Semiparametric Counterfactual Density Estimation

Edward H. Kennedy, Sivaraman Balakrishnan, Larry Wasserman

Department of Statistics & Data Science
Carnegie Mellon University

{edward, siva, larry} @ stat.cmu.edu

Abstract

Causal effects are often characterized with averages, which can give an incomplete picture of the underlying counterfactual distributions. Here we consider estimating the entire counterfactual density and generic functionals thereof. We focus on two kinds of target parameters. The first is a density approximation, defined by a projection onto a finite-dimensional model using a generalized distance metric, which includes f-divergences as well as L_p norms. The second is the distance between counterfactual densities, which can be used as a more nuanced effect measure than the mean difference, and as a tool for model selection. We study nonparametric efficiency bounds for these targets, giving results for smooth but otherwise generic models and distances. Importantly, we show how these bounds connect to means of particular non-trivial functions of counterfactuals, linking the problems of density and mean estimation. We go on to propose doubly robust-style estimators for the density approximations and distances, and study their rates of convergence, showing they can be optimally efficient in large nonparametric models. We also give analogous methods for model selection and aggregation, when many models may be available and of interest. Our results all hold for generic models and distances, but throughout we highlight what happens for particular choices, such as L_2 projections on linear models, and KL projections on exponential families. Finally we illustrate by estimating the density of CD4 count among patients with HIV, had all been treated with combination therapy versus zidovudine alone, as well as a density effect. Our results suggest combination therapy may have increased CD4 count most for high-risk patients. Our methods are implemented in the freely available R package *npcausal* on GitHub.

Keywords: causal inference, density estimation, influence function, model misspecification, semiparametric theory.

1 Introduction

It is very common in causal inference to quantify causal effects with means. The classic average treatment effect (ATE) parameter, for instance, measures the difference in mean outcome had all versus none in a population been treated. This can certainly be a useful summary, but it can also miss potentially important differences in the distributions of the counterfactual outcomes, beyond a simple mean shift. To illustrate, consider the densities in Figure 1, which all have exactly the same mean and variance. These would be indistinguishable with the ATE, or any other measure that did not look past the first two moments.

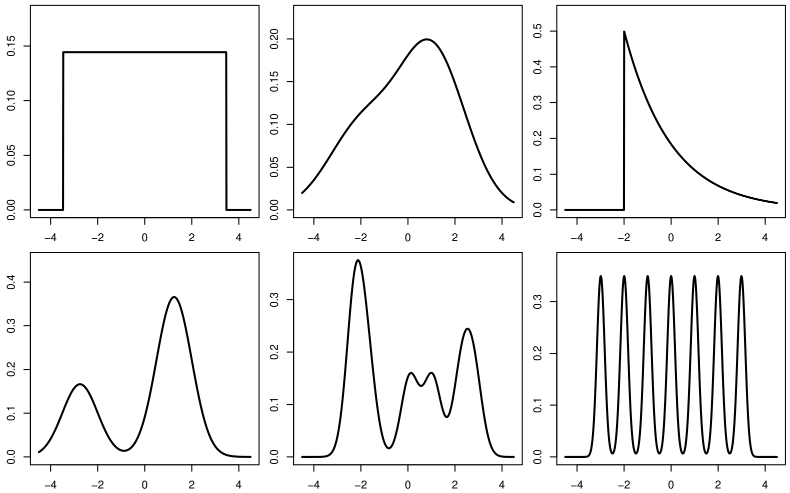

Figure 1: *Densities for six distributions, all with mean zero and variance four.*

In general, it can be very practically useful to know the shape of the counterfactual density. If the counterfactual density differed at all under treatment versus control (or under any other generic interventions), this would imply treatment had *some* effect, even if the ATE were zero. The presence of skew would indicate that some subjects have relatively extreme responses to treatment; next steps could include trying to understand who these subjects are, and why their responses are unusual. Similarly, multimodal structure could point to the existence of underlying subgroups with differential responses to treatment, which could be important for optimizing treatment policies. Contrasting the shape of the density under different interventions could inform hypotheses about how treatment works, e.g., perhaps it works by reducing variance, or driving up negative outcomes. This could help enhance future versions of treatment, or motivate the development of new treatments altogether.

There is a large literature on distributional treatment effects defined in terms of quantiles, or cumulative distribution functions (CDFs), with a similar goal of moving beyond simple mean summaries to study the entire counterfactual distribution [Abadie, 2002, Chernozhukov and Hansen, 2005, Chernozhukov et al., 2013, Díaz, 2017, Firpo, 2007, Fortin et al., 2011, Frölich and Melly, 2013, Machado and Mata, 2005, Melly, 2005, Rothe, 2010, Wang et al., 2018, Wang and Qin, 2010, Zhang et al., 2012]. However, the challenges and methods are substantially different for density estimation. This is largely a result of the fact that the CDF at y is the mean of the thresholded outcome $\mathbb{1}(Y \leq y)$, so that counterfactual CDF estimation mostly reduces to counterfactual mean estimation, after replacing the outcome with an indicator. A related difference is that the CDF is pathwise differentiable in a nonparametric model, whereas the density function is not [Bickel et al., 1993, van der Laan and Robins, 2003]. This is also true in the standard observational setup, where CDFs can be estimated at $n^{-1/2}$ rates with sample averages, while density estimation requires more careful balancing of bias and variance, with slower rates arising depending on underlying smoothness [Tsybakov, 2009, Wasserman, 2006]. Beyond this issue of statistical complexity, there are other trade-offs in targeting CDFs versus densities. One is that, although CDFs are easier to estimate nonparametrically, densities are arguably more visually appealing and interpretable to practitioners. We view CDFs and densities as complementary pieces of the distributional puzzle.

Unlike distribution function estimation, the literature on counterfactual density estimation appears much more sparse. In what was perhaps the first study of the problem, DiNardo et al. [1996] used a reweighted kernel estimator to estimate effects of US labor market factors on wages. However, the statistical properties of this proposed approach were not examined. Robins and Rotnitzky [2001] proposed a doubly robust version of the reweighted kernel estimator, and conjectured it would achieve usual density estimation rates under smoothness and other conditions. van der Laan and Dudoit [2003] and Rubin and van der Laan [2006] studied general cross-validation-based approaches for model selection in the presence of nuisance functions, and suggested minimizing counterfactual KL or L_2 loss for density estimation, but did not detail the statistical properties. More recently, Westling and Carone [2020] tackled the related problem of density estimation for right-censored outcomes, proposing new estimators that can attain $n^{-1/3}$ rates, but under an assumption that the density is monotone. Kim et al. [2018] analyzed a version of the doubly robust estimator from Robins and Rotnitzky [2001], showing its conjectured oracle properties, and used it to estimate the (nonsmooth) L_1 distance between counterfactual densities.

Somewhat surprisingly, none of the known work above on counterfactual density estimation considers a semiparametric approach, where the density is approximated with a finite-dimensional model. Our work aims to fill this gap in the literature, while also providing data-driven model selection and aggregation tools. A separate contribution is our study of generic density-based effects, which characterize the distance between counterfactual densities, using a generalized notion of distance that includes f-divergences as well as L_p norms.

The structure of our paper is as follows. After introducing some basics and causal assumptions in Section 2, in Section 3 we detail the different kinds of target parameters we consider. The first (described in Section 3.1) is an approximation of the density itself, defined by a projection onto a finite-dimensional model (3.1.1) using a generalized distance metric (3.1.2), which includes f-divergences as well as L_p norms. Importantly, we show in Section 3.1.3 that projection parameters for smooth models and distances can be framed as solutions to moment conditions, providing a link between counterfactual densities and means (of functions of counterfactuals). The second parameter we consider (described in Section 3.2) is the distance between counterfactual densities, which can be used as a new more nuanced effect measure, or as a tool for model selection (as in Section 3.3). In Section 4 we study nonparametric efficiency bounds, by characterizing the efficient influence functions of approximated density functions in Section 4.1, and density effects in Section 4.2. These follow from a master lemma in Section 4, which gives a von Mises expansion for generic integral functionals of the counterfactual density, and so may be of independent interest. In Section 5 we propose doubly robust-style estimators for the density approximations and distances, and study their rates of convergence, showing for example that they can be $n^{-1/2}$ consistent, asymptotically normal, and optimally efficient under weak high-level conditions on nuisance estimation error. All our results hold for smooth but otherwise arbitrary models and distances. However, in various corollaries, we also highlight specific expressions for typical choices of models and distances, such as L_2 projections on linear models, and KL projections on exponential families. Finally in Section 6 we use our proposed methods to estimate counterfactual densities and density effects of combination therapy (versus zidovudine alone) on CD4 count, among patients with HIV. Our results show treatment effects beyond a mean shift, suggesting that combination therapy may have increased CD4 count most for high-risk patients.

2 Setup

We assume access to an iid sample $(Z_1, ..., Z_n)$ of $Z = (X, A, Y) \sim \mathbb{P}$ where $X \in \mathbb{R}^d$ are covariates, $A \in \mathbb{R}$ is a treatment or exposure, and $Y \in \mathbb{R}$ is a continuous outcome. We let

$$\pi_a(x) = \mathbb{P}(A = a \mid X = x) \tag{1}$$

$$\int_{\mathcal{B}} \eta_a(y \mid x)\ dy = \mathbb{P}(Y \in \mathcal{B} \mid X = x, A = a) \ \text{ for measurable } \mathcal{B} \tag{2}$$

denote the propensity score (i.e., chance of being treated at level $A = a$ given covariates) and conditional outcome density, respectively. In this work we focus on discrete treatments, but in a companion paper we consider the continuous case.

We study "semiparametric" estimation of the covariate-adjusted marginal density

$$p_a(y) = \int \eta_a(y \mid x)\ d\mathbb{P}(x) \tag{3}$$

i.e., the conditional outcome density averaged over the covariates, as well as functionals thereof.

Remark 1. Although we refer to our work in this paper as semiparametric, in reality it is all done within a fully nonparametric model. As described in more detail starting in Section 3.1, the models we consider are only ever used as tools for defining nonparametric approximations, and corresponding projection parameters, and are never assumed to be correct descriptions of the underlying true data-generating process. Further, our results on estimating counterfactual density *functionals* (e.g., Sections 3.2 and 4.2) do not require any approximating models, and so are nonparametric in the usual sense.

We note that the density (3) is different from the marginal density of Y, since the treatment is fixed at $A = a$ in the conditioning; it is also not equal to the unadjusted conditional density $p(y \mid a)$. Instead, (3) is the density $p(y^a)$ of the counterfactual variable Y^a (i.e., the outcome that *would have been observed* if treatment were set to $A = a$), if the following assumptions hold:

Assumption 1 (Positivity). $\mathbb{P}\{\pi_a(X) \geq \epsilon\} = 1$ for some $\epsilon > 0$.

Assumption 2 (Consistency). $Y = Y^a$ if $A = a$.

Assumption 3 (Exchangeability). $A \perp\!\!\!\perp Y^a \mid X$.

Positivity ensures all subjects have some chance at receiving treatment level $A = a$. Consistency can be viewed as ruling out interference, for example, where a subject's counterfactual can depend not only on how they were treated, but how other subjects were treated as well. Exchangeability says the treatment is as good as randomized within levels of the observed covariates, and requires that sufficiently many relevant confounders are collected. Each of these assumptions can be weakened in various ways, at the expense of losing point identification of the marginal counterfactual distribution. Nonetheless, under only the positivity assumption, all our statistical results will hold relative to the observational quantity in (3), regardless of whether the causal Assumptions 2–3 are violated or not.

3 Target Parameters

In this section we detail the two kinds of quantities we consider estimating. The first is an approximation of the counterfactual density itself, defined via a projection in some distributional distance. The second is a distance measure, e.g., a density-based causal effect measuring the difference between counterfactual densities in terms of general f- or other divergences. The latter gives a more nuanced picture of how the counterfactual densities differ, compared to the usual ATE, for example. Finally in Section 3.3 we describe how these two kinds of target quantities can be adapted for the purposes of model selection and aggregation.

3.1 Density Functions

3.1.1 Models

First we consider approximations of the counterfactual density $p_a(y)$ based on some specified model $\{g(y;\beta) : \beta \in \mathbb{R}^d\}$. We mostly focus on the finite-dimensional parametric case with $\beta \in \mathbb{R}^d$, but more generally one could take β to be infinite-dimensional in some L^p space, or to belong to a subset of $\mathbb{R}^d$ such as the standard simplex. Note that $\beta(a)$ depends on a but for now we suppress this dependence in the notation and simply write β. Here are some examples.

Example 1a (Exponential family)**.** Let $b(y) = \{b_1(y), ..., b_d(y)\}^{\mathrm{T}}$ denote a vector of known basis functions. Then we can project onto the exponential family

$$g(y;\beta) = \exp\left\{\beta^{\mathrm{T}} b(y) - C(\beta)\right\} \tag{4}$$

where $C(\beta) = \log \int \exp\{\beta^{\mathrm{T}} b(y)\}\ dy$ so that $\int g(y;\beta)\ dy = 1$. Typical exponential family notation takes $\beta_1 = 1$ and sets $b_1(y) = \log h(y)$ for some known base measure h.

Although we refer to Example 1a as an exponential family, it can just as well be viewed as a truncated series expansion used together with a log link function. In the next example we consider a truncated series with an identity link function.

Example 1b (Truncated series)**.** Let $b(y) = \{b_1(y), ..., b_d(y)\}^{\mathrm{T}}$ denote a vector of known basis functions, and $q(y)$ a known base density (e.g., uniform). Then we can project onto the linear basis expansion

$$g(y;\beta) = q(y) + \sum_{j=1}^{d} \beta_j b_j(y)$$

where we can take $\int b_j(y)\ dy = 0$ so that the projection integrates to one. A natural choice when $Y \in [0,1]$ would be to take $q(y) = 1$ and $b(y)$ the cosine basis

$$b_j(y) = \sqrt{2}\cos(\pi j y) \tag{5}$$

which satisfies $\int b_j(y)\ dy = 0$ and $\int b_j(y) b_k(y)\ dy = \mathbb{1}(j = k)$ on the unit interval. One could alternatively take $q(y) = 0$ and let $b_j(y)$ be (the linear span of) a collection of d candidate densities, in which case the above could be viewed as a linear aggregation [Rigollet and Tsybakov, 2007]. Another related option would be to use a linear approximation for the square root of the density $\sqrt{g(y;\beta)} = \sum_j \beta_j b_j(y)$, so that $g(y;\beta) = \sum_j \sum_k \beta_j \beta_k b_j(y) b_k(y)$ [Chen et al., 2002, Pinheiro and Vidakovic, 1997]. Then the model would integrate to one if the basis functions were orthonormal ($\int b_j(y) b_k(y)\ dy = 0$ and $\int b_j(y)^2\ dy = 1$) and $\sum_j \beta_j^2 = 1$.

Example 1c (Gaussian mixture model)**.** Let $(\mu_1, ..., \mu_k)$ denote a vector of means, $(\sigma_1, ..., \sigma_k)$ a vector of positive standard deviations, $(\varpi_1, ..., \varpi_k)$ positive mixing proportions with $\sum_j \varpi_j = 1$, and ϕ the standard normal density. Then the standard Gaussian mixture model is

$$g(y;\beta) = \sum_{j=1}^{k} \varpi_j \left(\frac{1}{\sigma_j}\right) \phi\left(\frac{y-\mu_j}{\sigma_j}\right)$$

where $\beta = \{(\varpi_1, \mu_1, \sigma_1^2), ..., (\varpi_k, \mu_k, \sigma_k^2)\}$.

Now, based on the above approximations, a primary goal is to estimate the projection parameter

$$\beta_0 = \underset{\beta \in \mathbb{R}^p}{\arg\min} \ D_f\Big(p_a(y), g(y;\beta)\Big) \tag{6}$$

where D_f is a distributional distance measure of the form

$$D_f(p,q) = \int f(p,q)q(y)\ dy \tag{7}$$

for some given discrepancy function $f : \mathbb{R}^2 \to \mathbb{R}$.

Remark 2. In contrast to typical f-divergences [Ali and Silvey, 1966, Csiszár, 1967, Rényi et al., 1961, Sason and Verdú, 2016], we allow the function f to have two arguments, one for each distribution; this allows us to capture not only f-divergences but also other distances such as those based on L_p norms. (The usual f-divergence takes $f(p,q) = f(p/q)$ for some single-argument function, and so only depends on the density ratio). In a slight abuse of terminology we sometimes refer to (7) as a distance, even though for some of our choices of f it will be an asymmetric divergence not satisfying the triangle inequality.

Before giving examples of distances, we first discuss the interpretation of our projection parameter (6). Mathematically, β_0 is the parameter of the best-fitting model of the form $g(y;\beta)$, i.e., the parameter value that makes $g(y;\beta)$ closest (in corresponding distance) to the true density $p_a(y)$. If the model g is correctly specified, then $D_f(p_a(y), g(y;\beta_0)) = 0$ and so $g(y;\beta_0) = p_a(y)$ is simply the true counterfactual density; however, the projection (6) remains well-defined even under model misspecification. This is akin to the well-known concept of a *best linear predictor* in standard linear regression [White, 1980]. The projection approach, where a model is not assumed correct but instead only used for defining approximations, has been used widely throughout statistics [Beran, 1977, Buja et al., 2019a,b, Huber, 1967, Rakhlin et al., 2017, Rinaldo and Wasserman, 2010, Tsybakov, 2003, Wasserman, 2006, White, 1982, 1996] as well as in causal inference [Chernozhukov et al., 2018b, Cuellar and Kennedy, 2020, Kennedy et al., 2019, Neugebauer and van der Laan, 2007, Semenova and Chernozhukov, 2020, van der Laan, 2006], though not in the counterfactual density estimation context.

Remark 3. Since we only use models as tools to define approximations, all our results are formally nonparametric, as mentioned in Remark 1 and illustrated in subsequent theorems. This raises some interesting philosophical issues about the role of assumptions and corresponding bias-variance trade-offs. In particular, we can imagine a rough taxonomy of stances one might take in estimation problems like this one:

(i) model-ist: My finite-dimensional/parametric representation is *the* correct one.

(ii) model-agnostic: I may *use* a finite-dimensional model, but I do not know or require that it is a perfectly accurate picture of the truth.

(iii) anti-model-ist: No parametric model I can imagine contains the truth, and I do not care about approximations.

The model agnostic view is often captured by the famous quotes "All models are wrong but some are useful" (George Box) and "Use models but don't believe them" (possibly due to John Tukey). Of course, in practice, how much one relies on models is a continuum, and so any particular approach may not fall entirely in one of the three camps above. Similarly, our taxonomy uses parametric models as a benchmark, but one could just as well replace with a different assumption set (e.g., Hölder-smooth with index $s \geq 4$ versus $s < 4$). Nevertheless we find the above framing useful if imperfect. In this paper, we mostly take the stance of the model-agnostic, though we flirt with anti-model-ism in the data-driven model selection approaches of Sections 3.3 and 5.3 (and we are fully anti-model-ist in a companion paper). We also accept that each approach has advantages and disadvantages. The model-ist will do well when the model is correct, but could unknowingly suffer large bias otherwise. The anti-model-ist is most free from the constraints of human imagination (as they do not need to posit a parametric model), but with a more ambitious target can also suffer larger errors. The model-agnostic has a bit of the best of both worlds: when the model is correct, they may hope to do nearly as well as the model-ist, and when the model is wrong, their inference can still be valid for a still well-defined approximation. Of course, if the model is *very* wrong, the approximation may not be practically useful, no matter how well-defined it is; thus there can be important challenges in defining a useful approximating model and distance.

3.1.2 Distances

Now we give some examples of the distances we focus on in this paper:

Example 2a (L_2^2)**.** If $f(p,q) = (p-q)^2/q$ then $D_f(p,q) = \|p-q\|_2^2$ is the squared L_2 distance

$$\|p_a(y) - g(y;\beta)\|_2^2 = \int \Big(p_a(y) - g(y;\beta)\Big)^2 \, dy.$$

Example 2b (Kullback-Leibler)**.** If $f(p,q) = (p/q)\log(p/q)$ then $D_f(p,q) = \text{KL}(p,q)$ is the Kullback-Leibler divergence

$$\text{KL}\Big(p_a(y), g(y;\beta)\Big) = \int \log\left(\frac{p_a(y)}{g(y;\beta)}\right) p_a(y) \, dy.$$

Example 2c (χ^2)**.** If $f(p,q) = (p/q-1)^2$ then $D_f(p,q) = \chi^2(p,q)$ is the χ^2 divergence

$$\chi^2\Big(p_a(y), g(y;\beta)\Big) = \int \frac{\{p_a(y) - g(y;\beta)\}^2}{g(y;\beta)} \, dy.$$

Example 2d (Hellinger)**.** If $f(p,q) = (\sqrt{p/q}-1)^2$ then $D_f(p,q) = H^2(p,q)$ is the squared Hellinger divergence

$$H^2\Big(p_a(y), g(y;\beta)\Big) = \int \left(\sqrt{p_a(y)} - \sqrt{g(y;\beta)}\right)^2 \, dy.$$

Example 2e (Smoothed Total Variation)**.** If $f(p,q) = \frac{1}{2q}|p-q| = \frac{(p-q)}{2q}\,\mathrm{sgn}(p-q)$ then $D_f(p,q) = \mathrm{TV}(p,q) = \frac{1}{2}\|p-q\|_1$ is the total variation distance (and half the L_1 distance). Note $f(p,q)$ is not differentiable at $p/q = 1$. Smooth versions can be obtained by approximating the absolute value or sign functions in f. For example, let $\nu_t(y)$ be an approximation of the absolute value function $|y|$, with parameter t controlling the approximation error. For example one could use $\nu_t(y) = y\tanh(ty)$ or $\nu_t(y) = y\mathrm{erf}(ty)$ or a best polynomial approximation of degree t. Then taking $f(p,q) = f_t(p,q) = \frac{1}{2q}\nu_t(p-q)$ gives a smoothed total variation $D_f(p,q) = \mathrm{TV}^*(p,q)$ with

$$\mathrm{TV}^*\Big(p_a(y), g(y;\beta)\Big) = \frac{1}{2}\int \nu_t\Big\{p_a(y) - g(y;\beta)\Big\}\, dy.$$

There exist polynomial and rational approximations $\nu_t(y)$ of degree t ensuring that $|\mathrm{TV}(p,q) - \mathrm{TV}^*(p,q)|$ is of order t^{-1} and $\exp(-t)$, respectively [Newman et al., 1964]. We also note that the Hellinger divergence is closely related to total variation in the sense that $H^2(p,q)/2 \leq \mathrm{TV}(p,q) \leq H(p,q)$ for any densities p, q.

Figure 2 shows a few projections of a true density onto a truncated trigonometric series with six terms, using four different distances (L_2, Kullback-Leibler, χ^2, and Hellinger). The projections are all very similar in both cases. However, we note that, as discussed for example in Beran [1977], Hellinger projections should be more stable and robust to outliers or contamination, compared to for example KL. The projections are closer to the true density for the first simpler Gaussian mixture, and are more of a rough approximation for the second more complex mixture.

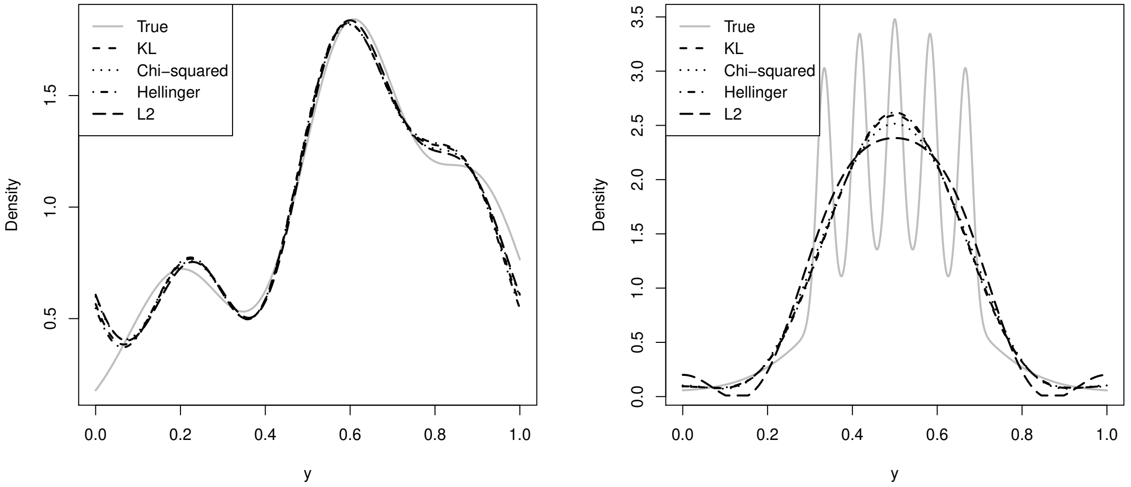

Figure 2: *Projections of a truncated Gaussian mixture (left) and the Bart Simpson density (right) onto a trigonometric basis with six terms, using L_2 distance, along with Kullback-Leibler, χ^2, and Hellinger divergences.*

3.1.3 Moment Conditions

The next proposition shows how, for smooth distances, the projection parameter β_0 can be defined more explicitly than in equation (6), as a solution to a population moment condition, involving derivatives of the model $g(y;\beta)$ and the function f. This links projection parameters to integral functionals of the counterfactual density (i.e., moments of transformations of counterfactuals), which is why our efficiency bounds and estimators in the next section resemble those for means of particular non-trivial functions of counterfactuals.

Proposition 1. *Assume g is differentiable in β, f is differentiable in its second argument with derivative $f_2'(q_1, q_2) = \frac{\partial}{\partial q_2} f(q_1, q_2)$, and that the minimizer in (6) is unique. Then the projection parameter*

$$\beta_0 = \underset{\beta \in \mathbb{R}^p}{\arg\min} \; D_f\Big(p_a(y), g(y;\beta)\Big)$$

can be expressed as a solution to the moment condition $m(\beta) = 0$, where

$$m(\beta) \equiv \int \frac{\partial g(y;\beta)}{\partial \beta} \Big\{ f\Big(p_a(y), g(y;\beta)\Big) + g(y;\beta) f_2'\Big(p_a(y), g(y;\beta)\Big) \Big\} \; dy. \tag{8}$$

The proof of Proposition 1 follows from the chain rule; all subsequent proofs are given in Appendix B. Throughout we assume there is a unique solution to $m(\beta) = 0$. Next we show how the moment condition defining β_0 simplifies for particular distances.

Corollary 1. *The quantity $f\Big(p_a(y), g(y;\beta)\Big) + g(y;\beta) f_2'\Big(p_a(y), g(y;\beta)\Big)$ in the integrand of the moment (8) equals*

$$\begin{cases} 2\Big\{g(y;\beta) - p_a(y)\Big\} & \text{if } D_f = L_2^2 \\ 1 - \dfrac{p_a(y)}{g(y;\beta)} & \text{if } D_f = KL \\ 1 - \left\{\dfrac{p_a(y)}{g(y;\beta)}\right\}^2 & \text{if } D_f = \chi^2 \\ 1 - \sqrt{\dfrac{p_a(y)}{g(y;\beta)}} & \text{if } D_f = H^2 \\ -\nu_t'\Big\{p_a(y) - g(y;\beta)\Big\}/2 & \text{if } D_f = TV^*. \end{cases}$$

Corollary 1 shows how the moment $m(\beta)$ essentially reduces to functionals of the counterfactual density for particular distances: simple means for L_2^2 and KL, a quadratic functional for χ^2, and a square root functional for H^2. For the smoothed TV distance, it depends on the form of the absolute value approximation (e.g., for ν_t a t degree polynomial approximation, the moment $m(\beta)$ would be an integral of a $t-1$ degree polynomial in the countef).

In the following corollaries we show how the form of the moment condition is particularly straightforward when based on L_2 or KL divergence with series models and exponential families, respectively.

Corollary 2. *If $D_f = L_2^2$ then*

$$m(\beta) = 2 \int \frac{\partial g(y;\beta)}{\partial \beta} \Big\{ g(y;\beta) - p_a(y) \Big\} \, dy.$$

Therefore if the support of Y is $[0,1]$, and $g(y;\beta) = 1 + \beta^{\mathrm{T}} b(y)$ is the truncated series in Example 1b then

$$\beta = \mathbb{E}\Big\{ b(Y^a) \Big\} \tag{9}$$

when $b(\cdot)$ is an orthogonal series with $\int b_j(y) \, dy = 0$ and $\int b_j(y) b_k(y) \, dy = \mathbb{1}(j = k)$.

Corollary 2 shows that when using orthogonal series with L_2^2 projections, there is a closed form for β, given by a simple mean of a known function of the counterfactual outcome. Estimation and inference for parameters like (9) is relatively well-understood [Robins et al., 2009, 2017], which allows exploiting existing theory and methods in the density estimation context.

Corollary 3. *If $D_f = KL$ then*

$$m(\beta) = -\mathbb{E}\left\{ \frac{\partial}{\partial \beta} \log g(Y^a;\beta) \right\},$$

and so if $g(y;\beta) = \exp\{\beta^{\mathrm{T}} b(y) - C(\beta)\}$ is the exponential family in Example 1a then

$$m(\beta) = \frac{\partial}{\partial \beta} C(\beta) - \mathbb{E}\Big\{ b(Y^a) \Big\} = \int b(y) \Big[\exp\{\beta^{\mathrm{T}} b(y) - C(\beta)\} - p_a(y) \Big] \, dy. \tag{10}$$

Similarly, for KL divergence, the moment $m(\beta)$ is simply the expected score under counterfactual density $g(y;\beta)$. Therefore, just as in the non-counterfactual setting, the parameter values that maximize a posited likelihood are also those that minimize KL divergence [Huber, 1967, White, 1982]. When one also uses an exponential family, the solution to $m(\beta) = 0$ corresponds to an intuitive "moment matching", i.e., finding the value of β that equates expectations of $b(\cdot)$ under g to those under the distribution of Y^a.

3.2 Distances & Density Effects

In addition to estimating projections of the counterfactual density onto a finite-dimensional model, in this section we also consider estimation of distributional distances themselves. The main focus is on density-based effects measuring the distance between counterfactual densities in terms of L_p^p and f-divergences. These effects can detect more nuanced disinctions between the distributions of Y^1 and Y^0, beyond simple differences-in-means captured by standard average treatment effects.

More specifically, we consider the distance between p_1 and p_0 given by

$$\psi_f = D_f\Big(p_1(y), p_0(y)\Big) = \int f\Big(p_1(y), p_0(y)\Big) p_0(y) \, dy \tag{11}$$

for discrepancy functions f as discussed in the previous subsection. In this setup we do not require approximating the densities $p_a(y)$ with finite-dimensional models, and instead consider estimating ψ_f in a fully nonparametric model.

3.3 Model Selection & Aggregation

In practice one may not have an approximating model such as (4) available *a priori*. In these cases it would be natural to instead set up a sequence of models, and use the data to choose among them. In standard regression and density estimation problems, simple cross-validation procedures are available for this task; however, because our goal is estimation of a more nuanced counterfactual density, these require some refinement, in the same spirit as van der Laan and Dudoit [2003]. Thus in this section we describe how the target quantities of Sections 3.1 and 3.2 can be adapted for the purposes of model selection and aggregation.

Specifically, for a set of estimators $\{\widehat{g}_k(y) : k = 1, ..., K\}$ of $p_a(y)$ (e.g., estimated from some initial training sample, with each projected onto the space of valid densities), we can define the risk for a given estimator as

$$R(\widehat{g}_k) = D_f\Big(p_a(y), \widehat{g}_k(y)\Big) \tag{12}$$

The minimum risk oracle estimator $\widehat{g}_{k_0}(y)$ can then be defined via

$$k_0 = \arg\min_k R(\widehat{g}_k) = \arg\min_k D_f\Big(p_a(y), \widehat{g}_k(y)\Big). \tag{13}$$

A model aggregation oracle can be defined more generally as $\widetilde{g}(y) = \sum_k \beta_{0k}\widehat{g}_k(y)$ where

$$\beta_0 = \arg\min_{\beta \in B} D_f\left(p_a(y), \sum_{k=1}^{K} \beta_k \widehat{g}_k(y)\right). \tag{14}$$

for some appropriate selection set, e.g., the standard simplex $B = \{(\beta_1, ..., \beta_K) \in \mathbb{R}^K : \beta_k \geq 0, \sum_k \beta_k = 1\}$ for convex aggregation [Rigollet and Tsybakov, 2007, Tsybakov, 2003]. If one takes $B = \mathbb{R}^K$ for linear aggregation, then f-divergences may not be well-defined, so this might naturally only be used in the $D_f = L_2^2$ setting.

Note that the proposed target parameters in Section 3.1 correspond to the aggregation target in (14) if we replace $\mathbb{R}^d$ with the relevant space B. However, since model selection as defined in Equation (13) does not satisfy the smoothness assumptions we relied on in Section 3.1.3, it can be useful in practice to estimate the risk separately for all K candidates; this is more akin to the effect estimation problem in Section 3.2, except where the density $p_0(y)$ in (11) is replaced with a candidate estimator $\widehat{g}_k(y)$ (e.g., which may be estimated on a separate independent sample/fold and conditioned upon, and so treated as fixed).

4 Efficiency Theory

In this section we present a crucial von Mises expansion (i.e., distributional Taylor expansion) for generic density functionals, which yields efficient influence functions for the projection parameters and density effects of interest, and thus nonparametric efficiency bounds [Bickel et al., 1993, van der Laan and Robins, 2003]. The latter can be further formalized as local minimax lower bounds [van der Vaart, 2002].

Throughout we make reference to the linear map $T \mapsto \phi_a(T;\mathbb{P})$ defined as

$$\phi_a(T;\mathbb{P}) = \frac{\mathbb{1}(A=a)}{\pi_a(X)}\Big\{T - \mathbb{E}(T \mid X, A=a)\Big\} + \mathbb{E}(T \mid X, A=a) - \mathbb{E}\{\mathbb{E}(T \mid X, A=a)\} \quad (15)$$

which takes a random variable T (and distribution $\mathbb{P}$) and outputs the efficient influence function for the functional $\mathbb{E}\{\mathbb{E}(T \mid X, A=a)\}$. Note we drop the dependence of $\phi_a(T;\mathbb{P})$ on (X, A) for simplicity; at times we also drop the dependence on $\mathbb{P}$ if the context is clear. In all our examples, $T = h(Y)$ will be a known or $\mathbb{P}$-dependent function of Y; the functionals we consider all have influence functions consisting of terms of the above form, but with different and non-standard choices of $T = h(Y)$, depending on the model and distance being used.

Recall that in Corollary 1 we showed the relevant moment $m(\beta)$ reduces to a functional of the counterfactual density for particular distances. Therefore our first result gives a von Mises-style expansion for generic smooth integral functionals of the counterfactual density. This result paves the way for later expansions and efficiency bounds, and may be of independent interest in other problems involving different counterfactual density functionals.

Lemma 1. *Let $\psi = \psi(\mathbb{P}) = \int h(p_a(y))\, dy$ for some twice continuously differentiable function h. Then ψ satisfies the von Mises expansion*

$$\psi(\overline{\mathbb{P}}) - \psi(\mathbb{P}) = \int \phi_a\left(h'\Big(p_a(Y)\Big); \overline{\mathbb{P}}\right)\ d(\overline{\mathbb{P}} - \mathbb{P}) + R_2(\overline{\mathbb{P}}, \mathbb{P}) \quad (16)$$

where

$$\begin{aligned} R_2(\overline{\mathbb{P}}, \mathbb{P}) = \int\int h'(\overline{p}_a(y)) \left\{\frac{\pi_a(x)}{\overline{\pi}_a(x)} - 1\right\} \Big\{\eta_a(y \mid x) - \overline{\eta}_a(y \mid x)\Big\}\ dy\ d\mathbb{P}(x) \\ + \frac{1}{2}\int h''(p_a^*(y))\Big\{\overline{p}_a(y) - p_a(y)\Big\}^2\ dy, \end{aligned}$$

where $p_a^*(y)$ lies between $p_a(y)$ and $\overline{p}_a(y)$.

Lemma 1 has several important consequences. First, it indicates how one can correct the first-order bias of a plug-in estimator $\psi(\widehat{\mathbb{P}})$ of counterfactual density functionals: by estimating the first term in the expansion and subtracting it off. This is how standard semiparametric estimators (particularly of the one-step variety) based on influence functions are constructed [Bickel et al., 1993, Chernozhukov et al., 2018a, van der Laan and Robins, 2003], and our proposed estimators in the next section do precisely this. Second, since the remainder term is quadratic in the nuisance functions, it implies that $\psi(\mathbb{P})$ is pathwise differentiable with efficient influence function $\phi_a(h'(p_a(Y)))$; for this fact we refer to Lemma 2 in the Appendix.

4.1 Density Functions

In this subsection we use Lemma 1 to detail the efficient influence function for the moment $m(\beta)$ at a fixed β, as well as the projection parameter β_0 and projected density $g(y;\beta_0)$. These efficient influence functions yield local minimax lower bounds, as well as estimators that can attain the nonparametric efficiency bounds under generic high-level rate conditions on nuisance estimators, which will be proved in Section 5.

Theorem 1. *Assume f is twice differentiable and denote partial derivatives as $f_j'(q_1, q_2) = \frac{\partial}{\partial q_j} f(q_1, q_2)$ and similarly $f_{jk}''(q_1, q_2) = \frac{\partial^2}{\partial q_j \partial q_k} f(q_1, q_2)$. Then, under an unrestricted nonparametric model, the efficient influence function for $m(\beta)$ is given by*

$$\phi_a\Big(\gamma_f(Y; \beta)\Big)$$

where

$$\gamma_f(y; \beta) \equiv \gamma_f(y; \beta, p_a) = \frac{\partial g(y; \beta)}{\partial \beta} \Big\{ f_1'\Big(p_a(y), g(y; \beta)\Big) + g(y; \beta) f_{21}''\Big(p_a(y), g(y; \beta)\Big) \Big\} .$$

The efficient influence functions for β_0 and $g(y; \beta_0)$ are similarly given by

$$-\frac{\partial m(\beta)}{\partial \beta}^{-1} \phi_a\Big(\gamma_f(Y; \beta)\Big) \Big|_{\beta=\beta_0} \quad \text{and} \quad - \frac{\partial g(y; \beta)}{\partial \beta^{\mathrm{T}}} \frac{\partial m(\beta)}{\partial \beta}^{-1} \phi_a\Big(\gamma_f(Y; \beta)\Big) \Big|_{\beta=\beta_0} \tag{17}$$

respectively.

The efficient influence functions given in Theorem 1 are analogous to those of usual ATE-type parameters, but with the crucial difference that they correspond to means of $\gamma_f(Y^a; \beta)$, not Y^a itself. This is what we should expect based on the result in Lemma 1, since the γ_f transformation is the derivative of the integrand in the moment condition (8) given in Proposition 1. Note also that the form of γ_f indicates that the efficiency bound for β_0 (i.e., the variance of the efficient influence function) will be adversely affected when the model g is sensitive to small changes in β, or when the distance is sensitive to small changes in its arguments, since then the derivatives in γ_f will be large.

In the next corollary, we give the particular form of the efficient influence functions when D_f is the L_2^2 and KL divergence, and the approximating models are a linear series and exponential family.

Corollary 4. *For L_2^2 and KL divergence the quantity γ_f from Theorem 1 reduces to*

$$\gamma_f(y; \beta) = \begin{cases} -2 \frac{\partial g(y;\beta)}{\partial \beta} & \text{if } D_f = L_2^2 \\ -\frac{\partial \log g(y;\beta)}{\partial \beta} & \text{if } D_f = KL. \end{cases}$$

Further, if either

1. *$D_f = L_2^2$ and $g(y; \beta) = q(y) + \beta^{\mathrm{T}} b(y)$ is the truncated series in Example 1b, or*

2. *$D_f = KL$ and $g(y; \beta) = \exp\{\beta^{\mathrm{T}} b(y) - C(\beta)\}$ is the exponential family in Example 1a*

then the efficient influence function for $m(\beta)$ is proportional to

$$\phi_a\Big(b(Y)\Big).$$

The proportionality constant is -2 for $D_f = L_2^2$, and -1 for $D_f = KL$.

Corollary 4 shows that the efficient influence functions are proportional for linear projections using L_2^2 distance, and for projections onto an exponential family using the KL divergence. Further, this efficient influence function simply corresponds to that of the counterfactual mean vector $\mathbb{E}\{b(Y^a)\}$, for b a known basis function vector. Thus the influence function conveniently reduces to that of the mean of a transformed version of the counterfactual outcome, with no dependence on β. As mentioned after Corollary 2, this allows for adapting existing theory and methods for average treatment effects to the density estimation context.

The following theorem summarizes the local minimax lower bound implied by the form of the efficient influence function in Theorem 1, as in Corollary 2.6 of van der Vaart [2002].

Corollary 5. *Let $\sigma^2 = \sigma^2_{\mathbb{P}}$ denote the variance of the efficient influence function from (17). The local minimax risk for β_0 is lower bounded as*

$$\inf_{\delta>0} \liminf_{n\to\infty} \sup_{TV(\overline{\mathbb{P}},\mathbb{P})<\delta} \mathbb{E}_{\overline{\mathbb{P}}}\left[\ell\left\{\sqrt{n}\Big(\widehat{\beta} - \beta_0(\overline{\mathbb{P}})\Big)\right\}\right] \geq \mathbb{E}\Big\{\ell(\sigma Z)\Big\}$$

for any estimator $\widehat{\beta}$, where $\ell : \mathbb{R}^p \mapsto [0,\infty)$ is any subconvex loss function.

Corollary 5 follows from Corollary 2.6 of van der Vaart [2002]. It shows that the worst-case mean squared error of any estimator, locally near the true $\mathbb{P}$, cannot be smaller than the efficiency bound, asymptotically and after scaling by $\sqrt{n}$. This gives an important benchmark for efficient estimation of projection parameters of the counterfactual density: no estimator can have mean squared error uniformly better than the variance of the efficient influence function (divided by n), without adding extra assumptions to the nonparametric model we consider.

4.2 Density Effects

Now we give the efficient influence function for the density effect parameters in (11). Unlike the projected densities in the previous subsection, the density effect parameters depend on both counterfactual densities of interest for comparison.

Theorem 2. *In an unrestricted nonparametric model, the efficient influence function for the density effect $\psi_f = \int f\left(p_1(y), p_0(y)\right) p_0(y)\, dy$ is given by*

$$\phi_1\Big(\lambda_1(Y)\Big) + \phi_0\Big(\lambda_0(Y)\Big)$$

where

$$\begin{aligned}
\lambda_1(y) &= p_0(y) f_1'\Big(p_1(y), p_0(y)\Big) \\
\lambda_0(y) &= f\Big(p_1(y), p_0(y)\Big) + p_0(y) f_2'\Big(p_1(y), p_0(y)\Big).
\end{aligned}$$

As with the result for β_0 in Theorem 1, the efficient influence function for ψ_f in Lemma 2 consists of inverse probability weighted residuals, plus a "plug-in"-type term, similar to ATE parameters. However, again this corresponds to the influence function for a transformed version of the outcome, depending on the counterfactual densities and choice of distance f. The efficient influence function simplifies somewhat for L_2^2 and KL divergence, as indicated in the following corollary. Expressions for other f-divergences are in Section B.1 in the Appendix.

Corollary 6. *If $D_f = L_2^2$, then the efficient influence function for ψ_f is*

$$2(\phi_1 - \phi_0)\Big(p_1(Y) - p_0(Y)\Big).$$

If $D_f = KL$, then the efficient influence function for ψ_f is

$$\phi_1\left(\log\left(\frac{p_1(Y)}{p_0(Y)}\right)\right) - \phi_0\left(\frac{p_1(Y)}{p_0(Y)}\right).$$

The fact that $\lambda_1 = -\lambda_0$ for L_2^2 projections simplifies the form of our proposed estimators, as we will detail further in the next section. We also note that the influence function reduces to zero when $p_1 = p_0$, which presents some complications for inference; this will be discussed in the next section as well.

As mentioned in Section 3.3, for the purposes of model selection and aggregation it is also useful to consider the distance between p_a and a fixed candidate g; we give the corresponding efficient influence function here.

Proposition 2. *In an unrestricted nonparametric model, the efficient influence function for $\Delta_f(g) = \int f\left(p_a(y), g(y)\right) g(y)\, dy$ for g fixed and known is given by*

$$\phi_a\left(g(Y) f_1'\Big(p_a(Y), g(Y)\Big)\right).$$

If $D_f = L_2^2$ then this influence function reduces to

$$2\phi_a\Big(p_a(Y) - g(Y)\Big).$$

5 Estimation and Inference

In this section we present doubly robust-style estimators of the proposed density functions and density effects, based on the functional expansions from Lemma 1 and the efficient influence function results in Theorems 1–2. We study their rates of convergence, and show they can be $n^{-1/2}$ consistent and asymptotically efficient under weak nonparametric conditions.

5.1 Density Functions

Here let $\widehat{\pi}_a(x)$ and $\widehat{\eta}_a(y \mid x)$ denote initial estimators of the propensity score and conditional density functions $\pi_a(x) = \mathbb{P}(A = a \mid X = x)$ and $\eta_a(y \mid x) = \frac{\partial}{\partial y}\mathbb{P}(Y \leq y \mid X = x, A = a)$, for example based on generic regression estimators and their numerical derivatives (or for the latter one can use a regression of a kernel transformed version of the outcome). Also let $\widehat{p}_a(y) = \mathbb{P}_n\{\widehat{\eta}_a(y \mid X)\}$ denote the plug-in estimator of the counterfactual density under $A = a$, where $\mathbb{P}_n\{h(Z)\} = n^{-1}\sum_i h(Z_i)$, and let

$$\widehat{m}(\beta) \equiv \int \frac{\partial g(y;\beta)}{\partial \beta}\left\{f\Big(\widehat{p}_a(y), g(y;\beta)\Big) + g(y;\beta) f_2'\Big(\widehat{p}_a(y), g(y;\beta)\Big)\right\} dy. \qquad (18)$$

denote the plug-in estimator of the moment condition $m(\beta)$, and similarly for ψ_f.

Remark 4. Although we suggest basing (18) on the plug-in estimator $\widehat{p}_a(y) = \mathbb{P}_n\{\widehat{\eta}_a(y \mid X)\}$ of the counterfactual density, one could just as well use other estimators (e.g., inverse-probability-weighted, or doubly robust, as in Kim et al. [2018]). Nonetheless, all results in this paper will only depend on high-level second-order rate conditions for estimating $p_a(y)$, which would be satisfied for the simple plug-in estimator as long as similar conditions hold for the underlying density estimator $\widehat{\eta}_a(y \mid x)$. We prove this in Appendix B.5, showing that the mean squared error of $\widehat{p}_a(y)$ is upper bounded by an integrated version of that of $\widehat{\eta}_a(y \mid x)$.

To ease notation we let $\widehat{\phi}_a(T) = \phi_a(T; \widehat{\mathbb{P}})$ denote the estimated version of the efficient influence function given in (15). Then our proposed projection estimators are given by approximate solutions in β (up to $o_{\mathbb{P}}(1/\sqrt{n})$ error) to

$$\widehat{m}(\beta) + \mathbb{P}_n \left\{ \widehat{\phi}_a\Big(\widehat{\gamma}_f(Y;\beta)\Big) \right\} = o_{\mathbb{P}}(1/\sqrt{n}) \tag{19}$$

In other words the estimators are one-step bias-corrected estimators (of the moment condition and the parameter itself, respectively), which take the plug-in estimator and add an estimate of the bias by averaging an estimate of the influence function.

Remark 5. For simplicity, in the following results we assume the various nuisance estimates in $\widehat{\mathbb{P}}$ are constructed from a single separate independent sample, of the same size n as the estimation sample on which $\mathbb{P}_n$ operates. Alternatively, if the same observations are used both for estimating nuisance functions and averaging estimates of the influence function, one generally needs to rely on empirical process conditions to avoid overfitting. In practice, with iid data, one can always obtain separate independent samples by randomly splitting the data in half (or in folds); further, to regain full sample size efficiency one can always swap the samples, repeat the procedure, and average the results, popularly called cross-fitting and used for example by Bickel and Ritov [1988], Chernozhukov et al. [2018a], Robins et al. [2008], Zheng and van der Laan [2010]. In this paper, to simplify notation we always analyze a single split procedure, with the understanding that extending to an analysis of an average across independent splits is straightforward.

Our first propositions give the form of the plug-in and bias-corrected projection estimators when using a linear series with L_2^2 distance, and an exponential family model with KL divergence, which take a particularly simple form.

Proposition 3. *If $D_f = L_2^2$, the support of Y is $[0,1]$, and $g(y;\beta) = 1 + \beta^{\mathrm{T}} b(y)$ is the truncated series in Example 1b, with $b(\cdot)$ an orthogonal series with $\int b_j(y)\, dy = 0$ and $\int b_j(y) b_k(y)\, dy = \mathbb{1}(j = k)$, then the plug-in estimator of β is*

$$\widehat{\beta} = \mathbb{P}_n\{\widehat{\mu}_a(X; b)\},$$

where $\widehat{\mu}_a(x; b)$ is an estimate of $\mu_a(x; b) = \mathbb{E}\{b(Y) \mid X = x, A = a)$. In contrast, the proposed one-step estimator in (19) is given by

$$\widehat{\beta} = \mathbb{P}_n \left[\frac{\mathbb{1}(A = a)}{\widehat{\pi}_a(X)} \Big\{ b(Y) - \widehat{\mu}_a(X; b) \Big\} + \widehat{\mu}_a(X; b) \right]. \tag{20}$$

Proposition 4. *If $D_f = KL$ and $g(y;\beta) = \exp\{\beta^{\mathrm{T}} b(y) - C(\beta)\}$ is the exponential family in Example 1a, then the plug-in estimator solving $\widehat{m}(\widehat{\beta}) = 0$ is the solution in β to*

$$\int \Big[b(y) - \mathbb{P}_n\{\widehat{\mu}_a(X;b)\}\Big] \exp\Big\{\beta^{\mathrm{T}} b(y)\Big\}\, dy = 0$$

where $\widehat{\mu}_a(x;b)$ is an estimate of $\mu_a(x;b) = \mathbb{E}\{b(Y) \mid X = x, A = a)$. In contrast, the proposed one-step estimator in (19) is given by the solution in β to

$$\int \left(b(y) - \mathbb{P}_n\left[\frac{\mathbb{1}(A=a)}{\widehat{\pi}_a(X)}\Big\{b(Y) - \widehat{\mu}_a(X;b)\Big\} + \widehat{\mu}_a(X;b)\right]\right) \exp\Big\{\beta^{\mathrm{T}} b(y)\Big\}\, dy = 0. \tag{21}$$

Propositions 3-4 shows that the plug-in and bias-corrected estimators for L_2^2 and KL projections solve simple estimating equations, which only require one to first estimate the components $\mathbb{E}\{\mu_a(X;b)\}$; importantly, straightforward doubly robust estimators as in (21) are available, and do not depend on the estimating equation parameter β. This is not necessarily true for other model/distance combinations; in general $\widehat{\gamma}_f$ would have to be estimated at each β in order to solve (19), which could be quite computationally intensive.

Next we give the main result of this section, which shows the rate of convergence for the proposed estimator. Importantly the rate involves products of nuisance estimation errors, allowing for $n^{-1/2}$ consistency and asymptotic normality in nonparametric models, and even when the nuisance estimators are generic and flexibly fit.

Theorem 3. *Let $\eta = (\pi_a, \eta_a)$, and $\varphi(Z;\beta,\eta) = m(\beta;\eta) + \phi_a(\gamma_f(Y;\beta),\eta)$. Assume:*

1. *The functions γ_f and $1/\overline{\pi}_a$ are bounded above by some constant, and γ_f is differentiable in $p_a(y)$, with derivative bounded uniformly above by δ.*
2. *The function class $\{\varphi(z;\beta,\eta) : \beta \in \mathbb{R}^p\}$ is Donsker in β for any fixed η.*
3. *The estimators are consistent in the sense that $\widehat{\beta} - \beta_0 = o_{\mathbb{P}}(1)$ and $\|\widehat{\eta} - \eta_0\| = o_{\mathbb{P}}(1)$.*
4. *The map $\beta \mapsto \mathbb{P}\{\varphi(Z;\beta,\eta)\}$ is differentiable at β_0 uniformly in η, with nonsingular derivative matrix $\frac{\partial}{\partial\beta}\mathbb{P}\{\varphi(Z;\beta,\eta)\}|_{\beta=\beta_0} = V(\beta_0,\eta)$, where $V(\beta_0,\widehat{\eta}) \xrightarrow{p} V(\beta_0,\eta_0)$.*

Then

$$\begin{aligned}\widehat{\beta} - \beta_0 = &-V(\beta_0,\eta_0)^{-1}(\mathbb{P}_n - \mathbb{P})\Big\{\phi_a\Big(\gamma_f(Y;\beta_0)\Big)\Big\}\\ &+ O_{\mathbb{P}}\left(\|\widehat{\pi}_a - \pi_a\|\|\widehat{\eta}_a - \eta_a\| + \delta\|\widehat{p}_a - p_a\|^2 + o_{\mathbb{P}}\left(\frac{1}{\sqrt{n}}\right)\right).\end{aligned}$$

Remark 6. In a slight abuse of notation, Theorem 3 holds when we define $\|\widehat{\eta}_a - \eta_a\|^2 = \|\zeta_a\|^2 \equiv \int \zeta_a(x)^2\, d\mathbb{P}(x)$ for integrated error $\zeta_a(x) = \int |\widehat{\eta}_a(y \mid x) - \eta_a(y \mid x)|\, dy$. This implies it also holds if we define $\|\widehat{\eta}_a - \eta_a\|^2 = \int\{\widehat{\eta}_a(y \mid x) - \eta_a(y \mid x)\}^2\, dy\, d\mathbb{P}(x)$, or $\|\widehat{\eta}_a - \eta_a\|^2 = \int\{\widehat{\eta}_a(y \mid x) - \eta_a(y \mid x)\}^2\, d\mathbb{P}(y,x)$ if $\eta_a(y \mid x)$ is bounded from below.

Importantly, Theorem 3 shows that $\widehat{\beta}$ attains substantially faster rates than its nuisance estimators $\widehat{\eta}$, and can be asymptotically efficient under weak nonparametric conditions, for example attaining the minimax lower bound in Corollary 5. First we give some description of the assumed conditions. The first condition ensures the influence function is not too complex as a function of β (though allowing arbitrary complexity in η). The second condition merely requires consistency of $(\widehat{\beta}, \widehat{\eta})$ at any rate. The third condition requires some smoothness in β, so as to allow a delta method argument. These conditions ensure $\widehat{\beta}$ has a rate of convergence that is second-order in the nuisance estimation error, thus attaining faster rates than the nuisance estimators. Thus, for example, under standard $n^{-1/4}$-type rate conditions on $\widehat{\eta}$, the estimator $\widehat{\beta}$ is $n^{-1/2}$-consistent, asymptotically normal, and efficient. Importantly, these rates can be attained under smoothness, sparsity, or other structural conditions (e.g., additive modeling or bounded variation assumptions, etc.). For instance, if it is assumed that all d-dimensional nuisance functions lie in a Holder class with smoothness index s (i.e., partial derivatives up to order s exist and are Lipschitz) then the assumption of Theorem 3 would be satisfied when $s > d/2$, i.e., the smoothness index is at least half the dimension. Alternatively, if the functions are s-sparse then one would need $s = o(\sqrt{n})$ up to log factors, as in Farrell [2015]. In these cases, asymptotically valid 95% confidence intervals can be constructed via the simple Wald form, $\widehat{\beta} \pm 1.96\sqrt{\text{diag}[\widehat{\text{cov}}\{\widehat{\phi}_a(\widehat{\gamma}_f(Y;\widehat{\beta}))\}/n]}$.

Remark 7. In some prominent cases (for example, L_2^2 and KL projections, as shown in Corollary 4), the function γ_f does not depend on the counterfactual density $p_a(y)$ at all, so its derivative is exactly zero and $\delta = 0$. In this case the second term in the second-order remainder in Theorem 3 drops out, making the proposed approach doubly robust in the usual sense, requiring no rate conditions on the initial pilot estimate of the counterfactual density.

5.2 Density Effects

Here we present doubly robust-style estimators of the density effects described in Section 3.2, and study their rate of convergence. As before we first construct initial estimators $\widehat{\pi}_a(x)$, $\widehat{\eta}_a(y \mid x)$, and $\widehat{p}_a(y) = \mathbb{P}_n\{\widehat{\eta}_a(y \mid X)\}$ of the propensity score and conditional and counterfactual densities. Estimated versions of $\phi_a(T)$ and λ_a defined in Theorem 2 follow accordingly.

Then the density effect estimators we propose are defined as

$$\widehat{\psi}_f = \int f\Big(\widehat{p}_1(y), \widehat{p}_0(y)\Big)\widehat{p}_0(y)\ dy + \mathbb{P}_n\left\{\widehat{\phi}_1\Big(\widehat{\lambda}_1(Y)\Big) + \widehat{\phi}_0\Big(\widehat{\lambda}_0(Y)\Big)\right\}, \tag{22}$$

which can again be viewed as one-step bias-corrected estimators, with plug-in bias estimated via an average of the estimated influence function. In practice, rather than estimating the conditional density η_a and integrating over its y argument, one could instead regress for example $\widehat{\lambda}_a$ on X for the integral terms in the estimated influence function.

Proposition 5. *If $D_f = L_2^2$ then the proposed density effect estimator can be written as*

$$\begin{aligned}\widehat{\psi}_f = 2\ \mathbb{P}_n\Bigg(\frac{2A-1}{\widehat{\pi}_A(X)}\bigg[\Big\{\widehat{p}_1(Y) - \widehat{p}_0(Y)\Big\} - \int \Big\{\widehat{p}_1(y) - \widehat{p}_0(y)\Big\}\widehat{\eta}_A(y \mid X)\ dy\bigg] \\ + \int \Big\{\widehat{p}_1(y) - \widehat{p}_0(y)\Big\}\Big\{\widehat{\eta}_1(y \mid X) - \widehat{\eta}_0(y \mid X)\Big\}\ dy\Bigg) - \int \Big\{\widehat{p}_1(y) - \widehat{p}_0(y)\Big\}^2\ dy.\end{aligned}$$

The estimator in Proposition 5 can be viewed as taking twice the doubly robust estimator of the mean of $(\widehat{p}_1(Y^1) - \widehat{p}_0(Y^1)) - (\widehat{p}_1(Y^0) - \widehat{p}_0(Y^0))$, which is $\int(\widehat{p}_1 - \widehat{p}_0)(p_1 - p_0)$, and subtracting a plug-in estimate of the L_2^2 distance. This is analogous to the standard one-step estimator of the expected (observational) density $\int p(x)^2\ dx$ [Bickel and Ritov, 1988], which takes twice an estimate of the mean of $\widehat{p}(X)$, i.e., $\int \widehat{p}p$, and subtracts the plug-in estimate $\int \widehat{p}^2$. For the expected density, the bias is just the integrated squared difference between $\widehat{p}$ and p; in contrast, in our setting, we show next that there is an additional doubly robust error term, due to the confounding adjustment required for estimating counterfactual densities.

Theorem 4. *Assume λ_a and $1/\widehat{\pi}_a$ are bounded above by some constant for $a = 0, 1$, and λ_a is differentiable in $p_a(y)$, with derivative bounded uniformly above by δ_a. Then*

$$\widehat{\psi}_f - \psi_f = (\mathbb{P}_n - \mathbb{P})\left\{\phi_1\Big(\lambda_1(Y)\Big) + \phi_0\Big(\lambda_0(Y)\Big)\right\} \\ + O_{\mathbb{P}}\left(\sum_{a=0}^{1}\Big(\|\widehat{\pi}_a - \pi_a\|\|\widehat{\eta}_a - \eta_a\| + \delta_a\|\widehat{p}_a - p_a\|^2\Big) + o_{\mathbb{P}}\left(\frac{1}{\sqrt{n}}\right)\right).$$

Theorem 4 (whose proof mimics that of Theorem 3) shows that $\widehat{\psi}_f$ can attain faster rates than its nuisance estimators, and can be asymptotically efficient under weak nonparametric conditions. The conditions and the form of the convergence rate are similar to those of Theorem 3, so we refer to our discussion there for more details. However we do comment on a few differences. First, for the density functions targeted in Theorem 3, the moment condition $m(\beta)$, and resulting influence functions and estimators, can have a complicated dependence on β; in contrast, this is not an issue for the density effect ψ_f since the influence function is linear in the parameter. Thus extra smoothness conditions on the influence function used in Theorem 3 are not required in Theorem 4. Second, although in Theorem 3 the derivative bound δ can be exactly zero in some prominent cases, in general in Theorem 4 this will not be the case (e.g., for L_2^2 distance the derivative of λ_a has absolute value equal to one). Therefore, for efficient estimation of density effects, we in general need an initial density estimator converging at $n^{-1/4}$ rate. However recall that, as described in Remark 4, there exist nonparametric counterfactual density estimators with error upper bounded by $\|\widehat{\pi}_a - \pi_a\|$ or $\|\widehat{\eta}_a - \eta_a\|$, so that $\|\widehat{p}_a - p_a\|^2$ would be of smaller or similar order compared to the product error preceding it.

There is a third distinction in density effect estimation. Under usual $n^{-1/4}$ rate conditions on the nuisance estimators, Theorem 4 suggests 95% confidence intervals of the form

$$\widehat{\psi}_f \pm 1.96\sqrt{\widehat{\text{cov}}\left\{\widehat{\phi}_1\Big(\widehat{\lambda}_1(Y)\Big) + \widehat{\phi}_0\Big(\widehat{\lambda}_0(Y)\Big)\right\}/n} \tag{23}$$

These intervals are asymptotically valid as usual when $p_1 \neq p_0$, but not when $p_1 = p_0$, since then the influence function of ψ_f reduces to zero, as mentioned in Section 4.2. This invalidates inference because the first sample average in Theorem 4 is no longer dominant, as with degenerate U-statistics or other estimators whose higher-order terms dominate their von Mises expansions (cf. Sections 12.3 and 20.1.1 of van der Vaart [2000]). However, the presence of nuisance functions complicates things substantially, as noted in other similarly complex functional estimation problems [Luedtke et al., 2019, Williamson et al., 2020], but we are not aware of a general solution. Thus we only recommend using the interval (23) in non-null settings when $p_1 \neq p_0$. A simple albeit ad-hoc fix is to use the interval $\widehat{\psi} \pm z_{\alpha/2}(s \vee 1/\sqrt{n})$ where $s = \sqrt{\widehat{\text{cov}}\{\widehat{\phi}_1(\widehat{\lambda}_1(Y)) + \widehat{\phi}_0(\widehat{\lambda}_0(Y))\}/n}$. This is valid but conservative near the null.

5.3 Model Selection & Aggregation

Here we briefly describe how the methods of the previous subsections can be used for the purposes of model selection and aggregation, in the same spirit as Tsybakov [2003], van der Laan and Dudoit [2003], and others. We leave technical details to future work.

First we consider the linear aggregation goal as defined in (14), where $B = \mathbb{R}^K$. In this setup the methods from Section 5.1 can be straightforwardly adapted, by adding an extra step of sample splitting. We focus on L_2^2 projections since f-divergences may not be well-defined for general linear combinations of candidate estimators. Our proposed approach is as follows:

Step 1. Randomly split the sample into a training set D_n^0 and test set D_n^1.

Step 2. On the training set D_n^0, estimate K different models (e.g., K different numbers of basis functions, or K different combinations of linear, exponential family, Gaussian mixture models, etc.), using the estimator in (19) to compute $\widehat{g}_k(y) = g(y; \widehat{\beta}_k)$, $k = 1, ..., K$.

Step 3. On the test set D_n^1, estimate the (L_2^2) projection onto an orthonormal basis of the linear span of $(\widehat{g}_1, ..., \widehat{g}_K)$, again using the estimator in (19), e.g., with the series model in Example 1b with $q(y) = 0$, to compute an aggregated estimator $\widehat{g}(y) = \sum_k \widehat{\theta}_k \widehat{g}_k(y)$.

Step 4. Reverse the roles of D_n^0 and D_n^1 and average the two resulting aggregates.

Note that inside Steps 2-3, another layer of sample splitting is required to avoid empirical process conditions in estimating the nuisance functions, as discussed in Remark 5. We also note that the cross-fitting in Step 4 could be considered optional if the corresponding efficiency loss was considered negligible, or alternatively one could instead implement Steps 1–4 with M different folds, at each step using $M - 1$ for training and the other fold for the test set. We conjecture that the above approach can attain the optimal K/n rates for linear density aggregation in the observational case [Rigollet and Tsybakov, 2007], under standard $n^{-1/4}$-type conditions on the nuisance estimators (or weaker, depending on how K scales with n).

For model selection and convex aggregation, we propose a similiar procedure, except where in Step 3 variants of the density effect estimators from Section 5.2 are used to estimate the distance between p_a and each of the k candidates estimated from the training split (after projecting each onto the space of valid densities). One can then pick the minimum distance candidate or an appropriately weighted combination, e.g., by finding the convex weights that minimize the estimated distance in the test split. For example, our proposed estimator of the L_2^2 error of a candidate g_k based on Proposition 2 is given by

$$\widehat{\Delta}_f(g_k) = \int \Big(\widehat{p}_a(y) - g_k(y)\Big)^2 \, dy + 2\mathbb{P}_n \Big\{\widehat{\phi}_a\Big(\widehat{p}_a(Y) - g_k(Y)\Big)\Big\}.$$

For the purposes of model selection, one can instead use the simpler pseudo-L_2^2 risk

$$\widehat{\Delta}_f^*(g_k) = -2\ \mathbb{P}_n\bigg[\frac{\mathbb{1}(A = a)}{\widehat{\pi}_a(X)}\Big\{g_k(Y) - \int g_k(y)\widehat{\eta}_a(y \mid X)\ dy\Big\} + \int g_k(y)\widehat{\eta}_a(y \mid X)\ dy\bigg] + \int g_k(y)^2\ dy, \tag{24}$$

based on the fact that the L_2^2 distance $\int (p_a - g_k)^2$ equals $\int g_k^2 - 2\int g_k p_a$ plus a term $\int p_a^2$ that does not depend on g_k. This is the estimator we use in the data analysis in the next section.

6 Illustration

Here we apply our proposed methods to analyze the effect of combined antiretroviral therapy for treating HIV. All code is given in Appendix A, and the methods are implemented in the *npcausal* R package on GitHub (https://github.com/ehkennedy/npcausal).

The data we use come from the ACTG 175 randomized trial [Hammer et al., 1996], and are available in the `speff2trial` R package. The treatment is whether patients received combination therapy ($A = 1$) versus zidovudine alone ($A = 0$), and the outcome Y is CD4 count at 96 weeks post-baseline. Baseline covariates X include age, weight, Karnofsky score, indicators for race, gender, hemophilia, homosexual activity, drug use, whether symptomatic, and previous zidovudine and antiretroviral use. There are a total of $n = 2319$ patients in the trial, 797 of which do not have outcome data (we use $R = 1$ to denote an observed outcome).

Since we are interested in the density of outcomes had all versus none been treated *in the entire population* (i.e., had all outcomes been measured), we can view the product indicator $\mathbb{1}(A = a, R = 1)$ as a joint "treatment" variable [van der Laan and Robins, 2003]. In other words our goal is to estimate counterfactual densities under $A = 1$ *and* $R = 1$, versus $A = 0$ and $R = 1$. Our methods therefore rely on no unmeasured confounding of A (which holds by design due to the experimental design) and missingness at random of Y (i.e., $R \perp\!\!\!\perp Y \mid X, A$), which is untestable regardless of whether treatment is randomized. For more details on the trial and data, we refer to Hammer et al. [1996] and Wang et al. [2018].

Throughout our analysis, we used 5-fold cross-fitting, with all nuisance functions estimated by random forests (via the R package `ranger` [Wright and Ziegler, 2015]). This includes conditional densities η_a, which we estimated by regressing a Gaussian kernel weighted outcome on covariates and treatment, on a grid of y values, with bandwidth chosen by Silverman's rule. Alternative approaches could also be used [Díaz and van der Laan, 2011, Hansen, 2004, Izbicki and Lee, 2017], potentially at the expense of some extra computational burden.

First we used the density effect methods from Section 5.2 to check for evidence of an effect of combination therapy on the density of CD4 count. Specifically, we used the cross-fit version of the estimator in Proposition 5 to estimate the L_2^2 distance between p_1 and p_0, with asymptotic variance estimated as usual, via the empirical variance of the estimated influence function. To ease interpretability we rescaled Y to be on the unit interval. The estimated L_2^2 distance was 0.279 with a 95% confidence interval of $[0.142, 0.415]$, indicating a statistically significant effect of combination therapy on CD4 count.

To more precisely understand how combination therapy impacted the CD4 distribution, we estimated the counterfactual densities using the methods of Sections 5.1 and 5.3. Specifically, we used L_2^2 projections onto the linear series in Example 1b with the cosine basis (5). We considered a range of models for both densities, including up to 15 basis terms (more than 15 terms did not improve fit). Figure 3 shows estimates of model fit via the pseudo-L_2^2 risk (24), along with confidence intervals, indicating that four basis terms does best for both counterfactual densities. Figure 4 shows the estimated counterfactual CD4 densities using four basis terms, along with pointwise CIs. Since the densities differ more substantially in the lowest CD4 range (e.g., 0-200), this suggests combination therapy may have increased CD4 count most for the high-risk patients with the lowest counts under control (zidovudine).

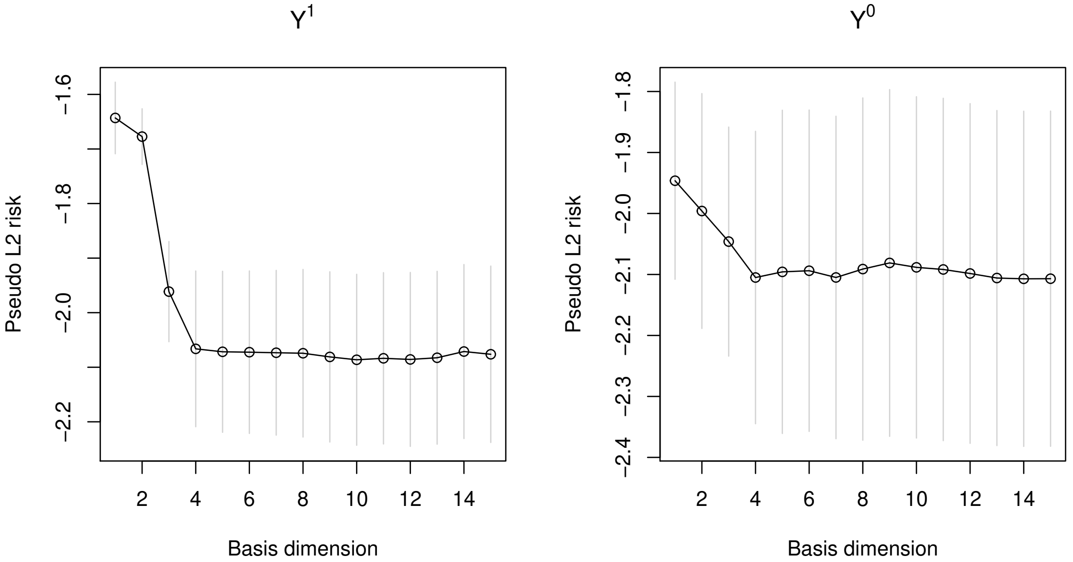

Figure 3: *Estimates of pseudo-L_2^2 risk for models of increasing dimension (using L_2 projections onto linear models with a cosine basis), with gray bars denoting confidence intervals.*

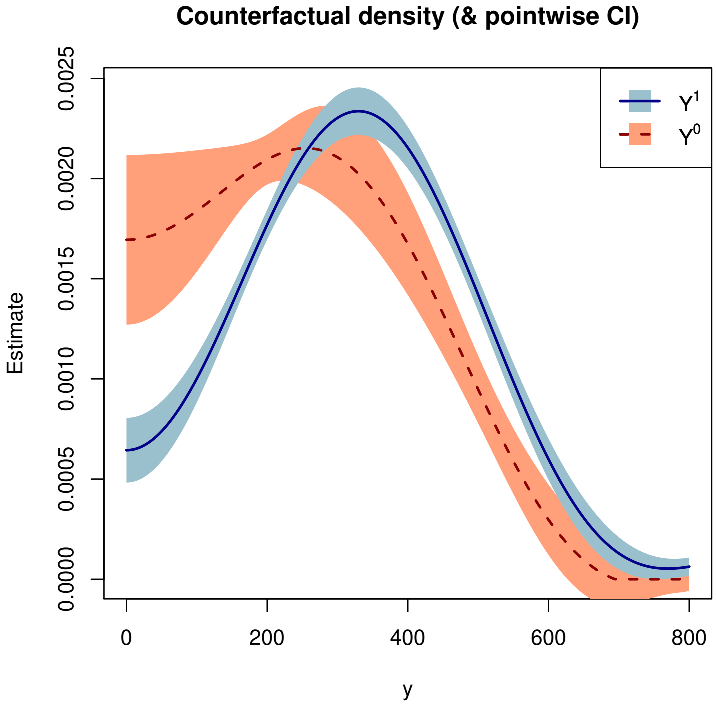

Figure 4: *Estimated counterfactual CD4 densities for combination therapy versus zidovudine.*

7 Discussion

In this paper we proposed methods for estimating counterfactual densities and corresponding distances and other functionals. We gave nonparametric efficiency bounds and flexible optimal estimators for a wide class of models and projection distances, and for new effects that quantify treatment impacts on the density scale. We also gave methods for data-driven model selection and aggregation in this context, and illustrated the ideas in an application studying effects of antiretroviral therapy on CD4 count.

There are many interesting avenues for future work. In upcoming companion papers, we consider the nonparametric version of the problem (where the target is the density p_a itself and not a projection) as well as non-discrete treatments (where A is for example a continuous dose). Much more work is needed on the computational side since, outside of L_2^2 projections on linear models and KL projections on exponential families, our methods require solving somewhat complicated estimating equations. Other extensions could involve time-varying treatments, instrumental variables, conditional effects, density-optimal treatment regimes, mediation, sensitivity analysis, and more. It is also of interest to apply the methods more broadly, to see if they bring any new insights about treatment mechanisms or ways to adapt treatment policies.

Acknowledgements

Edward Kennedy gratefully acknowledges support from NSF Grant DMS1810979, and Sivaraman Balakrishnan and Larry Wasserman from NSF Grant DMS1713003.

A Appendix: R Code

```
set.seed(100)

# install npcausal package
install.packages("devtools"); library(devtools)
install_github("ehkennedy/npcausal"); library(npcausal)

# load data
library(speff2trial); data(ACTG175); dat <- ACTG175[,c(2:17,19,21,23)]
x <- dat[,!(colnames(dat) %in% c("treat","cd496"))]

# create treatment*missing indicator
a1 <- dat$treat*(!is.na(dat$cd496)); a0 <- (1-dat$treat)*(!is.na(dat$cd496))
a <- a1; a[a0==0 & a1==0] <- -1; y <- dat$cd496; y[is.na(dat$cd496)] <- 0

# estimate pseudo-l2 risk for k=1:15
cv.cdensity(y,a,x, kmax=15, gridlen=50,nsplits=5)

# estimate densities at k=4
res <- cdensity(y,a,x, kmax=4, kforplot=c(4,4), gridlen=50,nsplits=5,ylim=c(0,800))
```

References

A. Abadie. Bootstrap tests for distributional treatment effects in instrumental variable models. *Journal of the American statistical Association*, 97(457):284–292, 2002.

S. M. Ali and S. D. Silvey. A general class of coefficients of divergence of one distribution from another. *Journal of the Royal Statistical Society: Series B (Methodological)*, 28(1):131–142, 1966.

R. Beran. Minimum hellinger distance estimates for parametric models. *The Annals of Statistics*, 5(3):445–463, 1977.

P. J. Bickel and Y. Ritov. Estimating integrated squared density derivatives: sharp best order of convergence estimates. *Sankhyā*, pages 381–393, 1988.

P. J. Bickel, C. A. Klaassen, Y. Ritov, and J. A. Wellner. *Efficient and Adaptive Estimation for Semiparametric Models*. Baltimore: Johns Hopkins University Press, 1993.

A. Buja, L. Brown, R. Berk, E. George, E. Pitkin, M. Traskin, K. Zhang, and L. Zhao. Models as approximations i: Consequences illustrated with linear regression. *Statistical Science*, 34 (4):523–544, 2019a.

A. Buja, L. Brown, A. K. Kuchibhotla, R. Berk, E. George, and L. Zhao. Models as approximations ii: A model-free theory of parametric regression. *Statistical Science*, 34(4):545–565, 2019b.

J. Chen, D. Zhang, and M. Davidian. A monte carlo em algorithm for generalized linear mixed models with flexible random effects distribution. *Biostatistics*, 3(3):347–360, 2002.

V. Chernozhukov and C. Hansen. An IV model of quantile treatment effects. *Econometrica*, 73(1):245–261, 2005.

V. Chernozhukov, I. Fernández-Val, and B. Melly. Inference on counterfactual distributions. *Econometrica*, 81(6):2205–2268, 2013.

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

V. Chernozhukov, M. Demirer, E. Duflo, and I. Fernandez-Val. Generic machine learning inference on heterogenous treatment effects in randomized experiments. Technical report, National Bureau of Economic Research, 2018b.

I. Csiszár. Information-type measures of difference of probability distributions and indirect observation. *studia scientiarum Mathematicarum Hungarica*, 2:229–318, 1967.

M. Cuellar and E. H. Kennedy. A non-parametric projection-based estimator for the probability of causation, with application to water sanitation in kenya. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 183(4):1793–1818, 2020.

I. Díaz. Efficient estimation of quantiles in missing data models. *Journal of Statistical Planning and Inference*, 190:39–51, 2017.

I. Díaz and M. J. van der Laan. Super learner based conditional density estimation with application to marginal structural models. 2011.

J. DiNardo, N. M. Fortin, and T. Lemieux. Labor market institutions and the distribution of wages, 1973-1992: A semiparametric approach. *Econometrica*, pages 1001–1044, 1996.

M. H. Farrell. Robust inference on average treatment effects with possibly more covariates than observations. *Journal of Econometrics*, 189(1):1–23, 2015.

S. Firpo. Efficient semiparametric estimation of quantile treatment effects. *Econometrica*, 75 (1):259–276, 2007.

N. Fortin, T. Lemieux, and S. Firpo. Decomposition methods in economics. *Handbook of Labor Economics*, 4:1–102, 2011.

M. Frölich and B. Melly. Unconditional quantile treatment effects under endogeneity. *Journal of Business & Economic Statistics*, 31(3):346–357, 2013.

S. M. Hammer, D. A. Katzenstein, M. D. Hughes, H. Gundacker, R. T. Schooley, R. H. Haubrich, W. K. Henry, M. M. Lederman, J. P. Phair, and M. Niu. A trial comparing nucleoside monotherapy with combination therapy in hiv-infected adults with cd4 cell counts from 200 to 500 per cubic millimeter. *New England Journal of Medicine*, 335(15):1081–1090, 1996.

B. E. Hansen. Nonparametric conditional density estimation. *Unpublished manuscript*, 2004.

P. J. Huber. The behavior of maximum likelihood estimates under nonstandard conditions. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 221–233. University of California Press, 1967.

R. Izbicki and A. B. Lee. Converting high-dimensional regression to high-dimensional conditional density estimation. *Electronic Journal of Statistics*, 11(2):2800–2831, 2017.

E. H. Kennedy, S. Lorch, and D. S. Small. Robust causal inference with continuous instruments using the local instrumental variable curve. *Journal of the Royal Statistical Society: Series B*, 81(1):121–143, 2019.

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

K. Kim, J. Kim, and E. H. Kennedy. Causal effects based on distributional distances. *arXiv 1806.02935*, 2018.

A. Luedtke, M. Carone, and M. J. van der Laan. An omnibus non-parametric test of equality in distribution for unknown functions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 81(1):75–99, 2019.

J. A. Machado and J. Mata. Counterfactual decomposition of changes in wage distributions using quantile regression. *Journal of Applied Econometrics*, 20(4):445–465, 2005.

B. Melly. Decomposition of differences in distribution using quantile regression. *Labour Economics*, 12(4):577–590, 2005.

R. Neugebauer and M. J. van der Laan. Nonparametric causal effects based on marginal structural models. *Journal of Statistical Planning and Inference*, 137(2):419–434, 2007.

D. J. Newman et al. Rational approximation to $|x|$. *Michigan Mathematical Journal*, 11(1): 11–14, 1964.

A. Pinheiro and B. Vidakovic. Estimating the square root of a density via compactly supported wavelets. *Computational Statistics & Data Analysis*, 25(4):399–415, 1997.

A. Rakhlin, K. Sridharan, A. B. Tsybakov, et al. Empirical entropy, minimax regret and minimax risk. *Bernoulli*, 23(2):789–824, 2017.

A. Rényi et al. On measures of entropy and information. *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, 1961.

P. Rigollet and A. B. Tsybakov. Linear and convex aggregation of density estimators. *Mathematical Methods of Statistics*, 16(3):260–280, 2007.

A. Rinaldo and L. Wasserman. Generalized density clustering. *The Annals of Statistics*, 38 (5):2678–2722, 2010.

J. M. Robins and A. Rotnitzky. Comments on: Inference for semiparametric models: Some questions and an answer. *Statistica Sinica*, 11:920–936, 2001.

J. M. Robins, L. Li, E. J. Tchetgen Tchetgen, and A. W. van der Vaart. Higher order influence functions and minimax estimation of nonlinear functionals. *Probability and Statistics: Essays in Honor of David A. Freedman*, pages 335–421, 2008.

J. M. Robins, E. J. Tchetgen Tchetgen, L. Li, and A. W. van der Vaart. Semiparametric minimax rates. *Electronic Journal of Statistics*, 3:1305–1321, 2009.

J. M. Robins, L. Li, R. Mukherjee, E. Tchetgen Tchetgen, and A. W. van der Vaart. Minimax estimation of a functional on a structured high dimensional model. *The Annals of Statistics*, 45(5):1951–1987, 2017.

C. Rothe. Nonparametric estimation of distributional policy effects. *Journal of Econometrics*, 155(1):56–70, 2010.

D. B. Rubin and M. J. van der Laan. Extending marginal structural models through local, penalized, and additive learning. *UC Berkeley Division of Biostatistics Working Paper Series*, 212:1–20, 2006.

I. Sason and S. Verdú. f-divergence inequalities. *IEEE Transactions on Information Theory*, 62(11):5973–6006, 2016.

V. Semenova and V. Chernozhukov. Debiased machine learning of conditional average treatment effects and other causal functions. *The Econometrics Journal*, 2020.

A. B. Tsybakov. Optimal rates of aggregation. *Learning theory and kernel machines*, pages 303–313, 2003.

A. B. Tsybakov. *Introduction to Nonparametric Estimation.* New York: Springer, 2009.

M. J. van der Laan. Statistical inference for variable importance. *The International Journal of Biostatistics*, 2(1), 2006.

M. J. van der Laan and S. Dudoit. Unified cross-validation methodology for selection among estimators and a general cross-validated adaptive epsilon-net estimator: Finite sample oracle inequalities and examples. *UC Berkeley Division of Biostatistics Working Paper Series*, Paper 130, 2003.

M. J. van der Laan and J. M. Robins. *Unified Methods for Censored Longitudinal Data and Causality.* New York: Springer, 2003.

A. W. van der Vaart. *Asymptotic Statistics.* Cambridge: Cambridge University Press, 2000.

A. W. van der Vaart. Semiparametric statistics. *In: Lectures on Probability Theory and Statistics*, pages 331–457, 2002.

L. Wang, Y. Zhou, R. Song, and B. Sherwood. Quantile-optimal treatment regimes. *Journal of the American Statistical Association*, 113(523):1243–1254, 2018.

Q. Wang and Y. Qin. Empirical likelihood confidence bands for distribution functions with missing responses. *Journal of Statistical Planning and Inference*, 140(9):2778–2789, 2010.

L. Wasserman. *All of Nonparametric Statistics.* Springer, 2006.

T. Westling and M. Carone. A unified study of nonparametric inference for monotone functions. *Annals of Statistics*, 48(2):1001, 2020.

H. White. Using least squares to approximate unknown regression functions. *International Economic Review*, pages 149–170, 1980.

H. White. Maximum likelihood estimation of misspecified models. *Econometrica*, pages 1–25, 1982.

H. White. *Estimation, inference and specification analysis.* Number 22. Cambridge University Press, 1996.

B. D. Williamson, P. B. Gilbert, N. R. Simon, and M. Carone. A unified approach for inference on algorithm-agnostic variable importance. *arXiv preprint arXiv:2004.03683*, 2020.

M. N. Wright and A. Ziegler. ranger: A fast implementation of random forests for high dimensional data in c++ and r. *arXiv preprint arXiv:1508.04409*, 2015.

Z. Zhang, Z. Chen, J. F. Troendle, and J. Zhang. Causal inference on quantiles with an obstetric application. *Biometrics*, 68(3):697–706, 2012.

W. Zheng and M. J. van der Laan. Asymptotic theory for cross-validated targeted maximum likelihood estimation. *UC Berkeley Division of Biostatistics Working Paper Series*, Paper 273:1–58, 2010.

B Appendix: Proofs

B.1 Proof of Corollaries 1–4 and 6

These corollaries all follow from the distance-specific form of f. For reference we list the relevant quantities here.

For L_2^2 distance we have

$$f(p,q) = \frac{(p-q)^2}{q} \qquad f_1'(p,q) = 2\left(\frac{p}{q} - 1\right)$$
$$f_2'(p,q) = 1 - \left(\frac{p}{q}\right)^2 \qquad f_{21}''(p,q) = -\frac{2p}{q^2}.$$

For KL divergence we have

$$f(p,q) = \left(\frac{p}{q}\right)\log\left(\frac{p}{q}\right) \qquad f_1'(p,q) = \frac{1}{q}\left\{\log\left(\frac{p}{q}\right) + 1\right\}$$
$$f_2'(p,q) = -\frac{p}{q^2}\left\{\log\left(\frac{p}{q}\right) + 1\right\} \qquad f_{21}''(p,q) = -\frac{1}{q^2}\left\{\log\left(\frac{p}{q}\right) + 2\right\}.$$

For χ^2 divergence we have

$$f(p,q) = \left(\frac{p}{q} - 1\right)^2 \qquad f_1'(p,q) = \frac{2(p-q)}{q^2}$$
$$f_2'(p,q) = -\frac{2p}{q^3}(p-q) \qquad f_{21}''(p,q) = \frac{2(q-2p)}{q^3}.$$

For Hellinger divergence we have

$$f(p,q) = \left(\sqrt{\frac{p}{q}} - 1\right)^2 \qquad f_1'(p,q) = \frac{1}{\sqrt{q}}\left(\frac{1}{\sqrt{q}} - \frac{1}{\sqrt{p}}\right)$$
$$f_2'(p,q) = \frac{\sqrt{p}}{q^2}\left(\sqrt{q} - \sqrt{p}\right) \qquad f_{21}''(p,q) = \frac{\sqrt{q/p} - 2}{2q^2}.$$

For TV* divergence we have

$$f(p,q) = \frac{1}{2q}\nu_t(p-q) \qquad f_1'(p,q) = \frac{1}{2q}\nu_t'(p-q)$$
$$f_2'(p,q) = \frac{-1}{2q}\left\{\frac{\nu_t(p-q)}{q} + \nu_t'(p-q)\right\} \qquad f_{21}''(p,q) = \frac{-1}{2q}\left\{\frac{\nu_t'(p-q)}{q} + \nu_t''(p-q)\right\}.$$

B.2 Proof of Lemma 1

Here we let $\psi = \psi(\mathbb{P}) = \int h(p_a(y))\ dy$, for some twice continuously differentiable function h. We will show that ψ satisfies the von Mises expansion given in Lemma 1.

Let $\overline{p}_a(y) = \int \overline{\eta}_a(y \mid x)\ d\overline{\mathbb{P}}(x)$ denote the marginal counterfactual density under $\overline{\mathbb{P}}$. Note for the posited influence function given by

$$\begin{aligned}\varphi(z;\mathbb{P}) = \frac{\mathbb{1}(A=a)}{\pi_a(X)}&\left\{h'(p_a(Y)) - \int h'(p_a(y))\eta_a(y \mid X)\ dy\right\}\\ &+ \int h'(p_a(y))\eta_a(y \mid X)\ dy - \int h'(p_a(y))\eta_a(y \mid x)\ dy\ d\mathbb{P}(x),\end{aligned}$$

we have, by iterated expectation, that it has mean under $\mathbb{P}$ equal to

$$\begin{aligned}\int \varphi(z;\overline{\mathbb{P}})\ d\mathbb{P} &= \int \Bigg[\frac{\mathbb{1}(A=a)}{\overline{\pi}_a(X)}\left\{h'(\overline{p}_a(Y)) - \int h'(\overline{p}_a(y))\overline{\eta}_a(y \mid X)\ dy\right\}\\ &\qquad + \int h'(\overline{p}_a(y))\overline{\eta}_a(y \mid X)\ dy - \int\int h'(\overline{p}_a(y))\overline{\eta}_a(y \mid x)\ dy\ d\overline{\mathbb{P}}(x)\Bigg]\ d\mathbb{P}\\ &= \int \frac{\pi_a(x)}{\overline{\pi}_a(x)} \int \Big\{h'(\overline{p}_a(y))\eta_a(y \mid x) - h'(\overline{p}_a(y))\overline{\eta}_a(y \mid x)\Big\}\ dy\ d\mathbb{P}(x)\\ &\qquad + \int\int h'(\overline{p}_a(y))\overline{\eta}_a(y \mid x)\ dy\ \Big\{d\mathbb{P}(x) - d\overline{\mathbb{P}}(x)\Big\}\\ &= \int\int h'(\overline{p}_a(y))\left\{\frac{\pi_a(x)}{\overline{\pi}_a(x)} - 1\right\}\Big\{\eta_a(y \mid x) - \overline{\eta}_a(y \mid x)\Big\}\ dy\ d\mathbb{P}(x)\\ &\qquad + \int h'(\overline{p}_a(y)) \int \Big\{\eta_a(y \mid x)\ d\mathbb{P}(x) - \overline{\eta}_a(y \mid x)\ d\overline{\mathbb{P}}(x)\Big\}\ dy\\ &= \int\int h'(\overline{p}_a(y))\left\{\frac{\pi_a(x)}{\overline{\pi}_a(x)} - 1\right\}\Big\{\eta_a(y \mid x) - \overline{\eta}_a(y \mid x)\Big\}\ dy\ d\mathbb{P}(x)\\ &\qquad + \int h'(\overline{p}_a(y))\Big\{p_a(y) - \overline{p}_a(y)\Big\}\ dy.\end{aligned}$$

Therefore the second-order remainder term in the von Mises expansion is

$$\begin{aligned}R_2(\overline{\mathbb{P}},\mathbb{P}) &\equiv \psi(\overline{\mathbb{P}}) - \psi(\mathbb{P}) - \int \varphi(z;\overline{\mathbb{P}})\ d(\overline{\mathbb{P}} - \mathbb{P}) = \psi(\overline{\mathbb{P}}) - \psi(\mathbb{P}) + \int \varphi(z;\overline{\mathbb{P}})\ d\mathbb{P}\\ &= \int\int h'(\overline{p}_a(y))\left\{\frac{\pi_a(x)}{\overline{\pi}_a(x)} - 1\right\}\Big\{\eta_a(y \mid x) - \overline{\eta}_a(y \mid x)\Big\}\ dy\ d\mathbb{P}(x)\\ &\qquad + \int h'(\overline{p}_a(y))\Big\{p_a(y) - \overline{p}_a(y)\Big\}\ dy + \int \Big\{h(\overline{p}_a(y)) - h(p_a(y))\Big\}\ dy\\ &= \int\int h'(\overline{p}_a(y))\left\{\frac{\pi_a(x)}{\overline{\pi}_a(x)} - 1\right\}\Big\{\eta_a(y \mid x) - \overline{\eta}_a(y \mid x)\Big\}\ dy\ d\mathbb{P}(x)\\ &\qquad + \frac{1}{2}\int h''(p_a^*(y))\Big\{\overline{p}_a(y) - p_a(y)\Big\}^2\ dy,\end{aligned}$$

where the last line follows by a Taylor expansion with remainder of the mean-value form, with $p_a^*(y)$ lying between $p_a(y)$ and $\overline{p}_a(y)$.

B.3 Proof of Theorem 1

First, for any fixed β, we have that each element of the p-vector

$$m(\beta) = \int \frac{\partial g(y;\beta)}{\partial \beta} \left\{ f\Big(p_a(y), g(y;\beta)\Big) + g(y;\beta) f_2'\Big(p_a(y), g(y;\beta)\Big) \right\} dy,$$

can be viewed as a density functional $\int h(p_a(y))\ dy$ for a specific function h. In particular, let $g_j'(y;\beta)$ denote the j^{th} element of $\frac{\partial g(y;\beta)}{\partial \beta}$ so that $h(p_a(y)) = \{h_1(p_a(y)), ..., h_d(p_a(y))\}^{\mathrm{T}}$ for

$$h_j(p_a(y)) = g_j'(y;\beta) \left\{ f\Big(p_a(y), g(y;\beta)\Big) + g(y;\beta) f_2'\Big(p_a(y), g(y;\beta)\Big) \right\}, \tag{25}$$

noting that, for a given β value, $g(y;\beta)$ is a known constant not depending on $\mathbb{P}$.
Now we apply Lemma 1 to each component of m. First note that

$$h_j'(p_a(y)) = g_j'(y;\beta) \left\{ f_1'\Big(p_a(y), g(y;\beta)\Big) + g(y;\beta) f_{21}''\Big(p_a(y), g(y;\beta)\Big) \right\},$$

by the chain rule, so that

$$\begin{aligned}\gamma_f(y;\beta) &= \frac{\partial g(y;\beta)}{\partial \beta} \left\{ f_1'\Big(p_a(y), g(y;\beta)\Big) + g(y;\beta) f_{21}''\Big(p_a(y), g(y;\beta)\Big) \right\} \\ &= \Big\{ h_1'(p_a(y)), ..., h_p'(p_a(y)) \Big\}^{\mathrm{T}}.\end{aligned}$$

Therefore Lemma 1 implies that $\overline{m}(\beta) = \int h(\overline{p}_a(y))\ dy$ satisfies the von Mises expansion

$$\overline{m}(\beta) - m(\beta) = \int \varphi_m(z; \overline{\mathbb{P}})\ d(\overline{\mathbb{P}} - \mathbb{P}) + R_2(\overline{\mathbb{P}}, \mathbb{P}), \tag{26}$$

where

$$\begin{aligned}\varphi_m(Z,\beta;\mathbb{P}) = \frac{\mathbb{1}(A=a)}{\pi_a(X)} &\left\{ \gamma_f(Y;\beta) - \int \gamma_f(y;\beta)\eta_a(y \mid X)\ dy \right\} \\ &+ \int \gamma_f(y;\beta)\eta_a(y \mid X)\ dy - \int\int \gamma_f(y;\beta)\eta_a(y \mid x)\ dy\ d\mathbb{P}(x),\end{aligned}$$

and where the j^{th} component of $R_2(\overline{\mathbb{P}}, \mathbb{P})$ is given by

$$\begin{aligned}R_{2,j}(\overline{\mathbb{P}}, \mathbb{P}) = \int\int h_j'(\overline{p}_a(y)) &\left\{ \frac{\pi_a(x)}{\overline{\pi}_a(x)} - 1 \right\} \Big\{ \eta_a(y \mid x) - \overline{\eta}_a(y \mid x) \Big\}\ dy\ d\mathbb{P}(x) \\ &+ \frac{1}{2} \int h_j''(p_a^*(y)) \Big\{ \overline{p}_a(y) - p_a(y) \Big\}^2\ dy. \end{aligned} \tag{27}$$

Now we give a lemma showing why finding a von Mises expansion like the above, with second-order remainder, is equivalent to finding the efficient influence function in a nonparametric model. This will prove φ_m is the efficient influence function for $m(\beta)$, and will also be useful for later results.

Lemma 2. *Let $\psi : \mathcal{P} \to \mathbb{R}$ denote some real-valued functional on a nonparametric model, so the set of distributions $\mathcal{P}$ does not constrain the tangent space. Assume the functional satisfies*

$$\psi(\overline{\mathbb{P}}) - \psi(\mathbb{P}) = \int \varphi(z; \overline{\mathbb{P}})\ (d\overline{\mathbb{P}} - d\mathbb{P}) + R_2(\overline{\mathbb{P}}, \mathbb{P})$$

for some mean-zero and finite variance function $\varphi(z;\mathbb{P})$. Then φ is the efficient influence influence function if $\frac{d}{d\epsilon} R_2(\mathbb{P}, \mathbb{P}_\epsilon)|_{\epsilon=0} = 0$ for any smooth parametric submodel.

Proof. Recall from Bickel et al. [1993] and van der Vaart [2002] that the efficient influence function is the mean-zero function whose variance equals the nonparametric efficiency bound, and is given by the unique function ϕ that is a valid submodel score (or limit of such scores) satisfying pathwise differentiability, i.e.,

$$\frac{d}{d\epsilon}\psi(\mathbb{P}_\epsilon)\Big|_{\epsilon=0} = \int \phi(z;\mathbb{P})\left(\frac{d}{d\epsilon}\log d\mathbb{P}_\epsilon\right)\Big|_{\epsilon=0} d\mathbb{P}(z) \tag{28}$$

for $\mathbb{P}_\epsilon$ any smooth parametric submodel. (e.g., differentiable in quadratic mean) In a nonparametric model only one such function ϕ satisfies the above. We will show that the above is satisfied by the function φ in the statement of the lemma.
First note that the assumed expansion implies

$$\psi(\mathbb{P}) - \psi(\mathbb{P}_\epsilon) = -\int \varphi(z;\mathbb{P})\ d\mathbb{P}_\epsilon + R_2(\mathbb{P},\mathbb{P}_\epsilon)$$

for any submodel $\mathbb{P}_\epsilon$. Differentiating with respect to ϵ gives

$$\begin{aligned}\frac{d}{d\epsilon}\psi(\mathbb{P}_\epsilon) &= \frac{d}{d\epsilon}\int \varphi(z;\mathbb{P})\ d\mathbb{P}_\epsilon + \frac{d}{d\epsilon}R_2(\mathbb{P},\mathbb{P}_\epsilon)\\ &= \int \varphi(z;\mathbb{P})\left(\frac{d}{d\epsilon}\log d\mathbb{P}_\epsilon\right)\ d\mathbb{P}_\epsilon + \frac{d}{d\epsilon}R_2(\mathbb{P},\mathbb{P}_\epsilon),\end{aligned}$$

where the second line follows from the dominated convergence theorem and uses the fact that $\frac{d}{d\epsilon}\log d\mathbb{P}_\epsilon = \frac{d}{d\epsilon}d\mathbb{P}_\epsilon / d\mathbb{P}_\epsilon$. Therefore evaluating at $\epsilon = 0$ we have

$$\frac{d}{d\epsilon}\psi(\mathbb{P}_\epsilon)\Big|_{\epsilon=0} = \int \varphi(z;\mathbb{P})\left(\frac{d}{d\epsilon}\log d\mathbb{P}_\epsilon\right)\Big|_{\epsilon=0}\ d\mathbb{P} + \frac{d}{d\epsilon}R_2(\mathbb{P},\mathbb{P}_\epsilon)\Big|_{\epsilon=0},$$

which yields the desired pathwise differentiability by the fact that $\frac{d}{d\epsilon}R_2(\mathbb{P},\mathbb{P}_\epsilon)|_{\epsilon=0} = 0$. □

Now we can immediately apply Lemma 2, noting that

$$\frac{d}{d\epsilon}R_2(\mathbb{P},\mathbb{P}_\epsilon)\Big|_{\epsilon=0} = 0$$

by virtue of the fact that the remainder $R_{2,j}(\mathbb{P},\mathbb{P}_\epsilon)$ in (27) consists of only second-order products of errors between $\mathbb{P}$ and $\mathbb{P}_\epsilon$. This follows since applying the product rule yields a sum of two terms, each of which is a product of a derivative term (which may not be zero at $\epsilon = 0$) and an error term involving differences of components of $\mathbb{P}_\epsilon$ and $\mathbb{P}$ (which will be zero at $\epsilon = 0$). Therefore φ_m is the efficient influence function for the parameter $m(\beta)$. The efficient influence functions for β_0 and $g(y;\beta_0)$ follow similarly, via the chain rule.

B.4 Proof of Theorem 2

From Lemmas 1 and 2, the efficient influence function of $\psi_f = \int f\left(p_1(y), p_0(y)\right) p_0(y)\ dy$ if $p_0(y)$ were known would be

$$\begin{aligned}\varphi_1(z;\mathbb{P}) = \frac{\mathbb{1}(A=1)}{\pi(1 \mid X)}&\left\{h_1'(Y) - \int h_1'(y)\eta_1(y \mid X)\ dy\right\}\\ &+ \int h_1'(y)\eta_1(y \mid X)\ dy - \int h_1'(y)\eta_1(y \mid x)\ dy\ d\mathbb{P}(x)\end{aligned}$$

where

$$h_1'(y) = h_1'(y; p_0, p_1) = p_0(y) f_1'(p_1(y), p_0(y)).$$

Similarly, if $p_1(y)$ were known, the efficient influence function of ψ_f would be

$$\begin{aligned}\varphi_0(z; \mathbb{P}) = \frac{\mathbb{1}(A=0)}{\pi(0 \mid X)} \left\{ h_0'(y) - \int h_0'(y) \eta_0(y \mid X) \, dy \right\} \\ + \int h_0'(y) \eta_0(y \mid X) \, dy - \int h_0'(y) \eta_0(y \mid x) \, dy \, d\mathbb{P}(x),\end{aligned}$$

where

$$h_0'(y) = h_0'(y; p_0, p_1) = f(p_1(y), p_0(y)) + p_0(y) f_2'(p_1(y), p_0(y)).$$

The result then follows from the fact that the influence function when p_1 and p_0 are both unknown is the sum of the two influence functions when p_1 and p_0 are known, separately.

B.5 Proof of Claim in Remark 4

Here we show why rates for estimating $p_a(y)$ with the plug-in estimator $\widehat{p}_a(y) = \mathbb{P}_n\{\widehat{\eta}_a(y \mid X)\}$ will not be slower than those for estimating $\widehat{\eta}_a(y \mid x)$, by bounding the mean squared error of the former in terms of the latter. To this end we denote the pointwise bias and variance of $\widehat{\eta}_a$ as $\mathbb{E}\{\widehat{\eta}_a(y \mid x)\} - \eta_a(y \mid x) = b(y \mid x)$ and $\text{var}\{\widehat{\eta}_a(y \mid x)\} = v(y \mid x)$, respectively. First note for the bias that

$$\begin{aligned}\mathbb{E}\{\widehat{p}_a(y)\} - p_a(y) &= \mathbb{E}\Big[\mathbb{P}_n\{\widehat{\eta}_a(y \mid X)\}\Big] - \int \eta_a(y \mid x) \, d\mathbb{P}(x) \\ &= \mathbb{E} \int \widehat{\eta}_a(y \mid x) \, d\mathbb{P}(x) - \int \eta_a(y \mid x) \, d\mathbb{P}(x) = \int b(y \mid x) \, d\mathbb{P}(x),\end{aligned}$$

where in the second line we used iterated expectation, conditioning on the training sample D^n used to construct $\widehat{\eta}_a$. For the variance we similarly have

$$\begin{aligned}\text{var}\{\widehat{p}_a(y)\} &= \text{var}\left(\mathbb{E}\Big[\mathbb{P}_n\{\widehat{\eta}_a(y \mid X)\} \mid D^n\Big]\right) + \mathbb{E}\left(\text{var}\Big[\mathbb{P}_n\{\widehat{\eta}_a(y \mid X)\} \mid D^n\Big]\right) \\ &= \text{var} \int \widehat{\eta}_a(y \mid x) \, d\mathbb{P}(x) + \frac{1}{n} \mathbb{E}\left[\text{var}\Big\{\widehat{\eta}_a(y \mid X) \mid D^n\Big\}\right].\end{aligned}$$

For the first term above, by Cauchy-Schwarz we have

$$\begin{aligned}\text{var} \int \widehat{\eta}_a(y \mid x) \, d\mathbb{P}(x) &= \mathbb{E}\left\{\int \Big[\widehat{\eta}_a(y \mid x) - \mathbb{E}\{\widehat{\eta}_a(y \mid x)\}\Big] \, d\mathbb{P}(x)\right\}^2 \\ &\leq \mathbb{E} \int \Big[\widehat{\eta}_a(y \mid x) - \mathbb{E}\{\widehat{\eta}_a(y \mid x)\}\Big]^2 \, d\mathbb{P}(x) = \int v(y \mid x) \, d\mathbb{P}(x)\end{aligned}$$

And for the second term note that

$$\begin{aligned}\mathbb{E}\left[\text{var}\Big\{\widehat{\eta}_a(y \mid X) \mid D^n\Big\}\right] &\leq \mathbb{E} \int \widehat{\eta}_a(y \mid x)^2 \, d\mathbb{P}(x) = \int \left(v(y \mid x) + \Big[\mathbb{E}\{\widehat{\eta}_a(y \mid x)\}\Big]^2\right) \, d\mathbb{P}(x) \\ &= \int \left(v(y \mid x) + \Big[\mathbb{E}\Big\{\widehat{\eta}_a(y \mid x) - \eta_a(y \mid x) + \eta_a(y \mid x)\Big\}\Big]^2\right) \, d\mathbb{P}(x) \\ &\leq \int \Big\{v(y \mid x) + 2b(y \mid x)^2 + 2\eta_a(y \mid x)^2\Big\} \, d\mathbb{P}(x).\end{aligned}$$

Therefore as long as $\int \eta_a(y \mid x)^2 d\mathbb{P}(x) \leq C$, we have

$$\mathbb{E}\left[\Big\{\widehat{p}_a(y) - p_a(y)\Big\}^2\right] \leq \left(1 + \frac{2}{n}\right) \int \mathbb{E}\left[\Big\{\widehat{\eta}_a(y \mid x) - \eta_a(y \mid x)\Big\}^2\right] d\mathbb{P}(x) + \frac{2C}{n}$$

B.6 Proof of Theorem 3

First we present a master lemma giving the rate of convergence of the solution to a sample-split estimating equation. The logic parallels that of Theorem 5.31 of van der Vaart [2000].

Lemma 3. *Let $\varphi(z;\theta,\eta)$ denote a vector estimating function for target parameter $\theta \in \mathbb{R}^p$ and nuisance functions $\eta \in H$ for some function space H. Suppose the true values (θ_0, η_0) satisfy $\mathbb{P}\{\varphi(Z;\theta_0,\eta_0)\} = 0$, and define the estimator $\widehat{\theta}$ as an approximate solution to the estimating equation satisfying*

$$\mathbb{P}_n\{\varphi(Z;\widehat{\theta},\widehat{\eta})\} = o_{\mathbb{P}}(1/\sqrt{n})$$

where $\widehat{\eta}$ is estimated on a separate independent sample. Assume:

1. *The function class $\{\varphi(z;\theta,\eta) : \theta \in \mathbb{R}^p\}$ is Donsker in θ for any fixed η.*
2. *The estimators are consistent, i.e., $\widehat{\theta} - \theta_0 = o_{\mathbb{P}}(1)$ and $\|\widehat{\eta} - \widehat{\eta}_0\| = o_{\mathbb{P}}(1)$.*
3. *The map $\theta \mapsto \mathbb{P}\{\varphi(Z;\theta,\eta)\}$ is differentiable at θ_0 uniformly in η, with nonsingular derivative matrix $\frac{\partial}{\partial\theta}\mathbb{P}\{\varphi(Z;\theta,\eta)\}|_{\theta=\theta_0} = V(\theta_0,\eta)$, where $V(\theta_0,\widehat{\eta}) \xrightarrow{p} V(\theta_0,\eta_0)$.*

Then

$$\widehat{\theta} - \theta_0 = -V(\theta_0,\eta_0)^{-1}(\mathbb{P}_n - \mathbb{P})\{\varphi(Z;\theta_0,\eta_0)\} + O_{\mathbb{P}}\Big(R_n\Big) + o_{\mathbb{P}}\left(\frac{1}{\sqrt{n}}\right)$$

for $R_n = \mathbb{P}\{\varphi(Z;\theta_0,\widehat{\eta}) - \varphi(Z;\theta_0,\eta_0)\}$.

Proof. First note that, since $(\widehat{\theta},\widehat{\eta})$ and (θ,η) are approximate and exact solutions of the empirical and population moment conditions, respectively, we have

$$\begin{aligned} o_{\mathbb{P}}(1/\sqrt{n}) &= \mathbb{P}_n\{\varphi(Z;\widehat{\theta},\widehat{\eta})\} - \mathbb{P}\{\varphi(Z;\theta_0,\eta_0)\} \\ &= (\mathbb{P}_n - \mathbb{P})\{\varphi(Z;\theta_0,\eta_0)\} + (\mathbb{P}_n - \mathbb{P})\{\varphi(Z;\widehat{\theta},\widehat{\eta}) - \varphi(Z;\theta_0,\widehat{\eta})\} \quad (29) \\ &\quad + (\mathbb{P}_n - \mathbb{P})\{\varphi(Z;\theta_0,\widehat{\eta}) - \varphi(Z;\theta_0,\eta_0)\} \quad (30) \\ &\quad + \mathbb{P}\{\varphi(Z;\widehat{\theta},\widehat{\eta}) - \varphi(Z;\theta_0,\widehat{\eta})\} + \mathbb{P}\{\varphi(Z;\theta_0,\widehat{\eta}) - \varphi(Z;\theta_0,\eta_0)\} \quad (31) \end{aligned}$$

where the second equality follows by simply adding and subtracting terms. The first term in (29) is a simple sample average of a fixed function and so will be asymptotically Gaussian by the central limit theorem. The second term in (29) and the term in (30) are empirical process terms. The first term in (31) will be linearized in $(\widehat{\theta} - \theta_0)$, while the second term in (31) captures the effect of the nuisance estimation error. We will tackle each of these in turn.
Under the Donsker and consistency conditions for $\widehat{\theta}$ in Assumptions 1 and 2, the second term in (29) is $o_{\mathbb{P}}(1/\sqrt{n})$ by Lemma 19.24 of van der Vaart [2000]. Under the consistency of $\widehat{\eta}$ in Assumption 2 and the sample splitting, the term in (30) is $o_{\mathbb{P}}(1/\sqrt{n})$ by Lemma 2 of Kennedy et al. [2020].
By the differentiability of the map $\theta \mapsto \mathbb{P}\{\varphi(Z;\theta,\eta)\}$ in Assumption 3, the first term in (31) can be expressed as

$$\begin{aligned} \mathbb{P}\{\varphi(Z;\widehat{\theta},\widehat{\eta}) - \varphi(Z;\theta_0,\widehat{\eta})\} &= V(\theta_0,\widehat{\eta})(\widehat{\theta} - \theta_0) + o_{\mathbb{P}}(\|\widehat{\theta} - \theta_0\|) \\ &= V(\theta_0,\eta_0)(\widehat{\theta} - \theta_0) + o_{\mathbb{P}}(\|\widehat{\theta} - \theta_0\|) \end{aligned}$$

where the last line follows by the consistency of $V(\theta_0,\widehat{\eta})$ in Assumption 3.

Therefore we have

$$o_{\mathbb{P}}(1/\sqrt{n}) = (\mathbb{P}_n - \mathbb{P})\{\varphi(Z;\theta_0,\eta_0)\} + V(\theta_0,\eta_0)(\widehat{\theta} - \theta_0) + R_n + o_{\mathbb{P}}(\|\widehat{\theta} - \theta_0\|)$$

where we let $R_n = \mathbb{P}\{\varphi(Z;\theta_0,\widehat{\eta}) - \varphi(Z;\theta_0,\eta_0)\}$ denote the second term in (31), or equivalently

$$\widehat{\theta} - \theta_0 = -V(\theta_0,\eta_0)^{-1}(\mathbb{P}_n - \mathbb{P})\{\varphi(Z;\theta_0,\eta_0)\} + O_{\mathbb{P}}(R_n) + o_{\mathbb{P}}(\|\widehat{\theta} - \theta_0\|) + o_{\mathbb{P}}(1/\sqrt{n})$$

by the nonsingularity of the derivative matrix in Assumption 3. This implies

$$\|\widehat{\theta} - \theta_0\|(1 + o_{\mathbb{P}}(1)) = O_{\mathbb{P}}(1/\sqrt{n} + R_n)$$

so that $\|\widehat{\theta} - \theta_0\| = O_{\mathbb{P}}(1/\sqrt{n} + R_n)$, which gives the result after noting that $o_{\mathbb{P}}(O_{\mathbb{P}}(1/\sqrt{n} + R_n)) = o_{\mathbb{P}}(1/\sqrt{n} + R_n)$ and that $O_{\mathbb{P}}(R_n) + o_{\mathbb{P}}(R_n) = O_{\mathbb{P}}(R_n)$. □

Now we can apply Lemma 3 to prove Theorem 3. First note that by definition the estimator satisfies

$$\mathbb{P}_n\{\phi(Z;\widehat{\beta},\widehat{\eta})\} = o_{\mathbb{P}}(1/\sqrt{n})$$

for $\phi(Z;\beta,\eta) = m(\beta;\eta) + \varphi(Z;\beta,\eta)$, and the true values (β_0,η_0) satisfy $\mathbb{P}\{\phi(Z;\beta_0,\eta_0)\} = 0$, again by definition.
Conditions 1–3 of Lemma 3 hold by Assumptions 1–4 of Theorem 3, so the result follows by virtue of the fact that

$$\begin{aligned}
R_n &\equiv \mathbb{P}\{\phi(Z;\beta_0,\widehat{\eta}) - \phi(Z;\beta_0,\eta_0)\} \\
&= m(\beta_0;\widehat{\eta}) + \mathbb{P}\{\varphi_m(Z;\beta_0,\widehat{\eta})\} \\
&= \mathbb{P}\int h'(\widehat{p}_a(y))\left\{\frac{\pi_a(X)}{\widehat{\pi}_a(X)} - 1\right\}\Big\{\eta_a(y \mid X) - \widehat{\eta}_a(y \mid X)\Big\}\, dy \\
&\qquad + \frac{1}{2}\int h''(\widehat{p}_a^*(y))\Big\{\widehat{p}_a(y) - p_a(y)\Big\}^2\, dy \\
&\lesssim \|\widehat{\pi} - \pi\|\|\widehat{\eta}_a - \eta_a\| + \delta\|\widehat{p}_a - p_a\|^2
\end{aligned}$$

where the second to last line follows from the result given in equation (27) of the proof of Theorem 1, for the vectors h and h' as defined in (25), and the last line follows by the Cauchy-Schwarz inequality and boundedness assumptions on h', h'', and $1/\widehat{\pi}$.