Supplementary Appendix for “Robustness, Heterogeneous Treatment Effects, and Covariate Shifts”

Pietro Emilio Spini

July 2024

1 General divergences

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 [2023]. The Kullback-Leibler divergence is a special case of a more general construction, known as φ-divergence. It is introduced below:

Definition 1 (φ-divergence)**.** *Consider the φ-divergence between F_X and F'_X given by:*

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

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 Theorem 8 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 2 (Fenchel-Conjugate)**.** *Given a topological vector space Z and convex function $\varphi : Z \to \mathbb{R}$, the Fenchel-conjugate $\varphi^* : Z^* \to \mathbb{R}$, defined on the dual space of*

Z, is defined by:

$$\varphi^* : z^* \mapsto \sup_{z\in Z}\langle z^*, z\rangle - \varphi(z)$$

Then we can have a generalization of the policy-maker's problem in Equations (4) and (5) for an arbitrary φ divergence in 1:

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

$$s.t. \int_{\mathcal{X}} \tau(x) dF'_X(x) \leq \tilde{\tau} \tag{2}$$

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

$$\sup_{\lambda\in\Lambda}\sup_{\xi}\left(\inf_{F'_X:\ P'_X \ll P_X; P'_X(\mathcal{X})=1} D_\varphi(F'_X||F_X) + \lambda\int_{\mathcal{X}}(\tau(x)-\tilde{\tau})dF'_X(x) + \xi\left(\int_{\mathcal{X}} dF'_X(x) - 1\right)\right) \tag{3}$$

where and ξ is the Lagrange multiplier for integration to 1 (i.e. it is a probability measure), 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 the 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 (7) the Lagrange multiplier λ is a 1-dimensional parameter. Notice that after fixing the experimental distribution F^*_X, the map $D_{KL}(\cdot||F_X)$ is convex in its first argument. Notice that, the Radon-Nikodym derivative $\frac{dF'_X}{dF_X}(x) \geq 0, \forall x \in \mathcal{X}$. We can express the inner problem as:

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

and recognize that, by rewriting the infimum as the supremum with a negative sign, we can substitute the expression for the Fenchel-conjugate of φ if we can interchange the supremum and integration[1]. Using the definition of Fenchel-conjugate:

$$\begin{aligned}
&\inf_{F'_X:\ P'_X \ll P_X; P'_X(\mathcal{X})=1} \int_{\mathcal{X}} \left(\varphi\left(\frac{dF'_X}{dF_X}(x)\right) - (-\lambda(\tau(x)-\tilde{\tau}) - \xi)\,\frac{dF'_X}{dF_X}(x)\right) dF_X(x) - \xi \\
&= - \sup_{F'_X:\ P'_X \ll P_X; P'_X(\mathcal{X})=1} \int_{\mathcal{X}} - \left(\varphi\left(\frac{dF'_X}{dF_X}(x)\right) - (-\lambda(\tau(x)-\tilde{\tau}) - \xi)\,\frac{dF'_X}{dF_X}(x)\right) dF_X(x) - \xi \\
&= - \int_{\mathcal{X}} \left(\sup_{x\in\mathcal{X}} z\left(-\lambda(\tau(x)-\tilde{\tau}) - \xi\right) - \varphi(z)\right) dF_X(x) - \xi \\
&= - \int_{\mathcal{X}} \varphi^*(-\lambda(\tau(x)-\tilde{\tau}) - \xi) dF_X(x) - \xi
\end{aligned}$$

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^*:

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

where $\dot{\varphi}^*(\cdot)$ is the derivative of $\varphi^*(\cdot)$ with respect to its argument. Observe that, for the KL divergence, the Fenchel-conjugate of $\varphi(t) = t\log(t)$ is given by $\varphi(t^*) = \exp(t^* - 1)$. Plugging this in and 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

$$\int_{\mathcal{X}} \dot{\varphi}^*(-\lambda^*(\tau(x)-\tilde{\tau}) - \xi^*)(\tau(x) - \tilde{\tau} + \dot{\xi}^*_\lambda) dF_X(x) - \dot{\xi}^*_\lambda = 0 \tag{4}$$

where $\dot{\xi}^*_\lambda$ is the derivative of ξ^* with respect to λ and λ^* is the value that implicitly solves the moment condition in Equation (4). Observe that plugging Equation (1) into Equation (4) allows to simplify it to:

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

[1] Hafsa and Mandallena [2003] contains a review of sufficient conditions for which this interchange is valid.

since the two terms in $\dot{\xi}^*_\lambda$ cancel out (by the envelope theorem). 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 (4) recovers Equation (7). In general, conditions that guarantee existence and uniqueness of the *least favorable distribution* given a φ-divergence may be subtle. For a review, consider Komunjer and Ragusa [2016].

2 Other Extensions

2.1 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:

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

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

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

Now consider 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:

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

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

2.2 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.

2.3 Distributional treatment effects

The treatment of this paper has focused on ATE as the main causal parameter of interest. The framework can be extended to other functionals. For example, consider the distributional treatment effect, at pre-selected level y', written as a functional of the covariate distribution:

$$\Delta(y') = \int_{\mathcal{X}} F_{Y_1|X=x}(y'|x) - F_{Y_0|X=x}(y'|x) dF_X(x)$$

Defining the conditional distributional treatment effect as $\Delta(y'|x) := F_{Y_1|X=x}(y'|x) - F_{Y_0|X=x}(y'|x)$. We can then formulate the distributional equivalent of the policy-maker's problem:

$$\inf_{dF'_X:\ dF'_X \ll dF_X; dF'_X(\mathcal{X})=1} D_{KL}(F'_X||F_X) \tag{5}$$

$$s.t. \int_{\mathcal{X}} \Delta(y'|x) dF'_X(x) \leq \tilde{\Delta}(y') \tag{6}$$

Here, $\tilde{\Delta}(y')$ fixes a particular conclusion on the distributional treatment effect at y', which may reflect the minimal threshold for the the policy to be cost effective. It is clear that here $\Delta(y'|x)$ plays the role of a summary of the heterogeneity of distributional treatment effects. In the the policy-maker's problem that targeted the ATE, the same role was played by $\tau(x)$, the CATE. Under Assumption 2, $\Delta(y'|x)$ does not depend on the distribution of F_X. One can give conditions under which $\Delta(y'|x)$ is identified in the experiment. One can obtain a result analogous to Theorem 8 to characterize the *least-favorable distribution* F^*_X and for the metric of robustness $\delta^*(\tilde{\Delta})$ for an experimental conclusion on distributional treatment effects. This is because the constrain set is again in Equation (6) is linear in the distribution of covariates, which allows the machinery of Theorem 8 to apply. On should note that the metric of robustness here considers a conclusion at a single point y'. That is, the *least-favorable distribution* F_X^* will re-weight the covariates to make the distributional effect at y' no larger than $\tilde{\Delta}(y')$. One may of course always specify finitely many points $y_1'), \cdots y_B'$ and control the respective distributional effects with thresholds $\tilde{\Delta}(y_1') \cdots \tilde{\Delta}(y_1')$, simultaneously. The solution F_X^*, if it exists, is characterized by B-many Lagrange multiplier per constraint, as the discussion in Section C illustrates. It is of course possible that, given the heterogeneity of distributional effects in $\Delta(y'|x)$ the B-many restrictions are incompatible, leading to no solution (and an arbitraly large value for the robustness metric). It is also possible to specify a continuum of restrictions, given by $\tilde{\Delta}(y')$. While the theoretical problem could leverage the results for existence of conditional information projections in Komunjer and Ragusa [2016], the computational cost would be much more burdensome.

3 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 Theorem 8 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 1. 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 Theorem 8. 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 1, the only sub-population that has $\tau(x) \geq \tilde{\tau}_b$ is $x = 0.6$, concentrated on a singleton. Distributions that put unit mass on singletons are not feasible in Equation (5). For $\tilde{\tau} = \tilde{\tau}_b$, the feasible set is empty so there is no solution. Looking 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?

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 Theorem 8. 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}.

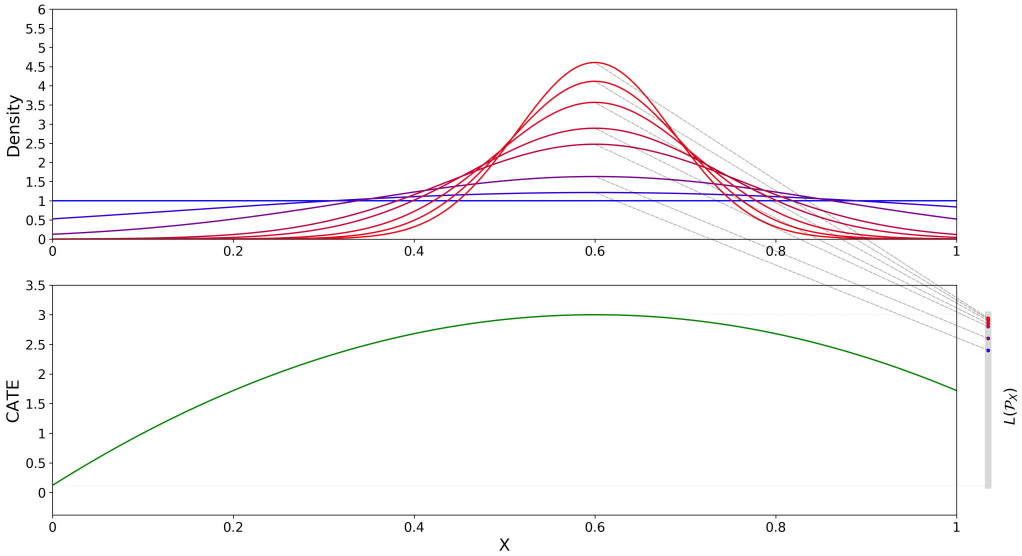

Figure 1: Local to boundary conditions. The lower panel displays the conditional average treatment effect, $\tau(x)$ for a univariate variable X. The experimental distribution is in blue: the uniform distribution. The gray segment on the left labelled $L(\mathcal{P}_X)$ is the image of all distributions 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, Theorem 8 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 $x = \arg\max \tau(x) = 0.6$.

Proposition 3 (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 with $\tilde{\tau}_m$, 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 $\mathcal{X}$.

The point-mass distribution δ_{x_b} is not a solution to Equation 4 with $\tilde{\tau}_b$ because the feasible set never includes point mass distributions unless $\mathcal{X}$ is discrete. Proposition 3 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}.[2]

[2] In fact, Posner [1975] showed that D_{KL} is lower-semicontinuous, that is, if $P_n \to P$ weakly, then

4 Proofs of additional results and lemmas

4.1 Proof of Lemma 21

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:

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

□

4.2 Proof of Fact 10

Proof. First, $F_X^* \ll F_X$ simply implies $p_1 = 0 \implies p_1^* = 0$. Aside such a trivial case, 10 characterizes $\frac{p_1^*}{p_1}$. We solve for the Lagrange multiplier λ in 10 noting that:

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

$\lim_{n\to\infty} D_{KL}(P_n || Q) \geq D_{KL}(P || Q)$. In this case we have $+\infty > 0$

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

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

Finally, replacing $-\lambda$ in 9 we have:

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

which, with $\tilde{\tau}=0$, is the solution obtained in Equation (9). □

Proposition 4. *Let $\epsilon > 0$. Then for $\tilde{\tau} > \inf_{\mathcal{X}} \tau(x) + \epsilon$, $\delta^*(\tilde{\tau})$ in Definition 4 is decreasing in $\tilde{\tau}$.*

4.3 Proof of Proposition 4

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 decreasing in $\tilde{\tau}$. This monotonicity is preserved if the reverse inequality is considered. □

4.4 Proof of Proposition 3

Proposition 3 (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 with $\tilde{\tau}_m$, 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 Theorem 8 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:

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

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:

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

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:

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

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. Also:

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

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 $\mathcal{X}_b$, rather than *any* distribution like in Figure 2. I leave this interesting case for future work.

4.5 Proof of Proposition 17

Proof. Suppose $\mathcal{X} = \mathbb{R}^k$, $X \sim \mathcal{N}(\mu, \sigma)$ and $\tau(x) = x^T A x + x^T \beta + c$. By Theorem 8 the Radon-Nikodym derivative of the least favorable distribution is given by Equation (6) so the distribution of F^*_X must have density:

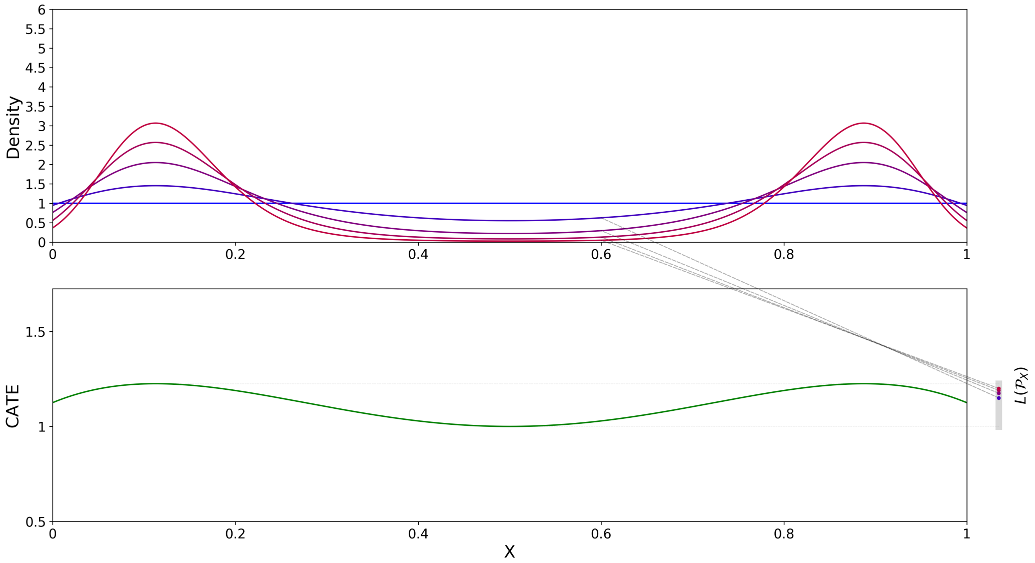

Figure 2: Here $\tau(x)$ is quadratic, experimental distribution is uniform and there are two peaks. It appears that the *least favorable distribution* concentrates around both peaks.

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

from which we can recognize the form of $\mathcal{N}(\mu^*,\Sigma^*)$. The steps above follow from completing the square and from the properties of $\exp(\cdot)$. □

4.6 Proof of Proposition 11

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. Then we have: $\mathbb{E}_{F_r}[Y_d|X] = \mathbb{E}_{F_r}[Y|D=d,X]$. We can obtain 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 $r=0$. Computing the derivative of the moment condition with respect to r and evaluating it at $r=0$ we have:

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

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 it with the derivative of the moment condition w.r.t γ we have:

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

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

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

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 $\mathbb{R}^2$.

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

4.7 Proof of Proposition 12

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 (13) 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

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

$$v(X) = \begin{bmatrix} \exp\left(-\lambda\cdot(\gamma_{1,F_r}(x)-\gamma_{0,F_r}(x)-\tilde{\tau})\right)\cdot(-\lambda) \\ \exp\left(-\lambda\cdot(\gamma_{1,F_r}(x)-\gamma_{0,F_r}(x)-\tilde{\tau})\right)\cdot(1-\lambda\cdot(\gamma_{1,F_r}(x)-\gamma_{0,F_r}(x)-\tilde{\tau})) \end{bmatrix}$$

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:

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

$$
\begin{aligned}
&= \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 $\delta_H := \frac{\partial\gamma_{1,F_r}}{\partial r}-\frac{\partial\gamma_{0,F_r}}{\partial r}\Big|_{r=0}\in L^2(\mathcal{X})$. 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. □

4.8 Proof of Lemma 22

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

4.9 Proof of Lemma 23

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:

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

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 [3] 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 $(\gamma_{1,0},\gamma_{0,0})$ and the directional derivative and the Frechet derivative coincide. A sufficient condition is given by the following.

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

which translates to

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

[3] 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 $L^2(\mathcal{X})^2\times L^2(\mathcal{X})^2\mathbb{R}^2$.

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:

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

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

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

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 5. *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.*

4.10 Proof of Lemma 24

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 top two are constants and the other two derivatives are, respectively:

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

Hence if $\mathbb{E}[\exp(-\lambda_0(\tau(x)-\tilde{\tau}))(\tau(x)-\tilde{\tau})^2] < \infty$ and $\mathbb{E}[\exp(-\lambda_0(\tau(x)-\tilde{\tau}))(\tau(x)-\tilde{\tau})^3] < \infty$. 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. □

4.11 Proof of Lemma 25

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 from Lemma 23 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 Lemma 23 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 23. 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:

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

□

4.12 Proof of Theorem 19

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

4.13 Proof of Theorem 20

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]. □

References

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. *Econometrica*, 91(1):263–298, 2023.

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.

O. A. Hafsa and J.-P. Mandallena. Interchange of infimum and integral. *Calculus of Variations and Partial Differential Equations*, 18(4):433–449, 2003.

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.

I. Komunjer and G. Ragusa. Existence and characterization of conditional density projections. *Econometric Theory*, 32(4):947–987, 2016.

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

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

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