (X_3 + 0.5) \cdot \Pi_{j=4}^{10}(0.1 \cdot X_j + 0.95) + U_1 - U_0$.

(U_1, U_0) are uncorrelated normals with $\mu = 0, \sigma = 0.25$.

robustness, keeping the same baseline ATE for the three DGPs is fundamental. I choose the shape of $\tau(x)$ to induce the same ATE across the three DGPs, regardless of the heterogeneity of treatment effects, when evaluated with respect to the experimental distribution. I consider $M = 1000$ replications for each DGP and a sample size of $N = 10,000$. The first step $\tau(x)$ is estimated through K-fold cross-fitting, using either boosting or random forest to estimate $\gamma_1(x)$, $\gamma_0(x)$ and the propensity score $\pi_X(x)$. The number of trees and splitting criteria are tuned to the sample size through heuristic criteria. In practice one would use *within-fold* cross-validation to tune hyper-parameters. I estimate the implied $\hat{\delta}^*(\tilde{\tau})$, with $\tilde{\tau} = 1.3$ and evaluate its bias, variance and MSE against the true value δ^*. Fixing the ATE and the experimental distribution of the covariates guarantees that a change in the population value for δ^* is only capturing the change in heterogeneity. I report the estimates of δ^* using both the plug-in GMM and de-biased GMM approach below. Note that, because of K-fold cross fitting, the own-observation bias in the plug-in GMM is attenuated. Still, the de-biased, GMM shows very good bias improvements over the plug-in approach.

Table 1: Monte Carlo Simulation reports the DGP, the population value for the robustness metrics, ML estimator used for the nonparametric components and MSE, Bias and Variance. Sample size $n = 10,000$, number of simulations $M = 1000$.

Data	$\delta^*(\tilde{\tau})$	Method	$\gamma(\cdot)$, $\alpha(\cdot)$ est	MSE	Bias^2	Variance
DGP1	0.4485	plug-in	Random Forest	$3.7568 \cdot 10^{-4}$	$0.1235 \cdot 10^{-4}$	$3.6334 \cdot 10^{-4}$
			Boosting	$1.6311 \cdot 10^{-3}$	$1.2056 \cdot 10^{-3}$	$0.4255 \cdot 10^{-3}$
		de-biased	Random Forest	$3.7148 \cdot 10^{-4}$	$0.1030 \cdot 10^{-4}$	$3.6117 \cdot 10^{-4}$
			Boosting	$1.5278 \cdot 10^{-3}$	$1.1038 \cdot 10^{-3}$	$0.4240 \cdot 10^{-3}$
DGP2	0.1344	plug-in	Random Forest	$5.0716 \cdot 10^{-3}$	$4.9474 \cdot 10^{-3}$	$0.1242 \cdot 10^{-3}$
			Boosting	$1.1218 \cdot 10^{-3}$	$1.0622 \cdot 10^{-3}$	$0.0597 \cdot 10^{-3}$
		de-biased	Random Forest	$3.6640 \cdot 10^{-3}$	$3.5616 \cdot 10^{-3}$	$0.1024 \cdot 10^{-3}$
			Boosting	$0.7309 \cdot 10^{-3}$	$0.6749 \cdot 10^{-3}$	$0.0560 \cdot 10^{-3}$
DGP3	0.1328	plug-in	Random Forest	$5.2825 \cdot 10^{-3}$	$5.1558 \cdot 10^{-3}$	$0.1267 \cdot 10^{-3}$
			Boosting	$1.4637 \cdot 10^{-3}$	$1.3991 \cdot 10^{-3}$	$0.0646 \cdot 10^{-3}$
		de-biased	Random Forest	$3.8369 \cdot 10^{-3}$	$3.7326 \cdot 10^{-3}$	$0.1043 \cdot 10^{-3}$
			Boosting	$0.9312 \cdot 10^{-3}$	$0.8716 \cdot 10^{-3}$	$0.0596 \cdot 10^{-3}$

Table 1 report the results. First, observe the reduction in the population value of $\delta^*(\tilde{\tau})$ as heterogeneity increases in the DGP. This is entirely driven by an increase in the heterogeneity of $\tau(x)$ since the ATE is the same across the three DGPs. This means that a smaller shift in the covariates is required to invalidate the policy-maker claim ($ATE \geq 1.3$). As a result, the robustness metric decreases. Moving from DGP1 to

DGP2 and DGP3 the population value of the robustness metric drops from 0.4485 to 0.1344 to 0.1328. The decrease is most accentuated between DGP1 and DGP2 because of the functional form of $\tau(x)$.

In DGP1 we can see that the heuristic choice for the hyper-parameters in boosting likely results in under-fitting the data, leading to a bias an order of magnitude higher than the variance. For DGP1, the de-biasing procedure results in approximately 20% squared bias reduction which drives the reduction of approximately the same percentage in the Mean Squared Error. Variances are comparable between plug-in and de-biased GMM. The random forest procedure is better overall for MSE criterion. In DGP2, the bias dominate the variance component, suggesting both random forest and boosting are under-fitting. This is likely do to the absence of a *within-fold* cross-validation step. In this case,the de-biased GMM reduces the squared bias by about 40% for both random forest and boosting methods. The variances are again very similar across plug-in and de-biased and boosting has about half of the variance of random forest. DGP3's heterogeneity increases slightly, reducing the associated $\delta^*(\tilde{\tau})$. Like in DGP2, the bias dominates the variance component regardless of the first-step estimation method. Similarly, the de-biased GMM approach results in substantial bias reduction in comparison to the plug-in GMM approach.

4 Empirical Application: How robust are the effect of the Oregon Medicaid expansion?

In this section, I apply my approach to study the robustness of health insurance policy with respect to shifts in the distribution of covariates. The key reference is Finkelstein et al. [2012], which uses experimental data to study the effect of the Oregon Medicaid expansion lottery on health-care consumption and financial outcomes. The positive results of the study are of great interest for any policy-maker who is potentially interested in implementing a similar intervention in their state. Because the populations of recipients are likely to differ across states, I propose to complement the experimental result in Finkelstein et al. [2012] with an estimate of my robustness metric δ^* to quantify the smallest shift in important experimental covariates needed to eliminate the positive effects of the insurance lottery.

4.1 Institutional context and heterogeneity

Between March and September 2008, the state of Oregon conducted a series of lottery draws that would award the selected individuals the option to enroll in the Oregon Health Plan (OHP) Standard. OHP Standard is a Medicaid expansion program available for Oregon adult residents that are between 19 and 64 years of age and have limited income and assets. Finkelstein et al. [2012] studies the effect of the insurance coverage on a set of metrics that include health-care utilization (number of prescription, inpatient, outpatient and ER visits), recommended preventive care (cholesterol and diabetes blood test, mammogram and pap-smear test) and measures of financial strain (outstanding medical debt, denied care, borrow/default). The study uses both administrative and survey data but only the survey data is publicly accessible through Finkelstein [2013]. The Online appendix of Finkelstein et al. [2012] discusses a variety of robustness concerns that center on external validity. For example they note that scaling up the experiment can induce a supply side change in providers' behavior. Additionally, they acknowledge substantial demographic differences between the study population in Oregon *versus* the potential recipients in other states. These differences include, for example, a smaller African American and a larger white sub-population in Oregon versus other states. From the survey data it appears that the Oregon lottery participants are older and their health metrics under-performs the national average. If these covariates are important in determining the treatment effects of the health insurance, the results of Oregon experiment may not be robust to a change in the distribution of covariates. This robustness is especially important to quantify if the experimental results are to be extrapolated for policy adoption in other states. I stress the fact that, in this context, the re-weighting procedure in Hartman [2020] or Hsu et al. [2020] is not applicable because it lacks the survey-specific health data that are likely to be most predictive of treatment effect heterogeneity. Absent full covariate data form other states, I proposed to study the robustness of the policy by augmenting each of the treatment effect estimators in Finkelstein et al. [2012] with my robustness metric, which can be computed by exploiting the heterogeneity in the publicly available survey data Finkelstein [2013].

4.2 Robustness in the Oregon Medicaid Experiment

For the robustness exercise I focus on the Intention to Treat Effect (ITT) of the Oregon Medicaid Experiment lottery. As noted in Finkelstein et al. [2012], not all recipients who were awarded the option to enroll in the insurance program actually enrolled. For this reason Finkelstein et al. [2012] estimates both an ITT and a LATE estimate. One could argue that the ITT is the key parameter for a policy-maker interested in offering the same intervention. To map my framework to the application, recall that the ITT effect can be considered as an ATE where the treatment D is simply the "the option to enroll in the health insurance" so the robustness approach discussed in the paper carries over to the ITT with only notational changes. I consider hypotheses of the form $ITT_j \geq \tilde{\tau}$ or $ITT_j \leq \tilde{\tau}$ (depending on the outcome measure of interest) where j indexes a health-care utilization or a financial strain outcome, following the notation convention in Finkelstein et al. [2012]. As noted in Finkelstein et al. [2012] all health-care utilization outcomes are defined consistently so that a positive sign for ITT means an increase in utilization. Similarly, all financial strain outcomes are defined so that a negative sign for the ITT means a decrease in financial strain. I focus on 2 value of interest for $\tilde{\tau}$ for each of the outcome measures. One of the values is $\tilde{\tau} = 0$ which reflects the claim that the ITT is non-negative (for health-care utilization outcomes), or non-positive (for financial strain outcomes). The second value is $\tilde{\tau} = t_j = z_\alpha \sigma_j$ where σ_j is the standard deviation of the ITT for outcome j. t_j is the critical value for the t-statistic of a one sided test with null hypothesis $ITT \leq 0$ for some pre-specified α. As a result $\delta(t_j)$ proxies for the magnitude of a change in the covariate distribution that would make the ITT statistically not distinguishable from a non-positive or non-negative outcome (respectively).[11] Because σ_j is in general not available, in the empirical procedure I use $\hat{\sigma}_j$ in place of σ_j. The researcher interested in different hypothesis may adapt the procedure easily by specifying a $\tilde{\tau}$ with a value different from the two discussed above.

For the application I group the outcome measure in three groups: measure of health-care utilization, measures of compliance with recommended preventive care and measures of financial strain. I replicate the estimates of the intention to treat effect (ITT) for outcome variables in each of the three groups in Finkelstein et al. [2012] from a reduced form regression of the outcome variable on the lottery indicator and controls. The regression includes survey waves indicators, household size indicators and interaction terms between the two as controls. Because the regression is fully saturated, the estimates for the ITT are nonparametric. In my robustness exercise I focus on covariates that

[11] This interpretation is heuristic, in the sense that the standard deviation of the ITT estimate can depend on the distribution of the covariates as well. It is possible to impose an additional constraint on optimization problem, requiring that the variance of the treatment effects about the ITT remains the same. Such a construction fall into the case discussed in Appendix C.

appear critical for external validity and are likely to differ across states. Among others, Finkelstein et al. [2012] identifies gender, age, race, credit access, education and proxies for health status. To capture the potential heterogeneity, I estimate a Conditional Intention to Treat effect (CITT) with the set of covariates listed above.[12] Finally I use the estimated CITT to compute the measure of robustness δ^* for each of the outcome variables in the three categories and report it, together with the original ITT estimate, for both values of $\tilde{\tau}$ discussed above.[13] All outcomes are measured on the survey data Finkelstein [2013].

In Table 2, column 2, 3 and 4 contain respectively the experimental ITT for each outcome variable, the estimates for $\delta^*(0)$ and the estimates for $\delta^*(t_j)$. Here $t_j = \pm 1.645 \cdot \sigma_j$ depending on whether the experimental ITT is positive or negative. As an example, consider a measure of financial strain, like whether a patient had to borrow or skip a payment because of medical debt. The intention to treat effect is equal to -0.0515 with standard error 0.0060. $\delta^*(0) = 0.367$ represents the smallest distributional shift of the covariates that can induce an ITT equal to 0. The $\delta^*(t_j) = 0.265$ represents the smallest distributional shift in the covariates that can result in an $ITT = -1.645 \cdot 0.0060 = -0.0118$ which leads to not rejecting the hypothesis $H_0 : ITT \geq 0$. For any distributional shift that is smaller than $\delta^*(t_j)$ the statistical claim $H_0 : ITT \geq 0$ would be rejected.

I highlight two benefits of this robustness metric. First, it allows a comparison of the robustness across outcomes because each δ^* has the same units and it is measured on the same covariate space. Second, the fourth column of Table 2 has a natural interpretation as a breakdown point: what is the smallest perturbation of the distribution of covariates that will break statistical significance of the ITT? A policy-maker may consider findings with larger δ^* as more readily applicable to her own policy setting. From the δ metrics reported in Table 2 I notice that among the health-care utilization metrics, the ITT on outpatient visits is the most robust while the ITT on ER visits is the least robust. For the measures of financial strain the ITT on out of pocket expenses is the most robust and the ITT on instances of refused care because of medical debt is the least robust. If one had access to census data, one could choose a set of census variables of interest and compute the KL divergence between the distribution of the Oregon census variables

[12] From a technical standpoint, the CITT estimated with a discrete set of covariates is still a parametric estimator. In practice, it can be obtained by a fully saturated regression where the lottery indicator is interacted with all possible combinations of dummies.

[13] Comparable (survey weighted) ITT estimates can be found in column 2 labelled Reduced form, of 2. Discrepancies with the (unweighted) ITT effects I compute are due to survey weights.

Table 2: δ^* robustness metric for the health-care utilization and financial strain outcomes in Finkelstein et al. [2012]. ITT for measures of preventive care are indistinguishable from 0 for the experimental distribution so the robustness metric is trivial in this case. The measure is evaluated at $\tilde{\tau} = 0$ and $\tilde{\tau} = t_j = \pm 1.645\sigma_j$ for each outcome, depending on the relevant sign of the estimated ITT. The third group of outcomes, preventive care measures, all have statistically insignificant ITT, leading to a 0 robustness for all $\delta^*(t_j)$. I omit them in this table.

Outcome	Experimental ATE	$\delta^*(0)$	$\delta^*(t_j)$
health-care Utilization			
Prescriptions	0.1296 (0.044)	0.380 (0.007)	0.068 (0.002)
Out-patient visits	0.2986 (0.039)	1.552 (0.022)	0.965 (0.014)
ER visits	0.0064 (0.013)	0.009 (0.001)	0 n/a
In-patient visits	0.0081 (0.005)	0.119 (0.003)	0 n/a
Financial Strain			
Out of pocket expenses	-0.0622 (0.0069)	0.462 (0.030)	0.346 (0.023)
Outstanding expenses	-0.0529 (0.0070)	0.290 (0.0231)	0.204 (0.016)
Borrow/Skip payments	-0.0515 (0.0060)	0.367 (0.019)	0.265 (0.014)
Refused care	-0.011 (0.0040)	0.063 (0.006)	0.013 (0.002)

and a target state's census variables. Then the researcher use this computed measure to benchmark the magnitude of the robustness metrics in Table 1 to assess whether the magnitude of each δ^* is high or low, relative to the observed differences in the census variables.

5 Conclusion

Robustness of (quasi)-experimental findings is an importance premise of evidence based policy-making. In this paper I propose a metric δ^* to quantify the robustness of (quasi)-experimental findings with respect to a shifts in the distribution of the covariates. I focus on claims on aggregate policy effects of the type $(ATE \geq \tilde{\tau})$. While I focus on ATE as a main object of interest, the extension to other linear policy parameters is straightforward. I characterize my robustness metric as the minimal distance, in terms of KL divergence, between the set of covariate distributions that invalidate the claim and the (quasi)-experimental covariates. My robustness metric gives a nonparametric, one-dimensional summary that links treatment effect heterogeneity, (quasi)-experimental findings and covariate shifts. Because the computation of the δ^* robustness metric for ATE requires computing CATE, I employ the debiased-GMM approach to allow for CATE to be estimated using a large collection of machine learning techniques, which only need to satisfy mild requirements on their L^2 norm convergence rates. These include, for example, lasso, random forest, boosting, neural nets.

I apply my framework to assess the robustness of the results in Finkelstein et al. [2012] about the Oregon Medicaid Experiment. I consider a set of covariates including gender, race and lottery timing and find that the increase in outpatient visits and the decrease in out-of-pocket expenses are, respectively the most robust findings among the measure of health-care utilization and financial strain. For most other measures, relatively small shifts in the covariate distributions appear to invalidate the results.

References

J. G. Altonji, T. E. Elder, and C. R. Taber. Selection on observed and unobserved variables: Assessing the effectiveness of catholic schools. *Journal of political economy*, 113(1):151–184, 2005.

I. Andrews, M. Gentzkow, and J. M. Shapiro. Measuring the sensitivity of parameter estimates to estimation moments. *The Quarterly Journal of Economics*, 132(4):1553–1592, 2017.

B. Antoine and P. Dovonon. Robust estimation with exponentially tilted hellinger distance. *Journal of Econometrics*, 2020.

T. B. Armstrong and M. Kolesár. Sensitivity analysis using approximate moment condition models. *Quantitative Economics*, 12(1):77–108, 2021.

S. Athey, G. W. Imbens, and S. Wager. Approximate residual balancing: De-biased inference of average treatment effects in high dimensions. *arXiv preprint arXiv:1604.07125*, 2016.

S. Bonhomme and M. Weidner. Minimizing sensitivity to model misspecification. *arXiv preprint arXiv:1807.02161*, 2018.

T. Broderick, R. Giordano, and R. Meager. An automatic finite-sample robustness metric: Can dropping a little data change conclusions? *arXiv preprint arXiv:2011.14999*, 2020.

N. Cartwright and J. Hardie. *Evidence-based policy: A practical guide to doing it better*. Oxford University Press, 2012.

V. Chernozhukov, S. Lee, and A. M. Rosen. Intersection bounds: estimation and inference. *Econometrica*, 81(2):667–737, 2013.

V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins. Double/debiased machine learning for treatment and structural parameters. *Econometrics Journal*, 21(1):C1–C68, 2018.

V. Chernozhukov, J. C. Escanciano, H. Ichimura, W. K. Newey, and J. M. Robins. Locally robust semiparametric estimation, 2020.

T. Christensen and B. Connault. Counterfactual sensitivity and robustness. *arXiv preprint arXiv:1904.00989*, 2019.

C. Cinelli and C. Hazlett. Making sense of sensitivity: Extending omitted variable bias. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(1): 39–67, 2020.

T. M. Cover. *Elements of information theory*. John Wiley & Sons, 1999.

I. Csiszár. Sanov property, generalized i-projection and a conditional limit theorem. *The Annals of Probability*, pages 768–793, 1984.

A. Deaton. Instruments, randomization, and learning about development. *Journal of economic literature*, 48(2):424–55, 2010.

M. D. Donsker and S. S. Varadhan. Asymptotic evaluation of certain markov process expectations for large time, i. *Communications on Pure and Applied Mathematics*, 28 (1):1–47, 1975.

A. Finkelstein. Oregon health insurance experiment public use data, 2013.

A. Finkelstein, S. Taubman, B. Wright, M. Bernstein, J. Gruber, J. P. Newhouse, H. Allen, K. Baicker, and O. H. S. Group. The oregon health insurance experiment: evidence from the first year. *The Quarterly journal of economics*, 127(3):1057–1106, 2012.

M. Gechter. Generalizing the results from social experiments: Theory and evidence from mexico and india. *manuscript, Pennsylvania State University*, 2015.

E. Hartman. Generalizing experimental results. In J. Druckman and D. Green, editors, *Advances in Experimental Political Science*. Cambridge University Press, 2020.

P. Ho. Global robust bayesian analysis in large models. 2020.

J. L. Horowitz and C. F. Manski. Identification and robustness with contaminated and corrupted data. *Econometrica: Journal of the Econometric Society*, pages 281–302, 1995.

Y.-C. Hsu, T.-C. Lai, and R. P. Lieli. Counterfactual treatment effects: Estimation and inference. *Journal of Business & Economic Statistics*, pages 1–16, 2020.

P. J. Huber. A robust version of the probability ratio test. *The Annals of Mathematical Statistics*, pages 1753–1758, 1965.

G. W. Imbens and D. B. Rubin. *Causal inference in statistics, social, and biomedical sciences.* Cambridge University Press, 2015.

S. Jeong and H. Namkoong. Robust causal inference under covariate shift via worst-case subpopulation treatment effects. In *Conference on Learning Theory*, pages 2079–2084. PMLR, 2020.

E. H. Kennedy, S. Balakrishnan, M. G'Sell, et al. Sharp instruments for classifying compliers and generalizing causal effects. *Annals of Statistics*, 48(4):2008–2030, 2020.

A. E. Kowalski. Reconciling seemingly contradictory results from the oregon health insurance experiment and the massachusetts health reform. Technical report, National Bureau of Economic Research, 2018.

D. G. Luenberger. *Optimization by vector space methods.* John Wiley & Sons, 1997.

M. A. Masten and A. Poirier. Inference on breakdown frontiers. *Quantitative Economics*, 11(1):41–111, 2020.

R. Meager. Understanding the average impact of microcredit expansions: A bayesian hierarchical analysis of seven randomized experiments. *American Economic Journal: Applied Economics*, 11(1):57–91, 2019.

W. K. Newey and D. McFadden. Chapter 36 large sample estimation and hypothesis testing. volume 4 of handbook of econometrics, 1994.

E. Oster. Unobservable selection and coefficient stability: Theory and evidence. *Journal of Business & Economic Statistics*, 37(2):187–204, 2019.

E. Posner. Random coding strategies for minimum entropy. *IEEE Transactions on Information Theory*, 21(4):388–391, 1975.

P. R. Rosenbaum and D. B. Rubin. Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American statistical Association*, 79(387):516–524, 1984.

M. Sanger-Katz. Oregon health study: The surprises in a randomized trial. *The New York Times*, 2014.

J. W. Tukey. A survey of sampling from contaminated distributions. *Contributions to probability and statistics*, pages 448–485, 1960.

A. W. Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

M. J. Williams. External validity and policy adaptation: From impact evaluation to policy design. *The World Bank Research Observer*, 35(2):158–191, 2020.

A Another look at the Lagrange multiplier λ

The formulation of the optimization problem in Equation (4) concerns a policy-maker who wishes to maintain the claim $ATE \geq \tilde{\tau}$ so that the constraint set in Equation (5) takes the opposite direction of the inequality. The formulation with the Lagrange multiplier in Equation (9) is without loss of generality. If the policy-maker is interested in maintaining a claim of the type $ATE \leq \tilde{\tau}$, the Lagrange multiplier would enter Equation (9) with a negative sign, or equivalently, if we want to preserve Equation (9), the value of the Lagrange multiplier would be negative.

The Lagrange multiplier λ in Equation (9) can give insight in what happens moving from the experimental distribution to the *least favorable distribution*. Note that λ has the opposite sign as the difference between the (quasi)-experimental ATE and the target ATE. To see this, we consider how the target ATE relates to the CATE. For each given $\tilde{\tau}$ there is a partition of the covariate support $\mathcal{X}$ into three sets, depending on what will be down-weighted or up-weighted by the *least favorable distribution*. The weight is given by:

$$w(x) = \frac{\exp(-\lambda(\tau(x) - \tilde{\tau}))}{\int_{\mathcal{X}} \exp(-\lambda(\tau(x) - \tilde{\tau}))dF_X}$$

so we see, after simplifying, that $w(x) = 1$, i.e dF_X^* and dF_X coincide, iff:

$$\exp(-\lambda\tau(x)) = \int_{\mathcal{X}} \exp(-\lambda\tau(x))dF_X$$

so the three sets are given by:

$$\begin{aligned}
\mathcal{X}^- &:= \{x \in \mathcal{X} \ s.t. \ \exp(-\lambda\tau(x)) - \mathbb{E}_{P_X}[\exp(-\lambda\tau(x))] < 0\} \\
\mathcal{X}^+ &:= \{x \in \mathcal{X} \ s.t. \ \exp(-\lambda\tau(x)) - \mathbb{E}_{P_X}[\exp(-\lambda\tau(x))] > 0\} \\
\mathcal{X}^0 &:= \{x \in \mathcal{X} \ s.t. \ \exp(-\lambda\tau(x)) - \mathbb{E}_{P_X}[\exp(-\lambda\tau(x))] = 0\}
\end{aligned}$$

For example, suppose that the researcher wants to support the claim $ATE \geq \tilde{\tau}$, which holds for the *experimental* ATE. Then, in order to achieve a lower ATE the *least favorable distribution* will have to shift weight from $\mathcal{X}^+$ to $\mathcal{X}^-$. These sets in the partition will in general not coincide with the sets $\{x \in \mathcal{X} \ s.t. \ \tau(x) - \tilde{\tau} < 0\}$, $\{x \in \mathcal{X} \ s.t. \ \tau(x) = \tilde{\tau}\}$

and $\{x \in \mathcal{X} \text{ } s.t. \text{ } \tau(x) < \tilde{\tau}\}$. One case when they coincide is when F_X follows the normal distribution.

B Relating parametric forms of least favorable distributions with assumptions on CATE

Lemma 7 gives a solution to the policy-maker's problem that does not depend on a specific functional form for CATE nor on a parametric assumption for the experimental distribution F_X. Leveraging the closed form solution I show that if the conditional treatment effect function does follow a particular form and the experimental distribution belongs to a certain parametric family, we can guarantee that the *least favorable distribution* belongs to the same parametric family, up to a shift in the parameters.

Definition 18. *We say that a class of parametric distributions indexed by θ, denoted F_X^θ is **least-favorable closed** with respect to a parametric class of Conditional Average Treatment Effects, $\tau(x)_\eta$, indexed by $\eta \in H$ if for any θ and η, the least favorable distribution $F_X^* = F_X^{\theta^*}$ for some $\theta^* \in \Theta$. The choice of θ^* will in general also depend on features of η as well.*

This means that the *least favorable distribution* belongs to the same parametric class as the original, experimental distribution. This idea is similar to the conjugate prior construction where the posterior distribution belongs to the same class of priors if the likelihood is within a conjugate parametric class. The distributional shift can then be thought of as a parameter shift.

Proposition 19 (Quadratic-Normal least favorable closed-ness)**.** *The parametric class $\mathcal{N}(\mu, \sigma^2)$ is **least favorable closed** for quadratic Conditional Average Treatment Effects. That is, if $X \in \mathbb{R}^k$ follows the multivariate normal distribution $X \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ where $\boldsymbol{\Sigma}$ is p.d. and $\tau(x) = \boldsymbol{x^T A x} + \boldsymbol{x^T \beta} + c$ for $\boldsymbol{\beta} \in \mathbb{R}^k$ then F_X^* is the measure induced by $X^* \sim \mathcal{N}(\boldsymbol{\mu^*}, \boldsymbol{\Sigma^*})$ with $\boldsymbol{\mu}^* = (\boldsymbol{\Sigma}^{-1} + 2\lambda \boldsymbol{A})^{-1}(\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} - \lambda\boldsymbol{\beta})$ and $\boldsymbol{\Sigma^*} = (\boldsymbol{\Sigma}^{-1} + 2\lambda\boldsymbol{A})^{-1}$, provided that $(\boldsymbol{\Sigma}^{-1} + 2\lambda\boldsymbol{A})^{-1}$ is p.d. The parameter λ is defined as in Equation (9).*

Proof. See Appendix I. □

Corollary 20 (Linear-Normal least favorable closed-ness)**.** *If* $\tau(x) = \boldsymbol{x}^{\boldsymbol{T}}\boldsymbol{\beta}$ *and* $X \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ *then* $X^* \sim \mathcal{N}(\boldsymbol{\mu}^*, \boldsymbol{\Sigma})$ *where* $\boldsymbol{\mu}^* = \boldsymbol{\mu} - \lambda\boldsymbol{\Sigma}\boldsymbol{\beta}$*.*

Proof. Follows from 19 when $\boldsymbol{A} = 0$. □

An extension of Proposition 19 could be shown to hold for the more general class of distributions in the exponential family given by $f(x|\theta) = g(\theta)h(x)\exp(\eta(\theta)^T T(x))$ but this is beyond the scope of this paper. The parametric example gives some additional insights in the geometry of the policy-maker's problem, which could also help to understand the analytical expression for the least favorable distribution above. Consider the univariate case ($d = 1$) where F_X is the normal distribution with mean μ and standard deviation σ. The policy-maker's desired claim is $ATE \geq 0$. The conditional average treatment effects are linear in the only covariate, that is $\tau(x) = \pi X$ for some $\pi \in \mathbb{R}$. Because CATE is linear in X, the ATE is only a function of the population mean μ. As a result, the feasible set of the policy-maker's problem in Figure 5 is the half space $\mu \leq 0$. Proposition 19 allows us to reduce the problem to a finite dimensional problem which we can solve with the usual KKT conditions. Observe that $D_{KL}(\mathcal{N}(\mu^*, \sigma^*)||\mathcal{N}(\mu, \sigma)) = \frac{1}{2}\left(\log\left(\frac{\sigma^2}{\sigma^{*2}}\right) + \frac{\sigma^{*2}}{\sigma^2} - 1 + \frac{1}{\sigma^2} \cdot (\mu - \mu^*)^2\right)$. In that case:

$$\begin{aligned} \min_{(\mu^*, \sigma^*) \in \mathbb{R} \times \mathbb{R}_+} \quad & \frac{1}{2}\left(\log\left(\frac{\sigma^2}{\sigma^{*2}}\right) + \frac{\sigma^{*2}}{\sigma^2} - 1 + \frac{1}{\sigma^2} \cdot (\mu - \mu^*)^2\right) \\ s.t. \quad & \pi\mu^* \leq \tilde{\tau} \end{aligned}$$

where the constraint is simplified because of the linear functional form of CATE and linearity of the expectation operator.

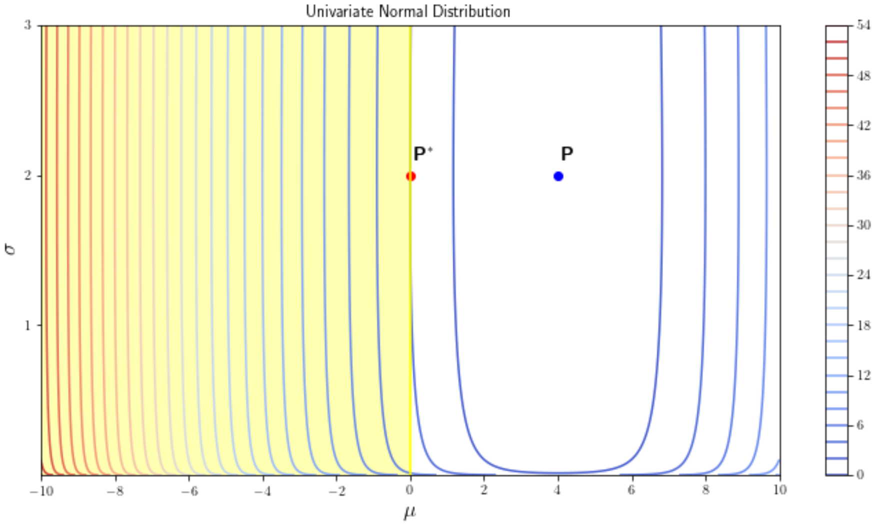

Figure 5: Univariate Normal Distribution, Linear CATE. Each point in the graph represents a normal distribution parametrized by its mean and standard deviation $\mathcal{N}(\mu, \sigma^2)$. The starting distribution, the experimental is taken to be $\mathrm{P} = \mathcal{N}(4, 2)$. The contour lines represent the KL divergence with respect to the experimental distribution. The policy-maker's desired claim is $ATE \geq 0$. The feasible set shaded in yellow represents all univariate normal distributions that satisfy $ATE \leq 0$. When CATE is linear (that is $\tau(x) = \pi X$), the only parameter that contributes to the ATE is the mean μ so the feasible set is parallel to the σ axis. As a result, the *least favorable* distribution, labelled as P^*, amounts to a mean shift from $\mu = 4$ to $\mu^* = \frac{\tilde{\tau}}{\pi} = 0$ and no shift in the σ parameter.

The KKT conditions imply:

$$\begin{aligned}
\mu^* &= \mu - \lambda \pi \sigma^2 \\
\sigma^* &= \sigma \\
\lambda &= \frac{1}{\pi \sigma^2}\left(\mu - \frac{\tilde{\tau}}{\pi}\right)
\end{aligned}$$

The *least favorable distribution* amounts to a mean shift of the prescribed magnitude and no change in the variance. Contrast the example above with the case where the CATE is allowed to be quadratic. Proposition 19 still applies, hence the problem can still be formulated as minimizing over the paramteric space (μ^*, σ^*). This time though, the variance of the covariate X matters in determining the ATE and the feasible set reflects this.

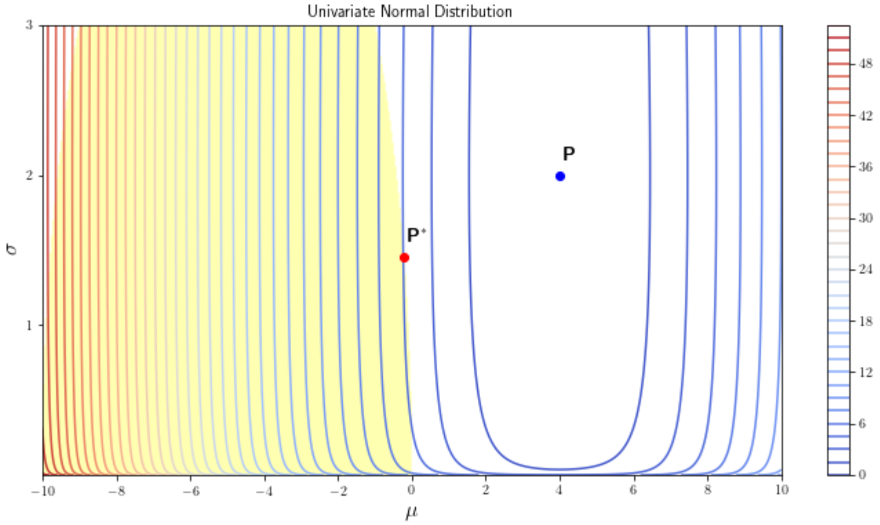

Figure 6: Univariate Normal Distribution: Quadratic CATE. The setting is identical as in Figure 5 but here is quadratic, $\tau(x) = 0.8 \cdot X^2 + 8 \cdot X$. As a result, $ATE(\mu, \sigma) = 0.8 \cdot (\mu^2 + \sigma^2) + 8\mu$ so both parameters of the covariate distribution contribute to determining the ATE. The feasible set in yellow has nonflat curvature. The *least favorable distribution*, labelled as P*, features a parameter shift in both the mean and the variance.

C Constrained Classes

An applied researcher may wish to restrict the class of distributions for problem 1 by imposing additional constraints. For example, they may want to fix the certain moments of the experimental distribution.[14] The computational price to pay for each additional constraint is one additional Lagrange multiplier per constraint as detailed out in Ho [2020]. For example, for a known moment function $q : \mathcal{X} \to \mathbb{R}$ we want:

$$\int_{\mathcal{X}} q(X) dF_X = \int_{\mathcal{X}} q(X) dF'_X$$

This requirement restricts the space of feasible probability distributions because it asks that the *least favorable distribution* preserves the additional moment. From the perspective of robustness, the value of the problem δ^* for the constrained problem must be

[14] Note that finitely many moment restrictions would still amount to searching the KL infimum within a infinite dimensional class of probability distributions, and, as such, the nonparametric nature of the problem persists.

larger or equal than the value for the unconstrained problem. That is:

$$\begin{aligned}
\inf_{dF'_X:\ dF'_X \ll dF_X; P'_X(\mathcal{X})=1} D_{KL}(F'_X||F_X) \quad &\leq \inf_{F'_X:\ F'_X \ll F_X; P'_X(\mathcal{X})=1} D_{KL}(F'_X||F_X) \\
s.t. \int_{\mathcal{X}} \tau(x) dF'_X(x) \leq \tilde{\tau} \quad & \qquad \int_{\mathcal{X}} \tau(x) dF'_X(x) \leq \tilde{\tau} \\
& \qquad \int_{\mathcal{X}} q(x) dF'_X(x) = q
\end{aligned}$$

Assumptions about moment preservation aides the robustness to external validity of a causal claim.

In case of additional constraints the solution to the KL problem takes the form:

$$\frac{dF^*_X}{dF_X} = \frac{\exp(-\lambda(\tau(x)-\tilde{\tau}))\prod_{l=1}^{L}\exp(-\mu_l(q(x)-\tilde{q}))}{\int_{\mathcal{X}}\exp(-\lambda(\tau(x)-\tilde{\tau}))\prod_{l=1}^{L}\exp(-\mu_l(q(x)-\tilde{q}))}$$

and each Lagrange multiplier can be solved by:

$$\int_{\mathcal{X}} \exp(-\mu_l(q(x)-\tilde{q}))(q(x)-\tilde{q}) dF_X = 0$$

For estimation, the additional restrictions result in L many additional parameters, one for each Lagrange multiplier that needs to be computed. One could adapt the estimation framework in Section 3 and have $\theta \in \Theta \subseteq \mathbb{R}^{L+2}$ gathers the original parameters α and λ as well as the Lagrange multipliers for the population optimization problem $\mu_1, \mu_2, \cdots, \mu_L$. At the cost of a more cumbersome notation, all the asymptotic results in Section 3 apply.

D Partial identification of CATE

In this section, I consider the case where the main ingredient needed to identify the robustness metric, $\tau(x)$ is only partially identified. This situation is important in practice. For example, with one-sided noncompliance $\tau(x)$ is only partially identified. In this section I will show that one can still recover bounds for $\delta^*(\tilde{\tau})$ that are robust to this partial identification.

In section 2.2, the covariate shift assumption allowed us to write the ATE as a linear functional of the covariate distribution, greatly simplifying the treatment. This linear functional is fixed because $\tau(x)$ is identifiable.

Suppose we can set identify $\tau \in \mathcal{T}$. For example $\tau(x)$ could be identified up to a finite dimensional parameter or one could have an identification region where any $\tau \in \tau$ satisfies $\underline{\tau}(x) \leq \tau(x) \leq \overline{\tau}(x)$, that is, there are identification bands bounding any $\tau \in \mathcal{T}$ above and below. Then we can compute a conservative version of the robustness metric define below:

$$\underline{\delta}^*(\tilde{\tau}) := \inf_{\tau \in \mathcal{T}} \inf_{dF'_X:\ dF'_X \ll dF_X; dF'_X(\mathcal{X})=1} D_{KL}(F'_X || F_X) \quad s.t. \ \int_{\mathcal{X}} \tau(x) dF'_X(x) \leq \tilde{\tau}$$

Because now $\tau(\cdot)$ is not identified, the problem above considers the least favorable among the ones in the set $\mathcal{T}$. Because τ controls the shape of the feasible set we can rewrite it as

$$\underline{\delta}^*(\tilde{\tau}) := \inf_{dF'_X:\ dF'_X \ll dF_X; dF'_X(\mathcal{X})=1} D_{KL}(F'_X || F_X) \quad s.t. \ \int_{\mathcal{X}} \tau(x) dF'_X(x) \leq \tilde{\tau} \text{ for some } \tau \in \mathcal{T}$$

Now regard the constraint set as a collection of $\mathcal{F}_\tau := \{F'_X : \int_{\mathcal{X}} \tau(x) dF'_X(x) \leq \tilde{\tau}\}$ for a given τ. It is immediate to notice that, if $\tau'(x) \leq \tau(x)$ point-wise, then $\mathcal{F}_\tau \subseteq \mathcal{F}_{\tau'}$. That is, if a CATE that is dominated point-wise (or in fact F_X almost everywhere) the constraint set admits a larger class of distributions. As a result, for $\underline{\tau}$ we have, for any $\tau \in \mathcal{T}$, $\mathcal{F}_\tau \subseteq \mathcal{F}_{\underline{\tau}}$. But this greatly simplifies the problem since now it is enough to write:

$$\underline{\delta}^*(\tilde{\tau}) := \inf_{dF'_X:\ dF'_X \ll dF_X; dF'_X(\mathcal{X})=1} D_{KL}(F'_X || F_X) \quad s.t. \ \int_{\mathcal{X}} \underline{\tau}(x) dF'_X(x) \leq \tilde{\tau}$$

so now the problem can be solved for the lower bound of the identified set. Again, this interpretation of delta amounts to considering robustness to the lack of identification the CATE. A similar argument applies for the reverse inequality ($ATE \leq \tilde{\tau}$) and $\overline{\tau}$.

E Re-evaluating policies over time

In the main paper, the policy-maker is concerned with extrapolating experimental results to different policy contexts. In the application, this takes the form of extrapolating the Medicaid extension program to other states. In this section I show that we can have an alternative interpretation that emphasizes changes over time rather than across regions. According to this interpretation, the measure of robustness δ^* captures the minimal change in demographic trends that is needed to invalidate a particular policy conclusion.

Consider a time horizon $t = -1, 0, 1, 2, \cdots, T$. Suppose that a policy is implemented at time 0. For the covariate distribution at time 0, $F_{X,0}$ the policy meets the target $\tilde{\tau}$, that is, $ATE_{F_{X,0}} \geq \tilde{\tau}$. Now, we may worry that over time, the covariate distribution might change from F_0 in such a way that does not justify the policy any longer.

How does the covariate shift assumption translate in thus context? It requires that the causal effect $\tau_{F_{X,0}}(\cdot) = \tau_{F_{X,t}}(\cdot)$ for all $t = 1, 2, \cdots, T$. That is, the CATE for whichever time horizon it is defined, does not change for new cohorts who are newly treated.

Here, a natural benchmark for comparison is given by the change between the reference point and the pre-policy period $t = -1$. This benchmark is given by $\delta_{benchmark} = D_{KL}(F_{X,-1}||F_{X,0})$. In this case, if one finds $\delta^*(\tau) > \delta_{benchmark}$ then the policy-maker may be comforted by observing that the amount of variation needed to invalidate the claim is larger than the natural variation that can be elicited from the time trends. Of course, one could decide to formalize this notion since we could seek to jointly characterize the asymptotic distribution of the vector of estimators $(\hat{\delta}^*(\tilde{\tau}), \hat{\delta}_{benchmark})^T$ which is beyond the scope of this paper.

F An interpretation of the robustness metric based on Sanov's theorem

In this section, I discuss some further details on the interpretation of $\delta^*(\tau)$ based on Sanov's theorem the I have introduced in Section F. The treatment in this section will be restricted to a covariate space $\mathcal{X}$ that is supported on finitely many points, reflecting the discussion of the method of *types* in Cover [1999]. Suppose there are k covariates, $X_1, \cdots, X_k$, each taking values in $\mathcal{X}_j$ with $|\mathcal{X}_j|$ finite. Let $m := \sum_{j=1}^{k} |\mathcal{X}_j|$. The set of probability distributions on $\mathcal{X} = \Pi_{j=1}^{k} \mathcal{X}_j$ can be identified with the $(D-1)$-dimensional simplex in $\mathbb{R}^m$. For a fixed sample size n consider the set of all sequences of data $x = (x_1, x_2, \cdots, x_n)$ taking values in $\mathcal{X}^n$ and define the *type* P_x of a sequence x as the relative proportion of each possible realization a in $\mathcal{X}$, that is, $P_x(a) = \frac{N(a|x)}{n}$ where $N(a|x)$ is the number of times realization a shows up is in sequence x. Let $\mathcal{P}_n$ be the collection of types. Cover [1999] version of Sanov theorem allows for a more general set E, not necessarily convex at the price of an additional multiplicative polynomial term $(n+1)^m$ in the number of observation. If the set E is a convex set, the upper bound can be tightened to $P^n(E \cap \mathcal{P}_n) \leq e^{-nD_{KL}(P^*||P)}$ and the polynomial term in n is dropped. Note that $E := \{Q : \int_{\mathcal{X}} \tau(x) dQ \leq \tilde{\tau}\}$ is obtained through imposing a linear restriction on Q and therefore E is convex. Sanov's theorem remains true for larger classes of probability distributions, not necessarily confined to finitely supported X like discussed in Csiszár [1984] but the method of types leans itself for a discussion on discrete spaces.

G Some additional results

Proposition 21. *Let $\epsilon > 0$. Then for $\tilde{\tau} > \inf_{\mathcal{X}} \tau(x) + \epsilon$, $\delta^*(\tilde{\tau})$ in Definition 3 is decreasing in $\tilde{\tau}$.*

Proof. First denote the feasible set $E(\tilde{\tau}) := \{F_X \in \mathcal{F} : \int_{\mathcal{X}} \tau(x) dF_X(x) \leq \tilde{\tau}\}$. Then, $G_X \in E(\tilde{\tau}) \iff \int_{\mathcal{X}} \tau(x) dF_X(x) \leq \tilde{\tau} < \tilde{\tau}'$ for any $\tilde{\tau}' > \tilde{\tau}$ so $G_X \in E(\tilde{\tau}')$. But then $E(\tilde{\tau}) \subseteq E(\tilde{\tau}')$. Hence, because we are minimizing on a larger set of distributions $\delta^*(\tilde{\tau}) := \inf_{G_X \in E(\tilde{\tau})} D_{KL}(G_X||F_X) \geq \inf_{G_X \in E(\tilde{\tau}')} D_{KL}(G_X||F_X) =: \delta^*(\tilde{\tau}')$. If the feasible set E has the reverse inequality, it follows immediately that $\delta^*(\tau)$ is increasing in $\tilde{\tau}$.

□

H General φ-divergence metrics and *least favorable* closed classes.

In this section I extend the theory of least favorable classes by considering different φ divergence measures. To this end I leverage the thorough treatment of φ divergences in Christensen and Connault [2019]. The Kullback-Leibler divergence is a special case of a more general construction, known as φ-divergence. It is introduced below:

Definition 22 (φ-divergence)**.** *Consider the φ-divergence between F_X and F'_X given by:*

$$D_\varphi(F'_X||F_X) := \int \varphi\left(\frac{dF'_X}{dF_X}\right) dF_X \tag{A.19}$$

where φ is a convex function with $\varphi(1) = 0$ and $\frac{dF'_X}{dF_X}$ is the Radon-Nikodym derivative of the probability distribution F'_X with respect to the probability distribution of F_X, provided that $P'_X \ll P_X$ for the respective measures. For example the choices $\varphi(t) = t \log t$ and $\varphi(t) = \frac{1}{2}|t-1|$ give rise to the KL-divergence and to the total variation divergence (TV) respectively.

There may be a reason to choose a different φ-divergence metric instead of the KL-divergence. Under suitable conditions, the construction of the proposed robustness metric will change in magnitude, since now the (pseudo)-metric on the space of distributions of the covariates is different. A closed form solution analogous to Lemma 7 is available. The characterization of the δ^* now depends on $\varphi(\cdot)$. In particular it is fully characterized in terms of the Fenchel-conjugate of φ and its derivative.

Definition 23 (Fenchel-Conjugate)**.** *Given a topological vector space X and convex function $\varphi : X \to \mathbb{R}$, the Fenchel-conjugate $\varphi^* : X^* \to \mathbb{R}$, defined on the dual space of X, is defined by:*

$$\varphi^* : x^* \mapsto \sup_{x \in X} \langle x^*, x\rangle - \varphi(x) \tag{A.20}$$

Then we can have a generalization of the policy-maker's problem in Equations (4) and (5) for an arbitrary φ divergence in 22:

$$\inf_{P'_X:\ P'_X \ll P_X; P'_X(\mathcal{X})=1} D_\varphi(F'_X||F_X) \tag{A.21}$$

$$s.t. \int_{\mathcal{X}} \tau(x) dF'_X(x) \leq \tilde{\tau} \tag{A.22}$$

From the KKT Theorem (Theorem 1, Ch.8, Sec. 3 in Luenberger [1997]) we can write the problem as:

$$\sup_{\lambda\in\Lambda}\sup_{\xi}\left(\inf_{P'_X:\ P'_X\ll P_X;P'_X(\mathcal{X})=1} D_\varphi(F'_X||F_X)+\lambda\int_{\mathcal{X}}(\tau(x)-\tilde{\tau})dF'_X(x)+\xi\left(\int_{\mathcal{X}}dF'_X-1\right)\right) \tag{A.23}$$

where and ξ is the Lagrange multiplier for integration to unity, λ is the Lagrange multiplier for the policy-maker's claim. The convexity conditions for Theorem 1, Ch.8, Sec. 3 in Luenberger [1997] are immediate to verify. The interior condition, analogous to a Slater condition, is satisfied by Assumption 4. Note that convex cone where the Lagrange multiplier takes values is $\mathbb{R}_+$ (or $\mathbb{R}_-$ if the policy-maker's claim is $ATE\leq\tilde{\tau}$ instead). In Equation (9) the Lagrange multiplier λ is a 1-dimensional parameter. Notice that after fixing the experimental distribution, $D_{KL}(\cdot||F_X)$ is convex in its first argument. With a careful rewriting we can express the inner problem as:

$$\inf_{P'_X:\ P'_X\ll P_X;P'_X(\mathcal{X})=1}\int_{\mathcal{X}}\left(\varphi\left(\frac{dF'_X}{dF_X}(x)\right)-(-\lambda(\tau(x)-\tilde{\tau})-\xi)\frac{dF'_X}{dF_X}(x)\right)dF_X(x)-\xi$$

and recognize that, if we can pass the infimum under the integral sign, we can substitute the expression for the Fenchel-conjugate of φ, switching the sign of the infimum.

$$\inf_{P'_X:\ P'_X\ll P_X;P'_X(\mathcal{X})=1}\int_{\mathcal{X}}\left(\varphi\left(\frac{dF'_X}{dF_X}(x)\right)-(-\lambda(\tau(x)-\tilde{\tau})-\xi)\frac{dF'_X}{dF_X}(x)\right)dF_X(x)-\xi$$
$$=-\int_{\mathcal{X}}\varphi^*(-\lambda(\tau(x)-\tilde{\tau})-\xi)dF_X(x)-\xi$$

Substituting this back into the outside problem one obtains:

$$\sup_{\lambda\in\Lambda}\sup_{\xi}\int_{\mathcal{X}}-\varphi^*(-\lambda(\tau(x)-\tilde{\tau})-\xi))dF_X(x)-\xi$$

which can be maximized with respect to ξ and delivers the first order condition, evaluated at ξ^*:

$$\int_{\mathcal{X}}\dot{\varphi}^*(-\lambda(\tau(x)-\tilde{\tau})-\xi^*)dF_X=1 \tag{A.24}$$

where $\dot{\varphi}^*(\cdot)$ is the derivative of $\varphi^*(\cdot)$ with respect to its argument. Observe that the Fenchel-conjugate of $\varphi(t) = t\log(t)$ is given by $\varphi(t^*) = \exp(t^* - 1)$. Solving for ξ^* here delivers:

$$\xi^* = \log\left(\int_{\mathcal{X}} \exp(-\lambda(\tau(x) - \tilde{\tau} - 1))dF_X(x)\right)$$

Now differentiating with respect to λ we obtain

$$\int_{\mathcal{X}} \dot{\varphi}^*(-\lambda^*(\tau(x) - \tilde{\tau}) - \xi^*)(\tau(x) - \tilde{\tau} + \dot{\xi}^*_\lambda)dF_X(x) - \dot{\xi}^*_\lambda = 0 \tag{A.25}$$

where $\dot{\xi}^*_\lambda$ is the derivative of ξ^* with respect to λ and λ^* is the value that implicitely solves the moment condition in Equation (A.25). Observe that plugging Equation (A.24) into Equation (A.25) allows to simplify it to:

$$\int_{\mathcal{X}} \dot{\varphi}^*(-\lambda^*(\tau(x) - \tilde{\tau}) - \xi^*)(\tau(x) - \tilde{\tau})dF_X = 0$$

since the two terms in $\dot{\xi}^*_\lambda$ cancel out. Moreover, if $\varphi(\cdot)$ is the KL divergence like in the main body of the paper, then

$$\int_{\mathcal{X}} \dot{\varphi}^*(-\lambda^*(\tau(x) - \tilde{\tau})(\tau(x) - \tilde{\tau})dF_X \cdot \exp(-\xi^*) = 0$$

so the additional term $\exp(-\xi^*) > 0$ can be dropped and Equation (A.25) recovers Equation (9).

I Proofs

First I introduce a few basic results for optimization problems like the one in Equations (4-5). Consider the set of probability distributions on $\mathcal{X}$, $\mathcal{P}_X := \{P_X : \int_{\mathcal{X}} dP_X = 1\}$. Under the L_1 norm, $\mathcal{P}_X$ is a complete metric space and it is convex. Namely, if $P_1, P_2 \in \mathcal{P}_X$ then $P_\alpha = \alpha P_1 + (1-\alpha)P_2 \in \mathcal{P}_X$ is a mixture distribution. Moreover, if there is a dominating measure μ such that $f_1 = \frac{dP_1}{d\mu}$ and $f_2 = \frac{dP_2}{d\mu}$ are the Radon-Nikodym derivatives then $\frac{dP_\alpha}{d\mu} = \alpha f_1 + (1-\alpha)f_2$. Now consider the constraint given in Equation (5). For any two P_1 and P_2 that satisfy the constraint, P_α for any $\alpha \in [0,1]$ will satisfy it as well. Hence the constraint set given by Equation (5) is a convex subset of $\mathcal{P}_X$. If such a set is non-empty, then, because $D_{KL}(\cdot||F_X)$ is a strictly convex function on a convex set, the infimization problem in Equation (4) has a unique solution (P_X-almost everywhere) and the infimum is achieved. Lemma 7 characterizes such a solution P_X-almost everywhere.

I.1 Proof of Lemma 7

The proof is based on a result that appeared first in Donsker and Varadhan [1975]. More recently Ho [2020] has used a similar argument to characterize global sensitivity in a Bayesian setting.

Lemma 7 (Closed form solution)**.** *Let Assumptions 1, 2, 3 and 4 hold. Then: i) The infimum in Equation (4) is achieved. Moreover F_X^*, is characterized, P_X-almost everywhere, by:*

$$\frac{dF_X^*}{dF_X}(x) = \frac{\exp(-\lambda(\tau(x) - \tilde{\tau}))}{\int_{\mathcal{X}} \exp(-\lambda(\tau(x) - \tilde{\tau}))dF_X(x)} \tag{8}$$

where $\frac{dF_X^*}{dF_X}$ is the Radon-Nikodym derivative of dF_X^* with respect to dF_X and λ is the Lagrange multiplier implicitly defined by the equation:

$$\int_{\mathcal{X}} \exp(-\lambda(\tau(x) - \tilde{\tau}))(\tau(x) - \tilde{\tau})dF_X(x) = 0 \tag{9}$$

ii) The value of the robustness metric $\delta^*(\tilde{\tau})$ is given by:

$$\delta^*(\tilde{\tau}) = D_{KL}(F_X^*||F_X) = -\log\left(\int_{\mathcal{X}} \exp(-\lambda(\tau(x) - \tilde{\tau}))dF_X(x)\right) \tag{10}$$

First note that, by the Radon-Nikodym theorem, $\frac{dF_X^*}{dF_X}$ exists and $\operatorname{supp}\left(\frac{dF_X'}{dF_X}\right) \subset \mathcal{X}$. Recall $\tau(x) = \mathbb{E}[Y_1|X = x] - \mathbb{E}[Y_0|X = x]$. Then:

$$\inf_{F_X':\ P_X' \ll P_X; P_X'(\mathcal{X})=1} D_{KL}(F_X'||F_X)$$
$$s.t. \int_{\mathcal{X}} \tau(x)dF_X'(x) = \tilde{\tau}$$

is equivalent to:

$$\begin{aligned} \inf_{F'_X:\ P'_X \ll P_X} & D_{KL}(F'_X || F_X) \\ s.t.\ & \int_{\mathcal{X}} \tau(x) \frac{dF'_X}{dF_X} dF_X(x) = \tilde{\tau} \\ & P'_X(\mathcal{X}) = 1 \end{aligned}$$

I adapt a lemma from Donsker and Varadhan [1975]:

Lemma 24. *Let* F^*_X *satisfy* $\frac{dF^*_X}{F_X} = \frac{\exp(-\lambda(\tau(x)-\tilde{\tau}))}{\int_{\mathcal{X}} \exp(-\lambda(\tau(x)-\tilde{\tau}))dF_X}$. *For any probability measure* $\tilde{F}_X$ *such that* $\tilde{F}_X \ll F_X$ *we have:*

$$\log\left(\int_{\mathcal{X}} \exp(-\lambda(\tau(x)-\tilde{\tau}))dF_X\right) = -\left[\int_{\mathcal{X}} \lambda(\tau(x)-\tilde{\tau})d\tilde{F}_X(x) + D_{KL}(\tilde{F}_X || F_X)\right] + D_{KL}(\tilde{F}_X || F^*_X)$$

Proof. First by definition of the KL-divergence we have:

$$\begin{aligned} D_{KL}(\tilde{F}_X || F^*_X) &= \int_{\mathcal{X}} \log\left(\frac{d\tilde{F}_X}{dF^*_X}\right) d\tilde{F}_X \\ &= \int_{\mathcal{X}} \log\left(\frac{\frac{d\tilde{F}_X}{dF_X}}{\frac{dF^*_X}{dF_X}}\right) d\tilde{F}_X \\ &= \int_{\mathcal{X}} \left(\log\left(\frac{d\tilde{F}_X}{dF_X}\right) - \log\left(\frac{dF^*_X}{dF_X}\right)\right) d\tilde{F}_X \\ &= \int_{\mathcal{X}} \log\left(\frac{d\tilde{F}_X}{dF_X}\right) d\tilde{F}_X - \int_{\mathcal{X}} \log\left(\frac{\exp(-\lambda(\tau(x)-\tilde{\tau})}{\int_{\mathcal{X}} \exp(-\lambda(\tau(x)-\tilde{\tau})dF_X}\right) d\tilde{F}_X \\ &= D_{KL}(\tilde{F}_X || F_X) + \int_{\mathcal{X}} \lambda(\tau(x)-\tilde{\tau})d\tilde{F}_X \\ &+ \int_{\mathcal{X}} \log\left(\int_{\mathcal{X}} \exp(-\lambda(\tau(x)-\tilde{\tau}))dF_X\right) d\tilde{F}_X \\ &= D_{KL}(\tilde{F}_X || F_X) + \int_{\mathcal{X}} \lambda(\tau(x)-\tilde{\tau})d\tilde{F}_X + \log\left(\int_{\mathcal{X}} \exp(-\lambda(\tau(x)-\tilde{\tau}))dF_X\right) \end{aligned}$$

since $\tilde{F}_X \ll F^*_X \ll F_X$ and simple algebra. Rearranging we get:

$$\log\left(\int_{\mathcal{X}} \exp(-\lambda(\tau(x)-\tilde{\tau})dF_X\right) = D_{KL}(\tilde{F}_X || F^*_X) - \left[\int_{\mathcal{X}} \lambda(\tau(x)-\tilde{\tau})d\tilde{F}_X + D_{KL}(\tilde{F}_X || F_X)\right]$$

□

Proof. i) From the lemma above we have:

$$\log\left(\int_{\mathcal{X}} \exp(-\lambda(\tau(x)-\tilde{\tau}))dF_X\right) = D_{KL}(\tilde{F}_X||F_X^*) - D_{KL}(\tilde{F}_X||F_X) - \int_{\mathcal{X}} \lambda(\tau(x)-\tilde{\tau})d\tilde{F}_X$$

Now observe that, since the term $\log\left(\int_{\mathcal{X}} \exp(-\lambda(\tau(x)-\tilde{\tau})dF_X\right)$ does not depend on $\tilde{F}_X$ we must have:

$$\begin{aligned}\arg\min_{\tilde{F}_X \ll F_X} D_{KL}(\tilde{F}_X||F_X^*) &= \arg\max_{\tilde{F}_X \ll F_X} -\int_X \lambda(\tau(x)-\tilde{\tau})d\tilde{F}_X - D_{KL}(\tilde{F}_X||F_X) \\ &= \arg\min_{\tilde{F}_X \ll F_X} \int_X \lambda(\tau(x)-\tilde{\tau})d\tilde{F}_X + D_{KL}(\tilde{F}_X||F_X)\end{aligned}$$

but clearly $F_X^* = \arg\min_{\tilde{F}_X \ll F_X} D_{KL}(\tilde{F}_X||F_X^*)$ so we must have

$$F_X^* = \arg\min_{\tilde{F}_X \ll F_X} D_{KL}(\tilde{F}_X||F_X) + \lambda\int_X (\tau(x)-\tilde{\tau})d\tilde{F}_X$$

which is the desired result. ii) Observe that $D_{KL}(F_X^*||F_X^*) = 0$ hence the value of the minimization problem:

$$\begin{aligned}&\min_{\tilde{F}_X \ll F_X} D_{KL}(\tilde{F}_X||F_X) + \lambda\int_X (\tau(x)-\tilde{\tau})d\tilde{F}_X \\ &= \min_{\tilde{F}_X \ll F_X} D_{KL}(\tilde{F}_X||F_X^*) - \log\left(\int_{\mathcal{X}} \exp(-\lambda(\tau(x)-\tilde{\tau}))dF_X\right) \\ &= -\log\left(\int_{\mathcal{X}} \exp(-\lambda(\tau(x)-\tilde{\tau}))dF_X\right)\end{aligned}$$

□

I.2 Proof of Fact 9

Proof. First, in this setting $F_X^* \ll F_X$ simply implies $p_1 = 0 \implies p_1^* = 0$. Excluding such a trivial case, 12 characterizes $\frac{p_1^*}{p_1}$. First we solve for the Lagrange multiplier λ in 12 by noting that:

$$\begin{aligned}\tilde{\tau} &= \int_{\mathcal{X}} \tau(x)dF_X^* \\ &= \frac{\exp(-\lambda(\tau(1)-\tilde{\tau}))\tau(1)p_1 + \exp(-\lambda(\tau(0)-\tilde{\tau}))\tau(0)(1-p_1)}{\exp(-\lambda(\tau(1)-\tilde{\tau}))p_1 + \exp(-\lambda(\tau(0)-\tilde{\tau}))(1-p_1)}\end{aligned}$$

rearranging the denominator and since $\tilde{\tau}$ is a constant, we obtain

$$\begin{aligned} &\exp(-\lambda(\tau(1)-\tilde{\tau}))\tau(1)p_1 + \exp(-\lambda(\tau(0)-\tilde{\tau}))\tau(0)(1-p_1) \\ &= \exp(-\lambda(\tau(1)-\tilde{\tau}))\tilde{\tau}p_1 + \exp(-\lambda(\tau(0)-\tilde{\tau}))\tilde{\tau}(1-p_1) \end{aligned}$$

which gives the condition:

$$\exp(-\lambda(\tau(1)-\tilde{\tau}))(\tau(1)-\tilde{\tau})p_1 + \exp(-\lambda(\tau(0)-\tilde{\tau}))(\tau(0)-\tilde{\tau})(1-p_1) = 0$$

And isolating each side and taking logs we obtain:

$$-\lambda(\tau(1)-\tau(0)) = \log\left(\frac{(\tilde{\tau}-\tau(0))(1-p_1)}{(\tau(1)-\tilde{\tau})p_1}\right)$$

so that

$$-\lambda = \frac{1}{(\tau(1)-\tau(0))}\log\left(\frac{(\tilde{\tau}-\tau(0))(1-p_1)}{(\tau(1)-\tilde{\tau})p_1}\right)$$

Finally, replacing $-\lambda$ in 11 we have:

$$\frac{p_1^*}{p_1} = \frac{\exp\left(\log\left(\frac{(\tilde{\tau}-\tau(0))(1-p_1)}{(\tau(1)-\tilde{\tau})p_1}\right)\frac{\tau(1)-\tilde{\tau}}{\tau(1)-\tau(0)}\right)}{\exp\left(\log\left(\frac{(\tilde{\tau}-\tau(0))(1-p_1)}{(\tau(1)-\tilde{\tau})p_1}\right)\frac{\tau(1)-\tilde{\tau}}{\tau(1)-\tau(0)}\right)p_1 + \exp\left(\log\left(\frac{(\tilde{\tau}-\tau(0))(1-p_1)}{(\tau(1)-\tilde{\tau})p_1}\right)\frac{\tau(0)-\tilde{\tau}}{\tau(1)-\tau(0)}\right)(1-p_1)}$$

Finally rearranging and combining terms we have:

$$
\begin{aligned}
p_1^* &= \frac{\exp\left(\log\left(\frac{(\tilde{\tau}-\tau(0))(1-p_1)}{(\tau(1)-\tilde{\tau})p_1}\right)\frac{\tau(1)-\tilde{\tau}}{\tau(1)-\tau(0)}\right)p_1}{\exp\left(\log\left(\frac{(\tilde{\tau}-\tau(0))(1-p_1)}{(\tau(1)-\tilde{\tau})p_1}\right)\frac{\tau(1)-\tilde{\tau}}{\tau(1)-\tau(0)}\right)p_1+\exp\left(\log\left(\frac{(\tilde{\tau}-\tau(0))(1-p_1)}{(\tau(1)-\tilde{\tau})p_1}\right)\frac{\tau(0)-\tilde{\tau}}{\tau(1)-\tau(0)}\right)(1-p_1)} \\
&= \frac{\left(\frac{(\tilde{\tau}-\tau(0))(1-p_1)}{(\tau(1)-\tilde{\tau})p_1}\right)^{\frac{\tau(1)-\tilde{\tau}}{\tau(1)-\tau(0)}}p_1}{\left(\frac{(\tilde{\tau}-\tau(0))(1-p_1)}{(\tau(1)-\tilde{\tau})p_1}\right)^{\frac{\tau(1)-\tilde{\tau}}{\tau(1)-\tau(0)}}p_1+\left(\frac{(\tilde{\tau}-\tau(0))(1-p_1)}{(\tau(1)-\tilde{\tau})p_1}\right)^{\frac{\tau(0)-\tilde{\tau}}{\tau(1)-\tau(0)}}(1-p_1)} \\
&= \frac{1}{1+\left(\frac{(\tilde{\tau}-\tau(0))(1-p_1)}{(\tau(1)-\tilde{\tau})p_1}\right)^{\frac{\tau(0)-\tilde{\tau}}{\tau(1)-\tau(0)}-\frac{\tau(1)-\tilde{\tau}}{\tau(1)-\tau(0)}}\frac{(1-p_1)}{p_1}} \\
&= \frac{1}{1+\left(\frac{(\tilde{\tau}-\tau(0))(1-p_1)}{(\tau(1)-\tilde{\tau})p_1}\right)^{-1}\frac{(1-p_1)}{p_1}} \\
&= \frac{1}{1+\frac{\tilde{\tau}-\tau(0)}{\tau(1)-\tau(0)}} \\
&= \frac{1}{\frac{\tau(1)-\tau(0)}{\tilde{\tau}-\tau(0)}} \\
&= \frac{\tilde{\tau}-\tau(0)}{\tau(1)-\tau(0)}
\end{aligned}
$$

which, with $\tilde{\tau}=0$, is the solution obtained in Equation (11). □

I.3 Proof of Proposition 10

Proposition 10 (Local to boundary $\tilde{\tau}$)**.** *Let Assumptions 1-3 hold and let $\tilde{\tau}_m \to \tilde{\tau}_b \in \partial L(\mathcal{P}_X)$. Assume that the pre-image $\tau^{-1}(\tilde{\tau}_b) = \mathcal{X}_b = \{x_b\} \in \mathcal{X}^\circ$ is a singleton. Further, let X be compactly supported, with density $f(x) < M$ on $\mathcal{X}$. Then the sequence of least favorable distributions for the policy-maker's problem with parameter $\tilde{\tau}_m$, denoted $F^*_{X,m}$, converges weakly to δ_{x_b}, the Dirac delta distribution with point mass at x_b, that is:*

$$\lim_{m\to\infty}\int_{\mathcal{X}} g(x)dF^*_{X,m}(x) \to \int_{\mathcal{X}} g(x)\delta_{x_b} := g(x_b)$$

for $g \in C_b(\mathcal{X})$, the space of all continuous, bounded functions on $\mathcal{X}$.

Proof. First observe that by Lemma 7 and the fact that each $\tau_m \in L^\circ(\mathcal{P}_X)$ we can construct the sequence of *least favorable distributions* $F^*_{m,X}$ satisfying:

$$\frac{dF^*_{m,X}}{dF_X}(x) = \frac{\exp(-\lambda_m(\tau(x) - \tilde{\tau}_m))}{\int_{\mathcal{X}} \exp(-\lambda_m(\tau(x) - \tilde{\tau}_m))dF_X}$$
$$\lambda_m : \quad \int_{\mathcal{X}} \exp(-\lambda_m(\tau(x) - \tilde{\tau}_m))(\tau(x) - \tilde{\tau}_m)dF_X = 0$$

□

Without loss of generality consider the case where $\tilde{\tau}_b = \max_{\mathcal{X}} \tau(x)$. First notice that the sequence of λ_m defined above is decreasing and unbounded below. To see that it's decreasing observe that implicitly differentiating $\lambda(\tilde{\tau})$:

$$\begin{aligned}
&\frac{\partial}{\partial \tilde{\tau}} \int_{\mathcal{X}} \exp(-\lambda_m(\tau(x) - \tilde{\tau}_m))(\tau(x) - \tilde{\tau}_m)dF_X(x) \\
&= -\frac{\partial \lambda}{\partial \tilde{\tau}}(\tilde{\tau}) \int_{\mathcal{X}} \exp(-\lambda(\tilde{\tau})(\tau(x) - \tilde{\tau}))(\tau(x) - \tilde{\tau})^2 dF_X \\
&+ \lambda(\tilde{\tau}) \int_{\mathcal{X}} \exp(-\lambda(\tilde{\tau})(\tau(x) - \tilde{\tau}))(\tau(x) - \tilde{\tau})dF_X \\
&- \int_{\mathcal{X}} \exp(-\lambda(\tilde{\tau})(\tau(x) - \tilde{\tau}))dF_X = 0
\end{aligned}$$

by the Dominated Convergence Theorem with envelope $g = \exp(2M) \cdot 2M$. Note that by the definition of $\lambda(\tilde{\tau})$ the second term is equal to 0. Isolating the derivative of λ with respect to $\tilde{\tau}$ we have:

$$\frac{\partial \lambda}{\partial \tilde{\tau}}(\tilde{\tau}) = -\frac{\int_{\mathcal{X}} \exp(-\lambda(\tilde{\tau})(\tau(x) - \tilde{\tau}))dF_X}{\int_{\mathcal{X}} \exp(-\lambda(\tilde{\tau})(\tau(x) - \tilde{\tau}))(\tau(x) - \tilde{\tau})^2 dF_X} < 0$$

so $\lambda(\tilde{\tau})$ is strictly decreasing on its domain. Suppose $\lambda_m \geq -B$ for all $m \in N$, with $B > 0$. Then:

$$\int_{\mathcal{X}} \exp(-\lambda_m(\tau(x) - \tilde{\tau}_m))(\tau(x) - \tilde{\tau}_m)dF_X \leq \quad \int_{\mathcal{X}} \exp(B(\tau(x) - \tilde{\tau}_m))(\tau(x) - \tilde{\tau}_m)dF_X$$

so taking the limit fro $m \to \infty$, if $P_X(\tau(x) \neq \tilde{\tau}_b) > 0$:

$$\begin{aligned} &\lim_{m\to\infty} \int_{\mathcal{X}} \exp(-\lambda_m(\tau(x) - \tilde{\tau}_m))(\tau(x) - \tilde{\tau}_m) dF_X \\ &\leq \lim_{m\to\infty} \int_{\mathcal{X}} \exp(B(\tau(x) - \tilde{\tau}_m))(\tau(x) - \tilde{\tau}_m) dF_X(x) \\ &\leq \int_{\mathcal{X}} \exp(B(\tau(x) - \tilde{\tau}_b))(\tau(x) - \tilde{\tau}_b) dF_X(x) < 0 \end{aligned}$$

Then, there exist $m^* \in \mathbb{N}$ such that $\int_{\mathcal{X}} \exp(\lambda_{m^*}(\tau(x) - \tilde{\tau}_{m^*}))(\tau(x) - \tilde{\tau}_{m^*}) dF_X < 0$ which is a contradiction. So λ_m must be unbounded below. Because it's a strictly decreasing, unbounded below sequence, it must be the case that $\lambda_m \to -\infty$ as $\tilde{\tau}_m \to \tilde{\tau}_b$. Now we show convergence in distribution to δ_{x_b}. Let $\varphi(\cdot) \in \mathcal{C}_b$. We want to show:

$$\lim_{m\to\infty} \int_{\mathcal{X}} \varphi(x) dF^*_{X,m}(x) \to \int_{\mathcal{X}} \varphi(x) \delta_{x_b}(x) = \varphi(x_b)$$

We have:

$$\begin{aligned} \int_{\mathcal{X}} \varphi(x) dF^*_{X,m}(x) &= \int_{\mathcal{X}} \varphi(x) \frac{\exp(-\lambda_m(\tau(x) - \tilde{\tau}_b)) dF_X(x)}{\int_{\mathcal{X}} \exp(-\lambda_m(\tau(x) - \tilde{\tau}_b)) dF_X(x)} \\ &= \int_{\mathcal{X}} \varphi(x) \frac{\exp(-\lambda_m(\tau(x) - \tilde{\tau}_b)) dF_X(x)}{\int_{\mathcal{X}} \exp(-\lambda_m(\tau(x) - \tilde{\tau}_b)) dF_X(x)} \end{aligned}$$

Noticing that $\lambda_m < 0$. Consider the change of variables $y = \sqrt{-\lambda_m}(x_b - x)$. Then $x = x_b - \frac{y}{\sqrt{-\lambda_m}}$, $dx = -\frac{dy}{\sqrt{-\lambda_m}}$. By the change of variable formula:

$$\begin{aligned} &\int_{\mathcal{X}} \varphi(x) \frac{\exp(-\lambda_m(\tau(x) - \tilde{\tau}_b)) f(x) dx}{\int_{\mathcal{X}} \exp(-\lambda_m(\tau(x) - \tilde{\tau}_b)) f(x) dx} \\ &= \frac{\int_{\mathbb{R}^k} \varphi\left(x_b - \frac{y}{\sqrt{-\lambda_m}}\right) \exp\left(-\lambda_m\left(\tau\left(x_b - \frac{y}{\sqrt{-\lambda_m}}\right) - \tau(x_b)\right)\right) f\left(x_b - \frac{y}{\sqrt{-\lambda_m}}\right) \mathbb{1}_{\mathcal{Y}(\lambda_m)}(y) \frac{1}{\sqrt{-\lambda_m}} dy}{\int_{\mathbb{R}^k} \exp\left(-\lambda_m\left(\tau\left(x_b - \frac{y}{\sqrt{-\lambda_m}}\right) - \tau(x_b)\right)\right) f\left(x_b - \frac{y}{\sqrt{-\lambda_m}}\right) \mathbb{1}_{\mathcal{Y}(\lambda_m)}(y) \frac{1}{\sqrt{-\lambda_m}} dy} \\ &= \frac{\int_{\mathbb{R}^k} \varphi\left(x_b - \frac{y}{\sqrt{-\lambda_m}}\right) \exp\left(-\lambda_m\left(\tau\left(x_b - \frac{y}{\sqrt{-\lambda_m}}\right) - \tau(x_b)\right)\right) f\left(x_b - \frac{y}{\sqrt{-\lambda_m}}\right) \mathbb{1}_{\mathcal{Y}(\lambda_m)}(y) dy}{\int_{\mathbb{R}^k} \exp\left(-\lambda_m\left(\tau\left(x_b - \frac{y}{\sqrt{-\lambda_m}}\right) - \tau(x_b)\right)\right) f\left(x_b - \frac{y}{\sqrt{-\lambda_m}}\right) \mathbb{1}_{\mathcal{Y}(\lambda_m)}(y) dy} \end{aligned}$$

Note that, if X is compactly supported then $f(x) = 0$ outside of a compact set $K \subseteq \mathbb{R}^k$

hence. Moreover, if $f(x) < M$ we have the dominating function given by:

$$\varphi\left(x_b - \frac{y}{\sqrt{-\lambda_m}}\right)\exp\left(-\lambda_m\left(\tau\left(x_b - \frac{y}{\sqrt{-\lambda_m}}\right) - \tau(x_b)\right)\right) f\left(x_b - \frac{y}{\sqrt{-\lambda_m}}\right)\mathbb{1}_{\mathcal{Y}(\lambda_m)}(y)dy$$
$$\leq \|\varphi\|_\infty M \mathbb{1}_K(y)$$

on $\mathbb{R}^k$ and $\int_{\mathbb{R}^k}\|\varphi\|_\infty M\mathbb{1}_K(x)dx = \|\varphi\|_\infty \cdot M \cdot \mathrm{vol}(K) < +\infty$. hence the assumptions of the Dominated Convergence theorem hold. Then we have:

$$\begin{aligned}
&= \lim_{m\to\infty}\int_{\mathbb{R}^k}\varphi\left(x_b - \frac{y}{\sqrt{-\lambda_m}}\right)\exp\left(-\lambda_m\left(\tau\left(x_b - \frac{y}{\sqrt{-\lambda_m}}\right) - \tau(x_b)\right)\right)\\
&\times f\left(x_b - \frac{y}{\sqrt{-\lambda_m}}\right)\mathbb{1}_{\mathcal{Y}(\lambda_m)}(y)dy\\
&= \int_{\mathbb{R}^k}\lim_{m\to\infty}\varphi\left(x_b - \frac{y}{\sqrt{-\lambda_m}}\right)\exp\left(-\lambda_m\left(\tau\left(x_b - \frac{y}{\sqrt{-\lambda_m}}\right) - \tau(x_b)\right)\right)\\
&\times f\left(x_b - \frac{y}{\sqrt{-\lambda_m}}\right)\mathbb{1}_{\mathcal{Y}(\lambda_m)}(y)dy
\end{aligned}$$

Now consider Taylor expanding $\tau(\cdot)$ around x_b. Because x_b is a maximizer, the Jacobian $J_\tau(x_b) : \mathbb{R}^k \to \mathbb{R}$ is the zero matrix, from first order conditions. Hence:

$$\begin{aligned}
&\exp\left(-\lambda_m\left(\tau\left(x_b - \frac{y}{\sqrt{-\lambda_m}}\right) - \tau(x_b)\right)\right)\\
&= \exp\left(-\lambda_m\left(\tau(x_b) - J_\tau(x_b)\left(\frac{y}{\sqrt{-\lambda_m}}\right) + \frac{1}{2}\cdot\frac{1}{-\lambda_m}y^T H_\tau(x_b)y - \tau(x_b)\right)\right)\\
&= \exp\left(\frac{1}{2}y^T H_\tau(x_b)y + o(1)\right)
\end{aligned}$$

where $H_\tau(x_b)$ is the $k\times k$ Hessian matrix of τ, evaluated at the maximizer x_b. Moreover:

$$\begin{aligned}
&\int_{\mathbb{R}^k}\lim_{m\to\infty}\varphi\left(x_b - \frac{y}{\sqrt{-\lambda_m}}\right)\exp\left(-\lambda_m\left(\tau\left(x_b - \frac{y}{\sqrt{-\lambda_m}}\right) - \tau(x_b)\right)\right) f\left(x_b - \frac{y}{\sqrt{-\lambda_m}}\right)\mathbb{1}_{\mathcal{Y}(\lambda_m)}(y)dy\\
&= \int_{\mathbb{R}^k}\varphi(x_b)\exp\left(\frac{1}{2}y^T H(x_b)y\right) f(x_b)dy\\
&= \varphi(x_b)\int_{\mathbb{R}^k}\exp\left(\frac{1}{2}y^T H(x_b)y\right) f(x_b)dy
\end{aligned}$$

Now the denominator can be treated identically to have:

$$\begin{aligned}
&\int_{\mathbb{R}^k} \lim_{m\to\infty} \exp\left(-\lambda_m \left(\tau\left(x_b - \frac{y}{\sqrt{-\lambda_m}}\right) - \tau(x_b)\right)\right) f\left(x_b - \frac{y}{\sqrt{-\lambda_m}}\right) \mathbb{1}_{\mathcal{Y}(\lambda_m)}(y)dy \\
&= \int_{\mathbb{R}^k} \exp\left(\frac{1}{2} y^T H(x_b) y\right) f(x_b) dy
\end{aligned}$$

Now because x_b is a maximizer, $H(x_b)$ is negative definite so the quantities above are finite and the numerator is greater than 0. Finally:

$$\begin{aligned}
&\lim_{m\to\infty} \int_{\mathcal{X}} \varphi(x) dF^*_{X,m}(x) \\
&= \lim_{m\to\infty} \frac{\int_{\mathcal{X}} \varphi(x)\exp(-\lambda_m(\tau(x) - \tilde{\tau}_b)) f(x)dx}{\int_{\mathcal{X}} \exp(-\lambda_m(\tau(x) - \tilde{\tau}_b)) f(x)dx} \\
&= \frac{\lim_{m\to\infty} \int_{\mathcal{X}} \varphi(x)\exp(-\lambda_m(\tau(x) - \tilde{\tau}_b)) f(x)dx}{\lim_{m\to\infty} \int_{\mathcal{X}} \exp(-\lambda_m(\tau(x) - \tilde{\tau}_b)) f(x)dx} \\
&= \frac{\varphi(x_b) \int_{\mathbb{R}^k} \exp\left(\frac{1}{2} y^T H(x_b) y\right) f(x_b)dy}{\int_{\mathbb{R}^k} \exp\left(\frac{1}{2} y^T H(x_b) y\right) f(x_b)dy} \\
&= \varphi(x_b)
\end{aligned}$$

Since $\varphi(\cdot) \in \mathcal{C}_b$ was arbitrary, by the Portmanteau theorem, $dF^*_{X,m} \xrightarrow{d} \delta_{x_b}$.

In the general case where $\mathcal{X}_b$ is not a singleton, it seems that the *least favorable distribution* still concentrates around the uniform distribution on the $\mathcal{X}_b$, rather than *any* distribution like the figure below suggests. I leave this interesting case for future work.

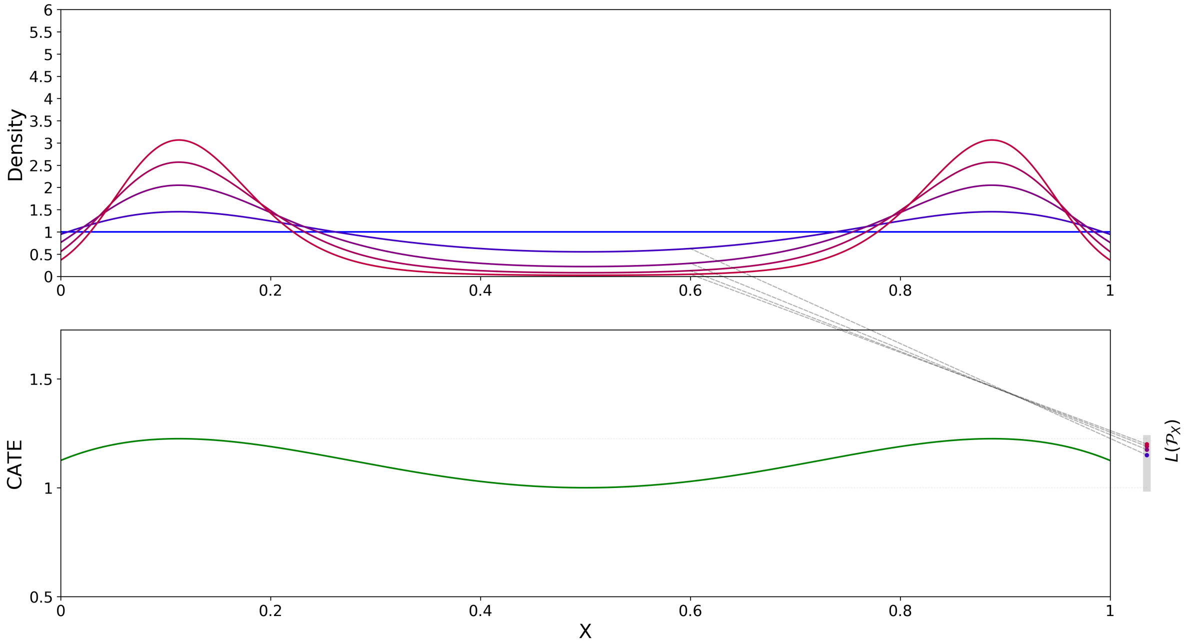

Figure 7: Here $\tau(x)$ is quadratic, experimental distribution is uniform and there are two peaks. It appears that the *least favorable distribution* concentrates around both peaks.

I.4 Proof of Proposition 19

Proposition 19 (Quadratic-Normal least favorable closed-ness)**.** *The parametric class $\mathcal{N}(\mu, \sigma^2)$ is* ***least favorable closed*** *for quadratic Conditional Average Treatment Effects. That is, if $X \in \mathbb{R}^k$ follows the multivariate normal distribution $X \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ where $\boldsymbol{\Sigma}$ is p.d. and $\tau(x) = \boldsymbol{x^T A x} + \boldsymbol{x^T \beta} + c$ for $\boldsymbol{\beta} \in \mathbb{R}^k$ then F_X^* is the measure induced by $X^* \sim \mathcal{N}(\boldsymbol{\mu^*}, \boldsymbol{\Sigma^*})$ with $\boldsymbol{\mu}^* = (\boldsymbol{\Sigma}^{-1} + 2\lambda \boldsymbol{A})^{-1}(\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} - \lambda\boldsymbol{\beta})$ and $\boldsymbol{\Sigma^*} = (\boldsymbol{\Sigma}^{-1} + 2\lambda \boldsymbol{A})^{-1}$, provided that $(\boldsymbol{\Sigma}^{-1} + 2\lambda \boldsymbol{A})^{-1}$ is p.d. The parameter λ is defined as in Equation (9).*

Proof. Suppose $\mathcal{X} = \mathbb{R}^k$, $X \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\sigma})$ and $\tau(x) = \boldsymbol{x^T A x} + \boldsymbol{x^T \beta} + c$. By Lemma 7 the Radon-Nikodym derivative of the least favorable distribution is given by Equation (8) so the distribution of F_X^* must have density:

$$
\begin{aligned}
d\mu_X^* &:= \frac{\exp(-\lambda(\tau(x)-\tilde{\tau}))\frac{\exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)}{\sqrt{(2\pi)^k\det(\boldsymbol{\Sigma})}}dx}{\displaystyle\int_{\mathcal{X}}\exp(-\lambda(\tau(x)-\tilde{\tau}))\frac{\exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)}{\sqrt{(2\pi)^k\det(\boldsymbol{\Sigma})}}dx}\\
&= \frac{\exp(-\lambda(\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x}+\boldsymbol{x}^T\boldsymbol{\beta}+c-\tilde{\tau}))\frac{\exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)}{\sqrt{(2\pi)^k\det(\boldsymbol{\Sigma})}}dx}{\displaystyle\int_{\mathcal{X}}\exp(-\lambda(\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x}+\boldsymbol{x}^T\boldsymbol{\beta}+c-\tilde{\tau}))\frac{\exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)}{\sqrt{(2\pi)^k\det(\boldsymbol{\Sigma})}}dx}\\
&= \frac{\frac{\exp(-\lambda(\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x}+\boldsymbol{x}^T\boldsymbol{\beta}+c-\tilde{\tau})-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}))}{\sqrt{(2\pi)^k\det(\boldsymbol{\Sigma})}}dx}{\displaystyle\int_{\mathcal{X}}\frac{\exp(-\lambda(\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x}+\boldsymbol{x}^T\boldsymbol{\beta}+c-\tilde{\tau})-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}))}{\sqrt{(2\pi)^k\det(\boldsymbol{\Sigma})}}dx}\\
&= \frac{\exp(-\lambda(\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x}+\boldsymbol{x}^T\boldsymbol{\beta}+c-\tilde{\tau})-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}))dx}{\displaystyle\int_{\mathcal{X}}\exp(-\lambda(\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x}+\boldsymbol{x}^T\boldsymbol{\beta}+c-\tilde{\tau})-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}))dx}\\
&= \frac{\exp(-\frac{1}{2}(\boldsymbol{x}-(\boldsymbol{\Sigma}^{-1}+2\lambda\boldsymbol{A})^{-1}(\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}-\lambda\boldsymbol{\beta}))(\boldsymbol{\Sigma}^{-1}+2\lambda\boldsymbol{A}))(x-(\boldsymbol{\Sigma}^{-1}+2\lambda\boldsymbol{A})^{-1}(\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}-\lambda\boldsymbol{\beta}))dx}{\displaystyle\int_{\mathcal{X}}\exp(-\frac{1}{2}(x-(\boldsymbol{\Sigma}^{-1}+2\lambda\boldsymbol{A})^{-1}(\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}-\lambda\boldsymbol{\beta}))(\boldsymbol{\Sigma}^{-1}+2\lambda\boldsymbol{A}))(x-(\boldsymbol{\Sigma}^{-1}+2\lambda\boldsymbol{A})^{-1}(\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}-\lambda\boldsymbol{\beta}))dx}\\
&\times\frac{\exp(\lambda c+\lambda\tilde{\tau}-\frac{1}{2}\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}-\frac{1}{2}(\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}-\lambda\boldsymbol{\beta})(\boldsymbol{\Sigma}^{-1}+2\lambda\boldsymbol{\beta})^{-1}(\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}-\lambda\boldsymbol{\beta}))}{\exp(\lambda c+\lambda\tilde{\tau}-\frac{1}{2}\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}-\frac{1}{2}(\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}-\lambda\boldsymbol{\beta})(\boldsymbol{\Sigma}^{-1}+2\lambda\boldsymbol{\beta})^{-1}(\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}-\lambda\boldsymbol{\beta}))}\\
&= \frac{\exp(-\frac{1}{2}(\boldsymbol{x}-(\boldsymbol{\Sigma}^{-1}+2\lambda\boldsymbol{A})^{-1}(\boldsymbol{\Sigma}^{-1}\mu-\lambda\boldsymbol{\beta}))(\boldsymbol{\Sigma}^{-1}+2\lambda\boldsymbol{A}))(x-(\boldsymbol{\Sigma}^{-1}+2\lambda\boldsymbol{A})^{-1}(\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}-\lambda\boldsymbol{\beta}))dx}{\displaystyle\int_{\mathcal{X}}\exp(-\frac{1}{2}(\boldsymbol{x}-(\boldsymbol{\Sigma}^{-1}+2\lambda\boldsymbol{A})^{-1}(\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}-\lambda\boldsymbol{\beta}))(\boldsymbol{\Sigma}^{-1}+2\lambda\boldsymbol{A}))(\boldsymbol{x}-(\boldsymbol{\Sigma}^{-1}+2\lambda\boldsymbol{A})^{-1}(\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}-\lambda\boldsymbol{\beta}))dx}
\end{aligned}
$$

from which we can recognize the form of the normal distribution with mean $\boldsymbol{\mu}^*$ and variance covariance matrix $\boldsymbol{\Sigma}^*$. The steps above follow from completing the square and from the properties of $\exp(\cdot)$. □

I.5 Proof of Proposition 13

Proposition 13. *The de-biased GMM nonparametric influence function based on moment function $g(\cdot)$ is:*

$$
\begin{aligned}
\phi(w,\theta,\gamma_0,\alpha_0) = &\begin{bmatrix} \exp\left(-\lambda\cdot(\gamma_{1,F_0}(x)-\gamma_{0,F_0}(x)-\tilde{\tau})\right)\cdot(-\lambda) \\ \exp\left(-\lambda\cdot(\gamma_{1,F_0}(x)-\gamma_{0,F_0}(x)-\tilde{\tau})\right)\cdot(1-\lambda\cdot(\gamma_{1,F_0}(x)-\gamma_{0,F_0}(x)-\tilde{\tau}))\end{bmatrix}\\
&\times\left(\frac{d(y-\gamma_{1,F_0}(x))}{\pi_{F_0}(x)}-\frac{(1-d)(y-\gamma_{0,F_0}(x))}{1-\pi_{F_0}(x)}\right)
\end{aligned}
$$

which could be written in the form:

$$\phi(w, \theta, \gamma_0, \alpha_0) = \begin{bmatrix} \exp\left(-\lambda \cdot (\gamma_{1,F_0}(x) - \gamma_{0,F_0}(x) - \tilde{\tau})\right) \cdot (-\lambda) \\ \exp\left(-\lambda \cdot (\gamma_{1,F_0}(x) - \gamma_{0,F_0}(x) - \tilde{\tau})\right) \cdot (1 - \lambda \cdot (\gamma_{1,F_0}(x) - \gamma_{0,F_0}(x) - \tilde{\tau})) \end{bmatrix}$$
$$\times \left(\begin{bmatrix} \alpha_{1,F_0}(x) \\ \alpha_{0,F_0}(x) \end{bmatrix}^T \begin{bmatrix} d(y - \gamma_{1,F_0}(x)) \\ (1-d)(y - \gamma_{0,F_0}(x)) \end{bmatrix} \right)$$

with $\alpha_{F_0}(x) := \begin{bmatrix} \alpha_{1,F_0}(x) \\ \alpha_{0,F_0}(x) \end{bmatrix} = \begin{bmatrix} \frac{1}{\pi_{F_0}(x)} \\ \frac{1}{1-\pi_{F_0}(x)} \end{bmatrix}$.

Proof. Let $F_r = (1-r)F_0 + rH$ for an arbitrary distribution H that satisfies unconfounded-ness. Then F_r is a distribution because it's a convex combination of two distributions, and it satisfies unconfounded-ness. Therefore we can refer to the identification results:

$$\mathbb{E}_{F_r}[Y_1|X] = \mathbb{E}_{F_r}[Y|D = 1, X]$$
$$\mathbb{E}_{F_r}[Y_0|X] = \mathbb{E}_{F_r}[Y|D = 0, X]$$

and derive the distributional derivative of $\mathbb{E}_{F_r}[Y|D = 1, X] - \mathbb{E}_{F_r}[Y|D = 0, X]$ with respect to r and evaluate it at $r = 0$. Alternatively one may start with the propensity score weighting identification result below:

$$\mathbb{E}_{F_r}\left[\frac{Y \cdot D}{\pi_{F_r}(X)} - \frac{Y \cdot (1-D)}{\pi_{F_r}(X)} \middle| X \right] = \mathbb{E}_{F_r}[Y_1 - Y_0|X]$$

and proceed as above to derive the distributional derivative of $\mathbb{E}_F[g(W, \theta, \gamma(F_r))]$. The second approach is more cumbersome so we present the proof for the regression adjustment method but note that both would be valid approaches to find the nonparametric influence function. Computing the derivative of the moment condition with respect to

r and evaluating it at $r = 0$ we have:

$$\begin{aligned}
\left.\frac{d\mathbb{E}[g(W,\theta,\gamma(F_r))]}{dr}\right|_{r=0} &= \frac{d}{dr}\mathbb{E}\left.\left[\begin{matrix}\exp(-\lambda_0(\gamma_{1,F_r}(x)-\gamma_{0,F_r}(x)-\tilde{\tau}))-\nu] \\ \exp\left(-\lambda_0(\gamma_{1,F_r}(x)-\gamma_{0,F_r}(x)-\tilde{\tau})\right)(\gamma_{1,F_r}(x)-\gamma_{0,F_r}(x)-\tilde{\tau})]\end{matrix}\right]\right|_{r=0} \\
&= \int_{\mathcal{X}} \frac{d}{dr}\left.\left[\begin{matrix}\exp\left(-\lambda_0\cdot(\gamma_{1,F_r}(x)-\gamma_{0,F_r}(x)-\tilde{\tau})\right) \\ \exp\left(-\lambda_0\cdot(\gamma_{1,F_r}(x)-\gamma_{0,F_r}(x)-\tilde{\tau})\right)(\gamma_{1,F_r}(x)-\gamma_{0,F_r}(x)-\tilde{\tau})\end{matrix}\right] f_0(x)dx\right|_{r=0} \\
&= \int_{\mathcal{X}} \left[\begin{matrix}\exp\left(-\lambda_0\cdot(\gamma_{1,F_r}(x)-\gamma_{0,F_r}(x)-\tilde{\tau})\right)\cdot(-\lambda_0) \\ \exp\left(-\lambda_0\cdot(\gamma_{1,F_r}(x)-\gamma_{0,F_r}(x)-\tilde{\tau})\right)\cdot\left(1-\lambda\cdot(\gamma_{1,F_r}(x)-\gamma_{0,F_r}(x)-\tilde{\tau})\right)\end{matrix}\right] \\
&\times\frac{\partial}{\partial r}(\gamma_{1,F_r}(x)-\gamma_{0,F_r}(x))f_0(x)dx
\end{aligned}$$

In order to characterize the contribution of the functional we have:

$$\begin{aligned}
&\frac{\partial}{\partial r}(\gamma_{1,F_r}(x)-\gamma_{0,F_r}(x)) \\
&= \frac{\partial}{\partial r}\int_{\mathcal{Y}} \frac{y}{\int_{\mathcal{Y}}(1-r)dF_0(y,1,x)+rdH(y,1,x)}((1-r)dF(y,1,x)+rdH(y,1,x)) \\
&- \frac{\partial}{\partial r}\int_{\mathcal{Y}} \frac{y}{\int_{\mathcal{Y}}(1-r)dF_0(y,0,x)+rdH(y,0,x)}((1-r)dF(y,0,x)+rdH(y,0,x)) \\
&= \frac{\int_{\mathcal{Y}} y\cdot[dH(y,1,x)-dF_0(y,1,x)]\int_{\mathcal{Y}}(1-r)dF_0(y,1,x)+rdH(y,1,x)}{\left(\int_{\mathcal{Y}}(1+r)dF_0(y,1,x)+rdH(y,1,x)\right)^2} \\
&- \frac{\int_{\mathcal{Y}} y[dH(y,1,x)-dF_0(y,1,x)]((1-r)dF_0(y,1,x)-dH(y,1,x))}{\left(\int_{\mathcal{Y}}(1+r)dF_0(y,1,x)+rdH(y,1,x)\right)^2} \\
&- \frac{\int_{\mathcal{Y}} y\cdot[dH(y,0,x)-dF_0(y,0,x)]\int_{\mathcal{Y}}(1-r)dF_0(y,0,x)+rdH(y,0,x)}{\left(\int_{\mathcal{Y}}(1+r)dF_0(y,0,x)+rdH(y,0,x)\right)^2} \\
&+ \frac{\int_{\mathcal{Y}} y[dH(y,0,x)-dF_0(y,0,x)]((1-r)dF_0(y,0,x)-dH(y,0,x))}{\left(\int_{\mathcal{Y}}(1+r)dF_0(y,0,x)+rdH(y,0,x)\right)^2}
\end{aligned}$$

Below $f_0(d,x) = \int_{\mathcal{Y}} dF_0(y,d,x)$ and the same holds for $h(\cdot)$. Evaluating this expression at $r = 0$ one obtains:

$$\int y\cdot\frac{dH(y,1,x)}{f_0(1,x)} - \int y\cdot\frac{h(1,x)\cdot dF_0(y,1,x)}{f_0(1,x)^2} - \int y\cdot\frac{dH(y,0,x)}{f_0(0,x)} + \int y\cdot\frac{h(0,x)\cdot dF_0(y,0,x)}{f_0(0,x)^2}$$

Combining this with the derivative of the moment condition with respect to the γ we have:

$$\begin{aligned}
\frac{d\mathbb{E}[g(W,\theta,\gamma(F_r))]}{dr} &= \int_{\mathcal{Y}\times\{0,1\}\times\mathcal{X}} \left[\begin{matrix}\exp\left(-\lambda\cdot(\gamma_{1,F_0}(x)-\gamma_{0,F_0}(x)-\tilde{\tau})\right)\cdot(-\lambda) \\ \exp\left(-\lambda\cdot(\gamma_{1,F_0}(x)-\gamma_{0,F_0}(x)-\tilde{\tau})\right)\cdot\left(1-\lambda\cdot(\gamma_{1,F_0}(x)-\gamma_{0,F_0}(x)-\tilde{\tau})\right)\end{matrix}\right] \\
&\times\left(\frac{d(y-\gamma_{1,F_0}(x))}{\pi_{F_0}(x)} - \frac{(1-d)(y-\gamma_{0,F_0}(x))}{1-\pi_{F_0}(x)}\right)dH(y,d,x)
\end{aligned}$$

or $\frac{d\mathbb{E}[g(W,\theta,\gamma(F_r))]}{dr} = \int_{\mathcal{Y}\times\{0,1\}\times\mathcal{X}} \phi(w,\theta,\gamma(F_0),\alpha(F_0))dH(w)$ for

$$
\begin{aligned}
\phi(w,\theta,\gamma,\alpha) &= \begin{bmatrix} \exp\left(-\lambda\cdot(\gamma_{1,F_0}(x)-\gamma_{0,F_0}(x)-\tilde{\tau})\right)\cdot(-\lambda) \\ \exp\left(-\lambda\cdot(\gamma_{1,F_0}(x)-\gamma_{0,F_0}(x)-\tilde{\tau})\right)\cdot(1-\lambda\cdot(\gamma_{1,F_0}(x)-\gamma_{0,F_0}(x)-\tilde{\tau})) \end{bmatrix} \\
&\quad\times\left(\frac{d(y-\gamma_{1,F_0}(x))}{\pi_F(x)}-\frac{(1-d)(y-\gamma_{0,F_0}(x))}{1-\pi_F(x)}\right) \\
&= \begin{bmatrix} \exp\left(-\lambda\cdot(\gamma_{1,F_0}(x)-\gamma_{0,F_0}(x)-\tilde{\tau})\right)\cdot(-\lambda) \\ \exp\left(-\lambda\cdot(\gamma_{1,F_0}(x)-\gamma_{0,F_0}(x)-\tilde{\tau})\right)\cdot(1-\lambda\cdot(\gamma_{1,F_0}(x)-\gamma_{0,F_0}(x)-\tilde{\tau})) \end{bmatrix} \\
&\quad\times\left(\begin{bmatrix}\alpha_{1,F_0}(x) \\ \alpha_{0,F_0}(x)\end{bmatrix}^T \begin{bmatrix} d(y-\gamma_{1,F_0}(x)) \\ (1-d)(y-\gamma_{0,F_0}(x)) \end{bmatrix}\right)
\end{aligned}
$$

and $\alpha_{F_0}(X) := \begin{bmatrix}\alpha_{1,F_0}(x) \\ \alpha_{0,F_0}(x)\end{bmatrix} = \begin{bmatrix}\frac{1}{\pi_{F_0}(X)} \\ \frac{1}{1-\pi_{F_0}(X)}\end{bmatrix}$. Note that above $\phi(\cdot)$ is the Riesz representer of the linear functional $\left.\frac{d\mathbb{E}[g(W,\theta,\gamma(F_r))]}{dr}\right|_{r=0} : \mathcal{H} \to \mathbb{R}^2$ which maps H to $\mathbb{R}^2$.

Observe that $\mathbb{E}_{F_0}[\phi(W,\theta,\gamma_0(X),\alpha_0(X)] = 0$ by the law of iterated expectations. Moreover, for any distribution F, $\mathbb{E}_F\left[\frac{D(Y-\mathbb{E}_F[Y|D=1,X])}{\pi_F(X)} - \frac{(1-D)(Y-\mathbb{E}_F[Y|D=0,X]}{1-\pi_F(X)}\middle| X\right] = 0$. □

I.6 Proof of Proposition 14

Proposition 14. *Equation (16) satisfies Neyman orthogonality.*

Proof. To show that they are Neyman orthogonal we verify the conditions for Theorem 1 in Chernozhukov et al. [2020] in the Appendix. Let $\gamma_{1,F}(X), \gamma_{0,0}(X)$ denote $\mathbb{E}_F[Y|D = 1, X], \mathbb{E}_F[Y|D = 0, X]$ respectively.

i) Equation (15) holds. This has been verified above.

ii) $\int_{\mathcal{Y}_0\times\mathcal{Y}_1\times\mathcal{X}} \phi(w,\gamma(F_r),\theta,\alpha(F_r))F_r(dw) = 0$ for all $r \in [0,\tilde{r})$:

This is immediate by the law of iterated expectations

$$\begin{aligned}
&\mathbb{E}_{F_r}[\phi(W,\gamma(F_r),\theta,\alpha(F_r)]\\
&=\mathbb{E}_{F_r}\left[\mathbb{E}_{F_r}[\phi(W,\gamma(F_r),\theta,\alpha(F_r)|X]\right]\\
&=\mathbb{E}_{F_r}\left[v(X)\cdot\mathbb{E}_{F_r}\left[\left(\frac{d(y-\gamma_{1,F_r}(X))}{\pi_{F_r}(X)}-\frac{(1-d)(y-\gamma_{1,F_r}(X))}{1-\pi_{F_r}(X)}\right)\middle|X\right]\right]\\
&=\mathbb{E}_{F_r}[v(X)\cdot 0]\\
&=0
\end{aligned}$$

for $v(X)=\begin{bmatrix}\exp\left(-\lambda\cdot(\gamma_{1,F_r}(x)-\gamma_{0,F_r}(x)-\tilde{\tau})\right)\cdot(-\lambda)\\ \exp\left(-\lambda\cdot(\gamma_{1,F_r}(x)-\gamma_{0,F_r}(x)-\tilde{\tau})\right)\cdot(1-\lambda\cdot(\gamma_{1,F_r}(x)-\gamma_{0,F_r}(x)-\tilde{\tau}))\end{bmatrix}$
iii) $\int_{\mathcal{Y}_0\times\mathcal{Y}_1\times\mathcal{X}}\phi(w,\gamma(F_r),\theta,\alpha(F_r))H(dw)$ and $\int_{\mathcal{Y}_0\times\mathcal{Y}_1\times\mathcal{X}}\phi(w,\gamma(F_r),\theta,\alpha(F_r))F_0(dw)$ are continuous at $r=0$.

For a given H, we show that function $b:r\mapsto\int_{\mathcal{Y}_0\times\mathcal{Y}_1\times\mathcal{X}}\phi(w,\gamma(F_r),\theta,\alpha(F_r))H(dw)$ is continuous at $r=0$. Take a sequence $r_m\to r=0$, then $\phi_n(w):=\phi(w,\gamma(F_{r_m}),\theta,\alpha(F_{r_m}))$ converges H-almost everywhere to $\phi_0(w):=\phi(w,\gamma(F_0),\theta,\alpha(F_0))$. Moreover we have $\phi_m(w)\leq F(w)$ for all $m\in\mathbb{N}$ with $F\in L^1(H)$. By the dominated convergence theorem we have: $b(r_m)\to b(0)$ which is the desired result.

An analogous argument applies to the integral with respect to F_0. As a consequence of Theorems 1,2 and 3 in Chernozhukov et al. [2020] $\psi(w,\gamma,\theta,\alpha)$ is Neyman orthogonal. We can also verify Neyman orthogonality directly from the form of the $\bar{\psi}$ function. In particular:

$$\begin{aligned}
&\frac{\partial}{\partial r}\mathbb{E}[\psi(W,\theta,\gamma_{F_r},\alpha_{F_r})]\Bigg|_{r=0} \\
&= \frac{\partial}{\partial r}\mathbb{E}[g(W,\theta,\gamma)+\phi(W,\theta,\gamma,\alpha)]\Bigg|_{r=0} \\
&= \mathbb{E}\Bigg[\frac{\partial}{\partial r}\begin{bmatrix}\exp\left(-\lambda_0\cdot(\gamma_{1,F_r}(X)-\gamma_{0,F_r}(X)-\tilde{\tau})\right) \\ \exp\left(-\lambda_0\cdot(\gamma_{1,F_r}(X)-\gamma_{0,F_r}(X)-\tilde{\tau})\right)(\gamma_{1,F_r}(X)-\gamma_{0,F_r}(X)-\tilde{\tau})\end{bmatrix} \\
&+ \frac{\partial}{\partial r}\Bigg(\begin{bmatrix}\exp\left(-\lambda\cdot(\gamma_{1,F_r}(X)-\gamma_{0,F_r}(X)-\tilde{\tau})\right)\cdot(-\lambda) \\ \exp\left(-\lambda\cdot(\gamma_{1,F_r}(X)-\gamma_{0,F_r}(X)-\tilde{\tau})\right)\cdot(1-\lambda\cdot(\gamma_{1,F_r}(X)-\gamma_{0,F_r}(X)-\tilde{\tau}))\end{bmatrix} \\
&\times\left(\frac{D(Y-\gamma_{1,F_r}(X))}{\pi_{F_r}(X)}-\frac{(1-D)(Y-\gamma_{0,F_r}(X))}{1-\pi_{F_r}(X)}\right)\Bigg)\Bigg] \\
&= \mathbb{E}\Bigg[\begin{bmatrix}\exp\left(-\lambda\cdot(\gamma_{1,F_0}(X)-\gamma_{0,F_0}(X)-\tilde{\tau})\right)\cdot(-\lambda) \\ \exp\left(-\lambda\cdot(\gamma_{1,F_0}(X)-\gamma_{0,F_0}(X)-\tilde{\tau})\right)\cdot(1-\lambda\cdot(\gamma_{1,F_0}(X)-\gamma_{0,F_0}(X)-\tilde{\tau}))\end{bmatrix} \\
&\times\left(\frac{\partial\gamma_{1,F_r}(X)}{\partial r}-\frac{\partial\gamma_{0,F_r}(X)}{\partial r}\right)\Bigg|_{r=0} \\
&- \begin{bmatrix}\exp\left(-\lambda\cdot(\gamma_{1,F_0}(X)-\gamma_{0,F_0}(X)-\tilde{\tau})\right)\cdot(-\lambda) \\ \exp\left(-\lambda\cdot(\gamma_{1,F_0}(X)-\gamma_{0,F_0}(X)-\tilde{\tau})\right)\cdot(1-\lambda\cdot(\gamma_{1,F_0}(X)-\gamma_{0,F_0}(X)-\tilde{\tau}))\end{bmatrix} \\
&\times\left(\frac{D}{\pi_{F_0}(X)}\cdot\frac{\partial\gamma_{1,F_r}(X)}{\partial r}\Bigg|_{r=0}-\frac{(1-D)}{1-\pi_{F_0}(X)}\cdot\frac{\partial\gamma_{0,F_r}(X)}{\partial r}\Bigg|_{r=0}\right) \\
&+ \begin{bmatrix}\exp\left(-\lambda\cdot(\gamma_{1,F_0}(X)-\gamma_{0,F_0}(X)-\tilde{\tau})\right)\cdot(\lambda)^2 \\ \exp\left(-\lambda\cdot(\gamma_{1,F_0}(X)-\gamma_{0,F_0}(X)-\tilde{\tau})\right)\cdot(-\lambda)\cdot(2-\lambda\cdot(\gamma_{1,F_0}(X)-\gamma_{0,F_0}(X)-\tilde{\tau}))\end{bmatrix} \\
&\times\left(\frac{\partial\gamma_{1,F_r}(X)}{\partial r}-\frac{\partial\gamma_{0,F_r}(X)}{\partial r}\right)\Bigg|_{r=0}\times\left(\frac{D(Y-\gamma_{1,F_0}(x))}{\pi_{F_0}(X)}-\frac{(1-D)(Y-\gamma_{0,F_0}(X))}{1-\pi_{F_0}(X)}\right)\Bigg] \\
&+ \begin{bmatrix}\exp\left(-\lambda\cdot(\gamma_{F_0}(X)-\tilde{\tau})\right)\cdot(-\lambda) \\ \exp\left(-\lambda\cdot(\gamma_{1,F_0}(X)-\gamma_{0,F_0}(X)-\tilde{\tau})\right)\cdot(1-\lambda\cdot(\gamma_{1,F_0}(X)-\gamma_{0,F_0}(X)-\tilde{\tau}))\end{bmatrix} \\
&\times\left(D(Y-\gamma_{1,F_0}(X))\cdot\frac{\partial}{\partial r}\left(\frac{1}{\pi_{F_r}(X)}\right)\Bigg|_{r=0}-(1-D)(Y-\gamma_{0,F_0}(X))\cdot\frac{\partial}{\partial r}\left(\frac{1}{1-\pi_{F_r}(X)}\right)\Bigg|_{r=0}\right) \\
&= 0
\end{aligned}$$

The last equality follows by the law of iterated expectations. The first and second term cancel out since $\mathbb{E}\left[\frac{D}{\pi_{F_0}(X)}\Big|X\right]=1,\mathbb{E}\left[\frac{1-D}{1-\pi_{F_0}(X)}\Big|X\right]=1$. The third term is 0 because the nonparametric influence function is centered at 0 conditional on X. Moreover, $\mathbb{E}\left[D(Y-\mathbb{E}[Y|D=1,X]\Big|X\right]=0$ and $\mathbb{E}\left[(1-D)(Y-\mathbb{E}[Y|D=0,X]\Big|X\right]=0$ so whenever $\frac{\partial}{\partial r}\left(\frac{1}{\pi_{F_r}(X)}\right)\Big|_{r=0}$ and $\frac{\partial}{\partial r}\left(\frac{1}{1-\pi_{F_r}(X)}\right)\Big|_{r=0}$ are integrable, the fourth term is also 0, since they are measurable with respect to $\sigma(X)$. So $\frac{\partial}{\partial r}\mathbb{E}[\psi(W,\theta,\gamma_{F_r},\alpha_{F_r})]\Big|_{r=0}=0$. Observe that this result implies Neyman orthogonality with respect to the γ and α functions separately as well. To show the Neyman orthogonality with respect to γ and to set up the further results contained in Theorem 3 in Chernozhukov et al. [2020], we build the following construction. Consider the linear space of square integrable functions of X (with respect to some dominating measure), denoted as $\Gamma = L^2(\mathcal{X})$. $\mathcal{H}$ is the closed set of distributions which is a closed subset of the Banach space $L^1(\mathcal{Y}_0 \times \mathcal{Y}_1 \times \mathcal{X}, \mu)$ under some appropriate dominating measure μ. Denote the Hadamard differential of the conditional mean function at F_0 as $\frac{\partial \gamma(F_r)}{\partial r} : \mathcal{H} \to \Gamma$. Denote the Hadamard differential for $\bar{\psi}(\gamma(F_r), \alpha_0, \theta)$ at F_0 as $\frac{\partial \mathbb{E}[\psi(W,\gamma(F_r),\alpha(F_r),\theta)]}{\partial r} : \mathcal{H} \to \mathbb{R}^2$. Finally denote the Hadamard differential of $\bar{\psi}(\gamma, \theta)$ with respect to γ as $\frac{\partial \bar{\psi}(\gamma,\alpha,\theta)}{\partial \gamma} : \Gamma \to \mathbb{R}^2$. Then the following diagram commutes by Proposition 20.9 in Van der Vaart [2000].

$$\begin{tikzcd}
& \Gamma \arrow[dr, "\frac{\partial \bar{\psi}(\gamma,\alpha_0,\theta)}{\partial \gamma}"] & \\
\mathcal{H} \arrow[ur, "\frac{\partial \gamma(F_r)}{\partial r}"] \arrow[rr, "\frac{\partial \mathbb{E}[\psi(W,\gamma(F_r),\alpha_0,\theta)]}{\partial r}"] & & \mathbb{R}^2
\end{tikzcd}$$

By Neymann orthogonality with respect to the distribution F_r, $\frac{\partial \mathbb{E}[\psi(W,\gamma(F_r),\alpha_0,\theta)]}{\partial r} \equiv 0$. $\frac{\partial \bar{\psi}(\gamma,\theta)}{\partial \gamma}$ is onto Γ which satisfies Chernozhukov et al. [2020] Theorem 3 condition iv). Then, by linearity of the Hadamard derivative and the commutativity of the above diagram it must be the case that $\frac{\partial \bar{\psi}(W,\gamma,\alpha_0,\theta)}{\partial \gamma} \equiv 0$. That is, the Hadamard derivative is the 0 function from $\Gamma \to \mathbb{R}^2$. Note that this is the case because $\frac{\partial \gamma(F_r)}{\partial r}$ is onto $L^2(\mathcal{X})$. According to the above calculations we have, for $\delta_H := \left.\frac{\partial \gamma_{1,F_r}}{\partial r} - \frac{\partial \gamma_{0,F_r}}{\partial r}\right|_{r=0} \in L^2(\mathcal{X})$. Then as specified above: $\frac{\partial \mathbb{E}[\bar{\psi}(\theta,\alpha_0,\gamma)]}{\partial \gamma}(\delta_H)$ is a linear map from $L^2(X) \to \mathbb{R}^2$ in δ_H. In particular it maps to $0 \in \mathbb{R}^2$ for any $\delta_H(X)$, so it's the 0 map. Hence we verified Neyman orthogonality with respect to γ directly. □

I.7 Proof of Theorem 15

Lemma 25. *For $\bar{\psi}(\theta, \gamma, \alpha) = \mathbb{E}[\psi(w, \theta, \gamma, \alpha)]$ we have:*

i) $\bar{\psi}(\gamma, \alpha_0, \theta_0)$ is twice continuously Frechet differentiable in a neighborhood of γ_0.

ii) If Λ is bounded then $\forall \theta \in \Theta$, $\bar{\psi}(\gamma, \alpha_0, \theta) \leq \bar{C}\|\gamma - \gamma_0\|^2_{L_2}$.

Proof. Endow the spaces Γ with the $L^2(\mathcal{X}, \mu)$ norm and $\mathbb{R}^2$ with the standard Euclidean norm $\|\cdot\|$. We directly compute the directional derivative of $\bar{\psi}(\theta, \gamma, \alpha)$ with respect to γ.

$$
\begin{aligned}
&\frac{\partial}{\partial r}\bar{\psi}(\gamma, \theta, \alpha_0) \\
=&\mathbb{E}\Bigg[\begin{bmatrix} \exp\Big(-\lambda \cdot ((1-r)(\gamma_{1,0}(X) - \gamma_{0,0}(X)) + r(\gamma_1(X) - \gamma_0(X)) - \tilde{\tau}\Big) \cdot (\lambda)^2 \\ \exp\Big(-\lambda \cdot ((1-r)(\gamma_{1,0}(X) - \gamma_{0,0}(X)) + r(\gamma_1(X) - \gamma_0(X)) - \tilde{\tau}\Big) \cdot (-\lambda) \cdot (2 - \lambda \cdot ((1-r)(\gamma_{1,0}(X) - \gamma_{0,0}(X)) + r(\gamma_1 - \gamma_0) - \tilde{\tau})) \end{bmatrix} \\
&\times \left(\frac{D(Y - (1-r)\gamma_{1,0}(X) - r\gamma_1(X))}{\pi_{F_0}(X)} - \frac{(1-D)(Y - (1-r)\gamma_{0,0}(X) - r\gamma_0(X))}{1 - \pi_{F_0}(X)}\right)[(\gamma_1 - \gamma_{1,0}) - (\gamma_0 - \gamma_{0,0})]\Bigg]
\end{aligned}
$$

where we emphasized linearity in $[(\gamma_1 - \gamma_{1,0}) - (\gamma_0 - \gamma_{0,0})]$, the discrepancy between the estimated CATE and the true one. The second order Frechet derivative, if it exists, is a bi-linear operator given below, obtained by differentiating the first order Frechet derivative with respect to r. Then:

$$
\begin{aligned}
&\frac{\partial}{\partial r}\frac{\partial\bar{\psi}(\gamma, \theta, \alpha_0)}{\partial r} \\
=&\mathbb{E}\Bigg[\Bigg\{\begin{bmatrix} \exp(-\lambda((1-r)(\gamma_{1,0}(X) - \gamma_{0,0}(X)) + r(\gamma_1(X) - \gamma_0(X)) - \tilde{\tau}))(-\lambda)^3 \\ \exp(-\lambda((1-r)(\gamma_{1,0}(X) - \gamma_{0,0}(X)) + r(\gamma_1(X) - \gamma_0(X)) - \tilde{\tau}))(-\lambda)^2(3 - (1-r)(\gamma_{1,0}(X) - \gamma_{0,0}(X)) + r(\gamma_1(X) - \gamma_0(X)) - \tilde{\tau})) \end{bmatrix} \\
&\times \left(\frac{D(Y - (1-r)\gamma_{1,0}(X) - r\gamma_1(X))}{\pi_{F_0}(x)} - \frac{(1-D)(Y - (1-r)\gamma_{0,0}(X) - r\gamma_0(X))}{1 - \pi_{F_0}(x)}\right) \\
&\times[(\gamma_1 - \gamma_{1,0}) - (\gamma_0 - \gamma_{0,0}); (\gamma_1 - \gamma_{1,0}) - (\gamma_0 - \gamma_{0,0})] \\
&+ \begin{bmatrix} \exp\Big(-\lambda \cdot ((1-r)(\gamma_{1,0}(X) - \gamma_{0,0}(X)) + r(\gamma_1(X) - \gamma_0(X)) - \tilde{\tau}\Big) \cdot (\lambda)^2 \\ \exp\Big(-\lambda \cdot ((1-r)(\gamma_{1,0}(X) - \gamma_{0,0}(X)) + r(\gamma_1(X) - \gamma_0(X)) - \tilde{\tau}\Big) \cdot (-\lambda) \cdot (2 - \lambda \cdot ((1-r)(\gamma_{1,0}(X) - \gamma_{0,0}(X)) + r(\gamma_1 - \gamma_0) - \tilde{\tau})) \end{bmatrix} \\
&\times[(\gamma_1 - \gamma_{1,0}) - (\gamma_0 - \gamma_{0,0})]\left(\frac{D}{\pi_{F_0}(X)}[\gamma_1(X) - \gamma_{1,0}(X)] - \frac{1-D}{1 - \pi_{F_0}(X)}[\gamma_0(X) - \gamma_{0,0}(X)]\right)\Bigg\}\Bigg]
\end{aligned}
$$

Evaluated at $r = 0$ the second order directional derivatives are:

$$
\begin{aligned}
\mathbb{E}\Bigg[&\begin{bmatrix} \exp(-\lambda \cdot ((\gamma_{1,0}(X) - \gamma_{0,0}(X)) - \tilde{\tau})) \cdot (\lambda)^2 \\ \exp(-\lambda \cdot (\gamma_{1,0}(X) - \gamma_{0,0}(X)) - \tilde{\tau})) \cdot (-\lambda) \cdot (2 - \lambda \cdot ((\gamma_{1,0}(X) - \gamma_{0,0}(X)) - \tilde{\tau})) \end{bmatrix} \\
&\times [(\gamma_1(X) - \gamma_{1,0}(X)) - (\gamma_0(X) - \gamma_{0,0}(X)); (\gamma_1(X) - \gamma_{1,0}(X)) - (\gamma_0(X) - \gamma_{0,0}(X))]\Bigg\}\Bigg]
\end{aligned}
$$

by the law of iterated expectations. We emphasized that the above expression, is bi-linear [15] in $(\gamma_1(X) - \gamma_{1,0}(X)) - (\gamma_0(X) - \gamma_{0,0}(X)$. If the bi-linear map is continuous at $(\gamma_{1,0}, \gamma_{0,0})$ with respect to the operator norm then $\bar{\psi}$ is Frechet differentiable at $(\gamma_{1,0}, \gamma_{0,0})$

[15] Denote the space of linear maps from Banach spaces X to Y as $B(X, Y)$. It is itself a Banach space. Then one may identify $B(L^2(\mathcal{X})^2, B(L^2(\mathcal{X})^2; \mathbb{R}^2))$ with $B(L^2(\mathcal{X})^2 \times L^2(\mathcal{X})^2; \mathbb{R}^2)$. Then the second order Frechet derivative is a bi-linear map from $L^2(\mathcal{X})^2 \times L^2(\mathcal{X})^2\mathbb{R}^2$.

and the directional derivative and the Frechet derivative coincide. A sufficient condition is given by:

$$\left\| \frac{\partial^2}{\partial r^2}\bar{\psi}(\gamma, \theta, \alpha_0) \right\|_{L_2} < \infty$$

which translates to

$$\begin{aligned}&\Bigg\| \Bigg[\Bigg[\begin{matrix} \exp\left(-\lambda \cdot ((\gamma_{1,0}(X) - \gamma_{0,0}(X)) - \tilde{\tau})\right) \cdot (\lambda)^2 \\ \exp\left(-\lambda \cdot (\gamma_{1,0}(X) - \gamma_{0,0}(X)) - \tilde{\tau})\right) \cdot (-\lambda) \cdot (2 - \lambda \cdot ((\gamma_{1,0}(X) - \gamma_{0,0}(X)) - \tilde{\tau})) \end{matrix} \Bigg] \\ &\times \left[(\gamma_1(X) - \gamma_{1,0}(X)) - (\gamma_0(X) - \gamma_{0,0}(X)); (\gamma_1(X) - \gamma_{1,0}(X)) - (\gamma_0(X) - \gamma_{0,0}(X))\right] \Bigg\} \Bigg] \Bigg\|_{L_2} < \infty\end{aligned}$$

Then Frechet differentiability follows from Holder's inequality with $p = q = 2$. Under a slightly stronger condition which holds uniformly over $r \in [0, 1]$ one can obtain stronger results. Then Theorem 3 ii) in Chernozhukov et al. [2020] can be applied and we have:

$$\bar{\psi}(\gamma, \alpha_0, \theta_0) \leq C\|\gamma_1(X) - \gamma_{1,0}(X) - (\gamma_0(X) - \gamma_{0,0}(X))\|_{L^2}^2 \leq C\left\| \begin{bmatrix} \gamma_1(X) - \gamma_{1,0}(X) \\ \gamma_0(X) - \gamma_{0,0}(X) \end{bmatrix} \right\|_{L^2, E}^2$$

where the E denotes the Euclidean norm on $\mathbb{R}^2$. More generally consider $C(\lambda)$ defined below:

$$\begin{aligned}C(\lambda) := \Bigg\| \sup_{r \in (0,1)} \Bigg\{ &\begin{bmatrix} \exp\left(-\lambda \cdot ((1-r)(\gamma_{1,0}(X) - \gamma_{0,0}(X)) + r(\gamma_1(X) - \gamma_0(X)) - \tilde{\tau})\right) \\ \exp\left(-\lambda \cdot ((1-r)(\gamma_{1,0}(X) - \gamma_{0,0}(X)) + r(\gamma_1(X) - \gamma_0(X)) - \tilde{\tau})\right) \end{bmatrix} \\ &\begin{bmatrix} (\lambda)^2 & 0 \\ 0 & (-\lambda)(2 - \lambda \cdot ((1-r)(\gamma_{1,0}(X) - \gamma_{0,0}(X)) + r(\gamma_1 - \gamma_0) - \tilde{\tau})) \end{bmatrix} \Bigg\} \Bigg\|_E\end{aligned}$$

For a general bound here the constant depends on $C(\lambda)$. If Λ is compact then we can afford a representation of the theorem which is uniform across values for λ_0 which gives a much stronger version of the approximating function in λ and gets rid of some terms. For $\bar{C} = \sup_{\lambda \in \Lambda} C(\lambda)$ then $\psi(\gamma, \theta, \alpha_0) \leq C\|\gamma - \gamma_0\|_{L_2}^2$ and Frechet differentiability in a neighborhood of λ_0 follows in a straightforward way from the continuity of $C(\lambda)$ and the compactness of Λ. □

Remark 26. *Compactness of Λ would follow, for example, from Assumption 4 which restricts λ to be finite. We note that a condition in the form of $\bar{C} < \infty$ is sufficient and* *does not require compactness of* Λ*.*

Lemma 27 ($\sqrt{n}$ - consistency)**.** *proposition Let Assumption 5 hold. Then*

$$\frac{1}{\sqrt{n}}\sum_{k=1}^{K}\sum_{i\in I_k} g(W_i,\theta,\hat{\gamma}_{-k}) + \phi(W_i,\tilde{\theta}_{-k},\hat{\gamma}_{-k},\hat{\alpha}_{-k}) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi(W_i,\theta,\gamma_0,\alpha_0) + o_P(1)$$

Proof. The proof mirrors the blueprint of Theorem 15 in Chernozhukov et al. [2020]. We have:

$$\begin{aligned}
&g(W_i,\theta_0,\hat{\gamma}_{-k}) + \phi(W_i,\hat{\gamma}_{-k},\tilde{\theta}_{-k},\hat{\alpha}_{-k}) - \psi(W_i,\gamma_0,\theta_0,\alpha_0)\\
&= \underbrace{g(W_i,\theta_0,\hat{\gamma}_{-k}) - g(W_i,\theta_0,\gamma_0)}_{\hat{R}_{1i,-k}}\\
&+ \underbrace{\phi(W_i,\theta_0,\hat{\gamma}_{-k},\alpha_0) - \phi(W_i,\theta_0,\gamma_0,\alpha_0)}_{\hat{R}_{2i,-k}}\\
&+ \underbrace{\phi(W_i,\tilde{\theta}_{-k},\gamma_0,\hat{\alpha}_{-k}) - \phi(W_i,\theta_0,\gamma_0,\alpha_0)}_{\hat{R}_{3i,-k}}\\
&+ \underbrace{\phi(W_i,\tilde{\theta}_{-k},\hat{\gamma}_{-k},\hat{\alpha}_{-k}) - \phi(W_i,\tilde{\theta},\gamma_0,\hat{\alpha}_{-k}) + \phi(W_i,\hat{\gamma}_{-k},\alpha_0,\theta_0) - \phi(W_i,\gamma_0,\alpha_0,\theta_0)}_{\hat{\Delta}_{i,-k}}\\
&+ g(W_i,\theta_0,\gamma_0) + \phi(W_i,\theta_0,\gamma_0,\alpha_0)\\
&- \psi(W_i,\theta_0,\gamma_0)\\
&= \hat{R}_{1i,-k} + \hat{R}_{i2,-k} + \hat{R}_{i3,-k} + \hat{\Delta}_{i,-k}
\end{aligned}$$

Conditioning on the set not used in the nonparametric estimation we have:

$$\begin{aligned}
\mathbb{E}[\hat{R}_{1i,-k} + \hat{R}_{2i,-k}|I_k^c] &= \int_{\mathcal{X}} (g(w,\theta_0,\hat{\gamma}_{-k},\alpha_0) + \phi(w,\theta_0,\hat{\gamma}_{-k},\alpha_0))dF_0(w)\\
&= \int_{\mathcal{X}} \psi(w,\theta_0,\hat{\gamma}_{-k},\alpha_0)dF_0(w)\\
&= \bar{\psi}(\theta_0,\hat{\gamma}_{-k},\alpha_0)
\end{aligned}$$

The third term's expected value, conditional on the subsample is given by $\mathbb{E}[\hat{R}_{i3,-k}|I_k] = \int_{\mathcal{X}} \phi(W_i,\tilde{\theta}_{-k},\gamma_0,\hat{\alpha}_{-k})dF_0(w) = 0$. Finally consider the term:

$$\frac{1}{\sqrt{n}}\sum_{i\in I_c} \hat{R}_{1i,-k} + \hat{R}_{i2,-k} + \hat{R}_{i3,-k} - \mathbb{E}[\hat{R}_{1,-k} + \hat{R}_{2,-k}|I_k^c] + \mathbb{E}[\hat{R}_{1,-k} + \hat{R}_{2,-k}|I_k^c]$$

Now by Kennedy et al. [2020] Lemma 2 we have:

$$\begin{aligned}\frac{1}{\sqrt{n}}\sum_{i\in I_c}\hat{R}_{1i,-k}+\hat{R}_{i2,-k}-\mathbb{E}[\hat{R}_{1,-k}+\hat{R}_{2,-k}|I_k^c] &= O_P(\|\psi(W_i,\theta_0,\hat{\gamma}_k,\alpha_0)-\psi(W_i,\theta_0,\gamma_0,\alpha_0)\|_L^2)\\ &= O_P(\|\hat{\gamma}_k-\gamma_0\|_L^2)\end{aligned}$$

where the last equality follows form proposition 25 ii).

Again by Kennedy et al. [2020] Lemma 2

$$\begin{aligned}\frac{1}{\sqrt{n}}\sum_{i\in I_k}\hat{R}_{i3,-k}-\mathbb{E}[\hat{R}_{i3,-k}|I_k] &= O_P(\|\phi(W_i,\tilde{\theta}_{-k},\gamma_0,\hat{\alpha}_{-k})-\phi(W_i,\theta_0,\gamma_0,\alpha_0)\|_{L^2})\\ &= O_P(\|\hat{\alpha}-\alpha_0|_L^2)+O_P(\|\tilde{\theta}-\theta_0\|_{\mathbb{R}^2})\end{aligned}$$

since $\phi(\cdot)$ is linear in α and differentiable in θ. Then Assumption 5 guarantees that these last two terms are $o_P(1)$. Furthermore, by Proposition 25 ii) for n sufficiently large we have:

$$\mathbb{E}[\hat{R}_{1,-k}+\hat{R}_{2,-k}|I_k]\leq\sqrt{n}C\|\hat{\gamma}_k-\gamma_0\|^2$$

for $\bar{C}$ given in proposition 25. A similar argument shows $\frac{1}{\sqrt{n}}\sum_{i\in I_k^c}\Delta_{i,-k}=o_P(1)$. If that's the case, we conclude that:

$$\frac{1}{\sqrt{n}}\sum_{i\in I_k}g(W_i,\theta_0,\hat{\gamma}_{-k})+\phi(W_i,\tilde{\theta}_k,\hat{\gamma}_k,\hat{\alpha}_{-k})=\frac{1}{\sqrt{n}}\sum_{i\in I_k}\psi(W_i,\gamma_0,\theta_0,\hat{\alpha}_0)+o_P(1)$$

□

Lemma 28 (Jacobian consistency)**.** *For Jacobian G of the debiased moment conditions:*

$$G=\mathbb{E}[D\psi(w,\theta_0,\gamma_0,\alpha_0)]=\mathbb{E}\left[\frac{\partial}{\partial\theta}\psi(w,\theta_0,\gamma_0,\alpha_0)\right]\tag{A.26}$$

and $\hat{\theta}\xrightarrow{p}\theta_0$ we have $\|\frac{\partial\hat{\psi}(\hat{\theta})}{\partial\theta}-G\|=o_P(1)$.

Proof. First observe that at $\gamma = \gamma_0$ and $\alpha = \alpha_0$:

$$\begin{aligned}\mathbb{E}\left[\frac{\partial}{\partial\theta}\psi(w,\theta,\gamma,\alpha)\right] &= \mathbb{E}\left[\frac{\partial}{\partial\theta}g(w,\theta,\gamma,\alpha)\right] + \mathbb{E}\left[\frac{\partial}{\partial\theta}\phi(w,\theta,\gamma,\alpha)\right] \\ &= \mathbb{E}\left[\frac{\partial}{\partial\theta}g(w,\theta,\gamma)\right] + 0 \\ &= \mathbb{E}\left[\frac{\partial}{\partial\theta}g(w,\theta,\gamma)\right]\end{aligned}$$

by the law of iterated expectations. (N.B: if α_0 is the propensity score than this holds in a neighborhood of the true F_0). Now, to show the result we verify the conditions in Lemma 17 of Chernozhukov et al. [2020]. First notice that for $\frac{\partial g(w,\theta,\gamma)}{\partial\theta}$, each of the functions:

$$\begin{aligned}&\theta \mapsto -1 \\ &\theta \mapsto 0 \\ &\theta \mapsto -\exp(-\lambda(\tau(x)-\tilde{\tau}))(\tau(x)-\tilde{\tau}) \\ &\theta \mapsto -\exp(-\lambda(\tau(x)-\tilde{\tau}))(\tau(x)-\tilde{\tau}))^2\end{aligned}$$

is continuously differentiable in θ at θ_0. The first two are constants and the other two derivatives are, respectively:

$$\begin{aligned}\exp(-\lambda(\tau(x)-\tilde{\tau}))(\tau(x)-\tilde{\tau})^2 \\ \exp(-\lambda(\tau(x)-\tilde{\tau}))(\tau(x)-\tilde{\tau}))^3\end{aligned}$$

Hence if $\mathbb{E}[\exp(-\lambda_0(\tau(x)-\tilde{\tau}))(\tau(x)-\tilde{\tau})^2] < \infty$ and $\mathbb{E}[\exp(-\lambda_0(\tau(x)-\tilde{\tau}))(\tau(x)-\tilde{\tau})^3] < \infty$. In particular Assumption 2 is a sufficient condition for locally bounded derivatives which satisfies Assumption 4 ii) in Chernozhukov et al. [2020]. Assumption 4 iii), namely $\int(\frac{\partial g_j}{\partial\theta_l}(w,\theta,\hat{\gamma}_k) - \frac{\partial g_j}{\partial\theta_l}(w,\theta,\gamma_0))dF_0(w)$ follows from the continuous mapping theorem and continuity of the the maps above with respect to $\gamma(\cdot) = \tau(\cdot)$ in the $\|\cdot\|_{L_2}$ norm. □

We are finally ready to prove 15 using the lemmas above.

Theorem 15 (Asymptotic normality of θ)**.** *Let Assumptions 1–5. For $\hat{\theta}$ defined in*

Equation (17)*:*

$$\begin{aligned}\sqrt{n}(\hat{\theta}-\theta_0) &\xrightarrow{d} \mathcal{N}(0,S)\\ S &:= (G)^{-1}\Omega(G')^{-1}\\ G &:= \mathbb{E}[D_\theta\psi(w,\theta,\gamma_0,\alpha_0)]\\ \Omega &:= \mathbb{E}[\psi(w,\theta_0,\gamma_0,\alpha_0)\psi(w,\theta_0,\gamma_0,\alpha_0)^T]\end{aligned}$$

and $D_\theta\psi(\cdot)$ *is the Jacobian of the augmented moment condition with respect to the parameters in* θ*.*

Denote $\hat{G} = \frac{\hat{g}(w,\hat{\theta},\hat{\gamma})}{\partial\theta}$. First note that by Lemma 28 we have $\|\hat{G}-G\| = o_P(1)$. Then, like in Chernozhukov et al. [2018] we have:

$$\begin{aligned}\hat{G}^{-1}-G^{-1} &= (G+\hat{\Delta}_n)^{-1} - G^{-1}\\ &= (G+\hat{\Delta}_n)^{-1}(GG^{-1}) - (G+\hat{\Delta}_n)G^{-1}\\ &= (G+\hat{\Delta}_n)^{-1}(G-(G+\hat{\Delta}_n))G^{-1}\\ &= (G+\hat{\Delta}_n)^{-1}\hat{\Delta}_n G^{-1}\end{aligned}$$

Then like in Chernozhukov et al. [2018] from the basic matrix inequalities we have:

$$\begin{aligned}\|\hat{G}^{-1}-G^{-1}\| &= \|(G+\hat{\Delta}_n)^{-1}\hat{\Delta}_n G^{-1}\|\\ &= \|(G+\hat{\Delta}_n)^{-1}\|\cdot\|\hat{\Delta}_n\|\cdot\|G^{-1}\|\\ &= O_P(1)\cdot o_P(1)\cdot O_P(1)\\ &= o_P(1)\end{aligned}$$

Now by the central limit theorem and Lemma 27 we have:

$$\begin{aligned}&\frac{1}{|K|}\sum_{k\in K}\left(\frac{1}{\sqrt{n}}\sum_{i\in I_k} g(W_i,\theta,\gamma_0) + \phi(W_i,\tilde{\theta}_{-k},\hat{\gamma}_{-k},,\hat{\alpha}_{-k})\right)\\ &= \frac{1}{|K|}\sum_{k\in K}\frac{1}{\sqrt{n}}\sum_{i\in I_k}\psi(W_i,\theta,\gamma_0,\alpha_0) + o_P(1) \xrightarrow{d} \mathcal{N}(0,\Omega)\end{aligned}$$

where $\Omega = \mathbb{E}[\psi(w,\theta_0,\gamma_0,\alpha_0)\psi(w,\theta_0,\gamma_0,\alpha_0)]$. Finally observe that a standard GMM

Taylor linearization gives:

$$\begin{aligned}\sqrt{n}\begin{bmatrix}\nu-\nu_0\\ \lambda-\lambda_0\end{bmatrix} =& \left\{\frac{\partial}{\partial\theta}\hat{\psi}(w,\theta_0,\hat{\gamma},\hat{\alpha})'V\frac{\partial}{\partial\theta}\hat{\psi}(w,\theta_0,\hat{\gamma},\hat{\alpha})\right\}^{-1}\frac{\partial}{\partial\theta}\hat{\psi}(w,\theta_0,\hat{\gamma},\hat{\alpha})'V\\ &\times\frac{1}{|K|}\sum_{k\in K}\left(\frac{1}{\sqrt{n}}\sum_{i\in I_k}g(W_i,\theta,\hat{\gamma}_{-k})+\phi(W_i,\tilde{\theta}_{-k},\hat{\gamma}_{-k})\right)\\ &=(G'VG)^{-1}G'V\left(\frac{1}{|K|}\sum_{k\in K}\frac{1}{\sqrt{n}}\sum_{i\in I_k}\psi(W_i,\theta,\gamma_0,\alpha_0)\right)+o_P(1)\xrightarrow{d}\mathcal{N}(0,S)\end{aligned}$$

which is the desired result.

I.8 Auxiliary Lemmas

Lemma 29. *(Kennedy et al. [2020]-Lemma 2)*

Let $\hat{g}(\cdot)$ be a function estimated from the I_k^c sample and evaluated on the I_k sample. Then $(\mathbb{P}_n-\mathbb{P})(\hat{g}-g_0)=O_P\left(\frac{|\hat{g}-g_0|}{\sqrt{n}}\right)$.

Proof. The proof follows from independence of I_k and I_k^c, the computation of conditional variance and Markov's inequality. See Kennedy et al. [2020] for a detailed treatment. □

J Additional Figures and Examples

In this section I include some additional visualizations and examples:

Example 30. *To visualize Corollary 20 consider the case where the dimension of the covariate space is $k=2$. The original data is normal $\mathcal{N}(\mu,\Sigma)$ with $\mu=(4,3)^T$ $\Sigma=\begin{bmatrix}2&0.5\\0.5&2\end{bmatrix}$. $\tau(x)=X^T\beta$ is linear with $\beta=(4,1)^T$. Experimental $ATE=18.98$. Target $ATE=15$.*

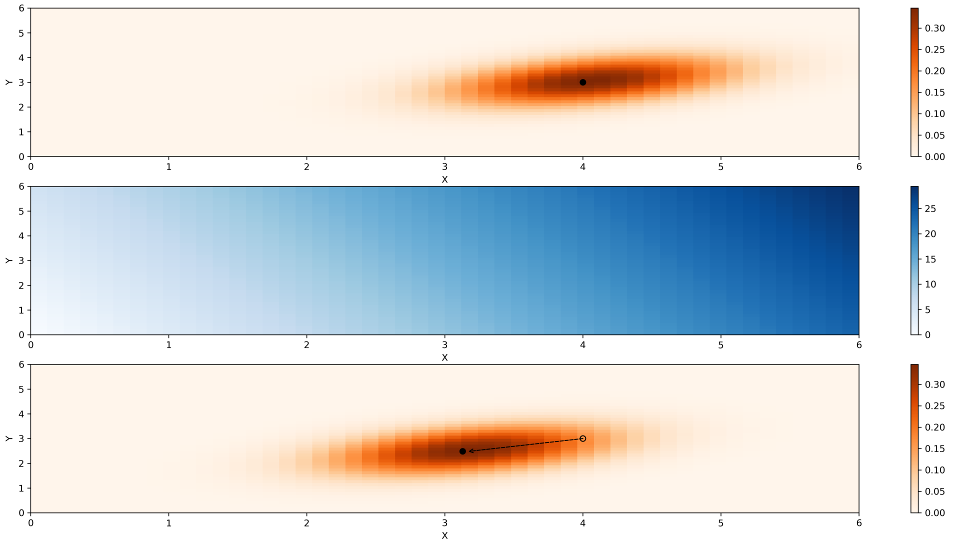

Figure 8: least favorable distribution for normally distributed data. First panel in red shows the density of $\mathcal{N}(\mu, \Sigma) \sim \mathcal{N}\left(\mu = \begin{bmatrix} 4 \\ 3 \end{bmatrix}; \Sigma = \begin{bmatrix} 2 & 0.5 \\ 0.5 & 2 \end{bmatrix}\right)$, the experimental distribution. The second panel shows $CATE = X^T\beta$, linear in X with $\beta = (4, 1)^T$. The third panel shows the parameter shift of the *least favorable distribution*.

Here $\lambda_0 = 0.396$*.* $\mu^* = (3.1288, 2.4852)$*. The KL divergence, for two multivariate normal distributions* $(\mu_1, \Sigma_1), (\mu_2, \Sigma_2)$ *is given by:*
$KL(X_1||X_2) = \frac{1}{2}\left[\log\left(\frac{|\Sigma_2|}{|\Sigma_1|}\right) - k + (\mu_2 - \mu_1)^T\Sigma_2^{-1}(\mu_2 - \mu_1) + tr(\Sigma_2^{-1}\Sigma_1)\right]$*. One could always compute the value of the KL divergence applying the nonparametric formula*

$$\delta^* = \int_{\mathcal{X}} \exp(-\lambda_0(\tau(X) - \tilde{\tau}))d\mu_X$$

or in this case, the "parametric" formula given by the KL divergence between two normal distributions.[16] *In this example both ways of computing the correspond to* $\delta^* = 0.789$ *corresponding to the mean shift illustrated above.*

[16] The "parametric" formula to compute the KL divergence would not be valid in general since the *least favorable distribution* may belong to a different class than the experimental distribution. Conversely, the "nonparametric" formula is always valid.

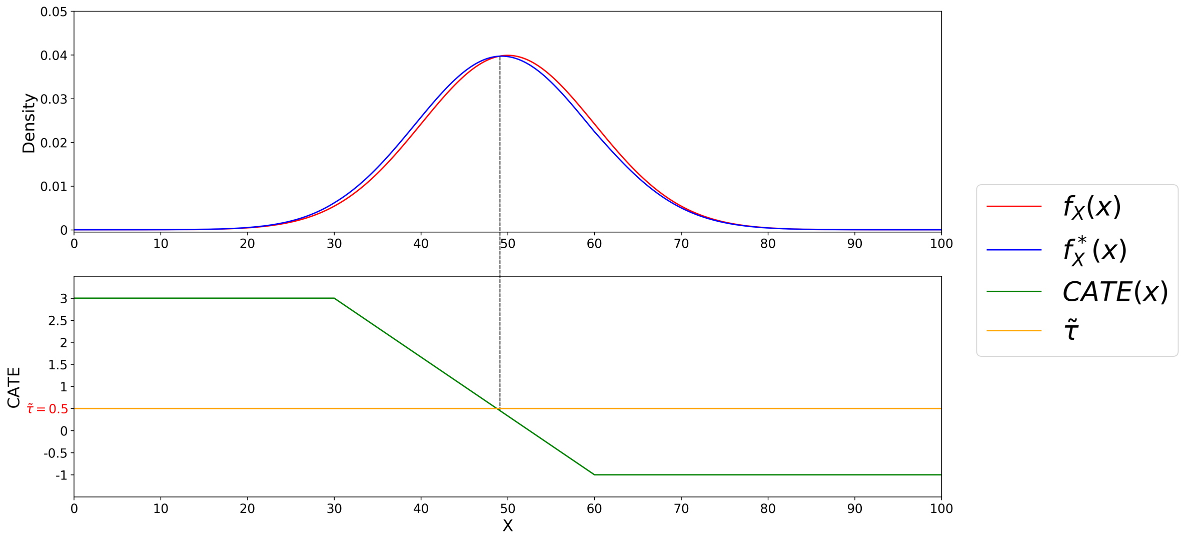

Figure 9: Piece-wise Linear CATE, experimental distribution is $\mathcal{N}(50, 10^2)$. Experimental ATE is 0.433, while $\tilde{\tau} = 0.5$. Because the experimental ATE is lower than the least favorable , F_X^* down-weights F_X on the subset of $\mathbb{R}$ where the $\tau(x)$ is greater than $\tilde{\tau}$ and up-weights it where it's lower. The blue curve is the closest curve to the red one, in KL-divergence, among the ones that satisfy $\tau \geq 0.5$.

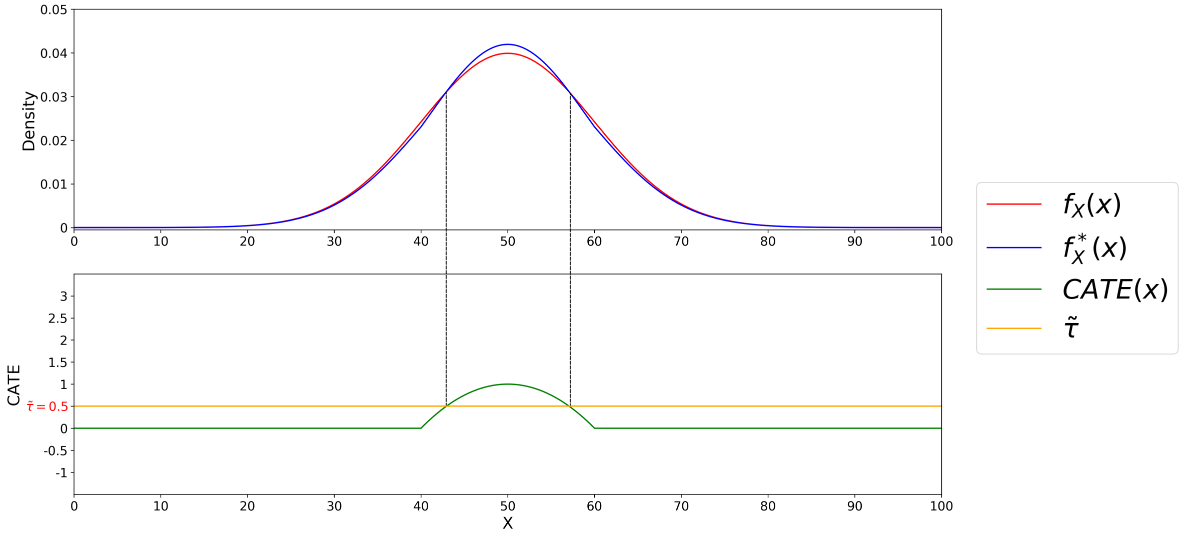

Figure 10: Piecewise Quadratic CATE, experimental distribution is $\mathcal{N}(50, 10^2)$. Experimental ATE is 0.484 while $\tilde{\tau} = 0.5$. Because the experimental ATE is lower than the least favorable , F_X^* down-weights F_X on the subset of $\mathbb{R}$ where the $\tau(x)$ is smaller than $\tilde{\tau}$ and up-weights it where it's greater. The blue curve is the closest curve to the red one, in KL-divergence, among the ones that satisfy $\tau \geq 0.5$.

Example 31. *Let's now see graphically how to construct an example for a one di-*

mensional continuous variable example. In Figures 9 and 10 conditional treatment effects, given the 1-dimensional variable X are in green, the experimental distribution is $\mathcal{N}(50, 10^2)$ is in red. Suppose that the policy-maker's wants to maintain the claim $ATE \leq 0.5$. The experimental ATE and the "least favorable" ATE are obtained by integrating the green curve $\tau(x)$ against the red curve $dF_X(x)$ (which has density $f_X(x)$) and the blue curve $dF_X^*(x)$ (which has density $f_X^*(x)$ respectively. The blue curve is the closest distribution to the experimental distribution in red, as measured by the KL divergence, that delivers the "least favorable " ATE $\tilde{\tau} = 0.5$.